Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Modelos para dados categorizados ordinais com efeito aleatório: uma aplicação à análise sensorial

Maíra Blumer Fatoretto

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

Piracicaba
2016

Maíra Blumer Fatoretto
Bacharel em Estatística

Modelos para dados categorizados ordinais com efeito aleatório: uma aplicação à análise sensorial

versão revisada de acordo com a resolução CoPGr 6018 de 2011

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

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

Piracicaba
2016

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

Fatoretto, Maíra Blumer
Modelos para dados categorizados ordinais com efeito aleatório: uma aplicação à análise sensorial / Maíra Blumer Fatoretto. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2016.
65 p. : il.

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

1. Dados categorizados 2. Modelos de logitos cumulativos 3. Modelos lineares generalizados mistos I. Título

CDD 519.53
F254m

DEDICATÓRIA

Dedico este trabalho …

A minha família e amigos que sempre estiveram presente nos momentos mais importantes da minha vida.

Aos excelentes professores que me fizeram amar a estatística.

AGRADECIMENTOS

Agradeço primeiramente a Deus por me dar forças, saúde, inteligência, por todo o aprendizado adquirido e por poder desenvolver este projeto.

O meu obrigado a minha mãe Ana Maria Blumer Fatoretto, meu pai Wilson Aparecido Fatoretto, minha irmã Laís Blumer Fatoretto, minha avó Silvina da Piedade Mateus Blumer e meu noivo Pedro Augusto Peres Spagnol, que sempre me apoiaram em minhas decisões e me ajudaram a enfrentar as dificuldades encontradas, com muito carinho, amor e compreensão.

A todos os meus colegas do departamento, que sempre me ajudaram nas disciplinas e trabalhos durante todo o mestrado, tornando-se grandes amigos, em especial a: Alessandra dos Santos, Vinícius Menarin, Christopher Pádua, Viviane Paulenas e Gislaine Pereira.

Agradeço a todos os funcionários e professores do departamento LCE/ESALQ/USP, por todo o saber ensinado, com muito comprometimento e dedicação, em especial ao meu orientador professor Dr. Idemauro Antonio Rodrigues de Lara que sempre teve paciência para me orientar e os professores Dra. Renata Alcarde, Dra. Taciana Villela Savian e Dr. Rafael Maia, que me auxiliaram em projetos, sempre com muita atenção, sendo grandes exemplos para minha carreira acadêmica.

Obrigada a professora Dra. Marta Helena Fillet Spoto e sua orientanda Ana Carolina Loro do departamento do Departamento de Agroindústria, Alimentos e Nutrição por cederem os dados utilizados neste trabalho.

À Capes pelo auxílio financeiro concedido nestes dois anos de mestrado.

Aos meus familiares, minha segunda família Peres Spagnol, minhas amigas estatísticas da Unicamp e minhas amigas de Limeira, que sempre entenderam meus momentos de ausência por motivos de estudo e por todas as palavras de apoio e perseverança.

Muito Obrigada.

“Que os vossos esforços desafiem as impossibilidades, lembrai-vos de que as grandes coisas do homem foram conquistadas do que parecia impossível".

Charles Chaplin

“Ninguém caminha sem aprender a caminhar, sem aprender a fazer o caminho caminhando, refazendo e retocando o sonho pelo qual se pôs a caminhar".

Paulo Freire

SUMÁRIO

RESUMO

Modelos para dados categorizados ordinais com efeito aleatório: uma aplicação à análise sensorial

Os modelos para dados categorizados ordinais são extensões dos Modelos Lineares Generalizados e suas suposições e inferências são fundamentadas por esta classe de modelos. Os Modelos de Logitos Cumulativos, em que a função de ligação é constituída de probabilidades acumuladas, são muito utilizados para este tipo de variável, sendo uma de suas simplificações, os Modelos de Chances Proporcionais, em que para todas as covaríaveis no modelo há um crescimento linear nas razões de chances, porém, neste caso, é necessária a verificação da suposição de paralelismo. Outros modelos como o Modelo de Chances Proporcionais Parciais, o Modelo de Categorias Adjacentes e o Modelo Logito de Razão Contínua também podem ser utilizados. Em diversos estudos deste tipo, é necessário a utilização de modelos mistos, seja pelo tipo de um fator ou a dependência entre observações da variável resposta. Objetivou-se, neste trabalho, o estudo de modelos para variável resposta ordinal com a inclusão de um ou mais efeitos aleatórios. Esses modelos são ilustrados com a utilização de dados reais de análise sensorial, cuja variável resposta é constituída de uma escala ordinal e deseja-se saber dentre duas variedades de tomates desidratados (Italiano e Sweet Grape), qual teve melhor aceitação pelos consumidores. Nesse experimento os provadores avaliaram uma única vez cada uma das variedades, sendo as repetições constituídas pelas avaliações dadas por diferentes provadores. Nesse caso, é necessária a inclusão de um efeito aleatório por provador, para que o modelo consiga capturar as diferenças entre esses provadores não treinados. O Modelo de Chances Proporcionais ajustou-se de maneira satisfatória aos dados, podendo-se fazer uso das estimativas de probabilidades e razões de chances para a interpretação dos resultados e concluindo-se que o sabor da variedade Sweet Grape foi o que mais agradou os provadores, independente do sexo.

Palavras-chave: Dados categorizados; Modelos de logitos cumulativos; Modelos lineares generalizados mistos

ABSTRACT

Models for ordinal categorical data with random effects: an application to the sensory analysis

Models for ordinal categorical data are extensions of the Generalized Linear Models and their assumptions and inferences are based on this class of models. The Cumulative Logit Models in wich the link function consists of accumulated probabilities are more used for this type of variable, with one of its simplifications are the Proportional Odds Model, in wich for all covariates in the model there is a linear growth in odds ratios, but in this case, checking the parallelism assumption is required. Other models such as the Partial Proportional Odds Model, the Adjacent-Categories Logits and Continuation-Ratio Logits model can also be used. In several of such studies, the use of mixed models is required, either by type of factor or dependence between the response variable observations. The aim of this work is studying models for ordinal variable response with the inclusion of one or more random effects. These models are illustrated by using real data of sensory analysis, the response variable consists of an ordinal scale and we want to know from two varieties of dried tomatoes, Italian and Sweet Grape, which had better acceptance by consumers. In this experiment, the panelists evaluated each variety once, and the repetitions constituted by the ratings given by different tasters. In this case, the inclusion of a random effect by taster is required so that the model can capture the difference between these untrained tasters. The Proportional Odds Model fitted satisfactorily to the data and it is possible to make use of the estimates of probabilities and odds ratios for the interpretation of results and concluding that the taste of the variety Sweet Grape was the one that most pleased the tasters regardless of sex.

Keywords: Categorical data; Cumulative Logit Models; Generalized Linear Mixed Models

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

1 INTRODUÇÃO

Estudos com variáveis respostas categorizadas aparecem frequentemente em diversas áreas de pesquisas. Por exemplo na área da medicina é muito comum o interesse em saber se um paciente teve melhora em uma doença (sim ou não), ou qual o grau da doença (leve, moderado ou avançado). Na área das ciências sociais esse tipo de variável pode ser encontrada em estudos de opinião, cujo interesse é saber o nível de concordância com um determinado assunto (discordo totalmente, discordo, nem discordo nem concordo, concordo ou concordo totalmente). Na agronômica, é comum encontrar experimentos cuja variável reposta é o grau de colonização por fungos, ou a severidade de doenças em frutos. Variáveis contínuas também podem ser categorizadas, como por exemplo, um indivíduo pode ser classificado de acordo com sua altura (baixo, médio ou alto) (GIOLO, 2012).

A classificação dessas variáveis ocorre de acordo com sua escala de mensuração: nominal, se suas categorias não são ordenadas, ou seja, podem ser permutáveis, não seguindo nenhuma estrutura e não influenciando a resposta do modelo; ordinal, cujas categorias seguem uma ordem natural, porém com distâncias desconhecidas entre categorias, não podendo ser permutáveis; intervalar, quando escores estão associados as categorias e existe uma distância numérica entre dois níveis. Métodos de análises para dados nominais não devem ser utilizados para dados ordinais, dado que estes não utilizam a informação de ordenação presente na variável resposta (AGRESTI, 2007; MCCULLAGH; NELDER, 1989).

Muitas vezes na análise desses dados, faz-se uso de transformações ou utilizam-se técnicas nas quais se supõe satisfeitas as pressuposições estatísticas do modelo para variável resposta normal. Porém, esse tipo de variável envolve distribuições multinomiais e técnicas específicas para este caso já foram desenvolvidas. O estudo desse tipo de variável resposta é realizado utilizando extensões dos Modelos Lineares Generalizados (MLG) proposto por Nelder e Wedderburn (1972). Os modelos para variáveis categorizadas ordinais não só descrevem a relação entre a variável resposta e as variáveis explicativas, mas também permitem medir esta relação, por meio das estimativas dos parâmetros e das razões de chances (SPYRIDES-CUNHA, 1998).

McCullagh (1980), propôs os modelos mais utilizados para dados categorizados ordinais, os Modelos de Chances Proporcionais (MCP), sendo simplificado e de fácil interpretação. Agresti (2002) descreve modelos para dados discretos, e é uma referência clássica para dados multicategóricos, em particular para dados ordinais. Ananth e Kleinbaum (1997), em uma revisão dos métodos utilizados para variável resposta ordinal, apresentam o Modelo de Chances Proporcionais Parciais (MCPP) (PETERSON; HARRELL, 1990), o Modelo de Razão Contínua (MRC) (FIENBERG, 1980) e o Modelo Logito de Categorias Adjacentes (MLCA) (AGRESTI,

1984).

Na prática, em diversificadas situações, além do erro, faz-se necessária a inclusão de um ou mais efeitos aleatórios no modelo, muitas vezes pelo experimento possuir um fator aleatório, ou seja, seus níveis são considerados como sendo uma amostra de uma certa população, ou os blocos de um delineamento em blocos casualizados serem sorteados dentre todos os possíveis blocos. Outra possibilidade é o fato de cada unidade experimental possuir mais de uma observação da variável resposta, ocorrendo as denominadas, medidas repetidas. Neste caso, essas observações podem estar correlacionadas, e essa estrutura de correlação precisa estar incluída no modelo. Umas das possibilidades nesta modelagem são os modelos mistos. Proposto inicialmente para o caso linear por Laird e Ware (1982), foi estendido posteriormente para os Modelos Lineares Generalizados (BRESLOW; CLAYTON, 1993). Tutz e Hennevogl (1996), Hedeker e Mermelstein (2000) dentre outros, apresentaram a inclusão de efeitos aleatórios em modelos para dados categorizados ordinais.

Este trabalho tem como objetivo apresentar a metodologia de modelos para dados categorizados ordinais, abordando a inclusão de efeitos aleatórios, com o desenvolvimento de uma aplicação utilizando dados de análise sensorial na avaliação de duas variedades de tomates.

O tomate é um produto muito comercializado no Brasil, além de possuir uma rica composição nutricional, composta por fibras, proteínas e antioxidantes. Duas de suas variedades, o Italiano e o Sweet Grape são muito produzidas e de grande aceitação pelos consumidores (KOH; CHAROENPRASERT; MITCHELL, 2012). Devido este produto ser altamente perecível, novas técnicas de desidratação estão sendo desenvolvidas, com o objetivo de diminuir o desperdício que ocorre na comercialização, utilizando como matéria prima o fruto *in natura* (ABREU *et al.*, 2011). Porém, estes novos produtos precisam ter aceitação no mercado, e por meio da análise sensorial é possível determinar esta aceitabilidade.

A análise sensorial, definida como disciplina científica (ABNT, 1993), é usada para analisar e medir características dos alimentos. Ela pode ser realizada por métodos discriminantes, fazendo o uso de testes para indicar diferenças entre amostras de produtos (TEIXEIRA, 2009). Os denominados teste de escalas podem ser utilizados para medir as principais características sensoriais, como o odor, o aroma e o sabor. Neste caso, geralmente utiliza-se uma escala de 9 pontos, denominada hedônica (IFT, 1981), sendo esta resposta enquadrada como variável categorizada ordinal. Neste tipo de análise os provadores avaliam sensorialmente mais de um produto. Nesta situação, a correlação entre as respostas do mesmo provador precisa ser considerada. Quando as repetições são constituídas por classificações de diferentes provadores, é necessário que cada provador seja considerado como um bloco, pelas diferenças existentes entre eles, sendo que este bloco será aleatório.

Esta dissertação está organizada da seguinte forma: uma revisão bibliográfica sobre a teoria dos Modelos Lineares Generalizados e que se estende para dados categorizados ordinais

apresentados na Seção 2, com enfoque para os Modelos Logitos Cumulativos. O Capítulo 3 apresenta uma aplicação e a metodologia utilizada na análise. No Capítulo 4 encontram-se os resultados e interpretações obtidas.

2 REVISÃO BIBLIOGRÁFICA

Nesta seção são apresentados aspectos da análise de dados categorizados, assim como uma revisão da teoria dos Modelos Lineares Generalizados, abordando sua extensão para modelos com variável resposta categorizada ordinal.

2.1 Dados Categorizados

A análise de dados discretos está associada a variáveis repostas categorizadas, cuja escala de medida consiste em um conjunto de categorias e distribuições discretas de probabilidade, como binominal, Poisson, Multinomial, entre outras, estão associadas a elas (GIOLO, 2012; AGRESTI, 2007). De acordo com McCullagh e Nelder (1989), estas variáveis podem ser classificadas de acordo com a quantidade de categorias que possuem, sendo dicotômica ou binárias quando apresentam duas categorias de respostas, ou politômica para três ou mais. Quando politômicas, podem ser classificadas em três escalas de mensuração:

i) **Escala Nominal**: As categorias não possuem uma ordem natural. Exemplos: espécies de animais, tipos de escolas, variedades da plantas, religiões;

ii) **Escala Ordinal**: As categorias não podem ser permutáveis por seguirem uma ordem natural, não existindo distância ou espaçamento entre pares de categorias de respostas. Exemplos: classe social, grau de severidade de doenças, nível de satisfação de um produto;

iii) **Escala Intervalar**: Escores são adicionados a cada uma das categorias, que geralmente são constituídas por classes numéricas, em que existe uma distância numérica entre os níveis. Exemplos: renda, faixa etária, escala de temperatura.

Por outro lado, as variáveis explicativas ou independentes, associadas à variável resposta ordinal, podem ser de qualquer natureza, e por meio da classificação cruzada destas variáveis é possível modelar as contagens ou frequências observadas para cada uma das categorias da variável resposta (SPYRIDES-CUNHA, 1998).

2.1.1 Tabelas de Contingência

Uma das possibilidades mais utilizadas para descrever os dados categorizados são as tabelas de contingências, termo introduzido por Karl Pearson (1904), sendo uma representação útil para medir o grau de associação entre as variáveis do estudo. Tais dados podem surgir de diferentes tipos de delineamentos e a interpretação dos resultados obtidos irá depender do objetivo que está sendo analisado do problema e do delineamento que gerou os dados.

A dimensão de uma tabela de contingência é determinada pelo número de variáveis, sendo o respectivo número de caselas determinado pelo produto do número de categorias ou níveis de cada uma das variáveis (PAULINO; SINGER, 2006; AGRESTI, 2007). Qualquer tabela de contingência pode ser descrita num formato bidimensional, sendo as subpopulações formadas pelas combinações dos níveis das variáveis explicativas e as categorias de resposta pela combinação dos níveis das variáveis respostas, como se indica na Tabela 2.1:

Tabela 2.1 – Exemplo de estrutura de uma tabela de continência com as notações de frequências e totais marginais

	Categorias de resposta						
Subpopulação	1	2	...	m	...	r	Total
1	n_{11}	n_{12}	...	n_{1m}	...	n_{1r}	$n_{1\bullet}$
2	n_{21}	n_{22}	...	n_{2m}	...	n_{2r}	$n_{2\bullet}$
.		.	...		...	.	
.		.	...		...	.	
.		.	...		...	.	
q	n_{q1}	n_{q2}	...	n_{qm}	...	n_{qr}	$n_{q\bullet}$
.		.	...		...	.	
.		.	...		...	.	
.		.	...		...	.	
s	n_{s1}	n_{s2}	...	n_{sm}	...	n_{sr}	$n_{s\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$		$n_{\bullet m}$		$n_{\bullet r}$	n

em que $n_{q\bullet} = \sum_{m=1}^{r} n_{qm}$, para $q = 1, \ldots, s$ com n_{qm} representa a frequência associada à casela correspondente à q-ésima linha e m-ésima coluna, sendo n o número de observações total da amostra.

Supondo X e Y duas variáveis categorizadas com i e j níveis respectivamente. A notação $\pi_{ij} = P(X = i, Y = j)$ denota a probabilidade de ocorrência na linha i e coluna j, ou seja, a distribuição de probabilidade conjunta de X e Y. Os totais das linhas e das colunas são as distribuições marginais, que resultam das somas das probabilidade conjuntas, de modo que $\sum_i \pi_{i\bullet} = \sum_j \pi_{j\bullet} = \sum_i \sum_j \pi_{ij} = 1$.

Em estudos em que Y é a uma variável resposta e X uma variável exploratória apresentando categorias fixas, é usual estudar como a probabilidade de Y muda de acordo com cada categoria de X. Dado que um indivíduo é classificado na linha i de X, a probabilidade de ser classificado na coluna j de Y é dada por $\pi_{j|i}$, isto é $P(Y = j|X = i)$, sendo $\sum_j \pi_{j|i} = 1$.

A notação $p_{ij} = P(X = i, Y = j)$ denota a distribuição amostral conjunta, sendo as frequências das caselas denotadas por n_{ij}. Então

$$p_{ij} = n_{ij}/n \quad \text{e} \quad p_{j|i} = n_{ij}/n_{i\bullet}.$$

2.1.2 Delineamentos e Modelos Probabilísticos Associados

Os delineamentos mais usuais para dados categorizados foram descritos por Fienberg(1980) e Dobson(1990). Os modelos probabilísticos são obtidos com base no delineamento amostral e explicam a ocorrência dos dados obtidos, fundamentando as inferências de interesse. Além do delineamento os objetivos da análise também influenciam na escolha dos modelos (PAULINO; SINGER, 2006; AGRESTI, 2002; GIOLO, 2012). Os três delineamentos comumente mais utilizados são:

i) **POISSON**

Este modelo probabilístico é utilizado em delineamentos em que somente o tempo do experimento é estabelecido. Neste caso, observa-se um processo de Poisson para cada uma das caselas da tabela de contingência. Porém, Paulino e Singer (2006) e Giolo (2012), apresentam algumas suposições que são necessárias para a definição deste modelo, como a independência do número de indivíduos em tempos disjuntos, a distribuição do número de indivíduos deve depender somente do comprimento do intervalo de tempo, a probabilidade de um individuo passar em um intervalo de tempo suficientemente pequeno, sendo proporcional ao intervalo e a probabilidade que dois indivíduos passem em um tempo muito pequeno sendo desprezível . Satisfeitas estas suposições, e admitindo que as contagens das caselas (n_{ij}) tem distribuições de Poisson independentes, com médias $\mu_{ij} = T\lambda_{ij}$, sendo λ_{ij} a taxa média por unidade de tempo e T a duração do experimento, o modelo probabilístico associado a este estudo é o produto das distribuições de Poisson independentes, tendo função de probabilidade dada por:

$$P(n_{ij}) = \prod_{ij} \frac{e^{-\mu_{ij}}(\mu_{ij})^{n_{ij}}}{n_{ij}!},$$

em que a frequência total $n = n_{\bullet\bullet}$ tem distribuição de Poisson com média $\mu_{\bullet\bullet} = \sum_{ij} \mu_{ij}$.

ii) **MULTINOMIAL**

Se em um experimento obtém-se uma determinada amostra de tamanho fixo n, em que, cada um dos indivíduos são selecionados de maneira aleatória de acordo com as categorias cruzadas da variável resposta, o vetor associado à tabela de contingência apresenta uma distribuição multinomial com função de probabilidade expressa por

$$P(n_{ij}/n) = n! \prod_{ij} \frac{\pi_{ij}^{n_{ij}}}{(n_{ij}!)},$$

em que $n_{ij} \geq 0, \sum_{i,j} n_{ij} = n$ e $\sum_{i,j} \pi_{ij} = 1$.

iii) **PRODUTO MULTINOMIAL**

Este modelo probabilístico corresponde a um delineamento com o processo de amostragem estratificada em que para cada nível de X se coleta, independentemente, uma amostra aleatória simples. Neste caso normalmente trata-se os totais de linhas como fixas, e para simplificar podemos usar a notação $n_i = n_{i\bullet}$. De acordo com Agresti (2002), supondo que as n_i observações em Y para cada nível de X são independentes, com distribuição de probabilidade $\{\pi_{1|i}, \ldots \pi_{J|i}\}$, satisfazendo $\sum_j n_{ij} = n_i$, tem-se

$$P(n_{ij}/n_i) = \prod_i \frac{n_i!}{\prod_j n_{ij}!} \prod_j \pi_{j|i}^{n_{ij}}.$$

2.2 Modelos de regressão para dados categorizados

2.2.1 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados (MLGs) foram propostos por Nelder e Wedderburn (1972) e proporcionaram opções para diversas distribuições da variável resposta, desde que esta pertença à família exponencial. Isto trouxe maior flexibilidade entre a média da variável resposta e as variáveis explicativas (DOBSON, 1990; PAULA, 2013).

Por definição os MLGs podem ser usados quando se tem uma única variável aleatória Y associada a um conjunto de variáveis explanatórias $x_1, \ldots, x_p$. Para uma amostra de n observações $(y_i, \boldsymbol{x_i})$, em que $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})^T$ é o vetor coluna de variáveis explanatórias, o MLG contém três componentes:

i) Componente aleatório - constituído pelas variáveis aleatórias independentes $Y_1, Y_2, \ldots, Y_n$ identicamente distribuídas e pertencentes à família exponencial da forma canônica, com médias $\mu_1, \mu_2, \ldots \mu_n$, ou seja,

$$f(y_i; \theta_i, \phi) = \exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i; \phi)\}, \tag{2.1}$$

sendo $b(.)$ e $c(.)$ funções conhecidas, θ_i o parâmetro canônico e $\phi > 0$ um parâmetro de dispersão, podendo-se demonstrar que

$$\mathrm{E}(Y_i) = \mu_i = b'(\theta_i) \text{ e } \mathrm{Var}(Y_i) = \phi b''(\theta_i) = \phi V_i,$$

em que $V_i = V(\mu_i) = d\mu_i/d\theta_i$ é denominada função de variância e depende unicamente da média μ_i;

ii) Componente sistemático - formado pelas variáveis explanatórias que entram na forma de uma soma linear de seus efeitos, que dão origem a um vetor de preditores lineares

$$\eta_i = \sum_{r=1}^{p} x_{ir}\beta_r = \boldsymbol{x_i^T\beta} \quad \text{ou} \quad \boldsymbol{\eta = X\beta},$$

em que $\boldsymbol{\eta}$ é o vetor de preditores lineares, $\boldsymbol{\beta}$ é um vetor de p parâmetros desconhecidos e $\boldsymbol{X} = (\boldsymbol{x_1, \ldots, x_n})^T$ é a matriz de delineamento;

iii) Função de ligação - relaciona o componente aleatório ao componente sistemático, conectando a média ao preditor linear

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

sendo $g(.)$ uma função real, monótona e diferenciável (NELDER; WEDDERBURN, 1972; MCCULLAGH; NELDER, 1989).

De acordo com Demétrio (2002), para a escolha do MLG é preciso definir a distribuição da variável reposta; a matriz do modelo que representa o delineamento experimental e a função de ligação.

2.2.1.1 Estimação dos parâmetros

Para o ajuste do modelo é necessária a estimação do vetor de parâmetros $\boldsymbol{\beta}$, sendo que os principais métodos para esta estimação são o Bayesiano, qui-quadrado mínimo e o método de máxima verossimilhança, tendo estes propriedade ótimas, tais como, consistência e eficiência assintótica (COLLETT, 2002; DEMÉTRIO; CORDEIRO; MORAL, 2014). Considerando uma amostra aleatória $\boldsymbol{y} = (y_1, y_2 \ldots y_n)$ com n observações de uma distribuição pertencente à família exponencial 2.1, a função de verossimilhança é dada por:

$$L = L(\boldsymbol{\theta}, \phi; \boldsymbol{y}) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi) = \exp\left\{\sum_{i=1}^{n}[\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i; \phi)]\right\},$$

cujo logaritmo é

$$l = l(\boldsymbol{\theta}; \phi, \boldsymbol{y}) = \log L(\boldsymbol{\theta}; \phi, \boldsymbol{y}) = \sum_{i=1}^{n}\left\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i; \phi)\right\}. \tag{2.2}$$

Da expressão 2.2 pode-se obter o vetor $\boldsymbol{\beta}$ de parâmetros, que maximizam $l(\boldsymbol{\beta}, \phi, \mathbf{y})$, pela regra da cadeia, derivando-se a função escore

$$U_j = \frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \phi^{-1}\sum_{i=1}^{n}(y_i - \mu_i)\frac{1}{V_i}\frac{d\mu_i}{d\eta_i}x_{ij}. \tag{2.3}$$

A estimativa de máxima verossimilhança (EMV) $\hat{\boldsymbol{\beta}}$ é calculada igualando-se U_j a zero para $j = 1, \ldots, p$. Porém em geral estas equações não são lineares, podendo serem resolvidas por processos iterativos como o de Newton-Raphson ou método escore de Fisher.

Pelo método de Newton-Raphson, tem-se:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + [\mathbf{K}^{(m)}]^{-1} U^{(m)},$$

sendo $U^{(m)}$ o vetor escore, com elementos $\left(\frac{\partial l}{\partial \beta_j}\right)$ e $\mathbf{K}^{(m)}$ a matriz de informação observada com os elementos $\left(-\frac{\partial^2 l}{\partial \beta_j \partial \beta_j'}\right)$ provadores em $\boldsymbol{\beta} = \boldsymbol{\beta^{(m)}}$.

Substituindo a matriz de informação observada *I* pela matriz de informação de esperada de Fisher é possível obter

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

sendo $\boldsymbol{X}$ a matriz do modelo; $\boldsymbol{W} = \text{diag}\{w_1, \ldots, w_n\}$ uma matriz diagonal de pesos que capta a informação sobre a distribuição, em que $w_i = V_i^{-1}(d\mu_i/d\eta_i)^2$ denominada função peso e $\boldsymbol{z}^{(m)} = \boldsymbol{X}\boldsymbol{\beta}^{(m)} + \Delta^{(m)}(\boldsymbol{y} - \boldsymbol{\mu})^{(m)} = \eta^{(m)} + \Delta^{(m)}(\boldsymbol{y} - \boldsymbol{\mu})^{(m)}$ é o vetor da variável dependente ajustada no passo m, com Δ = diag$\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$. No passo inicial pode-se tomar $\hat{\boldsymbol{\beta}}^{(0)} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T} \hat{\boldsymbol{\eta}}$, sendo que $\hat{\eta}_i = g(\hat{\mu}_i) = g(y_i)$.

Para as distribuições binomial e Poisson, estudadas neste trabalho, tem-se o parâmetro de dispersão $\phi = 1$. Porém quando ϕ é desconhecido é necessário estimá-lo para realizar as inferências sobre o modelo (DOBSON, 1990; COSTA, 2003; DEMÉTRIO; CORDEIRO; MORAL, 2014).

2.2.1.2 Teste de hipóteses

Três estatísticas são úteis para testar as hipóteses relativas aos parâmetros $\beta' s$, sendo estas deduzidas da distribuição assintótica dos seus estimadores de máxima verossimilhança. A estatística escore (RAO, 1973) é obtida a partir da função definida pela expressão 2.3. A estatística de Wald (1943) é baseada na distribuição normal assintótica de $\hat{\boldsymbol{\beta}}$ e é mais utilizada nos testes relativos a um único coeficiente β_j. A razão de verossimilhança é o critério que define o teste uniformemente mais poderoso, podendo ser obtida como uma diferença de desvios de modelos encaixados (MCCULLOCH; NEUHAUS, 2001; DEMÉTRIO; CORDEIRO; MORAL, 2014).

Suponha que o interesse esteja em testar $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0}$ e ϕ conhecido, versus a hipótese alternativa $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta_0}$, em que $\boldsymbol{\beta_0}$ é um vetor especificado para o vetor $\boldsymbol{\beta}$ de parâmetros desconhecidos. Temos como estatística de Wald, da razão de verossimilhança e da estatística escore, respectivamente:

$$W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})^T \hat{\mathbf{K}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \sim \chi^2_{p,1-\alpha}, \tag{2.4}$$

$$w = 2[l(\hat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta_0})] \sim \chi^2_{p,1-\alpha} \tag{2.5}$$

e

$$S_R = \mathbf{U}(\boldsymbol{\beta_0})^{\mathbf{T}}\mathbf{K_0^{-1}}\mathbf{U}(\boldsymbol{\beta_0}) \sim \chi^2_{p,1-\alpha},$$

em que $\mathbf{U}(\boldsymbol{\beta_0})$ e $\mathbf{K_0}$ são o vetor escore e matriz de informação avaliada em $\boldsymbol{\beta_0}$; $l(\hat{\boldsymbol{\beta}})$ e $l(\boldsymbol{\beta_0})$ são os valores do logaritmo da função de verossimilhança (2.2) em $\hat{\boldsymbol{\beta}}$ e $\boldsymbol{\beta_0}$ e $\hat{\mathbf{K}}$ a matriz de informação avaliado na EMV de $\hat{\boldsymbol{\beta}}$ (DOBSON, 1990; PAULA, 2013).

Quando deseja-se testar um subconjunto de parâmetros, é feita a comparação dos valores do logaritmo da função de verossimilhança maximizada sem restrição $(l(\hat{\boldsymbol{\beta}}_\mathbf{1}, \hat{\boldsymbol{\beta}}_\mathbf{2}))$ e sob H_0 $(l(\boldsymbol{\beta}_{\mathbf{1,0}}, \tilde{\boldsymbol{\beta}}_\mathbf{2}))$ em que $\tilde{\boldsymbol{\beta}}_\mathbf{2}$ é o Estimador de Máxima verossimilhança de $\boldsymbol{\beta}_\mathbf{2}$ sob H_0. Como estatística para o teste temos

$$w = 2[l(\hat{\boldsymbol{\beta}}_\mathbf{1}, \hat{\boldsymbol{\beta}}_\mathbf{2}) - l(\boldsymbol{\beta}_{\mathbf{1,0}}, \tilde{\boldsymbol{\beta}}_\mathbf{2})] \sim \chi^2_{q,1-\alpha}.$$

2.2.1.3 Estatística de Pearson e Função desvio

O objetivo da seleção de variáveis é encontrar um modelo intermediário, entre uum modelo muito complicado e um modelo muito simples, que consiga descrever corretamento os dados.

Por meio da função desvio ou *deviance* proposta por Nelder e Wedderburn (1972) é possível fundamentar a permanência de um parâmetro e, a falta de ajuste causada por sua omissão, assim como a adequabilidade do modelo (MCCULLAGH; NELDER, 1989; COLLETT, 2002; DOBSON, 1990). A função desvio ou *deviance* é definida a partir de

$$S_p = 2(\hat{l}_n - \hat{l}_p),$$

sendo $\hat{l}_n$ e $\hat{l}_p$ os máximos do logaritmo da função de verossimilhança para os modelos saturados(número de parâmetros igual ao número de observações), e sob pequisa respectivamente. Sendo $\tilde{\theta}_i = q(y_i)$ e $\hat{\theta}_i = q(\hat{\mu}_i)$ as EMVs do parâmetros canônico sob os modelos saturado e corrente, então tem-se:

$$S_p = \phi^{-1}D_p = 2\phi^{-1}\sum_{i=1} n[y_i(\tilde{\theta}_i - \hat{\theta}_i) + b(\hat{\theta}_i) - b(\tilde{\theta}_i)], \tag{2.6}$$

em que S_p e D_p são denominados de desvio escalonado e desvio, respectivamente.

A verificação da discrepância do ajuste de um modelo também pode ser realizada, no caso assintótico pela estatística χ^2 de Pearson generalizada, dada por

$$\chi^2_p = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

sendo $V(\hat{\mu}_i)$ a função de variância estimada para a distribuição em estudo (MCCULLAGH; NELDER, 1989).

2.2.1.4 Análise de desvio e Seleção de Modelos

Um dos principais objetivos no ajuste de um MLG é determinar quantos e quais termos são necessários na estrutura linear, para que este consiga descrever os dados de forma razoável, em que seja possível substituir $\boldsymbol{y}$ por um conjunto de valores estimados $\hat{\boldsymbol{\mu}}$. A seleção das variáveis do modelos deve-se iniciar de maneira cuidadosa, realizando-se primeiramente uma análise univariada no caso de muitas variáveis explicativas. O teste da razão de verossimilhança, assim como o teste de qui-quadrado de Pearson podem ser utilizados para verificar a relação entre a covariável e a variável dependente. Após a analise univariada é possível selecionar variáveis para a análise multivariada (HOSMER; LEMESHOW, 2004).

A análise de desvio (ANODEV), generalização da análise de variância (ANOVA), utiliza modelos encaixados e diferenças entre *deviances* para analisar os efeitos das covariáveis, fatores e interações, permitindo selecionar assim o melhor modelo para o conjunto de dados.

Comparando-se dois modelos M_q e M_p $(M_q \subset M_q, q < p)$ encaixados, com q e p parâmetros respectivamente, a estatística $D_q - D_p$ com $(p - q)$ graus de liberdade é uma medida da variação dos dados que pode ser explicada pelos termos que estão em M_p e não estão M_q (PAULA, 2013; DEMÉTRIO; CORDEIRO; MORAL, 2014). Assintoticamente, para ϕ conhecido, temos a estatística semelhante a da razão de verossimilhanças

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

Outra forma de realizar a seleção dos modelos é por meio de algoritmos de seleção de variáveis como o *stepwise*, muito utilizado em modelos de regressão linear. Dentre estes métodos estão o $Forward$ que acrescenta termos até a última variável adicionada não melhorar o ajuste, o $Backward$ que começa com um modelo completo e sequencialmente remove os termos, e o $Stepwise$ que utilizada os dois métodos anteriores (AGRESTI, 2002; PAULA, 2013; DEMÉTRIO; CORDEIRO; MORAL, 2014). Estes métodos podem ser aplicados utilizando como critério a $Deviance$ ou outros critérios como o Critério de Informação de Akaike (AIC) (AKAIKE, 1974) e o Critério Bayesiano de Schwarz (BIC) (SCHWARZ, 1978), dados em MLG por:

$$AIC = S_p + 2p - 2\hat{l}_n$$

e

$$BIC = S_p + p\log(n) - 2\hat{l}_n,$$

sendo l_n o máximo do logaritmo da função de verossimilhança, p o número de parâmetros do modelo e n o número de observações da amostra. O modelo com menor valor destes critério é que mais se aproxima do modelo verdadeiro.

2.2.2 Variável aleatória multinomial

A distribuição multinomial, é uma generalização da distribuição binomial para variável politômica. Seja o vetor de respostas denotado por $\mathbf{Y}$, em que $\mathbf{Y} = (Y_1, \ldots, Y_J)^T$, com J categorias de resposta $(j = 1, \ldots, J)$, em uma amostra aleatória de tamanho n, então

$$P(Y_1 = y_1, \ldots, Y_J = y_J; n, \pi) = \binom{n}{\boldsymbol{y}} \pi_1^{y_1} \ldots \pi_J^{y_J}$$
$$= \frac{n!}{y_1! \ldots y_J!} \pi_1^{y_1} \ldots \pi_J^{y_J} = \frac{n!}{\prod_{j=1}^{J} y_j!} \prod_{j=1}^{J} \pi_j^{y_j},$$

com $\pi = (\pi_1, \ldots, \pi_J)$ são as probabilidades de pertencer em cada uma das categorias, tais que $\sum_{j=1}^{J} \pi_j = 1$ (FREEMAN, 1987; AGRESTI, 2007).

Aplicando o logaritmo, temos que:

$$l(n, \pi) = \log\left(\frac{n!}{\prod_{j=1}^{J} y_j!}\right) + \sum_{j=1}^{J} y_j \log \pi_j$$
$$= \sum_{j=1}^{J} y_j \log \pi_j + \log n! - \sum_{j=1}^{J} \log y_j!.$$

Portanto, essa distribuição está na família exponencial multiparamétrica.

2.3 Modelos para dados categorizados ordinais

Uma extensão dos MLGs para o caso multivariado são os modelos para dados categorizados ordinais, sendo utilizados quando se tem interesse no comportamento de uma variável resposta ordinal em relação às variáveis explanatórias do estudo, que constituirão o componente sistemático do modelo. Nesta seção, são apresentados os modelos mais referenciados na literatura para este tipo de variável.

2.3.1 Modelos de Logitos Cumulativos

Variáveis respostas em escala ordinal são as que frequentemente mais aparecem em diversos tipos de estudos, porém a escolha das categorias de respostas são subjetivas e é essencial que as conclusões obtidas não sejam influenciadas pelo número de categorias de respostas. Por outro lado, uma característica importante no caso ordinal é a possibilidade de trabalhar com probabilidades acumuladas.

Esta importante consideração leva aos modelos mais utilizados para dados categorizados ordinais, o qual utiliza logitos de probabilidades de respostas acumuladas, chamados logitos cumulativos. A probabilidade acumulada de Y é a probabilidade de que Y seja igual ou inferior a um determinado ponto, isto é

$$P(Y \leq j|\mathbf{x}) = \gamma_j(\mathbf{x}) = \pi_1(\mathbf{x}) + ... + \pi_j(\mathbf{x}), \quad j = 1, ..., J,$$

em que $\pi_1(\mathbf{x}) = P(Y = 1|\mathbf{x}), \pi_2(\mathbf{x}) = P(Y = 2|\mathbf{x}), ..., \pi_j(\mathbf{x}) = P(Y = j|\mathbf{x})$.

Como as classes são ordenadas, as probabilidades acumuladas refletem a ordem natural com $P(Y \leq 1|\mathbf{x}) \leq P(Y \leq 2|\mathbf{x}) \leq ... \leq P(Y \leq J-1|\mathbf{x})$, não englobando a última classe pois $P(Y \leq J|\mathbf{x}) = 1$.

Os logitos para as probabilidades acumuladas são definidos como

$$\text{logito}[P(Y \leq j|\mathbf{x})] = \log\left[\frac{P(Y \leq j|\mathbf{x})}{1 - P(Y \leq j|\mathbf{x})}\right].$$

Quando a variável resposta apresenta número de categoria maior do que 2, $J > 2$, Tem-se $J - 1$ logitos que podem ser expressos pelo modelo

$$\text{logito}[P(Y \leq j|\mathbf{x})] = \alpha_j + \sum_{k=1}^{p} \beta_{jk} x_k = \alpha_j + \boldsymbol{\beta_j'}\mathbf{x}, \qquad j = 1, .., J-1,$$

em que p é o número de covariáveis, $\boldsymbol{\beta_j} = (\beta_{j1}, \ldots, \beta_{jp})'$ é o vetor de parâmetros e $\mathbf{x} = (x_1, \ldots, x_p)'$ representa o conjunto de valores das p-covariáveis no modelo.

A partir destes logitos é possível obter as razões de chances de acordo com o modelo selecionado, sendo esta uma medida de associação entre uma variável explanatória e a variável resposta. No caso ordinal ela possui diversas formas, local, local-global e global. A forma local é utilizada nos Modelos log-lineares, a forma global é utilizada em Modelos de Razão de Chances Globais e a forma local-global é utilizada nos modelos em discussão nesta seção, os Modelos Logitos (AGRESTI, 2002; TUTZ; SCHOLZ, 2003).

A diferença entre dois logitos cumulativos constitui a razão de chances denominado Local-Global:

$$\frac{P(Y \leq j|\mathbf{x_2})/P(Y > j|\mathbf{x_2})}{P(Y \leq j|\mathbf{x_1})/P(Y > j|\mathbf{x_1})}. \tag{2.7}$$

Como uma simplificação destes modelos, afim de poder comparar as respostas entre duas subpopulações e analisar simultaneamente todos os logitos cumulativos, McCullagh(1980) propôs os Modelos de Chances Proporcionais.

2.3.1.1 Modelos de chances proporcionais

Os modelos mais simples nesta classe envolvem regressões paralelas na escala escolhida e assumem proporções equivalentes a supor $\boldsymbol{\beta}_j = \boldsymbol{\beta}$ para todo j, o que simplifica o modelo.

Quando utiliza-se o logito como função de ligação, temos os Modelos de chances proporcionais (MCP):

$$\text{logit}[P(Y \leq j|\mathbf{x})] = \alpha_j + \sum_{k=1}^{p} \beta_k x_k = \alpha_j + \boldsymbol{\beta}'\mathbf{x}, \qquad j = 1, .., J-1, \tag{2.8}$$

sendo os efeitos de x idênticos para todos os $J-1$ logitos cumulativos, ou seja o modelo assume que os efeitos das variáveis independentes sobre o logito são idênticos para todas as classes da variável dependente, e que a resposta observada em cada classe se desloca apenas em função de α_j (MCCULLAGH; NELDER, 1989; AGRESTI, 2002; AGRESTI, 2007). Por este motivo o nome de chances proporcionais, pois assume haver uma idêntica razão de chances nos $J-1$ pontos de corte, ou suposição de regressão paralela (Figura 2.1).

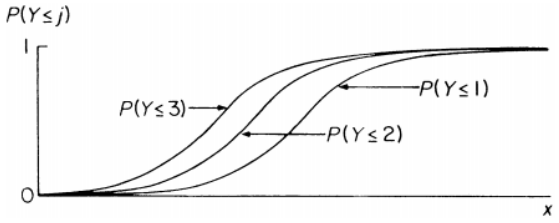

Figura 2.1 – Probabilidades acumuladas nos Modelos de Chances Proporcionais

Fonte: Agresti, 2007, p.181

Uma maneira de verificar a suposição de proporcionalidade é assumir o Modelo Logito Cumulativo

$$\text{logito}[P(Y \leq j|\mathbf{x})] = \alpha_j + \boldsymbol{\beta}_j'\mathbf{x}, \tag{2.9}$$

e aplicar o teste da razão de verossimilhanças (TRV) para a seguinte hipótese:

$$\begin{cases} H_0 : \boldsymbol{\beta_j} = \boldsymbol{\beta} \\ H_1 : \boldsymbol{\beta_j} \neq \boldsymbol{\beta} \qquad j = 1, \ldots, J-1. \end{cases}$$

A estatística deste teste é

$$\Lambda = -2\mathbf{log}\left[\frac{L_{H_0}}{L_{H_1}}\right], \tag{2.10}$$

em que L_{H_0} é o logaritmo da função de verossimilhança sob a hipótese H_0, que é igual ao MCP, e L_{H_1} representa o logaritmo da função de verossimilhança sob a hipótese H_1, ou seja, um modelo geral de logito cumulativo. Se $\Lambda < \chi^2_{(m,95\%)}$, em que m é o grau de liberdade correspondem a diferença do número de parâmetros dos modelos sob H_1 e H_0, a hipótese H_0 não é rejeitada.

A pressuposição de chances proporcionais pode ser verificada de maneira global, contendo todas as covariáveis no modelo e individual, covariável por covariável. Outra maneira é

utilizando o teste escore (RAO, 1948), pela maximização do logaritmo da função de verossimilhança sob a hipótese nula de proporcionalidade e tendo distribuição assintótica qui-quadrado com $p(J-2)$ graus de liberdade, sendo p o número de parâmetro no modelo e J o número de classes da variável resposta (STIGER; BARNHART; WILLIAMSON, 1999).

Não rejeitada esta pressuposição, os parâmetros do modelo apresentado na eq. (2.8) podem ser estimados pelo método da máxima verossimilhança, utilizando o processo iterativo de Newton Raphson. O teste para a significância dos parâmetros estimados do modelo pode ser realizado utilizando a estatística da razão de verossimilhanças apresentada na eq. (2.5) ou pela estatística de Wald que pode ser obtida por

$$z^2 = \left(\frac{\hat{\beta}}{EP}\right)^2,$$

em que z tem aproximadamente uma distribuição normal padrão, z^2 tem aproximadamente uma distribuição qui-quadrado com 1 g.l e EP é o erro padrão de $\hat{\beta}$. O intervalo de 95% de confiança de Wald para β será

$$\hat{\beta} \pm 1,96(EP). \tag{2.11}$$

Uma vez estimados os parâmetro do modelo é possível estimar as probabilidades acumuladas dadas por:

$$\hat{P}(Y \leq j|\mathbf{x}) = \hat{\gamma}_j(\mathbf{x}) = \frac{\exp(\hat{\alpha}_j + \hat{\boldsymbol{\beta}}'\mathbf{x})}{1 + \exp(\hat{\alpha}_j + \hat{\boldsymbol{\beta}}'\mathbf{x})}$$

Sob o modelo (2.9) as probabilidades marginais associada a $j-$ésima categoria são obtidas por:

$$\begin{cases} \hat{\pi}_1(\mathbf{x}) = \hat{\gamma}_1(\mathbf{x}) \\ \hat{\pi}_2(\mathbf{x}) = \hat{\gamma}_2(\mathbf{x}) - \hat{\gamma}_1(\mathbf{x}) \\ \hat{\pi}_3(\mathbf{x}) = \hat{\gamma}_3(\mathbf{x}) - \hat{\gamma}_2(\mathbf{x}) \\ \dots \\ \hat{\pi}_j(\mathbf{x}) = 1 - \hat{\gamma}_{j-1}(\mathbf{x}). \end{cases}$$

O logaritmo da razão de chances apresentada na eq. (2.7) é a diferença entre os logitos cumulativos em dois valores de $\mathbf{x}$, sendo para os MCPs igual a $\boldsymbol{\beta}'(\mathbf{x}_2 - \mathbf{x}_1)$, proporcional à distância entre eles. Se considerarmos um modelo com apenas uma covaríavel, sempre que $x_2 - x_1 = 1$ as chances de resposta sob alguma categoria é multiplicada por $\dfrac{\exp(\alpha_j + \beta)}{\exp(\alpha_j)} = \exp(\beta)$ para cada aumento de uma unidade em x (STOKES; DAVIS; KOCH, 2001; AGRESTI,

2002). É possível também encontrar um intervalo de confiança de Wald para esta razão de chances, sendo obtido utilizando o intervalo de confiança estimado para os parâmetros do modelo (2.11) este intervalo de 95% de confiança será dado por:

$$\exp(\hat{\beta} \pm 1,96(EP)).$$

McCullagh (1980) apresentou além do Modelo de Chances Proporcionais, diferentes funções de ligações geralmente utilizadas para dados binários, como a probito ($\phi^{-1}[P(Y \leq j|\mathbf{x})]$), a log-log (log{-log$[P(Y \leq j|\mathbf{x})]$}), e o completamento log-log. Quando se utiliza como ligação o complemento log-log este modelo passa a ser denominado Modelo de Risco Proporcional dado por:

$$\log[-\log\{1 - P(Y \leq j|\mathbf{x})\}] = \alpha_j + \boldsymbol{\beta'}\mathbf{x}, \qquad j = 1, .., J-1.$$

Uma motivação para estas probabilidades é assumir que existe uma variável latente continua subjacente Y^* que não é possível medir diretamente, e que Y resulta do corte da variável latente em J classes ordinais, mutuamente exclusivas. Esta variável latente é utilizada somente para facilitar a interpretação e os cálculos em modelos cujos rótulos das categorias da variável resposta é subjacente (ANDERSON; PHILIPS, 1981; LIU; AGRESTI, 2005). Agresti (2007) cita que uma interpretação neste caso seria verificar se é plausível um modelo de regressão ordinal, com seu respectivo preditor linear, descrever bem os efeitos de uma variável latente subjetiva, com isto seria correto utilizar um MCP.

2.3.1.2 Modelo de chances proporcionais parciais

Caso o pressuposto de chances proporcionais seja rejeitado no teste global, utiliza-se os testes individuais para verificar qual ou quais covariáveis estão violando esta suposição e pode-se utilizar os Modelos de Chances Proporcionais Parciais (MCPP) (LEMOS *et al.*, 2015) como extensão dos MCP. Introduzido por Peterson e Harrel (1990), existem dois tipos de MCPP, o modelo sem e com restrição.

O modelo "sem restrição"é o de chances proporcionais, no entanto, para as variáveis em que esta suposição não é satisfeita, há um incremento de um coeficiente (ϱ_i), que é o efeito de não proporcionalidade associado a cada $j-$ésimo logito cumulativo, ajustado pelas demais variáveis explanatórias (LALL *et al.*, 2002; ABREU; SIQUEIRA; CAIAFFAI, 2009). Para este modelo são estimados $J-1$ interceptos, p coeficientes $\beta's$, considerando p variáveis preditoras, que são independentes das categorias comparadas e $q(J-1)$ parâmetros (ϱ_j), sendo q as covariáveis em que a propriedade de chances proporcionais não foi válida

$$\text{logito}[P(Y \leq j|\mathbf{x})] = \alpha_j + \boldsymbol{\beta'}\mathbf{x} + \boldsymbol{\varrho_j'}\tilde{\mathbf{x}}, \qquad j = 1, .., J-1,$$

em que $\tilde{\mathbf{x}}$ é vetor que contém $q(q \leq p)$ variáveis preditoras do modelo para quais o pressuposto de proporcionalidade não é satisfeito e (ϱ_j) é um vetor de q parâmetros. Se os parâmetros $\varrho_j' s$ são nulos, o modelo se reduz ao modelo de chances proporcionais.

Algumas restrições podem ser incorporadas ao modelo para explicar as tendências que frequentemente ocorrem quando uma covariável e a variável resposta não são proporcionais. Incorporando estas restrições o modelo é chamado de Modelos de Chances Proporcionais parciais restritos (PETERSON; HARRELL, 1990; TUTZ; SCHOLZ, 2003).

2.3.2 Modelo Logito de Razão Contínua

Este modelo foi proposto por Feinberg (1980) e é utilizado em aplicações em que as categorias de resposta são sequenciais, ou seja as categorias são ordenadas de tal forma que possam ser alcançadas apenas sucessivamente, por exemplo sobrevivência em vários períodos de idade (ANANTH; KLEINBAUM, 1997; AGRESTI, 2007). O logito neste modelo é dado por:

$$\log\left[\frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x}) + \ldots + \pi_J(\mathbf{x})}\right] = \alpha_j + \boldsymbol{\beta}_{\boldsymbol{j}}'\mathbf{x}, \quad j = 1, ..., J-1.$$

Este modelo compara a probabilidade de uma resposta igual a uma determinada categoria $Y = j$ com a probabilidade de resposta no agrupamento de categorias de respostas de ordem mais elevadas $Y > j$, que podem ser influenciadas pela direção escolhida para modelar a variável resposta, ou seja depende da codificação. Uma de suas vantagens é que eles podem ser ajustados por meio de j modelos de regressão logística.

2.3.3 Modelo Logito de Categorias Adjacentes

Os Modelos Logitos de Categorias Ajacentes (MLCA) são muito utilizados quando se deseja comparar cada categoria de resposta com uma categoria de interesse. Estes utilizam como logito a razão das frequências de uma categoria de resposta em relação à categoria adjacente. Este logito é definido por

$$L_j = \log\left[\frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})}\right], \quad j = 1, ..., J-1.$$

Sendo estes logitos equivalentes aos logitos de categoria *baseline* e podem ser expressos também por:

$$L_j^* = \log\left[\frac{\pi_j(\mathbf{x})}{\pi_J(\mathbf{x})}\right] = \log\left[\frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})}\right] + \log\left[\frac{\pi_{j+1}(\mathbf{x})}{\pi_{j+2}(\mathbf{x})}\right] + \ldots + \log\left[\frac{\pi_{J-1}(\mathbf{x})}{\pi_J(\mathbf{x})}\right].$$

O modelo de categorias adjacentes e o respectivo modelo de categorias *baseline* são definidos por:

$$\log\left[\frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})}\right] = \alpha_j + \boldsymbol{\beta}'\mathbf{x}, \quad j = 1, ..., J-1$$

e

$$\log\left[\frac{\pi_j(\mathbf{x})}{\pi_J(\mathbf{x})}\right] = \sum_{k=j}^{J-1}\alpha_k + \boldsymbol{\beta}'(J-j)\mathbf{x}, \quad j = 1, ..., J-1.$$

Spyrides-Cunha (1998) também apresenta uma revisão destas classes de modelos.

2.3.4 Diagnóstico

Selecionado o modelo é necessário verificar a qualidade deste ajuste, ou seja quão próximo os valores preditos se encontram dos correspondentes valores observados. Com a *deviance* residual é possível verificar a qualidade do ajuste, sendo que grandes valores indicam problemas com o modelo.

Utilizando a deviance apresentada na eq. (2.6), segundo Agresti (2007) esta qualidade de ajuste em tabelas de contingências, também conhecida como qui-quadrado *deviance*, pode ser simplificada para

$$G^2 = 2[\sum_{ij} n_{ij}\log n_{ij} - \sum_{ij} n_{ij}\log e_{ij}] = 2\sum_i\sum_j n_{ij}\log\left(\frac{n_{ij}}{e_{ij}}\right), \tag{2.12}$$

que segue uma distribuição χ^2 com $((r-1)(c-1)-q)$ graus de liberdade, sendo r o número de linhas; c o número de colunas da tabela e q o número de parâmetros associados as variáveis do modelo (FREEMAN, 1987; GIOLO, 2012; SPYRIDES-CUNHA, 1998).

2.3.4.1 Resíduos

Alguns dos resíduos mais utilizados em MLG também podem ser utilizados para respostas ordinais, porém com algumas adaptações e modificações de requisitos. Li e Shepherd (2012), sugeriram um novo resíduo, baseado na soma dos resíduos cumulativos, satisfazendo estas propriedades e preservando a ordem sem atribuir pontuações arbitrárias às categorias.

Considerando um conjunto de J categorias ordenadas , $j = 1, \ldots, J$. Para a categoria y em j e a distribuição F sobre j o resíduo é definido como

$$r(y, F) = E\{\text{sinal}(y, Y)\} = P(y > Y) - P(y < Y),$$

em que Y é a variável aleatória com distribuição F, e sinal(a, b) é -1,0 e 1 para $a < b$, $a = b$ e $a > b$. Dados os valores (y_i, x_i) e o modelo ajustado com o vetor de parâmetros estimados $\hat{\boldsymbol{\theta}}$, o resíduo para a i-ésima observação será $\hat{r}_i = r(y_i, F_{x_i,\hat{\boldsymbol{\theta}}})$.

Estes resíduos são úteis na verificação de possíveis faltas de ajustes para covariáveis quantitativas no modelo. A partir de um gráfico dos $\hat{\boldsymbol{r}}$ versus uma das variáveis explanatórias

incluídas no modelo pode-se verificar se ainda existe uma relação sistemática entre os resíduos e esta variável, necessitando na adição de um termo de maior grau no modelo.

2.4 Modelos Lineares Generalizados Mistos

Muito frequentemente os indivíduos podem ser classificados em um conjunto de sub-populações, tais como as características demográficas ou, de acordo com o experimento, serem observadas medidas repetidas da variável resposta no mesmo indivíduo, podendo estas, serem longitudinais ou não. Neste caso, os modelos apresentados até agora não incorporam esta dependência entre observações e uma das possibilidades é a utilização das equações de estimações generalizadas (LIANG; ZEGER, 1986; LIPSITZ; KIM; ZHAO, 1994), que utiliza uma matriz de correlação para modelar a matriz de covariâncias intra-unidades amostrais. Outra opção é a utilização dos Modelos Lineares Generalizados Mistos (MLGM) que é abordada nesta seção.

Modelos que possuem efeitos fixos e aleatórios, além do erro, são denominados modelos mistos. Estes modelos são principalmente utilizados para descrever relações entre a variável resposta e as covariáveis quando existe dependência entre observações, que podem surgir, segundo Pinheiro e Bates (2006) geralmente em medidas repetidas, no tempo ou no espaço, dados multiníveis e delineamentos em blocos. Estes modelos conseguem flexibilizar uma estrutura para a matriz de variâncias-covariâncias, introduzindo assim, a possível estrutura de correlação presente nos dados (PINHEIRO; BATES, 2006; GALWEY, 2006).

Segundo Brien (2015), os efeitos de unidades individuais, por exemplo, animais e pessoas estão previstos para ocorrer de forma aleatória e estes efeitos podem seguir uma distribuição normal, por esse motivo um termo de variação deve ser incluído no modelo, e neste caso também sugere-se a utilização dos modelos mistos.

O modelo linear misto é definido como

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\mathbf{u} + \boldsymbol{\epsilon},$$

em que $\boldsymbol{Y}$ é o vetor de observações; $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos; $\mathbf{u}$ é o vetor de variáveis aleatórias não observadas; $\boldsymbol{X}$ e $\boldsymbol{Z}$ são as matrizes de delineamento para os efeitos fixos e aleatórios e $\boldsymbol{\epsilon}$ o vetor de erros aleatórios. É usualmente assumido que $\mathbf{U}$ e $\boldsymbol{\epsilon}$ são independentes e normalmente distribuídos com vetor de médias $\mathbf{0}$, tendo $\mathbf{U} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{G})$ e $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ (LAIRD; WARE, 1982).

Para os MLGs é necessária a inclusão de um efeito aleatório $\mathbf{u}$ no preditor linear, modelando-se uma função de η do vetor de médias condicionais $\boldsymbol{\mu} = \mathbf{E}[\boldsymbol{Y}|\mathbf{u}]$ por meio de um preditor (MCCULLOCH; NEUHAUS, 2001; SUESSE; LIU, 2013). Neste contexto os elementos de $\boldsymbol{Y}$, correspondente ao vetor de valores da variável resposta de elementos condicionalmente independentes , condicionados nos efeitos aleatório, $\mathbf{u}$, são independentes e identicamente distri-

buídos, com qualquer distribuição pertencente a família exponencial. Assim, os MLGMs podem ser definidos por:

$$g(\mu_i) = \mathbf{x_i'}\boldsymbol{\beta} + \mathbf{z_i'}\mathbf{u}, \tag{2.13}$$

em que

$$\begin{aligned} \mathrm{E}[Y_i|\mathbf{u}] &= \mu_i \\ \mathrm{Var}[Y_i|\mathbf{u}] &= \phi V(\mu_i), \end{aligned}$$

sendo $g(.)$ a função de ligação; $\mathbf{x_i}$ a i-ésima linha da matriz de delineamento associada aos efeitos fixos; $\boldsymbol{\beta}$ o vetor de parâmetros de feitos fixo; $\mathbf{z_i}$ a i-ésima linha da matriz associada aos efeitos aleatório; $\mathbf{u}$ o vetor de efeitos aleatórios, independentes e com distribuição $f_{\mathbf{U}}(\mathbf{u})$, usualmente N$(\mathbf{0}, \mathbf{G})$, em que $\mathbf{0}$ é o vetor de zeros e $\mathbf{G}$. (MCCULLOCH; NEUHAUS, 2001; HARTZEL; AGRESTI; CAFFO, 2001).

Pelas propriedades da esperança condicional temos que:

$$\mathrm{E}[Y_i] = \mathrm{E}[\mathrm{E}(Y_i|\mathbf{u})] = \mathrm{E}[g^{-1}(\mathbf{x_i'}\boldsymbol{\beta} + \mathbf{z_i'}\mathbf{u})]$$

$$\mathrm{Var}[Y_i] = \mathrm{Var}[\mathrm{E}(Y_i|\mathbf{u})] + \mathrm{E}[\mathrm{Var}(Y_i|\mathbf{u})]$$

$$= \mathrm{Var}[g^{-1}(\mathbf{x_i'}\boldsymbol{\beta} + \mathbf{z_i'}\mathbf{u})] + \mathrm{E}\{\phi^{-1}V[g^{-1}(\mathbf{x_i'}\boldsymbol{\beta} + \mathbf{z_i'}\mathbf{u})]\}.$$

Os efeitos aleatórios introduzem uma correlação entre observações que tenham algum efeito em comum. Assumindo-se independência condicional dos elementos de $\mathbf{y}$, as covariâncias marginais podem ser definidas por:

$$\mathrm{Cov}[Y_i, Y_{i'}] = \mathrm{Cov}[\mathrm{E}(Y_i|\mathbf{u}), \mathrm{E}(Y_{i'}|\mathbf{u})] + \mathrm{E}[\mathrm{Cov}(Y_i, Y_{i'}|\mathbf{u})]$$

$$= \mathrm{Cov}[g^{-1}(\mathbf{x_i'}\boldsymbol{\beta} + \mathbf{z_i'}\mathbf{u}), g^{-1}(\mathbf{x_{i'}'}\boldsymbol{\beta} + \mathbf{z_{i'}'}\mathbf{u})] + \mathrm{E}(0).$$

Por meio do teste da razão de verossimilhança também é possível verificar se o efeito de $\mathbf{u}$ é significativo. Para este teste, consideramos como hipótese nula $H_0 : \sigma_u^2 = 0$, em que σ_u^2 é a variância de $\mathbf{u}$. Se este teste não foi rejeitado a inclusão do efeito aleatório não é necessária para o modelo. Este teste possui distruibuição assintótica χ^2 com 1 grau de liberdade, porém sendo conservador ao testar na fronteira do espaço paramétrico, em que é aconselhavél considerar uma mistura de qui-quadrados $(\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2)$ (MCCULLOCH; NEUHAUS, 2001; HARTZEL; AGRESTI; CAFFO, 2001; SILVA, 2013).

2.4.1 Estimação por máxima verossimilhança

Hedeker (2005) cita que a estimação dos parâmetros dos MLGMs envolvem caracteristicamente o método da máxima verossimilhança. Chen, Cai e Zhou (2004) referenciam outras abordagens citadas na literatura, como o método de aproximação Bayesiana (STIRATELLI; LAIRD; WARE, 1984), o método de estimação de máxima verossimilhança hierárquica (LEE; NELDER, 1996) e a abordagem de pseudo verossimilhança (WOLFINGER; O'CONNELL, 1993). Uma outra alternativa é o método de quase-verossimilhança penalizada (PQL), que aplica uma penalidade na função de quase-verossimilhança (BRESLOW; CLAYTON, 1993), sendo que este método não requer integração numérica e é computacionalmente mais viável para grandes bancos de dados (AGRESTI, 2002).

Sendo a distribuição dos efeitos aleatórios $\mathbf{u} \sim f_{\mathbf{U}}(\mathbf{u})$, de acordo com a eq. (2.13) a função de verossimilhança marginal de Y é dada por

$$L = \int \prod_i f_{Y_i|\mathbf{u}}(y_i|\mathbf{u}) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u}. \tag{2.14}$$

Contudo a resolução desta integral não possui forma analítica e fechada quando $Y_i|\mathbf{u}$ não segue uma distribuição normal, sendo necessária a utilização de métodos de aproximação numérica (PINHEIRO; BATES, 1995; LIU; AGRESTI, 2005).

De acordo com McCulloch (2001) a estimação dos parâmetros correspondentes aos efeitos fixos por meio das equações de verossimilhança é numericamente difícil de calcular. Porém a eq. (2.14) pode ser simplificada por

$$l = \log \int f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) f_{\mathbf{u}}(\mathbf{u}) d\mathbf{u} = f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} = \log f_{\mathbf{Y}}(\mathbf{y}).$$

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \left[\log \int f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) f_{\mathbf{u}}(\mathbf{u}) d\mathbf{u} \right]$$

$$\frac{1}{f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) f_{\mathbf{u}}(\mathbf{u}) d\mathbf{u}} \left[\frac{\partial}{\partial \boldsymbol{\beta}} \int f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) f_{\mathbf{u}}(\mathbf{u}) d\mathbf{u} \right]$$

Como $f_{\mathbf{U}}(\mathbf{u})$ não envolve $\boldsymbol{\beta}$, podemos reescrever

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{f_{\mathbf{Y}}(\mathbf{y})} \int \left[\frac{\partial}{\partial \boldsymbol{\beta}} f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) \right] f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} \tag{2.15}$$

então

$$\frac{\partial}{\partial \boldsymbol{\beta}} f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) = \frac{1}{f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u})} \left[\frac{\partial}{\partial \boldsymbol{\beta}} f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) \right] f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u})$$

$$= \left[\frac{\partial}{\partial \boldsymbol{\beta}} \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u})\right] f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u})$$

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \int \left[\frac{\partial}{\partial \boldsymbol{\beta}} \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u})\right] \frac{f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) f_{\mathbf{u}}(\mathbf{u}) d\mathbf{u}}{f_{\mathbf{y}}(\mathbf{y})}$$

$$= \int \left[\frac{\partial}{\partial \boldsymbol{\beta}} \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u})\right] f_{\mathbf{U}|\mathbf{y}}(\mathbf{u}|\mathbf{y}) d\mathbf{u}.$$

Sabendo que

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \mathbf{X}'\mathbf{W}\boldsymbol{\partial}(\mathbf{y} - \boldsymbol{\mu}) \text{ com } \mathbf{W} = \text{diag}\{W_i\},$$

em que

$$\mathbf{W} = \text{diag}\{W_i\} = \text{diag}\left[a_i(\phi) V(\mu_i) \frac{\partial \eta_i}{\partial \mu_i}\right]^2 \text{ e } \Delta = \text{diag}\left\{\frac{\partial \eta_i}{\partial \mu_i}\right\},$$

portanto,

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \int \mathbf{X}'\mathbf{W} * \boldsymbol{\Delta}(\mathbf{y} - \boldsymbol{\mu}) f_{\mathbf{U}|\mathbf{y}}(\mathbf{u}|\mathbf{y}) d\mathbf{u} = \mathbf{X}'\mathbf{y}\mathrm{E}[\mathbf{W} * |\mathbf{y}] - \mathbf{X}'\mathrm{E}[\mathbf{W} * \boldsymbol{\mu}|\mathbf{y}],$$

em que $W^* = \text{diag}\left[a(\phi) V(\mu_i) \frac{\partial \eta}{\partial \mu}\right]^{-1}$, então,

$$\mathbf{X}'\mathbf{y}\mathrm{E}[\mathbf{W}^*|\mathbf{y}] = \mathbf{X}'\mathrm{E}[\mathbf{W}^*\boldsymbol{\mu}|\mathbf{y}].$$

Seja $\boldsymbol{\varphi}$ referente aos parâmetros de efeito aleatório, para a estimação destes efeitos, podemos encontrar

$$\frac{\partial l}{\partial \boldsymbol{\varphi}} = \int \frac{\partial \log f_{\mathbf{U}}(\mathbf{u})}{\partial \boldsymbol{\varphi}} f_{\mathbf{U}|\mathbf{y}}(\mathbf{u}|\mathbf{y}) \mathbf{du}$$

$$= \mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\varphi}} \log f_{\mathbf{U}}(\mathbf{u})|\mathbf{y}\right].$$

Segundo Agresti (1996) esta modelagem é mais simples para os modelos ordinais do que nominais, pois é frequentemente utilizado o mesmo efeito aleatório para todos os logitos. O logito cumulativo para o caso de MLGM é definido como

$$\text{logito}[P(Y_i) \leq j|\mathbf{u}] = \alpha_j + \mathbf{x}_\mathbf{i}'\boldsymbol{\beta} + \mathbf{z}_\mathbf{i}'\mathbf{u}, \qquad j = 1, \ldots J - 1.$$

Tendo como função de verossimilhança

$$L = \prod_{i=1}^{n} \int_{\mathbf{u}} \left[\prod_{j=1}^{J} (P(Y_i \leq j|\mathbf{u}) - P(Y_i \leq j-1|\mathbf{u}))^{y_{ij}} \right] f(\mathbf{u}) d\mathbf{u}. \tag{2.16}$$

Maximizando a eq. (2.16) e utilizando-se aproximações numéricas, obtém-se as estimativas dos parâmetros. Os métodos de aproximação numérica mais utilizados são o de Laplace ou de Quadratura gaussiana, porém não são eficientes computacionalmente se o número de efeitos aleatórios for grande (MCCULLOCH; NEUHAUS, 2001; SILVA, 2013).

Como já salientado, no ajuste deste modelo é de interesse a estimação dos parâmetros $\beta's$ e do parâmetro de dispersão σ^2. Um dos métodos numéricos utilizado na estimação por máxima verossimilhança é o de quadratura (MOLENBERGHS; VERBEKE, 2006). A integral descrita na eq. (2.16) tem dimensão que depende da estrutura do efeito aleatório e quando esta dimensão é pequena, métodos de integração numérica podem aproximar esta função de verossimilhança. O método de quadratura Gauss-Hermite pode ser utilizado quando se tem apenas um efeito aleatório e no caso de efeitos aleatórios normais é definido por

$$\int_{-\infty}^{\infty} f(x)w(x)dx \approx \sum_{l=1}^{Q} w_l f(z_l),$$

em que $f(x)$ é uma função conhecida, $w(x)$ uma função densidade de probabilidade, w_l são os pesos e z_l os pontos de quadratura, sendo Q a ordem da aproximação que melhora, a medida que o número de pontos da quadratura aumenta (LIU; PIERCE, 1994; AGRESTI, 2002).

3 MATERIAL E MÉTODOS

3.1 Material

O tomate é um produto muito consumido no Brasil, tanto na forma *in natura* quanto na forma de produtos processados, como polpa, extrato, *ketchup*, molhos, dentre outros. A variedade Italiano é altamente produzida, destacando-se pelo sabor e pela polpa espessa. O Sweet Grape, mini tomate, trazido para o Brasil em 2000, também apresenta uma boa aceitação entre os consumidores brasileiros, sendo muito resistente a doenças e de alto valor nutricional (LORO, 2015).

Pelo fato de ser um produto altamente perecível, o processo de desidratação vem sendo uma das alternativas para as altas perdas do produto (ABREU *et al.*, 2011). Porém este processo para a variedade Sweet Grape ainda é pouco estudado. Loro (2015), conduziu um experimento no Departamento de Agroindústria, Alimentos e Nutrição (LAN) da Escola Superior de Agricultura "Luiz de Queiroz"(ESALQ), propondo modificações nas técnicas de desidratação destas duas variedades. Amostras de tomates Italiano e Sweet Grape foram adquiridas e foi realizado um processo diferenciado de desidratação, considerando uma nova etapa, o congelamento prévio dos tomates, desenvolvendo um novo produto, preservando as características nutricionais e sensoriais do alimento.

Os resultados das análises laboratoriais mostraram que o Sweet Grape manteve a qualidade nutricional quando realizado este processo de desidratação. Outro interesse está na aceitação deste produto pelos consumidores (LORO, 2015).

Para este propósito foi realizada a análise sensorial afim de comparar a aceitação destes dois produtos e para este estudo foi selecionada como variável reposta o sabor. Foram utilizados 50 provadores não treinados, que apreciam o produto tomate desidratado, sendo 40 mulheres e 10 homens, representados por discentes, docentes e funcionários da ESALQ. A variável sexo foi aqui considerada como covariável, por poder exercer influência sobre a classificação do sabor atribuída por cada provador.

As amostras foram servidas na forma de fatias de tomate desidratado e cada provador provou duas amostras, sendo uma de cada variedade de tomate. No total 100 amostras foram degustadas. As categorias de resposta são as avaliações do sabor de acordo com a escala hedônica numérica de 9 pontos (IFT, 1981), conforme descrito na Tabela 3.1.

Tabela 3.1 – Escala hedônica utilizada na análise sensorial para definir a preferência pelo produto

Categorias de resposta
1 - Desgostei extremamente
2 - Desgostei muito
3 - Desgostei moderadamente
4 - Desgostei ligeiramente
5 - Indiferente
6 - Gostei ligeiramente
7 - Gostei moderadamente
8 - Gostei muito
9 - Gostei extremamente

3.2 Métodos

Seja a variável $\mathbf{Y}$ a classificação do sabor que cada provador atribuiu a cada uma das variedades do tomate, então $\mathbf{Y}$ tem distribuição multinomial conforme descrita na subseção 2.2.2. Como a variável possui uma escala ordinal, assumindo valores no conjunto {1,2,3,4,5,6,7,8,9}, a resposta do c-ésimo provador é denotada por $\mathbf{y}_c = (\mathrm{I}_{c1}, \ldots, \mathrm{I}_{c9})'$, sendo I_{cj} variáveis indicadoras, em que $\mathrm{I}_{cj} = 1$ quando provador escolheu a $j-$ésima classificação ou $\mathrm{I}_{cj} = 0$, caso contrário.

Como não observou-se resposta na categoria "1-Desgostei extremamente", para os dados apresentados neste estudo existem 8 categorias de resposta, sendo γ_{jlk} as probabilidades acumuladas da j-ésima categoria, referente a l-ésima variedade (Italiano, l=1; Sweet Grape, l=2) dos e ao $k-$ésimo sexo (Homem, k=1; Mulher k=2). Neste caso, tem-se 7 logitos, expressos por

$$\text{logito}(\gamma_{2lk}) = \log\left(\frac{\pi_{2lk}}{1-\pi_{2lk}}\right), \ldots, \text{logito}(\gamma_{8lk}) = \log\left(\frac{\pi_{2lk} + \ldots \pi_{8lk}}{1 - [\pi_{2lk}) + \ldots \pi_{8lk}]}\right).$$

Nesta análise sensorial os provadores experimentaram uma única vez as duas variedades de tomate, sendo que neste tipo de delineamento, as repetições serão as avaliações dadas por diferentes provadores. Neste caso, podem existir variações não medidas pelo modelo nas avaliações de cada diferente provador, como a preferência pessoal por determinado alimento. Para capturar todos os possíveis erros entre estas avaliações o provador deve ser colocado no modelo com o efeito de um bloco. Este efeito do bloco será aleatório pelo fato destas diferenças não serem sistemáticas e podendo variar de acordo com cada provador que participar do experimento. Neste contexto foram utilizados no ajuste do modelo os MLGMs apresentados na seção 2.4.

Assumindo-se como função de ligação o logito cumulativo, tendo o fator variedade com dois níveis, a covariável sexo, a interação entre variedade e sexo, e o efeito aleatório correspondente aos 50 provadores constituindo o preditor linear, o modelo completo pode ser

expresso como

$$\log\left[\frac{\gamma_{jlkc}}{1-\gamma_{jlkc}}\right] = \alpha_j + \beta_{lj} + \tau_{kj} + (\beta\tau)_{lkj} + u_c, \quad j = 2,\dots 8, \quad l,k = 1,2 \quad \text{e} \quad c = 1\dots 50, \tag{3.1}$$

sendo α_j o intercepto no j-ésimo logito , β_{lj} o efeito associado a l-ésima variedade de tomate no j-ésimo logito, τ_{kj} associado ao k-ésimo sexo no j-ésimo logito, u_c o efeito aleatório associado ao c-ésimo individuo em que $u_c \sim N(0, \sigma_u^2)$ e $(\beta\tau)_{lkj}$ o efeito associado à interação entre variedade e sexo. Segundo Paula (2013), a parametrização casela de referencia é padrão para os casos de MLGs e assumindo esta parametrização teremos como restrições $\beta_{j1} = 0$ e $\tau_{j1} = 0$.

De acordo com Giolo (2012), para a verificação da suposição de proporcionalidade descrita na subseção 2.3.1.1 utilizam-se apenas o modelo com efeitos principais, e caso esta suposição não seja rejeitada verifica-se a presença da interação no MCP e adiciona-se o efeito aleatório. O modelo de logito cumulativo utilizado para este teste é definido como

$$\log\left[\frac{\gamma_{jlk}}{1-\gamma_{jlk}}\right] = \alpha_j + \beta_{lj} + \tau_{kj}. \tag{3.2}$$

em que α_j o intercepto no j-ésimo logito , β_{jl} o efeito associado a l-ésima variedade de tomate no j-ésimo logito, τ_{kj} associado ao k-ésimo sexo no j-ésimo logito.

Tendo como estatística do teste a eq. (2.10), o modelo eq. (3.2) é comparado ao MCP

$$\log\left[\frac{\gamma_{jlk}}{1-\gamma_{jlk}}\right] = \alpha_j + \beta_l + \tau_k. \tag{3.3}$$

sendo α_j o intercepto no j-ésimo logito , β_l o efeito associado a l-ésima variedade de tomate, τ_k associado ao k-ésimo sexo.

Sob a suposição de chances proporcionais cujos os efeitos dos fatores são os mesmo para todos os logitos, os modelos testados a partir deste passo, serão todos MCP. Para a seleção das covariáveis no modelo utilizou-se o processo de natureza sequencial, eliminando variáveis até o modelo mais adequado. Nesta seleção foi utilizado como critério o teste da razão de verossimilhanças descrito na subseção 2.2.1.4 e os seguintes modelos:

Modelo 1

$$\log\left[\frac{\gamma_{jlkc}}{1-\gamma_{jlkc}}\right] = \alpha_j + \beta_l + \tau_k + (\beta\tau)_{lk} + u_c, \tag{3.4}$$

sendo α_j o intercepto no j-ésimo logito , β_l o efeito associado a l-ésima variedade de tomate, τ_k o efeito associado ao k-ésimo sexo, u_c o efeito aleatório associado ao c-ésimo individuo em que $u_c \sim N(0, \sigma_u^2)$ e $(\beta\tau)_{lk}$ o efeito associado a interação.

Modelo 2

$$\log\left[\frac{\gamma_{jlkc}}{1-\gamma_{jlkc}}\right] = \alpha_j + \beta_l + \tau_k + u_c, \tag{3.5}$$

sendo α_j o intercepto no j-ésimo logito , β_l o efeito associado a l-ésima variedade de tomate, τ_k o efeito associado ao k-ésimo sexo e u_c o efeito aleatório associado ao c-ésimo individuo.

Modelo 3

$$\log\left[\frac{\gamma_{jkc}}{1-\gamma_{jkc}}\right] = \alpha_j + \tau_k + u_c, \tag{3.6}$$

sendo α_j o intercepto no j-ésimo logito , τ_k o efeito associado ao k-ésimo sexo, u_c o efeito aleatório associado ao c-ésimo individuo.

Modelo 4

$$\log\left[\frac{\gamma_{jlc}}{1-\gamma_{jlc}}\right] = \alpha_j + \beta_l + u_c, \tag{3.7}$$

sendo α_j o intercepto no j-ésimo logito , β_l o efeito associado a l-ésima variedade de tomate, u_c o efeito aleatório associado ao c-ésimo individuo.

Modelo 5

$$\log\left[\frac{\gamma_{jc}}{1-\gamma_{jc}}\right] = \alpha_j + u_c, \tag{3.8}$$

sendo α_j o intercepto no j-ésimo logito , u_c o efeito aleatório associado ao c-ésimo individuo.

3.2.0.1 Recursos computacionais

A verificação da hipótese do modelo de chances proporcionais foi realizada utilizando a função **clm**(.) da biblioteca **ordinal** (CHRISTENSEN, 2011), sendo também verificada pela função PROC LOGISTIC do *software* SAS versão 9.3 (SAS, 2011) que utiliza o teste Escore.

Para o ajuste do modelo de *odds* proporcionais misto, comparação dos modelos encaixados, seleção dos modelos e qualidade do ajuste foi utilizada a função **olmm** (.) do pacote **vcrpart** (BUERGIN, 2015). Esta função estima os parâmetros fixos utilizando a máxima verossimilhança, com o método numérico de quadratura de Gauss-Hermite, e na otimização a matriz de informação esperada de Fisher. Este ajuste também pode ser obtido pela função **clmm**(.) da biblioteca **ordinal** (CHRISTENSEN, 2011) que utiliza como método numérico a quadratura de Gauss-Hermite adaptada e a matriz Hessiana no algoritmo de estimação, porém esta função não possui resíduos implementados. Os ajustes foram feitos no *software* R versão 3.2.1 (R Development Core Team, 2014).

4 RESULTADOS E DISCUSSÃO

4.1 Análise Descritiva

Inicialmente uma análise descritiva dos dados foi realizada para visualização da classificação recebida em função da variedade de tomate e do sexo do provador. A Tabela 4.1 apresenta as frequências observadas em cada umas das caselas e os totais marginais apresentam a quantidade de homens e mulheres que provaram as duas variedades, no qual observa-se que a maior parte, 80% dos provadores são do sexo feminino.

A Figura 4.1 apresenta as proporções de respostas para as duas variedades, de acordo com o sexo dos provadores. Este gráfico sinaliza um comportamento semelhante referente ao gosto de homens e mulheres, dado que ambos apreciam mais a variedade Sweet Grape, sendo que 83% das mulheres e 80% do homens classificaram o sabor desta variedade como "8-Gostei muito" ou "9-Gostei extremamente", enquanto para a variedade Italiano, houve mais avaliações "7-Gostei moderadamente". É possível concluir também que 20% da mulheres desapreciaram o tomate Italiano desidratado, avaliando-o com notas menores ou iguais a "4-Desgostei ligeiramente".

Tabela 4.1 – Frequências observadas para cada classificação, por variedade de tomate e sexo do avaliador

		Categorias								
Variedade	Sexo	2	3	4	5	6	7	8	9	Total
Italiano	Homem	0	0	1	0	2	3	3	1	10
Italiano	Mulher	2	1	5	2	8	10	6	6	40
Sweet Grape	Homem	0	0	0	0	1	1	3	5	10
Sweet Grape	Mulher	0	0	0	2	2	3	17	16	40

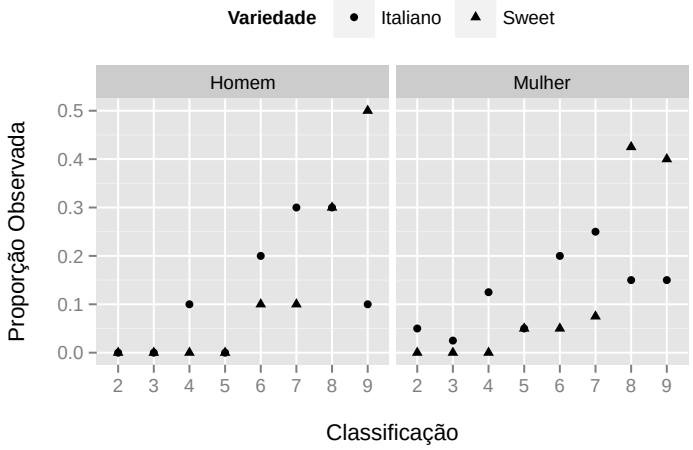

Figura 4.1 – Proporções observadas para cada classificação por sexo e variedade de tomate

4.2 Ajuste do Modelo

Como análise preliminar foi realizada a verificação da suposição de proporcionalidade pelo TRV conforme apresentado na eq. (2.10). Este teste compara o modelo logito cumulativo (eq 3.2) e o modelo de chances proporcionais (eq 3.3), utilizando como hipótese nula ($\boldsymbol{\beta_j} = \boldsymbol{\beta}$). Como estatística do TRV, utilizando o *software* R, obteve-se o valor de $12,943$, com 12 de liberdade, o que apresenta um nível descritivo $0,3732$. Neste caso, a hipótese nula não foi rejeitada, indicando que o MCP se ajustaria bem aos dados e pelo princípio da parcimônia os demais modelos ajustados também são MCPs.

Os modelos sequenciais, descritos na seção 3.2, foram ajustados e procedeu-se a uma seleção levando-se em conta os critérios descritos na subseção 2.2.1.4. O número de parâmetros de cada um dos modelos é apresentado na Tabela 4.2, sendo este obtido pela soma do número de interceptos, totalizando 7 em um total de 8 categorias de resposta, do parâmetro correspondente ao efeito aleatório e dos parâmetros estimados para as covariáveis do modelo.

Tabela 4.2 – Modelos sequenciais testados na seleção das covariáveis e número de parâmetro de cada modelo

Modelo	Número de parâmetros
1 : Efeito aleatório + Variedade + Sexo + Variedade : Sexo (3.4)	11
2 : Efeito aleatório + Variedade + Sexo (3.5)	10
3 : Efeito aleatório + Sexo (3.6)	9
4 : Efeito aleatório + Variedade (3.7)	9
5 : Efeito aleatório (3.8)	8

Os resultados dos testes dos modelos encaixados são apresentados na Tabela 4.3, em que pode-se observar, na comparação dos modelos 1 e 2 (nível descritivo $= 0,79$), que a interação não é significativa ao nível de significância de 5%. O efeito do sexo também não foi significativo, produzindo um nível descritivo$= 0,4019$ na comparação dos modelos 2 e 4. A comparação dos modelos 4 e 5 apresentou um nível descritivo$< 0,0001$, mostrando que a variedade é significativa no modelo. Com estes resultados, conclui-se que apenas o tipo do tomate está influenciando na classificação recebida ou no gosto dos provadores.

Tabela 4.3 – Graus de liberdade (g.l), estatística do teste, valor p e modelos selecionados nos testes da razão de verossimilhança para modelos encaixados

Teste	g.l	Razão de verossimilhança	valor p	Modelo selecionado
Modelo 1 versus Modelo 2	1	$0,071$	$0,7900$	Modelo 2
Modelo 2 versus Modelo 3	1	$25,512$	$< 0,0001$	Modelo 2
Modelo 2 versus Modelo 4	1	$0,703$	$0,4019$	Modelo 4
Modelo 4 versus Modelo 5	1	$25,820$	$< 0,0001$	Modelo 4

De acordo com a Tabela 4.3 pode-se observar que o modelo mais adequado é o de número 4, conforme descrito na eq. (3.7). A Tabela 4.4 mostra as estimativas, erros padrões e

significância dos interceptos e do parâmetro estimado para o efeito da variedade deste modelo selecionado. A significância dos interceptos não deve ser interpretada e é colocada aqui apenas como complemento da tabela.

Tabela 4.4 – Estimativas dos parâmetros dos efeitos fixos do modelo 4 selecionado, com os respectivos erros padrões, estatística do teste e valor p

Parâmetro	Estimativa	Erro padrão	z_{calc}	valor p
α_2 : Intercepto 2	$-3,319$	$0,754$	$-4,400$	$< 0,0001$
α_3 : Intercepto 3	$-2,895$	$0,657$	$-4,408$	$< 0,0001$
α_4 : Intercepto 4	$-1,680$	$0,393$	$-4,268$	$< 0,0001$
α_5 : Intercepto 5	$-1,239$	$0,341$	$-3,635$	$0,0001$
α_6 : Intercepto 6	$-0,298$	$0,273$	$-1,087$	$0,2770$
α_7 : Intercepto 7	$0,646$	$0,319$	$2,026$	$0,0427$
α_8 : Intercepto 8	$2,158$	$0,436$	$4,953$	$< 0,0001$
β : Sweet Grape	$-1,927$	$0,455$	$-4,758$	$< 0,0001$

O desvio estimado para o efeito aleatório deste modelo foi de $\sigma_u = 2,732$ x 10^{-7}. Uma estimativa baixa, indicando que provavelmente não existe correlação significativa entre as respostas do mesmo provador. Utilizando o TRV verificou-se a significância deste efeito, comparando o modelo 4 com o modelo

$$\log\left[\frac{\gamma_{jl}}{1-\gamma_{jlc}}\right] = \alpha_j + \beta_l, \tag{4.1}$$

Este teste não foi rejeitado, produzindo um nível descritivo de $0,50$, ou seja, a variância do efeito aleatório não difere de zero. Mesmo este não sendo significativo, optou-se em mantê-lo no modelo pelo delineamento em bloco em que foi realizado o experimento. Porém por apresentar um valor muito baixo, a inclusão ou não deste efeito não muda as estimativas dos parâmetros apresentadas na Tabela 4.4, interferindo apenas no erros padrões destas estimativas.

Após selecionado o modelo, verificou-se a qualidade deste ajuste. Para isto foi utilizada a estatística *deviance* descrita na eq. (2.12). Observou-se $G^2 = 13,099$, seguindo uma distribuição χ^2 com $((4-1)(8-1)-1)$ graus de liberdade, obtendo um nível descritivo $= 0,8730$. Desse modo, não se rejeita a hipótese nula, a nível de significância de 5%, de que o modelo escolhido se ajusta satisfatoriamente aos dados.

A Figura 4.2 apresenta as proporções observadas para os dois sexos e as proporções estimadas por variedade de tomate, que visualmente nos indica que os valores estimados estão próximos dos observados, indicando que o modelo está bem ajustado. A Figura 4.3 mostra a correlação entre as proporções observadas e estimadas pelo modelo, em que se pode observar que estes valores estão correlacionados.

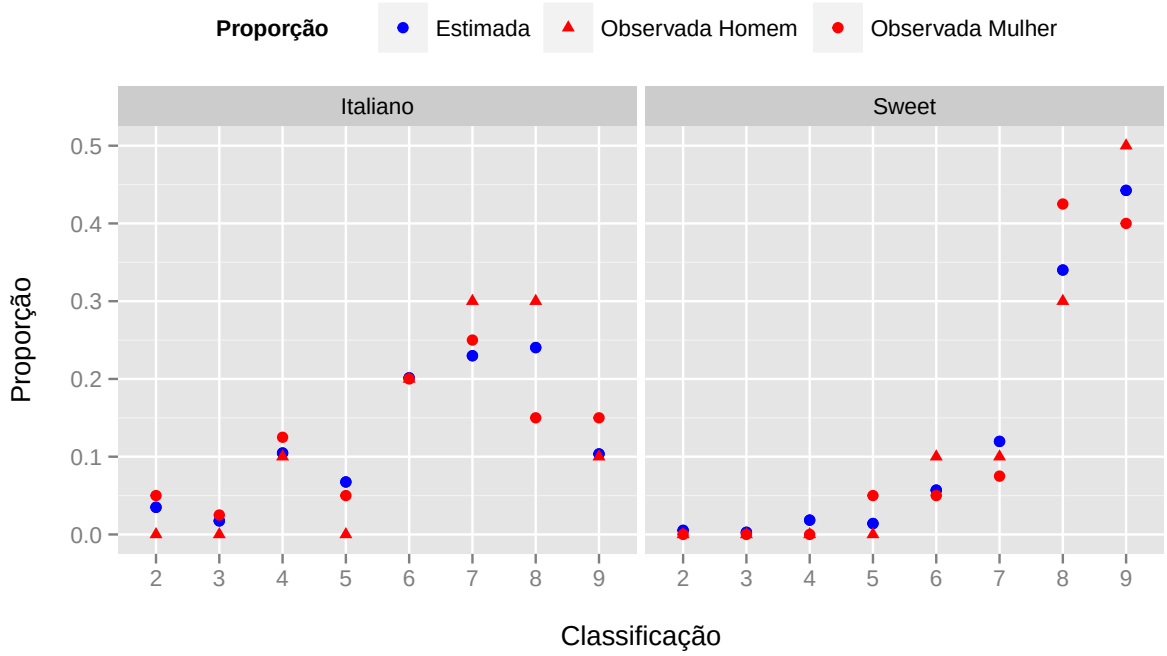

Figura 4.2 – Proporções observadas para cada sexo e proporções estimadas pelo modelo 4 por variedade de tomate

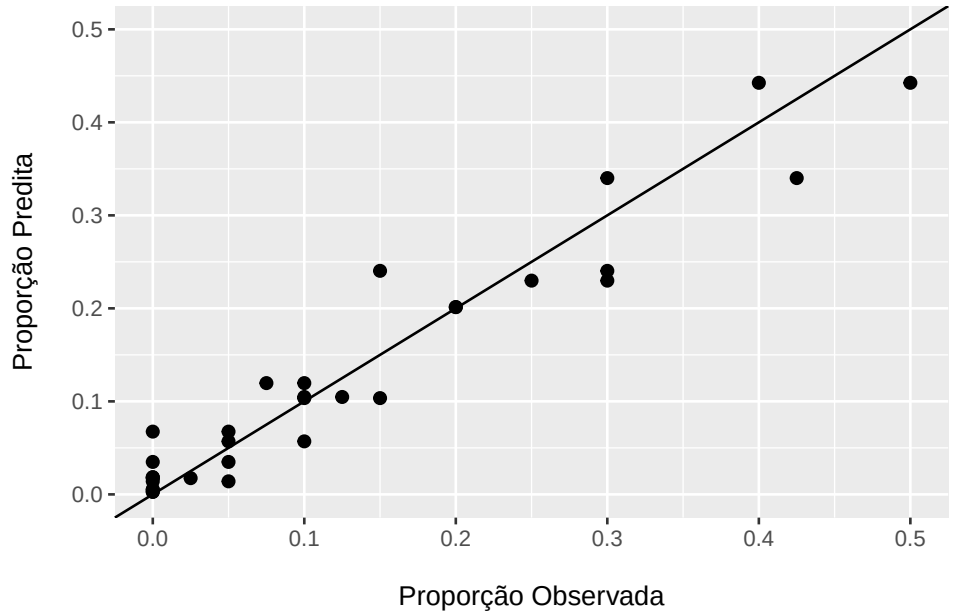

Figura 4.3 – Correlação entre as proporções observadas e estimadas pelo modelo 4 selecionado

Adicionalmente, foi realizado o ajuste de modelos com diferentes funções de ligação. Para comparação da melhor função utilizou-se o Critério Bayesiano de Schwarz (BIC) e o Critério de Informação de Akaike (AIC). Apesar da proximidade destes valores pode-se observar na Tabela 4.5 que a função de ligação logito$[P(Y \leq j|\mathbf{x})]$ produziu os menores valores nos dois critérios, indicando um melhor ajuste.

Tabela 4.5 – Comparação pelos critérios AIC e BIC para os modelos com diferentes funções de ligação

Função de Ligação	AIC	BIC
logito$[P(Y \leq j\|\mathbf{x})]$	333,63	357,08
$\phi^{-1}[P(Y \leq j\|\mathbf{x})]$	333,84	357,29
log{-log$[1 - P(Y \leq j\|\mathbf{x})]$}	334,00	357,45
log{-log$[P(Y \leq j\|\mathbf{x})]$}	339,47	362,92

Verificada a qualidade do ajuste deste modelo é de interesse inferir resultados sobre estes parâmetros.

Pode-se observar na Tabela 4.4 que o coeficiente associado ao parâmetro β é negativo, indicando probabilidades de notas mais altas da escala hedônica para a variedade Sweet Grape. Na Tabela 4.6 apresenta-se a razão de chances para as duas variedades e o respectivo intervalo de confiança. Pode-se interpretar que o tomate Italiano tem $6,89$ vezes mais chances de ser classificado com menores escores em relação ao tomate Sweet Grape. Ou seja existe uma associação entre estas variáveis e a variedade Italiano tende a receber categorias de respostas mais baixas do que a variedade Sweet Grape.

Tabela 4.6 – Intervalo de 95% de confiança de Wald para a razão de chances das variedades

Efeito	Estimativa	Razão de chances	Limite Inferior	Limite Superior
Sweet Grape vs Italiano	-1,927	6,869	2,817	16,667

Na Tabela 4.7 são apresentadas as probabilidades estimadas pelo modelo 4 para cada uma das categorias de resposta para as duas variedades. Por ela observa-se que a variedade Sweet Grape tem probabilidade de $0,4425$ de receber uma classificação 9-Gostei extremamente, enquanto que a variedade Italiano possui somente $0,1035$. É possível observar também que o tomate Italiano tem maiores probabilidades de receber classificações baixas do que o Sweet Grape, indicando a boa aceitação pelos provadores deste produto.

Tabela 4.7 – Probabilidade Estimadas pelo modelo 4 para cada categoria por variedade de tomate

	Categorias							
Variedade	2	3	4	5	6	7	8	9
Italiano	0,0349	0,0175	0,1047	0,0675	0,2014	0,2298	0,2403	0,1035
Sweet	0,0052	0,0027	0,0184	0,0140	0,0570	0,1196	0,3401	0,4425

5 CONCLUSÕES

Este trabalho apresenta uma abordagem para análise de dados categorizados ordinais, com ênfase nos Modelos Logitos Cumulativos e Chances Proporcionais.

Estes modelos são úteis na análise sensorial pelo fato da variável resposta de interesse ser usualmente categorizada em alguma escala conhecida, na maioria das vezes de natureza ordinal. A partir destes modelos é possível verificar as relações entre covariáveis ou fatores que influenciam esta variável resposta de interesse, bem como fazer previsões das razões de chances e probabilidades para cada uma das categorias observadas. Utilizando os Modelos Logitos Cumulativos na análise sensorial é possível inferir qual marca ou produto é mais apreciado e quais covariáveis podem estar influenciando no gosto de diferentes tipos de consumidores. Este estudo é feito na maioria das vezes antes deste produto começar a ser comercializado, evitando assim prejuízos e desperdícios.

Para a análise sensorial apresentada neste trabalho selecionou-se um Modelo de Chances Proporcionais, em que se pode concluir que existe diferença entre o Sabor das duas variedades de tomate desidratado. A variedade que teve melhor aceitação ao paladar dos provadores, foi a variedade Sweet Grape, sendo esta preferência independente do gênero. Esta análise também pode ser realizada utilizando outras características sensoriais como variável resposta. Um estudo englobando todas estas características pode ser realizado utilizando extensões multivariadas destes modelos.

Outras áreas das Ciências também realizam estudos em que a variável resposta é categorizada ordinal, e os modelos descritos neste trabalho também podem ser utilizados. Contudo, este campo de pesquisa (dados discretos) e em particular para variável ordinal ainda é emergente na Literatura, como, por exemplo, a análise de resíduos.

Como previsão para trabalhos futuros, pretende-se estudar e buscar a construção de resíduos mais apropriados para a análise de dados discretos desta natureza, bem como a implementação de gráficos semi-normais de probabilidades ("half normal plots"), para a verificação da qualidade do ajuste dos modelo aqui apresentados.

REFERÊNCIAS

ABREU, M. N. S.; SIQUEIRA, A. L.; CAIAFFAI, W. T. Regressão logística ordinal em estudos epidemiológicos. **Revista Saúde Pública**, São Paulo, v. 43, p. 183–94, 2009.

ABREU, W. C. D.; BARCELOS, M. D. F. P.; LOPES, C. D. O.; MALFITANO, B. F.; PEREIRA, M. C. D. A.; BOAS, E. V. D. B. V. Características físicas e químicas de tomates secos em conserva. **Boletim do Centro de Pesquisa de Processamento de Alimentos**, Curitiba, v. 31, p. 237–244, 2011.

AGRESTI, A. **Analysis of ordinal categorical data**. New York: John Wiley & Sons, 1984. 287 p.

AGRESTI, A. **Categorical data analysis**. 2ª. ed. New York: John Wiley & Sons, 2002. 710 p.

AGRESTI, A. **An introduction to categorical data analysis**. 2ª. ed. New York: John Wiley & Sons, 2007. 372 p.

AKAIKE, H. A new look at the statistical model identification. **Automatic Control, IEEE Transactions on**, Berlin, v. 19, p. 716–723, 1974.

ANANTH, C. V.; KLEINBAUM, D. G. Regression models for ordinal responses: a review of methods and applications. **International journal of epidemiology**, Bristol, v. 26, p. 1323–1333, 1997.

ANDERSON, J.; PHILIPS, P. Regression, discrimination and measurement models for ordered categorical variables. **Applied statistics**, Chicago, v. 30, p. 22–31, 1981.

ASSOCIAÇÃO BRASILEIRA DE NORMAL TÉCNICAS - ABNT. **Análise sensorial dos alimentos e bebidas: terminologia**. 1993. 8 p.

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. **Journal of the American Statistical Association**, Michigan, v. 88, p. 9–25, 1993.

BRIEN, C. **Notas de aula-IV Randomized Complete Block Design**. 2015. Disponível em: <http://chris.brien.name/ee2/>. Acesso em: 20 set. 2015.

BUERGIN, R. **Tree-Based Varying Coefficient Regression for Generalized Linear and Ordinal Mixed Models**. 2015. Disponível em: <https://cran.r-project.org/web/packages/vcrpart/vcrpart.pdf>. Acesso em: 15 out. 2015.

CHEN, J.; CAI, J.; ZHOU, H. Two-step estimation for a generalized linear mixed model with auxiliary covariates. **Statistica Sinica**, San Diego, v. 14, p. 361–376, 2004.

CHRISTENSEN, R. **Analysis of ordinal data with cumulative link models estimation with the R-package ordinal**. 2011. Disponível em: <http://cran.r-project.org/web/packages/ordinal/vignettes/clm_intro.pdf>. Acesso em: 1 maio 2015.

COLLETT, D. **Modelling binary data**. London: Chapman and Hall/CRC, 2002. 408 p.

COSTA, S. C. da. **Modelos lineares generalizados mistos para dados longitudinais**. 2003. 108 p. Tese (Doutorado em Agronomia) — Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2003.

DAHL, D. B. **xtable: Export Tables to LaTeX or HTML**. Provo, 2015. R package version 1.8-0. Disponível em: <https://CRAN.R-project.org/package=xtable>.

DEMÉTRIO, C. G.; CORDEIRO, G. M.; MORAL, R. A. **Modelos lineares generalizados e Extensões**. Piracicaba: ESALQ, Departamento de Ciências Exatas, 2014. 298 p.

DOBSON, A. J. **An introduction to generalized linear models**. 1ª. ed. London: Chapman and Hall/CRC, 1990. 176 p.

FIENBERG, B. **Analysis of Cross-Classified Data**. 2ª. ed. Cambridgeshire: MIT Press, 1980. 198 p.

FREEMAN, D. H. **Applied categorical data analysis**. New York: Chapman and Hall/CRC, 1987. 336 p.

GALWEY, N. W. **Introduction to mixed modelling: beyond regression and analysis of variance**. London: Wiley, 2006. 376 p.

GIOLO, S. R. **Introdução à análise de dados categóricos com aplicações**. 2012. Disponível em: <www.ufpr.br/~giolo>. Acesso em: 10 jun. 2015.

HARTZEL, J.; AGRESTI, A.; CAFFO, B. Multinomial logit random effects models. **Statistical Modelling**, North Lauderdale, v. 1, p. 81–102, 2001.

HEDEKER, D.; MERMELSTEIN, R. J. Analysis of longitudinal substance use outcomes using ordinal random-effects regression models. **Addiction**, Chicago, v. 95, p. 381–394, 2000.

HOSMER, D. W.; LEMESHOW, S. **Applied logistic regression**. New York: Wiley, 2004. 672 p.

INST FOOD TECHNOLOGISTS-IFT. Sensory evaluation guide for testing food and beverage products. **Food technology**, Chicago, v. 35, p. 50–59, 1981.

KOH, E.; CHAROENPRASERT, S.; MITCHELL, A. E. Effects of industrial tomato paste processing on ascorbic acid, flavonoids and carotenoids and their stability over one-year storage. **Journal of the Science of Food and Agriculture**, Davis, v. 92, p. 23–28, 2012.

LAIRD, N. M.; WARE, J. H. Random-effects models for longitudinal data. **Biometrics**, New Yokr, v. 38, p. 963–974, 1982.

LALL, R.; CAMPBELL, M.; WALTERS, S.; MORGAN, K.; CO-OPERATIVE, M. C. A review of ordinal regression models applied on health-related quality of life assessments. **Statistical methods in medical research**, Lewis, v. 11, p. 49–67, 2002.

LEE, Y.; NELDER, J. A. Hierarchical generalized linear models. **Journal of the Royal Statistical Society. Series B (Methodological)**, London, v. 58, p. 619–678, 1996.

LEMOS, T. D. O.; RODRIGUES, M. D. C. P.; LARA, I. A. R. D.; ARAÚJO, A. M. S. D.; LEMOS, T. L. G. D.; PEREIRA, A. L. F.; PAULA, L. V. T. D. Modeling the acceptability of cashew apple nectar brands using the proportional odds model. **Journal of Sensory Studies**, v. 30, p. 136–144, 2015.

LI, C.; SHEPHERD, B. E. A new residual for ordinal outcomes. **Biometrika**, Nashville, v. 99, p. 473–480, 2012.

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

LIPSITZ, S. R.; KIM, K.; ZHAO, L. Analysis of repeated categorical data using generalized estimating equations. **Statistics in medicine**, Stamford, v. 13, p. 1149–1163, 1994.

LIU, I.; AGRESTI, A. The analysis of ordered categorical data: An overview and a survey of recent developments. **Test**, Gainesville, v. 14, p. 1–73, 2005.

LIU, Q.; PIERCE, D. A. A note on gauss—hermite quadrature. **Biometrika**, North Lauderdale, v. 81, p. 624–629, 1994.

LORO, A. C. **Caracterização quimica e funcional de tomates "Sweet Grape" e Italiano submetidos á desidratação osmótica e adiabática**. 2015. 89 p. Dissertação (Mestrado em Ciências) — Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2015.

MCCULLAGH, P. Regression models for ordinal data. **Journal of the royal statistical society. Series B (Methodological)**, London, v. 42, p. 109–142, 1980.

MCCULLAGH, P.; NELDER, J. A. **Generalized linear models**. 2ª. ed. London: Chapman and Hall/CRC, 1989. 532 p.

MCCULLOCH, C. E.; NEUHAUS, J. M. **General, linear and mixed models**. [S.l.]: Wiley Online Library, 2001.

MOLENBERGHS, G.; VERBEKE, G. **Models for discrete longitudinal data**. New York: Springer, 2006. 687 p.

NELDER, J. A.; WEDDERBURN, W. Generalized linear models. **Journal of the Royal Statistical Society Series A**, New York, v. 135, p. 370–384, 1972.

PAULA, G. A. **Modelos de regressão: com apoio computacional**. São Paulo: Universidade de São Paulo, 2013. 428 p.

PAULINO, C. D. M.; SINGER, J. da M. **Análise de dados categorizados**. São Paulo: Edgard Blucher, 2006. 648 p.

PEARSON, K. Mathematical contributions to the theory of evolution xiii: On the theory of contingency and its relation to association and normal correlation. **Biometric Series**, London, v. 135, p. 1–47, 1904.

PETERSON, B.; HARRELL, F. E. Partial proportional odds models for ordinal response variables. **Applied statistics**, Durham, v. 39, p. 205–217, 1990.

PINHEIRO, J.; BATES, D. **Mixed-effects models in S and S-PLUS**. New York: Springer, 2006. 528 p.

PINHEIRO, J. C.; BATES, D. M. Approximations to the log-likelihood function in the nonlinear mixed-effects model. **Journal of computational and Graphical Statistics**, Madison, v. 4, p. 12–35, 1995.

R Development Core Team. R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria. 2014.

RAO, C. R. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. **Mathematical Proceedings of the Cambridge Philosophical Society**, Cambridgeshire, v. 44, p. 50–57, 1948.

RAO, C. R. **Linear statistical inference and its applications**. 2ª. ed. New York: Wiley-Interscience, 1973. 656 p.

SAS Institute INC.SAS/STAT **SAS user's guide for windows environment**. Cary: SAS Institute, 2011.

SCHWARZ, G. Estimating the dimension of a model. **The annals of statistics**, Jerusalem, v. 6, p. 461–464, 1978.

SILVA, N. B. da. **Diferentes estratégias para modelagem de respostas politômicas ordinais em estudos longitudinais**. 2013. Belo Horizonte, 2013. 66 p. Dissertação (Mestrado em Estatística) — Universidade Federal de Minas Gerais, Belo Horizonte, 2013.

SPYRIDES-CUNHA, M. H. **Modelos para dados categorizados ordinais: aplicações em agropecuária**. 1998. 80 p. Dissertação (Mestrado em Agronomia) — Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 1998.

STIGER, T. R.; BARNHART, H. X.; WILLIAMSON, J. M. Testing proportionality in the proportional odds model fitted with gee. **Statistics in medicine**, Groton, v. 18, p. 1419–1433, 1999.

STIRATELLI, R.; LAIRD, N.; WARE, J. H. Random-effects models for serial observations with binary response. **Biometrics**, New York, v. 40, p. 961–971, 1984.

STOKES, M. E.; DAVIS, C. S.; KOCH, G. G. **Categorical data analysis using SAS**. 2ª. ed. Cary: SAS Institute Inc, 2001. 626 p.

SUESSE, T.; LIU, I. Modelling strategies for repeated multiple response data. **International Statistical Review**, Wollongong, v. 81, p. 230–248, 2013.

TEIXEIRA, L. V. Análise sensorial na indústria de alimentos. **Revista do Instituto de Laticínios Cândido Tostes**, Juiz de Fora, v. 64, p. 12–21, 2009.

TUTZ, G.; HENNEVOGL, W. Random effects in ordinal regression models. **Computational Statistics & Data Analysis**, Chicago, v. 22, p. 537–557, 1996.

TUTZ, G.; SCHOLZ, T. **Ordinal regression modelling between proportional odds and non-proportional odds**. Munich. 2003, 31 p. Relatório Técnico, Institute of Statistics, University of Munich.

WALD, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. **Transactions of the American Mathematical society**, Raleigh, v. 54, p. 426–482, 1943.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. Disponível em: <http://had.co.nz/ggplot2/book>.

WICKHAM, H. The split-apply-combine strategy for data analysis. **Journal of Statistical Software**, v. 40, n. 1, p. 1–29, 2011. Disponível em: <http://www.jstatsoft.org/v40/i01/>.

WOLFINGER, R.; O'CONNELL, M. Generalized linear mixed models a pseudo-likelihood approach. **Journal of statistical Computation and Simulation**, Blacksburg, v. 48, p. 233–243, 1993.

ANEXOS

ANEXO A - PROGRAMAÇÃO R

```
#pacote utilizados
require(ggplot2)
require(plyr)
require(ordinal)
require(vcrpart)
require(xtable)
dados=read.csv("tomate.csv")
#variáveis: variedade,sexo,sabor,avaliador e classificação.

#Análise descritiva
dados$tomate=as.factor(dados$tomate)
dados$sexo=as.factor(dados$sexo)
dados$avaliador=as.factor(dados$avaliador)
dados$Sabor=as.ordered(dados$Sabor)


aux <-with(dados, paste(dados$tomate, dados$sexo))
xtable(table(aux,dados$Sabor))

Homem=subset(dados, sexo=="homem")
Mulher=subset(dados, sexo=="mulher")

#Mulher
df0 = data.frame(table(Mulher$tomate,Mulher$Sabor))

df1=rename(df0, replace = c("Var1" = "Variedade","Freq" = "Mulher"))

ggplot(df1, aes(x=Var2,y=Freq/40, stat="identity") +
aes(shape = Variedade)+ geom_point(size=3)
+xlab('\nClassificação\n (a)')+ ylab('Proporção Observada\n')
+theme(axis.text=element_text(size=13),axis.title=element_text(size=12))
+scale_y_continuous(limits=c(0, 0.5))

#Homem
df3= data.frame(table(Homem$tomate,Homem$Sabor))

df4=rename(df3, replace = c("Var1" = "Variedade","Freq" = "Homem"))

ggplot(df4, aes(x=Var2,y=Freq/10, stat="identity")) +
aes(shape = Variedade)+ geom_point(size=3)
+xlab('\nClassificação\n (b)')+ ylab('Proporção Observada\n')
 +theme(axis.text=element_text(size=12),axis.title=element_text(size=12))+
scale_shape_discrete(breaks=c("Sweet", "Italiano"),labels=c("Sweet Grape","Italiano"))
```

```
+scale_y_continuous(limits=c(0, 0.5))

#Teste de odds proporcional R
mod.1=clm(Sabor~tomate+sexo,data=dados)
mod.2=clm(Sabor~1,nominal=~tomate+sexo,data=dados)
anova(mod.1,mod.2)

#Teste de odds proporcional SAS


#Seleção das variáveis
m1 <- olmm(Sabor ~ ge(Tomate)+ge(sexo)+ge(Tomate*sexo) + re(1|avaliador), data = dados
m2 <- olmm(Sabor ~ ge(Tomate)+ge(sexo) + re(1|avaliador), data = dados)
m3 <- olmm(Sabor ~ ge(sexo)+ re(1|avaliador), data = dados)
m4 <- olmm(Sabor ~ ge(Tomate)+ re(1|avaliador), data = dados)
m5 <- olmm(Sabor ~ 1+ re(1|avaliador), data = dados)

anova(m1,m2)
anova(m2,m3)
anova(m2,m4)
anova(m4,m5)

summary(m4)

#Cálculo Deviance

p1 <- predict(m4, ranef = TRUE, type = "prob")
estima=aggregate(p1 ~ tomate + sexo, data=dados, FUN=sum)
estima<-estima[,-c(1,2)]

pi.hat <- tab / estima

tabela=tab*log(pi.hat)
tabela=tabela[ !is.na(tabela) ]
D=2*sum(tabela)

pchisq(D,20,lower.tail = FALSE)



#Gráfico probabilidade observadas e estimadas

(dados2 = data.frame(tomate = dados$tomate, sexo = dados$sexo,
 Level = dados$Level, valor = c(dados$Estimada,dados$Observada),
 tipo=rep(c('Estimada','Observada'),each=length(dados$tomate))))
dados2$Level=as.factor(dados2$Level)
```

```
ggplot(dados2, aes(x=Level, y=valor, colour=tipo)) +
geom_point(size=2) + facet_grid(.~Tomate) +
xlab('\nClassificação\n')+ ylab('Proporção\n')
+scale_colour_manual(name="Proporção",breaks=c('Estimada','Observada'),
 values=c('blue','red'))+ theme(legend.position="top")

ggplot(dados1, aes(x=Observada, y=Estimada)) + geom_point(size=2)+
xlab('\nProporção Observada\n')+ ylim(0,0.5)+ ylab('Proporção Predita\n')
 +scale_colour_manual(name="Proporção",breaks=c('Estimada','Observada'),
values=c('blue','red'))+ theme(legend.position="top")


#Probabilidade estimadas.
newData=expand.grid(sexo=levels(dados$sexo),Tomate=levels(dados$Tomate))
predict(m4,newData,type="prob")


#Comparações funções de ligação
mod4.1=clmm(Sabor~Tomate+(1|avaliador),link="logit",nAGQ = -7,data=dados)
mod4.2=clmm(Sabor~Tomate+(1|avaliador),link="probit",nAGQ = -7,data=dados)
mod4.3=clmm(Sabor~Tomate+(1|avaliador),link="cloglog",nAGQ = -7,data=dados)
mod4.4=clmm(Sabor~Tomate+(1|avaliador),link="loglog",nAGQ = -7,data=dados)
```

ANEXO B- PROGRAMAÇÃO SAS

```
data dados;
input tomate $ avaliador sexo $ Sabor;
datalines;
Sweet    1 mulher 8
Italiano 1 mulher 6
Sweet    2 mulher 8
Italiano 2 mulher 8
Sweet    3 mulher 8
Italiano 3 mulher 7
...
;

proc logistic data=dados;
class tomate sexo;
model Sabor=tomate sexo/scale=none agreggate;
run;
```