Universidade Federal de São Carlos
Centro de Ciências Exatas e de tecnologia
Departamento de Estatística

MODELOS LINEARES GENERALIZADOS MULTINOMIAIS UNIVARIADO E BIVARIADO

Leticia Caroline da Silva David

Trabalho de Conclusão de Curso

Leticia Caroline da Silva David

Modelos Lineares Generalizados Multinomiais Univariado e Bivariado

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Leticia Caroline da Silva David e aprovado pela banca examinadora.

São Carlos, 10 de Junho de 2021.

Banca Examinadora

- Prof. Dr. Francisco Antônio Rojas Rojas
- Prof. Dr. Afrânio Márcio Corrêa Vieira
- Profª. Drª. Vera Lucia Damasceno Tomazella

Agradecimentos

Agradeço a meus pais, Inês e Edson, pelo amor e apoio financeiro dedicado nessa minha jornada. Principalmente a minha mãe por todos os sacrifícios.

Agradeço a minha irmã, Bianca, por sempre acreditar em mim e puxar minha orelha quando foi necessário.

Agradeço a minha vó, Cleuza, por sempre me receber nas férias com a minha comida favorita.

Agradeço ao Profº Dr. Francisco Rojas por aceitar ser meu orientador e dedicar inúmeras horas na construção deste trabalho.

Gostaria de agradecer também aos professores da banca Profº Dr. Afrânio Vieira e Profª. Drª. Vera Tomazella pelas correções que auxiliaram na finalização deste trabalho.

E, não menos importante, agradeço a Aline Rizo por sempre me apoiar e acreditar no meu potencial.

.

"Apenas que... busquem conhecimento"

(Bilu, ET)

Resumo

Este Trabalho de Graduação apresenta os fundamentos teóricos básicos dos modelos lineares generalizados e as características das análises de regressão logística binomial e multinomial univariadas. Aplicações práticas da regressão binomial e da multinomial foram realizadas. As aplicações dizem com relação a avaliação da eficácia das doses de um herbicida para controle de certa planta daninha; da eficácia do nível de radiação para controle de moscas domésticas; e para avaliação do grau de retinopatia como função da condição de diabetes e outras características dos indivíduos. Os códigos das implementações dos procedimentos analíticos, executados pelo software R, são disponibilizados.

São descritos, também, três modelos de análise de regressão logística multinomial bivariada: condicional de efeitos fixos, marginal e de efeitos mistos. Aplicações para cada uma dessas análises são apresentadas. As aplicações se referem: à avaliação do grau de retinopatia nos dois olhos dos indivíduos como função da condição de diabetes e outras características dos indivíduos; à severidade da artrite reumatoide como função de condições dos pacientes. Os códigos das implementações dos procedimentos analíticos, executados pelos softwares R, MATLAB e SAS são disponibilizados.

Palavras-chave: *modelo de efeitos fixos, modelo linear generalizado, modelo linear generalizado misto, modelo marginal, regressão logística multinomial bivariada, regressão logística multinomial univariada.*

Sumário

Lista de Tabelas

Capítulo 1

Introdução

Encontramos na literatura estudos que condicionam uma variável resposta a uma ou mais variáveis explicativas. Em muitos dos estudos a variável resposta é do tipo contínuo e as variáveis explicativas são do tipo quantitativo ou do tipo qualitativo. Essas situações têm sido extensivamente analisadas mediante o uso do chamado Modelo Linear Simples (ML) de maneira bastante simplificada e já bem conhecida e compreendida pelos pesquisadores. Entretanto, nas situações em que a variável resposta é quantitativa discreta ou é qualitativa (categórica), a sua relação com as variáveis explicativas não é descrita de forma apropriada por um ML.

No caso em que a variável resposta tem distribuição Normal, o ML permite a estimação e interpretação dos parâmetros do modelo de maneira simples. Para análise de alguns experimentos que possuem resposta contínua não normal é possível obter uma transformada da família das transformações de Box e Cox (1964), cuja distribuição seja normal e, então, utilizar o ML.

Modelos de regressão mais adequados para variáveis com respostas normais são os Modelos Lineares Generalizados propostos por Nelder e Wedderburn (1972) no artigo "Generalized Linear Models" em que apresentaram uma classe geral de que incorpora um conjunto de metodologias de análise que já haviam sido utilizadas anteriormente, mas de forma isolada.

Os autores mostraram que se a distribuição da variável resposta não for normal, mas pertencer a família exponencial linear, as análises são similares a análise por ML, sem a necessidade de aplicar qualquer transformação na variável resposta para adequá-la a suposição de normalidade e de homocedasticidade. Para tanto, basta escolher uma função de ligação que melhor se adeque a $E(Y|X) : R, R^+$.

Em estudos que relacionam uma única variável resposta quantitativa discreta ou uma variável resposta binária, ou até uma variável resposta do tipo multinomial, a um conjunto de variáveis independentes quantitativas e qualitativas, não pode ser analisada de maneira adequada pelo uso de modelos lineares simples. Os Modelos Lineares Generalizados são uma alternativa aceita como bem apropriada para analisar conjuntos de dados cujas variáveis resposta tenham essas características.

Algumas pesquisas podem ter o interesse de analisar duas variáveis resposta categóricas, por exemplo, relacionadas a um conjunto de variáveis independentes quantitativas e/ou qualitativas. Os modelos maginais e mistos podem ser utilizados como uma extensão natural do caso univariado, MLG, para analisar conjuntos de dados com estas características.

1.1 Objetivos do Trabalho

Esse Trabalho de Graduação se propôs a:

1. Revisar a metodologia de Modelos Lineares Generalizados para o caso em que a variável resposta é categórica;

2. Estudar metodologias de regressão logística binomial e multinomial univariadas, e de regressão logística multinomial bivariada;

3. Levantar possíveis aplicações desses modelos e conjuntos de dados para essa aplicações;

4. Levantar disponibilidade de funções, procedimentos ou pacotes para a execução das análises estatísticas correspondentes a essas aplicações.

1.2 Organização do Trabalho

No capítulo 2 são expostos os fundamentos gerais dos modelos lineares generalizados, de acordo com Cordeiro e Demétrio (2008) e Paula (2013). Os modelos de regressão logística binomial e multinomial univariados são descritos segundo Cordeiro e Demétrio (2008). Uma aplicação do modelo de regressão logístico binomial, para analisar a eficácia de doses de um herbicida sobre o controle de plantas, segundo Turner *et al.* (1992), é descrita. Duas aplicações do modelo de regressão logístico multinomial são apresentadas. A primeira analisa a eficácia de concentração de níveis de radiação sobre o controle de

moscas domésticas, segundo Itepan (1995), é descrita. A segunda analisa o grau de severidade da retinopatia como função das variáveis idade, uso de insulina, níveis de hemoglobina e de proteína na urina, e da duração do diabetes.

No capítulo 3 são apresentados os fundamentos de três modelos alternativos de regressão logística multinomial bivariada. O primeiro é o Modelo Condicional de Efeitos Fixos é descrito segundo Uddin e Begum (2018). O segundo é chamado de Modelo Marginal é descrito de acordo com Liang e Zeger (1986) e Touloumis *et al.* (2013), e o terceiro é o denominado Modelo de Efeitos Mistos é descrito de acordo com Hartzel *et al.* (2001). Duas aplicações para os modelos são apresentadas, a primeira analisa o grau de severidade da retinopatia de pacientes diabéticos. E, a segunda aplicação, de Bombardier *et al.* (1986), analisa o grau de severidade da artrite reumatoide como função das variáveis sexo, idade, aplicação de um tratamento e medida de um escore inicial.

E no capítulo 4 são apresentadas as conclusões finais deste trabalho. E finalmente, no Apêndice A são apresentados os códigos dos procedimentos de análise para as aplicações pelos diferentes modelos.

Capítulo 2

Regressão Multinomial Univariada

Na modelagem estatística, uma das metodologias mais utilizadas é a análise de regressão. O modelo mais simples dessa metodologia busca descrever a relação de dependência entre uma variável resposta e uma ou mais variáveis explicativas.

A partir de uma amostra de valores observados dessas variáveis em um conjunto de indivíduos ou elementos, obtém-se um modelo que permite predizer o valor da variável resposta, com base nos valores das variáveis explicativas.

A equação do modelo de regressão linear univariado é da forma:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i, \tag{2.1}$$

sendo que para o indivíduo ou elemento i, y_i é o valor da variável resposta, x_{ij} é o valor da variável explicativa x_j, ϵ_i é o valor do erro aleatório e β_k é o parâmetro que será estimado, com i=1,...,n , j=1,...,p e k=0,...,p.

Quando têm se apenas uma variável explicativa, $y_i = \beta_0 + \beta_1 x_1 + \epsilon_i$, o modelo é conhecido como regressão linear simples. O modelo para p ≥ 2 é conhecido como regressão linear múltipla.

No modelo linear (ML) o comportamento da resposta é dito linear porque, para cada unidade que aumentar na variável explicativa x_j, resulta no acréscimo de $\beta * 1$ unidade no valor da resposta y_i.

A análise de regressão tem como pressupostos do modelo que os erros sejam independentes, identicamente e normalmente distribuídos, com média zero e com variância constante e desconhecida. Ou seja, os ϵ_i são *i.i.d.* $N(\mu, \sigma^2)$.

Há diversas situações em que o modelo de regressão linear pode ser utilizado, e em

diversas áreas como em áreas da saúde, economia, educação, entre outras. Porém, considerando as suposições do modelo de regressão linear há muitas situações em que o modelo não é o adequado, como o caso em que a variável resposta observada é uma contagem de indivíduos ou elementos, ou no caso em que a variável é qualitativa organizada de forma categórica. Em ambos esses casos, a variável não é do tipo contínuo e, portanto, sua distribuição não pode ser considerada como normal.

Quando a variável não tem distribuição normal pode se tentar adequar os dados observados ao ML mediante transformações na variável resposta, como as transformações propostas por Box e Cox (1964). No entanto, transformar a variável resposta não resolve a não adequação das suposições do modelo de regressão linear em alguns casos. Além disso, a transformação em muitos casos fornece valores de parâmetros de difícil interpretação. Uma estratégia que contempla as dificuldades ou restrições do modelo linear simples, é o denominado Modelo Linear Generalizado (MLG), proposto por Nelder e Wedderburn (1972).

2.1 Modelo Linear Generalizado

O Modelo Linear Generalizado possui a vantagem sobre o modelo linear geral de não possuir as restrições ou as exigências de um ML, que são as suposições de normalidade e/ou homocedasticidade dos erros.

A equação do Modelo Linear Generalizado é da forma:

$$g(E[y_i]) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}, \tag{2.2}$$

sendo que para o indivíduo ou elemento i, y_i é o valor da variável resposta, x_{ij} é o valor da variável explicativa x_j, β_k é o parâmetro que será estimado, com i=1,...,n , j=1,...,p e k=0,...,p e g uma função invertível chamada função de ligação.

Os MLG's são definidos por uma distribuição de probabilidade, membro da família exponencial linear de distribuições, e são formados pelas componentes: aleatória, sistemática e função de ligação. Para a aplicação de um modelo linear generalizado, é necessário que a função de distribuição de probabilidade, f.d.p., de Y possa ser escrita na forma seguinte, chamada família exponencial linear de densidades:

$$f(y_i; \theta_i, \phi) = \exp\left\{\phi[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)]\right\}, \tag{2.3}$$

sendo $b(.)$ e $c(.)$ funções dependentes da distribuição a ser estudada e $\phi^{-}1 > 0$ ($\phi > 0$) conhecido como parâmetro de dispersão ou como parâmetro de precisão. As funções escritas na forma da equação 2.3 possuem $\mathrm{E}(Y_i) = \mu_i = \mathrm{b'}(\theta_i)$ e $\mathrm{Var}(Y_i) = \phi^{-}1\ \mathrm{V}\ (\mu_i)$, em que $V_i = \mathrm{V}\ (\mu_i) = \mathrm{d}\mu_i/\mathrm{d}\theta_i$ é a função de variância e ϕ é o parâmetro de dispersão. (Paula, 2013)

Quando a variável resposta analisada possui distribuição normal com média μ e variância σ^2, a sua f.d.p. escrita como na equação 2.3 é,

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\},$$

$$f(y;\mu,\sigma^2) = \exp\left\{\frac{1}{\sigma^2}\left(y\mu - \frac{\mu^2}{2}\right) - \frac{1}{2}\log(2\sigma^2) - \frac{y^2}{2\sigma^2}\right\},$$

$$\theta = \mu, \quad \phi = \sigma^2, \quad b(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2} \text{ e } \quad c(y,\phi) = \frac{-1}{2}\left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right].$$

Segundo Paula (2013), são três os componentes que estruturam um MLG: o componente aleatório, o componente sistemático e a função de ligação. O componente aleatório é composto pelas variáveis aleatórias Y_i independentes, i=1,...n, que são provenientes de uma mesma distribuição que possui a função de densidade de probabilidade dada como em (2.3) e que possuem médias μ_i.

O componente sistemático estabelece a forma em que as variáveis explicativas irão compor o modelo. Na forma matricial:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

sendo:

$\boldsymbol{\eta}$: o vetor dos i preditores lineares, i=1,...,n;

$\mathbf{X}$: a matriz do modelo composta pelas m variáveis explicativas; e

$\boldsymbol{\beta}$: o vetor de (p+1) parâmetros. No caso em que há parâmetros que possuam valores conhecidos eles serão chamados de parâmetros *offset*.

A função de ligação $g(\mu_i) = \eta_i$ é o elemento que liga o componente sistemático do modelo ao seu componente aleatório, relacionando a média da variável resposta à combinação linear das variáveis explicativas. A função $g(.)$ seja estritamente monótona e duplamente diferenciável, não há restrições quanto ao seu formato.

Para as distribuições pertencentes a família exponencial linear há funções de ligação

especificas em que o preditor linear modela diretamente o parâmetro θ_i. Essas funções são chamadas de funções de ligação canônica. Utilizar a função de ligação canônica para o modelo traz como vantagem a existência de estatísticas suficientes para os parâmetros, o que conduz simplificações no algoritmo de estimação. Além disso, os parâmetros obtidos poderão ser interpretados de forma direta. As funções canônicas para as principais distribuições pertencentes a família exponencial linear são apresentadas na tabela 2.1:

Tabela 2.1: Funções de ligação canônicas

Distribuição	Função de ligação canônica
Normal	$\eta = \mu$
Poisson	$\eta = \log(\mu)$
Binomial	$\eta = \log\left(\frac{\pi}{1-\pi}\right)$
Gama	$\eta = \frac{1}{\mu}$
Normal Inversa	$\eta = \frac{1}{\mu^2}$

Para o caso da distribuição normal padrão e a função de ligação identidade, tem se o modelo de regressão linear apresentado anteriormente em (3.4).

2.2 Estimação dos Parâmetros

Para a estimação dos parâmetros do MLG podem ser utilizadas vários métodos de estimação como qui-quadrado mínimo, o Bayesiano e o de máxima verossimilhança. Utilizando o método de máxima verossimilhança, a função de verossimilhança para modelos lineares generalizados é dada por

$$L(\boldsymbol{\beta};\phi) = \prod_{i=1}^{n} \exp\left\{\phi[y_i\theta_i - b(\theta_i)] + c(y_i,\phi)]\right\}. \tag{2.4}$$

O logaritmo da função de verossimilhança é

$$l(\boldsymbol{\beta};\phi) = \log(L(\boldsymbol{\beta};\phi)) = \sum_{i=1}^{n} \left\{\phi[y_i\theta_i - b(\theta_i)] + c(y_i,\phi)]\right\}. \tag{2.5}$$

O vetor $\mathbf{U}(\gamma)$ que contém as k+2 derivadas de primeira ordem de β e do parâmetro de dispersão ϕ é chamado de vetor escore.

$$\mathbf{U}(\gamma) = \mathbf{U}(\boldsymbol{\beta};\phi) = \begin{pmatrix} \mathbf{U}_{\boldsymbol{\beta}} \\ \mathbf{U}_{\phi} \end{pmatrix}.$$

Para encontrar os estimadores β_i é necessário resolver o sistema de equações $\mathbf{U}_r$=0. Obtidos a partir da aplicação da regra da cadeia em $\mathbf{U}_{\boldsymbol{\beta}}$ sobre o r-ésimo parâmetro

$$\mathbf{U}_r = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_r} = \sum_{i=1}^{n} \frac{dl_i}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{d\eta_i}{d\beta_r} = \phi^{-1}\sum_{i=1}^{n}(y_i-\mu_i)\frac{1}{V_i}\frac{d\mu_i}{d\eta_i}x_{ir}.$$

Essas equações nem sempre são lineares, sendo necessário que sejam aplicados métodos numéricos como o Método de Newton-Raphson ou o Método Scoring de Fisher para encontrar a resolução.

Para estimação de $\boldsymbol{\beta}$, Cordeiro e Demétrio (2008) sugerem o algoritmo que considera os 4 passos seguintes:

1. Obter as estimativas

$$\boldsymbol{\eta_i^{(m)}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{x_{ir}\beta_r^{(m)}} \text{ e } \boldsymbol{\mu_i^{(m)}} = \boldsymbol{g^{-1}(\eta_i^{(m)})};$$

2. Obter a variável dependente ajustada

$$\boldsymbol{z_i^{(m)}} = \boldsymbol{\eta_i^{(m)}} + \boldsymbol{(y_i - \mu_i^{(m)})g'(\mu_i^{(m)})},$$

   e os pesos

$$\boldsymbol{w_i^{(m)}} = \frac{\mathbf{1}}{\boldsymbol{V(\mu_i^{(m)})[g'(\mu_i^{(m)})]^2}};$$

3. Calcular

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

   sendo $\mathbf{W}$= $\boldsymbol{diag\{w_1, ..., w_n\}}$, com $\boldsymbol{w_i} = \boldsymbol{V_i^{-1}}(\frac{\boldsymbol{d\mu_i}}{\boldsymbol{d\eta_i}})^{\mathbf{2}}$ e $\mathbf{z} = \boldsymbol{\eta} + \mathbf{G(y-\boldsymbol{\mu})}$, sendo a matriz diagonal $\mathbf{G}$ formada pelas derivadas de primeira ordem da função de ligação; e,

4. Voltar ao passo 1 com $\boldsymbol{\beta^{(m)}} = \boldsymbol{\beta^{(m+1)}}$ e repetir o processo até convergir.

Uma das vantagens da utilização da função de ligação canônica pode ser vista no cálculo de $\boldsymbol{w_i}$, pois após algumas simplificações, $\boldsymbol{w_i} = \boldsymbol{V_i}$. Para o caso em que a distribuição é normal e função de ligação identidade, o $\mathbf{W}$ é uma matriz identidade de dimensão n e $\mathbf{z}$=$\mathbf{y}$, então a estimativa de $\boldsymbol{\beta}$ é $\hat{\boldsymbol{\beta}} = \mathbf{(X^t X)^{-1} X^t y}$, que coincide com o estimador do ML.

Em alguns casos, como para as distribuições binomial e Poisson, o parâmetro de dispersão $\boldsymbol{\phi}$ é conhecido e não precisa ser estimado. Para algumas outras distribuições pertencentes a família exponencial linear é necessário estimar esse parâmetro. Por máxima verossimilhança o estimador de $\boldsymbol{\phi}$ é obtido ao resolver $\mathbf{U}_{\boldsymbol{\phi}} = 0$, sendo,

$$\boldsymbol{U_\phi(\gamma) = \sum_{i=1}^{n}(y_i\theta - b(\theta_i)) + \sum_{i=1}^{n}\frac{\partial}{\partial\phi}c(y_i,\phi).}$$

Por outro lado, o estimador de $\boldsymbol{\phi}$ também pode ser obtido de forma simples utilizando o Método dos Momentos, considerando que a distribuição assintótica de $\boldsymbol{\phi = \sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{V(\mu_i)}}$ é qui-quadrado com (n-k-1) graus de liberdade. O estimador resultante é:

$$\boldsymbol{\hat{\phi}_{MM} = \frac{n-k-1}{\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{V(\hat{\mu}_i)}}.}$$

A seguir são apresentados os modelos para os casos em que a variável resposta é categórica, podendo a variável resposta ser modelada pelos modelos lineares binomial ou multinomial.

2.3 Regressão Binomial

Em uma ampla variedade de aplicações algumas variáveis resposta são binárias, com apenas dois resultados ou categorias possíveis, comumente chamados de "sucesso" e "falha" (ou "fracasso"). Por exemplo, no caso de inspeção de peças em uma fábrica, as peças são classificadas como defeituosas ou não-defeituosas, o resultado de exames laboratoriais e o resultado da aplicação de herbicidas ou inseticidas.

Por outro lado, também há situações em que algumas variáveis resposta, originalmente contínuas, são categorizadas a partir de um valor de referência e assim são transformadas em binárias, considerando as categorias de valores acima e valores abaixo de algum nível critico. O resultado de interesse é chamado de sucesso enquanto o outro resultado é chamado de fracasso.

Considerando essas situações, a regressão adequada não é a regressão linear, pois a distribuição da variável resposta não obedece a suposição de normalidade, que é uma das três suposições do modelo de regressão linear. A regressão adequada nesses casos é o MLG com resposta binomial.

Estudos em que se deseja testar uma determinada droga, um herbicida, ou um inseticida são conhecidos como ensaios do tipo dose-resposta. Para estes testes serão administrados k diferentes doses $\boldsymbol{d_i = d1, \ldots, dk}$, a $\boldsymbol{m_i}$ diferentes amostras de indivíduos, com $\boldsymbol{m_i}$=$\boldsymbol{m_1, m_2, \ldots, m_k}$. Para cada uma das diferentes amostras observadas a resposta y do individuo, sucesso ou fracasso, possui distribuição de Bernoulli dada por

$$\boldsymbol{f(y; \pi_i) = \pi_i(1 - \pi_i).} \tag{2.6}$$

Em uma amostra de tamanho m, o interesse do pesquisador é a proporção de sucessos $\boldsymbol{\pi_i}$ com distribuição binomial. Então a f.d.p. da variável resposta é,

$$\boldsymbol{f(y; m_i, \pi_i) = \binom{m_i}{y} \pi_i^y (1 - \pi_i)^{m_i - y},} \text{ para y = 0, 1, 2, ..., m.} \tag{2.7}$$

Colocando no formato da família exponencial linear como na equação (2.3),

$$\boldsymbol{f(y; m_i, \pi_i) = \exp\left[y \log\left(\frac{\pi_i}{1 - \pi_i}\right) + m_i \log(1 - \pi_i)\right] + \log\binom{r}{m_i}.}$$

O interesse nesses experimentos é encontrar quais doses serão as mais efetivas e para isso é modelada a proporção de sucesso $\boldsymbol{\pi_i}$. Para isso é necessário considerar dois aspectos que são a intensidade da dose e o estímulo que ela produzirá no indivíduo. Na farmacologia, o nível de intensidade com que acontece a reação de interesse é denominado de tolerância.

Se U é a variável aleatória que representa a tolerância para cada individuo, com função de densidade definida como *f(u)*. Então a probabilidade, $\boldsymbol{\pi}$, do indivíduo responder a dose que foi aplicada a ele é dada por,

$$\boldsymbol{\pi = P(U) = F(d) = \int_{-\infty}^{d} f(u)du.} \tag{2.8}$$

Essa distribuição é representada por curvas simétricas ou assimétricas, e a probabilidade de ocorrer sucesso depende das quantidades da dose aplicadas no individuo. Quanto maior a dose mais provável o sucesso. Assim, para valores de d pequenos a probabilidade de sucesso estará mais próxima de zero. As transformações Probito e a Logito permitem estabelecer um a relação linear da probabilidade de sucesso $\boldsymbol{\pi_i}$ com a dose $\boldsymbol{d_i}$.

2.3.1 A função Probito

Seja U com distribuição normal padrão com média $\mu \in R$ e variância $\sigma^2 > 0$, com f.d.p. definida como,

$$f(u;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(u-\mu)^2}{2\sigma^2}\right\}.$$

Então, tomando π_i como a probabilidade acumulada da distribuição normal padrão, tem-se,

$$\pi_i = P(U \leq d_i) = P\left(Z \leq -\frac{\mu}{\sigma} + \frac{1}{\sigma}d_i\right) = P(Z \leq \beta_0 + \beta_1 d_i),$$

sendo $\beta_0 = \frac{-\mu}{\sigma}$, $\beta_1 = \frac{1}{\sigma}$ e Z $= \frac{U-\mu}{\sigma}$ com distribuição N(0,1).

Então,

$$\pi_i = \Phi(\beta_0 + \beta_1 d_i),$$

sendo Φ a função de distribuição acumulada da normal padrão. E, aplicando a inversa da função de distribuição acumulada normal, chamada de Probito, à probabilidade π_i obtém-se uma função linear da dose d_i, (Cordeiro e Demétrio, 2008)

$$probit(\pi_i) = \Phi^{-1} = \beta_0 + \beta_1 d_i.$$

2.3.2 A função Logito

Se a variável U possui distribuição logística com a seguinte f.d.p.,

$$f_U(u;\mu,\tau) = \frac{\exp\left(\frac{u-\mu}{\tau}\right)}{\tau\left[1+\exp\left(\frac{u-\mu}{\tau}\right)\right]^2} \; ; \mu \in R, \tau > 0,$$

com $E(U) = \mu$ e $Var(U) = \pi^2\tau^2/3$.

Fazendo, $\beta_0 = -\mu/\tau$ e $\beta_1 = 1/\tau$, tem-se,

$$f_U(u;\beta_0,\beta_1) = \frac{\beta_1 e^{\beta_0+\beta_1 u}}{(1+e^{\beta_0+\beta_1 u})^2}.$$

Tomando π_i como a probabilidade acumulada da distribuição logística tem-se:

$$\pi_i = P(U_i) = F(d_i) = \frac{e^{\beta_0+\beta_1 d_i}}{1+e^{\beta_0+\beta_1 d_i}}.$$

Aplicando a função logito a π_i, obtém-se uma função linear da dose d_i: (Cordeiro e Demétrio, 2008)

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 d_i.$$

2.3.3 A função Complemento Log-log

Considerando que U possui distribuição Gumbel com f.d.p. dada por,

$$f_U(u;,\tau) = \frac{1}{\tau}\exp\left(\frac{u-\alpha}{\tau}\right)\exp\left[-\exp\left(\frac{u-\alpha}{\tau}\right)\right],\ \alpha \in R, \tau > 0,$$

com média $E(U) = \alpha + \gamma\tau$ e variância = $Var(U) = \pi^2\tau^2/6$, sendo γ o número de Euler.

Fazendo $\beta_0 = -\alpha/\tau$ e $\beta_1 = -1/\tau$, tem-se,

$$f_U(u, \beta_0\beta_1) = \beta_1 \exp\left\{\beta_0 + \beta 1u - e^{\beta_0+\beta_1 u}\right\}.$$

Aplicando tomando π_i como a probabilidade acumulada da distribuição Gumbel, obtém-se uma função da dose d_i, (Cordeiro e Demétrio, 2008)

$$\pi_i = P(U_i) = F(d_i) = 1 - \exp\left[-\exp\left(\beta_0 + \beta_1 d_i\right)\right].$$

As funções de ligação para a regressão de uma variável com distribuição binomial são Probito, Logito e Complemento log-log, apresentadas a seguir na tabela 2.2. Dentre essas três funções de ligação, a logito é canônica para a distribuição binomial.

Tabela 2.2: Funções de ligação do modelo binomial

	Função de ligação
Probito	$g(\mu_i) = \phi^-1(\pi_i)$
Logito	$g(\mu_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$
Complemento log-log	$g(\mu_i) = \log(-\log(1-\pi_i))$

A função de ligação logito tem a propriedade de produzir parâmetros interpretáveis. Assim, a função logito é a mais utilizada na regressão binomial. A regressão logística

é aplicada no presente trabalho para as variáveis resposta com distribuição binomial e multinomial. Quando o experimento tem interesse em predizer uma combinação linear das k variáveis preditoras, com k = 0, . . . , p, a regressão logística é,

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p. \tag{2.9}$$

O interesse nesse modelo é estimar a probabilidade de sucesso, ou seja, $P(Y = 1|x) = \pi(x)$, que pode ser escrita como,

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}. \tag{2.10}$$

Na regressão logística a interpretação do valor da E[Y]=π não é feita como no ML, em que o aumento em uma unidade na variável preditora resultava em um aumento linear na E[Y]. Na regressão logística, o aumento provocado por determinada variável preditora p é obtido de,

$$\exp(\beta_p) = \frac{\dfrac{\pi(x+1)}{1-\pi(x+1)}}{\dfrac{\pi(x)}{1-\pi(x)}}. \tag{2.11}$$

O valor em $\exp(\beta_p)$ é conhecido como razão de chances e compara a probabilidade de sucesso para x=1 com a probabilidade de sucesso para x.

É apresentada a seguir, uma aplicação da regressão logística em um experimento que visava obter a eficácia de um herbicida.

2.3.4 Aplicação a dados de Sobrevida de Plantas

Uma pequisa que visava determinar a eficácia do herbicida Picloram contra a planta Delphinum, popularmente conhecida como Esporinha-gigante, foi realizada na região central de Utah (EUA) nos anos de 1986 e 1987, foi descrita por Turner *et al.* (1992).

O experimento considerou o fator dosagem em quatro níveis: 0, 1.1, 2.2 e 4.5 kg/ha, sendo 0 o nível de referência. O experimento foi alocado a três localidades geográficas utilizando Blocos Completamente Aleatorizados (RCB) de forma desbalanceada, com tamanho amostral de 313 plantas. A frequência de cada observações dos dados pode ser observadas na tabela 2.3.

Tabela 2.3: Plantas mortas por dose

Bloco	Dose			
	0	1.1	2.2	4.5
1	0	11	28	24
2	0	11	18	31
3	0	9	24	27

Para análise do efeito da concentração do herbicida e dos blocos na sobrevivência ou não das plantas foi considerado o modelo logístico binomial. Foi escolhida a função de ligação logito para análise dos dados pela possibilidade interpretabilidade dos parâmetros estimados. O código utilizado para a estimação do modelo pode ser visto no Apêndice A. O modelo utilizado foi,

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_{1i}x_1 + \beta_{2j}x_2 + (\beta_{12})_{ij}x_1x_2,$$

$$Y_{ij} = \begin{cases} 1, \text{ se a planta morreu e} \\ 0, \text{ se a planta sobreviveu.} \end{cases}$$

β_0: proporção global;

β_{1i}: efeito do bloco i, com i= 1, 2 e 3;

β_{2j}: efeito da dosagem no nível j, com j= 1, 2, 3 e 4;

$(\beta_{12})_{ij}$: efeito de interação entre o bloco i e a dosagem j; i= 1, 2 e 3 ; j= 1, 2, 3 e 4;

A proporção de sucesso, ou seja, de plantas mortas pode ser obtida de,

$$\pi(x) = \frac{\exp(\beta_0 + \beta_{1i}x_1 + \beta_{2j}x_2 + (\beta_{12})_{ij}x_1x_2)}{1 + \exp(\beta_0 + \beta_{1i}x_1 + \beta_{2j}x_2 + (\beta_{12})_{ij}x_1x_2)}.$$

Da análise do ajuste do modelo acima o efeito de interação entre os fatores resultou não significativo. A seguir, foi ajustado uma regressão binomial sem considerar a interação dos fatores e o resultado está apresentado na tabela 2.4.

Tabela 2.4: Ajuste do modelo logístico binomial

	Estimado	$\exp(Estimado)$	Erro padrão
Intercepto	-20.1236	1.8216	2034.6408
Bloco 2	-0.5290	5.8919	0.5078
Bloco 3	-0.7150	4.8921	0.4700
Dose 1.1	20.1625	5.7078	2034.6408
Dose 2.2	22.6071	6.5780	2034.6408
Dose 4.5	41.1525	7.4529	2815.8384

Da tabela 2.4 pode-se concluir que, mantendo a demais variável preditora constante, estima-se que ao ser mudado o bloco em que a planta está localizada da região 1 para a 2 a chance da morte da planta aumente 489.19%.

Mantendo a demais variável preditora constante, estima-se que ao ser mudado o bloco em que a planta está localizada da região 1 para a 3 a chance da morte da planta aumente 389.21%.

Mantendo a demais variável preditora constante, estima-se que ao aumentar a dose de herbicida Picloram aplicado de 0 para $\mathbf{1.1}\boldsymbol{kg/ha^{-1}}$ a chance da morte da planta seja de 470.78%.

Mantendo a demais variável preditora constante, estima-se que ao aumentar a dose de herbicida Picloram aplicado de 0 para $\mathbf{2.2}\boldsymbol{kg/ha^{-1}}$ a chance da morte da planta seja de 557.80%.

Mantendo a demais variável preditora constante, estima-se que ao aumentar a dose de herbicida Picloram aplicado de 0 para $\mathbf{4.5}\boldsymbol{kg/ha^{-1}}$ a chance da morte da planta seja de 645.29%.

Pode ser concluído que as maiores diferenças de chance de morte da planta comparados aos níveis de referência ocorrem quando a planta está alocada na região 2 ou quando é aplicada a maior dose de herbicida.

2.4 Regressão Multinomial

O Modelo Multinomial Logístico, também chamado de Regressão Logística Multinomial ou Regressão Logística Politômica é uma extensão da Regressão Logística Binomial, que permite três ou mais categorias na variável resposta. A regressão logística multinomial é usada para prever a probabilidade de associação de cada categoria da variável resposta com o conjunto das variáveis explicativas. As variáveis explicativas podem ser dicotômicas ou contínuas.

Como exemplo das muitas utilizações da regressão logística multinomial temos: mostrar os fatores que influenciam na escolha aeroportuária (Abreu *et al.*, 2016), analisar dados de sinistralidade rodoviária (Vieira, 2019), analise de determinantes de votos (Nicolau, 2014), entre outros.

Enquanto que no modelo logístico binomial para a estimação de proporção de sucessos apenas uma regressão é necessária, na regressão multinomial com c categorias, $\boldsymbol{c-1}$

regressões devem ser ajustadas, para a estimação das $\boldsymbol{c-1}$ proporções de sucesso. Estas $\boldsymbol{c-1}$ regressões são estimadas como no modelo logístico binomial, sendo considerado como fracasso uma categoria pré-escolhida, sendo esse chamado valor de referência. A categoria que será utilizada como referência não interfere nas probabilidades de cada nível, porém as estimativas dos parâmetros serão afetadas nessa escolha.

Considerando como categoria de referência a categoria de valor 0, cada uma das respostas $\boldsymbol{Y_c}$ um modelo logístico multinomial é definido como,

$$\log\left[\frac{\boldsymbol{P(Y=c|x)}}{\boldsymbol{P(Y=0|x)}}\right] = \boldsymbol{\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_px_p} \text{ , c= 1, 2, ..., (c-1).} \quad (2.12)$$

Assim como no modelo logístico binomial o interesse é em estimar a probabilidade de sucesso $\boldsymbol{\pi}$ para cada categoria, que pode ser escrita como,

$$\boldsymbol{\pi_c(x)} = \frac{\boldsymbol{\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_px_p}}{\boldsymbol{1+\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_px_p}}. \quad (2.13)$$

Na regressão logística multinomial a interpretação do valor da E[Y]=$\boldsymbol{\pi}$ não é feita como no modelo de regressão logística binomial, por meio das razões de chance. O aumento provocado por determinada variável preditora p é obtido de,

$$\boldsymbol{\exp(\beta_p)} = \frac{\dfrac{\boldsymbol{\pi_c(x+c)}}{\boldsymbol{1-\pi_c(x+c)}}}{\dfrac{\boldsymbol{\pi_c(x)}}{\boldsymbol{1-\pi_c(x)}}}. \quad (2.14)$$

O valor em $\boldsymbol{\exp(\beta_p)}$ compara a probabilidade de sucesso para x=c com a probabilidade de sucesso para o nível de referência.

A seguir, são apresentados três aplicações da regressão logística multinomial. A primeira em dados de sobrevida de moscas doméstica, e a segunda e a terceira foram feitas em dados sobre pacientes diabéticos que possuíam níveis de retinopatia diabética.

2.4.1 Aplicação a dados de fase de Sobrevida de Moscas

Um experimento proposto por Itepan (1995), que visava verificar o nível de radiação gama que determinaria o tempo de aceitabilidade de moscas domesticas, considerando três períodos: se a pupa não foi aberta, se a mosca morreu antes de emergir completamente da pupa e se a mosca emergiu da pupa.

Foram considerados sete níveis de concentração de raios gama e para cada nível foram observadas 500 pupas, totalizando uma amostra de 3500 pupas. Os números de pupas abertas por cada nível de concentração nos períodos $\boldsymbol{P_i}$ são apresentados na tabela 2.5

Tabela 2.5: Número de pupas abertas por período

Nível	P_1	P_2	P_3	Total
80	62	5	433	500
100	94	24	382	500
120	179	60	261	500
140	335	80	85	500
160	432	46	22	500
180	487	11	2	500
200	498	2	0	500

A categoria tomada como referência para a estimação dos parâmetros foi a categoria 3. Os modelos que foram utilizados para ser analisado o efeito entre a concentração da radiação gama e a consequência na sobrevivência das larvas das moscas foi,

$$\log\left[\frac{\boldsymbol{P(Y=1|x)}}{\boldsymbol{P(Y=3|x)}}\right] = \boldsymbol{\beta_0 + \beta_1 x_1},$$

$$\log\left[\frac{\boldsymbol{P(Y=2|x)}}{\boldsymbol{P(Y=3|x)}}\right] = \boldsymbol{\beta_0 + \beta_1 x_1},$$

$$\boldsymbol{Y} = \begin{cases} 1, \text{ se a pupa não foi aberta} \\ 2, \text{ se a mosca morreu antes de emergir da pupa, e} \\ 3, \text{ se a mosca emergiu da pupa.} \end{cases}$$

$\boldsymbol{\beta_0}$: proporção global;

$\boldsymbol{\beta_{1i}}$: efeito do nível de concentração i da radiação gama, com i = 1, 2, 3, 4, 5, 6 e 7.

A proporção de sucesso para categoria 1, ou seja, se a pupa não foi aberta e a proporção de sucesso para categoria 2, ou seja, se a mosca morreu antes de emergir, são mostradas abaixo,

$$\boldsymbol{\pi_1(x)} = \frac{\boldsymbol{\exp(\beta_0 + \beta_{1i} x_1)}}{\boldsymbol{1 + \exp(\beta_0 + \beta_{1i} x_1)}},$$

$$\boldsymbol{\pi_2(x)} = \frac{\boldsymbol{\exp(\beta_0 + \beta_{1i} x_1)}}{\boldsymbol{1 + \exp(\beta_0 + \beta_{1i} x_1)}}.$$

Para a estimação do modelo foi feita utilizando a função *vglm* do pacote *VGAM* do software R Core Team (2019). O código utilizado para a estimação do modelo pode ser visto no Apêndice A. Os valores obtidos podem ser vistos na tabela 2.6 a seguir.

Tabela 2.6: Ajuste da regressão modelo multinomial

	Estimado	$\exp(\boldsymbol{Estimado})$	Erro padrão	Valor z	p-valor
Intercepto$_1$	-1.9436	1.4319	0.1358	-14.313	<0.0001
Concentração 100$_1$	0.5415	1.7185	0.1780	3.041	0.0023
Concentração 120$_1$	1.5665	4.7900	0.1669	9.386	<0.0001
Concentração 140$_1$	3.3151	2.7525	0.1822	18.197	<0.0001
Concentração 160$_1$	4.9210	1.3714	0.2573	19.125	<0.0001
Concentração 180$_1$	7.4387	1.7006	0.7215	10.311	<0.0001
Concentração 200$_1$	22.0421	3.7392	628.8092		
Intercepto$_2$	-4.4613	1.1547	0.4498	-9.919	<0.0001
Concentração 100$_2$	1.6939	5.4408	0.4966	3.411	<0.0001
Concentração 120$_2$	2.9911	1.9908	0.4720	6.337	<0.0001
Concentração 140$_2$	4.4007	8.1506	0.4760	9.245	<0.0001
Concentração 160$_2$	5.1989	1.8107	0.5191	10.015	<0.0001
Concentração 180$_2$	6.1660	4.7630	0.8906	6.923	<0.0001
Concentração 200$_2$	19.0424	1.8621	628.8098	0.030	0.975841

Da tabela 2.6 pode-se concluir que, estima-se que ao ser aumentado o nível da radiação de 80 para 100, a chance da mosca não emergir aumente em 71.85% e da mosca morrer antes de emergir completamente aumente em 444.08%.

Estima-se que ao ser aumentado o nível da radiação de 80 para 120, a chance da mosca não emergir aumente em 379% e da mosca morrer antes de emergir completamente aumente em 99.08%.

Estima-se que ao ser aumentado o nível da radiação de 80 para 140, a chance da mosca não emergir aumente em 175.25% e da mosca morrer antes de emergir completamente aumente em 715.06%.

Estima-se que ao ser aumentado o nível da radiação de 80 para 160, a chance da mosca não emergir aumente em 37.14% e da mosca morrer antes de emergir completamente aumente em 81.07%.

Estima-se que ao ser aumentado o nível da radiação de 80 para 180, a chance da mosca não emergir aumente em 70.06% e da mosca morrer antes de emergir completamente aumente em 376.30%.

Estima-se que ao ser aumentado o nível da radiação de 80 para 200, a chance da mosca não emergir aumente em 273.92% e da mosca morrer antes de emergir completamente aumente em 86.21%.

Pode ser concluído que a maior razão de chance da mosca não emergir da pupa ocorre na transição do nível de radiação de 80 para 120 e a maior razão de chance da mosca morrer antes de emergir completamente ocorre na transição de 80 para 140.

2.4.2 Aplicação a dados de Retinopatia Diabética

A diabetes é uma doença comum que acomete pessoas independente da idade ou do gênero. Ela é causada pela ausência ou má absorção de insulina nas células, o que provoca o desequilíbrio da glicose no sangue. O tratamento incorreto ou não realizado faz com que a doença progrida e provoque outras doenças tais como, lesão nervosa, lesão renal, infecções, danos hepáticos e retinopatia. Esta última doença é a temática da próxima aplicação.

O banco de dados utilizado é uma amostra de tamanho 2049, do estudo *Wisconsin Epidemiologic Study of Diabetic Retinopatia*. Esse estudo foi realizado em Wisconsin (EUA) na década de 80 com 10135 pacientes diagnosticados com diabetes. Dos pacientes foram compiladas as características: idade, tempo em anos desde que foi diagnosticado com diabetes, percentual hemoglobina glicada, quantidade de proteína na urina e uso ou não de insulina. A variável resposta considera a ausência ou o grau de retinopatia em cada um dos olhos.

No presente capítulo foi realizada duas regressões multinomial logísticas para esses dados, considerando como resposta o nível de retinopatia em um determinado olho para cada paciente. Estas duas análises consideram que duas observações obtidas no mesmo paciente são independentes. No próximo capítulo a utilização desse conjunto de dados será feita considerando as respostas de ambos os olhos e, como consequência incorporarão a correlação na estimação do parâmetro.

A categoria tomada como referência para a estimação dos parâmetros foi a categoria 3. As regressões logísticas multinomial utilizada para esse conjunto de dados consideraram os seguinte modelos,

$$\log\left[\frac{\boldsymbol{P(Y=1|\mathrm{x})}}{\boldsymbol{P(Y=3|\mathrm{x})}}\right] = \boldsymbol{\beta_0 + \beta_{1i}x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5},$$

$$\log\left[\frac{\boldsymbol{P(Y=2|\mathrm{x})}}{\boldsymbol{P(Y=3|\mathrm{x})}}\right] = \boldsymbol{\beta_0 + \beta_{1i}x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5},$$

em que,

$$\boldsymbol{Y} = \begin{cases} 1, \text{ para a ausência de retinopatia} \\ 2, \text{ para a retinopatia moderada, e} \\ 3, \text{ para a retinopatia severa.} \end{cases}$$

$\boldsymbol{\beta_0}$: proporção global;

$\boldsymbol{\beta_1}$: efeito da idade do paciente;

$\boldsymbol{\beta_2}$: efeito da duração em anos da diabete no paciente;

$\boldsymbol{\beta_3}$: efeito do percentual de hemoglobina glicada no paciente;

$\boldsymbol{\beta_4}$: efeito de proteína na urina;

$\boldsymbol{\beta_5}$: efeito do uso de insulina.

A proporção de sucesso para categoria 1, ou seja, a ausência de retinopatia,

$$\boldsymbol{\pi_1(x)} = \frac{\boldsymbol{\exp(\beta_0 + \beta_{1i}x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5)}}{\boldsymbol{1 + \exp(\beta_0 + \beta_{1i}x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5)}}.$$

A proporção de sucesso para categoria 2, ou seja, a retinopatia moderada,

$$\boldsymbol{\pi_2(x)} = \frac{\boldsymbol{\exp(\beta_0 + \beta_{1i}x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5)}}{\boldsymbol{1 + \exp(\beta_0 + \beta_{1i}x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5)}}.$$

Considerando as duas variáveis resposta, a primeira regressão ajustada foi referente a resposta obtida no olho esquerdo do paciente e a segunda regressão para a resposta referente ao olho direito. Para estimação do modelo de cada uma das variáveis resposta foi utilizada a função *vglm* do pacote *VGAM* do software R Core Team (2019). O código utilizado para a estimação do modelo pode ser visto no Apêndice A. Os valores obtidos podem ser vistos na tabela 2.7 a seguir.

Tabela 2.7: Ajuste da regressão multinomial para o olho esquerdo

	Estimado	$\boldsymbol{\exp(Estimado)}$	Erro padrão	Valor z	p-valor
Intercepto_1	8.0059	29998.5687	0.68630	11.665	<0.0001
Idade_1	0.0190	1.0192	0.0054	3.5320	0.0004
Diabetes_1	-0.2197	0.8028	0.0126	-17.3770	<0.0001
Hemoglobina_1	-0.1597	0.8523	0.0384	-4.1560	<0.0001
Proteína_1	-0.9914	0.3710	0.0837	-11.8450	<0.0001
Insulina_1	-1.0875	0.3370	0.3518	-3.0910	0.0020
Intercepto2	4.0723	58.6957	0.6317	6.4470	<0.0001
Idade2	0.0170	1.0172	0.0049	3.4560	0.0005
Diabetes2	-0.0903	0.9136	0.0097	-9.2580	<0.0001
Hemoglobina2	-0.0290	0.9714	0.0344	-0.8420	0.3998
Proteína2	-0.5652	0.5683	0.0592	-9.5490	<0.0001
Insulina2	-0.4510	0.6370	0.3407	-1.3240	0.1856

Da tabela 2.7 pode-se concluir que, mantendo as demais variáveis preditoras constantes, estima-se que a cada aumento de 1 ano de idade, a chance de o nível de retinopatia passar de ausente para severa aumente em 1.92% e de passar de moderada para severa

aumente em 1.72%.

Mantendo as demais variáveis preditoras constantes, estima-se que a cada aumento de 1 ano de duração da diabetes, a chance de o nível de retinopatia passar de ausente para severa diminua em 20% e de passar de moderada para severa diminua em 8.64%.

Mantendo as demais variáveis preditoras constantes, estima-se que a cada aumento de 1 porcento da e hemoglobina glicada, a chance de o nível de retinopatia passar de ausente para severa diminua em 15% e de passar de moderada para severa diminua em 3%.

Mantendo as demais variáveis preditoras constantes, estima-se que a cada aumento de 1 porcento de proteína na urina, a chance de o nível de retinopatia passar de ausente para severa diminua em 67% e de passar de moderada para severa diminua em 44%.

Mantendo as demais variáveis preditoras constantes, estima-se que ao paciente tomar insulina, a chance de o nível de retinopatia passar de ausente para severa diminua em 65% e de passar de moderada para severa diminua em 66%.

Tabela 2.8: Ajuste da regressão multinomial para o olho direito

	Estimado	**exp(*Estimado*)**	Erro padrão	Valor z	p-valor
$\text{Intercepto}_{\mathbf{1}}$	7.7932	2424.1064	0.6664	11.6940	<0.0001
$\text{Idade}_{\mathbf{1}}$	0.0140	1.0141	0.0052	2.6750	0.0075
$\text{Diabetes}_{\mathbf{1}}$	-0.2113	0.8095	0.0123	-17.0730	<0.0001
$\text{Hemoglobina}_{\mathbf{1}}$	-0.1469	0.8633	0.0377	-3.9000	<0.0001
$\text{Proteína}_{\mathbf{1}}$	-0.9205	0.3983	0.0810	-11.359	<0.0001
$\text{Insulina}_{\mathbf{1}}$	-1.1061	0.3308	0.3328	-3.3240	0.0009
$\text{Intercepto}_{\mathbf{2}}$	3.9217	50.4876	0.6134	6.3930	<0.0001
$\text{Idade}_{\mathbf{2}}$	0.0141	1.0141	0.0048	2.9330	0.0033
$\text{Diabetes}_{\mathbf{2}}$	-0.0750	0.9277	0.0094	-7.9970	<0.0001
$\text{Hemoglobina}_{\mathbf{2}}$	-0.0354	0.9652	0.0339	-1.0460	0.2955
$\text{Proteína}_{\mathbf{2}}$	-0.5648	0.5684	0.0582	-9.6950	<0.0001
$\text{Insulina}_{\mathbf{2}}$	-0.3838	0.6812	0.3222	-1.1910	0.2336

Da tabela 2.8 pode-se concluir que, mantendo as demais variáveis preditoras constantes, estima-se que a cada aumento de 1 ano de idade, a chance de o nível de retinopatia ir de ausente para severa aumente em 1.41% e de ir de moderada para severa aumente em 1.41%

Mantendo as demais variáveis preditoras constantes, estima-se que a cada aumento de 1 ano de duração da diabete, a chance de o nível de retinopatia ir de ausente para severa diminua em 20% e de ir de moderada para severa diminua em 8%.

Mantendo as demais variáveis preditoras constantes, estima-se que a cada aumento de 1 porcento da e hemoglobina glicada, a chance de o nível de retinopatia ir de ausente

para severa diminua em 14% e de ir de moderada para severa diminua em 7%.

Mantendo as demais variáveis preditoras constantes, estima-se que a cada aumento de 1 porcento de proteína na urina, a chance de o nível de retinopatia ir de ausente para severa diminua em 61% e de ir de moderada para severa diminua em 44%.

Mantendo as demais variáveis preditoras constantes, estima-se que ao paciente tomar insulina, a chance de o nível de retinopatia ir de ausente para severa diminua em 67% e de ir de moderada para severa diminua em 32%.

A significância das estimativas dos parâmetros do olho esquerdo e do olho direito foram similares, com as variáveis preditoras Hemoglobina e Insulina não significantes para a categoria referente a retinopatia moderada. Apesar de em ambas as regressões as variáveis significativas serem as mesmas, não é adequado estimar dois modelos diferentes pois as respostas são correlacionadas por serem do mesmo paciente.

2.5 Considerações

Da regressão binomial aplicada nos dados de análise do herbicida Picloram, foi obtido que a regressão binomial com função logística é adequada para esse dados e, a partir dela, pode ser observado que o herbicida é eficaz para seu propósito. Níveis de doses mais altos produzem maiores valores de razões de chance de morte de planta. E além disso, plantas alocadas na região 2 obtiveram maior valores de razão de chances.

Da aplicação da regressão multinomial nos dados de sobrevida das moscas, foi obtido que maiores valores de razões de chance da mosca não emergir da pupa ocorre na transição do nível de radiação de 80 para 120 e a maior razão de chance da mosca morrer antes de emergir completamente ocorre na transição de 80 para 140.

Para os dados de pacientes diabéticos a significância das estimativas dos parâmetros do olho esquerdo e do olho direito foram similares, com as variáveis preditoras Hemoglobina e Insulina não significantes para a categoria referente a retinopatia moderada.

Capítulo 3

Regressão Multinomial Bivariada

Em alguns estudos podem ser observadas duas ou mais variáveis resposta para cada unidade experimental. Como exemplos temos os estudos genéticos (Cuyabano *et al.*, 2010) e estudos epidemiológicos tais quais o Estudo de Wiscosin sobre retinopatia (Klein *et al.*, 1989).

Em analogia ao caso univariado, os modelos lineares generalizados multivariados podem ser considerados como tendo os mesmos componentes: o componente aleatório que é baseados em uma distribuição pressuposta, o componente sistemático e uma função de ligação.(Tutz, 2011)

No modelo bivariado para cada indivíduo há duas respostas, e o objetivo é encontrar a relação de ambas as variáveis repostas com as variáveis explicativas. As variáveis respostas podem ser escritas da seguinte forma,

$$\boldsymbol{Y_{hi}} = \boldsymbol{\eta} = \boldsymbol{X^t \beta},$$

com,

$$\boldsymbol{Y_{hi}} = \begin{cases} 1, \text{ se a varável receber a categoria resposta de interesse do estudo;} \\ 0, \text{ caso contrário. com i=1,..,n e h=1,2} \end{cases}$$

Nas seções a seguir são apresentados o Modelo Linear Condicional com Efeitos Fixos, o Modelo Marginal e o Modelo Linear Generalizado Misto. Essas três metodologias conseguem modelar a correlação entre as variáveis resposta na estimativa dos seus parâmetros.

A aplicação dos modelos apresentados nas seções a seguir será feita em dois bancos de dados diferentes, um que possui as informações sobre pacientes diagnosticados com artrite reumatoide e o outro que possui as informações sobre dados visuais de pacientes

diabéticos. Ambos os bancos de dados podem ser encontrados no software R Core Team (2019), sendo que o primeiro se encontra no pacote *gee* e o segundo no pacote *gss*.

3.1 Modelo Linear Condicional com Efeitos Fixos

Seja a variável resposta $\boldsymbol{Y_1}$, para i indivíduos com i=1,...n, com distribuição multinomial com $\boldsymbol{J_1}$ categorias e a variável $\boldsymbol{Y_2}$, também uma variável resposta para os mesmos i indivíduos, porém com $\boldsymbol{J_2}$ categorias. Podemos denotar os vetores das variáveis resposta como $\boldsymbol{Y_1} = (\boldsymbol{y_{11}}, \boldsymbol{y_{12}}, ..., \boldsymbol{y_{1J_1}})^t$ e $\boldsymbol{Y_2} = (\boldsymbol{y_{21}}, \boldsymbol{y_{22}}, ..., \boldsymbol{y_{2J_2}})^t$.

De acordo com Sun (2014), podemos mostrar as probabilidades conjuntas das variáveis como apresentada na tabela 3.1

Tabela 3.1: Probabilidades Conjuntas das $\mathrm{Y_1} e \boldsymbol{Y_2}$

$\mathrm{Y_1}\vert\boldsymbol{Y_2}$	1	$\cdots$	j	$\cdots$	$\mathrm{J_2}$
1	$\boldsymbol{\pi_{i11}}$	$\cdots$	$\boldsymbol{\pi_{i1j}}$	$\cdots$	$\boldsymbol{\pi_{i1J_2}}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
j	$\boldsymbol{\pi_{ij1}}$	$\cdots$	$\boldsymbol{\pi_{ijj}}$	$\cdots$	$\boldsymbol{\pi_{ijJ_2}}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$\mathrm{J_1}$	$\boldsymbol{\pi_i}$	$\cdots$	$\boldsymbol{\pi_i}$	$\cdots$	$\boldsymbol{\pi_i}$

E a probabilidade conjunta das variáveis resposta $\boldsymbol{Y_1}$ e $\boldsymbol{Y_2}$ resultam ser:

$$\boldsymbol{\pi_{j_1 j_2}} = \boldsymbol{P}(\boldsymbol{y_{i1}} = \boldsymbol{y_{i1}}, \boldsymbol{y_{i2}} = \boldsymbol{y_{i2}}) = \frac{\boldsymbol{exp}(\boldsymbol{\alpha_{j_1}} + \boldsymbol{\beta_{j_2}} + \boldsymbol{\gamma_{j_1 j_2}})}{\sum_{\boldsymbol{j_1}=1}^{\boldsymbol{J_1}} \sum_{\boldsymbol{j_2}=1}^{\boldsymbol{J_2}} \boldsymbol{exp}(\boldsymbol{\alpha_{j_1}} + \boldsymbol{\beta_{j_2}} + \boldsymbol{\gamma_{j_1 j_2}})},$$

com:

$\boldsymbol{\beta_{j_2}}$ o efeito da $\boldsymbol{j_2}$-ésima categoria; $\boldsymbol{\alpha_{j_1}}$ o efeito da $\boldsymbol{j_1}$-ésima categoria; e $\boldsymbol{\gamma}$ denota a correlação entre as variáveis X e Y.

Segundo Uddin e Begum (2018), tem-se que a função de verossimilhança para o caso de m respostas e com d categorias pode ser escrita como,

$$\boldsymbol{P}(\boldsymbol{Y_1} = \boldsymbol{y_1}, ..., \boldsymbol{Y_m} = \boldsymbol{y_m}, \boldsymbol{X} = \boldsymbol{x}) = \prod_{\boldsymbol{i_1}=\boldsymbol{0}}^{\boldsymbol{d-1}} \cdots \prod_{\boldsymbol{i_m}=\boldsymbol{0}}^{\boldsymbol{d-1}} \boldsymbol{P}_{\boldsymbol{i_1,...,i_m}}^{\prod_{\boldsymbol{k}=\boldsymbol{1}}^{\boldsymbol{b}} (\boldsymbol{q(w)})}, \tag{3.1}$$

com q(w)=$\boldsymbol{1 - w_k^{(i_1,...,i_m)}} + \boldsymbol{(-1)^{1-w_k^{(i_1,...,i_m)}}} \boldsymbol{z_k}$.

Ainda, os autores mostraram que quando tem-se 2 variáveis respostas cada uma com

3 categorias, a distribuição conjunta é,

$$\boldsymbol{P(Y_1 = y_1, Y_2 = y_2) = P_{00}^{(1-z_1)(1-z_2)(1-z_3)(1-z_4)} P_{01}^{(1-z_1)(1-z_2)(1-z_3)(1-z_4)} *}$$

$$\boldsymbol{* P_{02}^{(1-z_1)(1-z_2)(1-z_3)(1-z_4)} P_{10}^{(1-z_1)(1-z_2)(1-z_3)(1-z_4)} *} \tag{3.2}$$

$$\boldsymbol{* P_{11}^{(1-z_1)(1-z_2)(1-z_3)(1-z_4)} P_{12}^{(1-z_1)(1-z_2)(1-z_3)(1-z_4)}.}$$

A dependência condicional das variáveis $\boldsymbol{Y_1}$ e $\boldsymbol{Y_2}$ pode ser dada pelo modelo,

$$\boldsymbol{P(Y_{2,x}|y_1, x) = \frac{\exp(x^t\beta + \alpha y_1)}{1 + \exp(x^t\beta + \alpha y_1)}}, \tag{3.3}$$

onde $\boldsymbol{y_1}$; são os valores observados da variável resposta $\boldsymbol{Y_{1,x}}$, X são as variáveis explicativas, $\boldsymbol{\beta}$ é um vetor de parâmetros a ser estimado, e $\boldsymbol{\alpha_1}$ parâmetro do populacional.

A estimação dos parâmetros $\boldsymbol{\beta}$ 3.3 pode ser obtida a partir de rotina criada por Uddin e Begum (2018) no software MATLAB. A seguir é apresentado a a plicação dessa metodologia nos dados de dos pacientes diabéticos apresentados no Capítulo 2 .

3.1.1 Aplicação a dados de Retinopatia Diabética

Para a aplicação do modelo condicional de efeitos fixos considerou-se a mesma pesquisa sobre retinopatia diabética, referida no capítulo anterior na seção 2.4,que registrou a variável resposta para os dois olhos, sendo $\boldsymbol{Y_1}$ para o olho esquerdo e $\boldsymbol{Y_2}$ a variável para o olho direito. Com as categorias das variáveis resposta:

$$\boldsymbol{Ym} = \begin{cases} 1, \text{ para a ausência de retinopatia naquele olho} \\ 2, \text{ para a retinopatia moderada naquele olho, e} \\ 3, \text{ para a retinopatia severa naquele olho.} \end{cases}$$

Os valores estimados para os parâmetros do modelo linear condicionado de efeitos fixos foram obtidos mediante da rotina proposta por Uddin e Begum (2018) para o software MATLAB e podem ser vistos na tabela 3.2. O código utilizado para a estimação do modelo pode ser visto no Apêndice A.

Tabela 3.2: Parâmetros estimados para o modelo condicional

	Idade	Diabetes	Glicose	Upro	Insulina	y_j
Y_{1}	6,5852	2,5511	2,3442	1,0641	1,0845	-
Y_{2}	4,1302	1,8158	1,7363	1,0370	1,0445	-
Y_{01}	5,7283	2,2580	2,1149	1,0524	1,0686	0
Y_{11}	1,8047	1,2828	1,2096	1,0093	1,0149	1
Y_{21}	1,0530	1,0240	1,0131	1,0015	1,0010	2
Y_{02}	5,9170	2,4614	2,1948	1,0710	1,0763	0
Y_{12}	-0,1823	0,6429	0,7069	0,9878	0,9807	1
Y_{22}	1,4864	1,2503	1,1389	1,0187	1,0107	2

Considerando como exemplo o caso em que o paciente diabético tenha como resposta no olho esquerdo a categoria 2, considerada como retinopatia severa, e no olho direito também a categoria 2, considerada como retinopatia severa, a proporção estimada pode ser obtida de,

$$\boldsymbol{log\left(\frac{\pi_2}{1-\pi_2}\right) = 1,4864x_1 + 1,2503x_2 + 1,1389x_3 + 1,0187x_4 + 1,0107x_5.}$$

Considerando como exemplo o caso em que o paciente diabético tenha como resposta no olho esquerdo a categoria 1, considerada como retinopatia moderada, e no olho direito na categoria 0 considerada como a ausência de retinopatia, a resposta estimada para a variável resposta 2 é dada por

$$\boldsymbol{log\left(\frac{\pi_0}{1-\pi_0}\right) = 5,7283x_1 + 2,2580x_2 + 2,1149x_3 + 1,0524x_4 + 1,068x_5.}$$

3.2 Modelo Marginal

Modelos marginais foram propostos por Liang e Zeger (1986) como uma extensão dos MLG's, para permitir que nessa modelagem a correlação entre as variáveis resposta seja levada em consideração.

A proposta inicial dessa abordagem era a utilização em dados longitudinais. Nesse tipo de modelagem o objetivo é a utilização de um MLG para a distribuição marginal de $\mathbf{Y_i}$. A estimação utilizada é diferente da apresentada no capítulo 2 e é conhecida como Equações de Estimação Generalizadas (EEG), sendo a extensão multivariada do método de estimação por quase-verossimilhança, técnica proposta por Wedderburn em 1974. Para o caso em que as variáveis respostas possuem distribuição normal, a estimação utilizando o método EEG se reduz ao método de Máxima Verossimilhança do MLG apresentado no

Capítulo 2.

A equação do Modelo Marginal é da forma:

$$g(E[y_{im}]) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}, \tag{3.4}$$

Seja y_{im} a m-ésima resposta observada para o i-ésimo individuo, e considerando a matrix X com as q variáveis preditoras, um modelo marginal E(Y_{im}) = μ_{im}, Var(Y_{im}) $=\frac{V(\mu_{im})}{\phi\omega_{it}}$, em que V_i = V (μ_{im}) é a função de variância e $\phi^{-}1 > 0$ ($\phi > 0$) é o parâmetro de dispersão. Além dessas propriedades, a correlação entre Y_{im} e Y_{il}, com m $\neq$ l, é uma função do valor médio marginal α, o vetor paramétrico que descreve o padrão de correlação. (Fitzmaurice *et al.*, 2012)

O método EEG possui suposições fracas e fornece estimadores consistentes de β e da matriz de covariância para a regressão marginal mesmo que α não seja especificado adequadamente. (Liang e Zeger, 1986)

3.2.1 Estimação dos Parâmetros

Para a estimação do modelo marginal são utilizadas EEG's. Tem se em Touloumis *et al.* (2013) que o estimador $\hat{\beta}_G$ é a solução das seguintes equações de estimação,

$$U(\beta, \alpha) = \frac{1}{n}\sum_{i=1}^{n} D_i V_i^{-1}(Y_i - \pi_i) = 0, \tag{3.5}$$

onde $D_i = \frac{\partial \pi_i}{\partial \beta}$ e $V_i = V_i(\beta, \alpha)$ é a matriz de variância e covariâncias das variáveis Y.

E $R_i(\alpha)$ é a matriz que recebe o nome matriz de correlação de trabalho, (*working correlation matrix*). A matriz $R_i(\alpha)$ tem como objetivo oferecer uma estrutura de correlação entre as observações Y_1 e Y_2 de um mesmo indivíduo. Ela é da forma,

$$R_i(\alpha) = \begin{bmatrix} 1 & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & 1 & \dots & \alpha_{21} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{n1} & \alpha_{n2} & \dots & 1 \end{bmatrix}$$

Quando α é desconhecido, Liang e Zeger (1986) mostraram que $\hat{\alpha}$ é um estimador $\sqrt{n}$-consistente. Na tabela 3.3, são apresentados os valores mais comuns de α para

$\boldsymbol{R_i(\alpha)}$.

Tabela 3.3: Valores $\boldsymbol{\alpha}$ para tipos de matriz correlação de trabalho

Estrutura de R	$\boldsymbol{\alpha_{mo} = corr(y_{im}, y_{il})}$
Independente	0
Uniforme	$\boldsymbol{\alpha}$
Autoregressiva (1)	$\boldsymbol{\alpha^{\lvert m-l \rvert}}$
Não estruturada	$\boldsymbol{\alpha_{m-l}}$

O algoritmo de estimação de $\boldsymbol{\beta}$ pode ser resumido nos seguintes quatro passos: (Galdino, 2015)

1. Ajustar um MLG considerando que as observações são independentes para obter a estimativa inicial de $\boldsymbol{\beta}$;

2. Calcular os valores da correlação $\boldsymbol{\alpha}$, da matriz de correlação $\boldsymbol{R_i(\alpha)}$, a variância $\boldsymbol{V_i(a)}$, o resíduo de Pearson $\boldsymbol{e_{im}}$ e o parâmetro de dispersão $\boldsymbol{\phi}$, sendo:

$$\boldsymbol{\alpha} = \left(\frac{\mathbf{1}}{\boldsymbol{n\hat{\phi}}}\right) \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{e_{im}e_{il}},$$

$$\boldsymbol{V_i(a)} = \boldsymbol{\phi(A_i^{1/2} R_i(\alpha) A_i^{1/2})},$$

$$\boldsymbol{e_{im}} = \frac{\boldsymbol{(y_{im} - \hat{\pi}_{im})}}{\sqrt{\boldsymbol{var(\hat{\pi}_{im})}}}$$

$$\boldsymbol{\phi} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \frac{\mathbf{1}}{\boldsymbol{M_i}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{M_i}} \boldsymbol{e_{im}^2};$$

3. Calcular a estimativa $\hat{\boldsymbol{\beta}}^{(\boldsymbol{m+1})}$

$$\hat{\boldsymbol{\beta}}^{(\boldsymbol{m+1})} = \hat{\boldsymbol{\beta}}^{(\boldsymbol{m})} - \left[\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{\hat{D}_i V_i^{-1} \hat{D}_i^t}\right]^{-1} * \left[\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{\hat{D}_i V_i^{-1} (Y_i - \hat{\pi}_i)}\right];$$

4. Voltar ao passo 2 e repetir o processo até convergir.

3.2.2 Modelo Marginal Multinomial

Considerando variáveis respostas categóricas ordinais, o modelo marginal logístico cumulativo é apresentado a seguir:

$$\boldsymbol{log}\left(\frac{\boldsymbol{P(Y_{im} \leq j | X_{im})}}{\boldsymbol{P(Y_{im} > j | X_{im})}}\right) = \boldsymbol{\alpha_j + X_{im}\beta}, \qquad \boldsymbol{j = 1, \ldots, J-1.} \tag{3.6}$$

Por meio de simulações, mostra-se que a utilização de estrutura de associação entre as respostas multinomiais correlacionadas por meio das odds ratio locais é um método mais flexível do que os métodos até então existentes. As estruturas de odds ratios locais podem ser vistas na tabela 3.4.

Tabela 3.4: Estruturas de odds ratios locais marginalizadas

$\log \boldsymbol{\theta}_{tjt'j'}$	Estrutura
$\boldsymbol{\phi}$	Uniforme
$\boldsymbol{\phi}^{(t,t')}$	Permutabilidade de categoria
$\boldsymbol{\phi}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_{j+1})(\boldsymbol{\mu}_{j'} - \boldsymbol{\mu}_{j'+1})$	Permutabilidade de tempo
$\boldsymbol{\phi}(\boldsymbol{\mu}_j^{(t,t')} - \boldsymbol{\mu}_{j+1}^{(t,t')})(\boldsymbol{\mu}_{j'}^{(t,t')} - \boldsymbol{\mu}_{j'+1}^{(t,t')})$	RC

Para respostas categóricas ordinais todas as quatro estruturas podem ser aplicadas. Em contrapartida, quando as respostas são nominais apenas as estruturas de permutabilidade de tempo e RC podem ser empregadas.

Além disso, há recomendações da utilização das estruturas permutabilidade de tempo para amostra de tamanhos pequenos, para demais tamanhos de amostras é necessário a utilização dos parâmetros intrínsecos estimados de L para a escolha da melhor estrutura de Odds Ratios Locais. (Touloumis *et al.*, 2013)

Para a estimação dos por meio de Odds Ratios Locais (Touloumis, 2014) apresentou o pacote do pacote *multgee* do software R Core Team (2019). Este pacote é utilizado para a aplicação apresentada na próxima seção.

3.2.3 Aplicação a dados de Artrite Reumatoide

Artrite reumatoide é uma doença inflamatória crônica comum que afeta o revestimento das articulações. É uma enfermidade sem cura e a única forma de ajudar a adiar a progressão da doença é por meio de fisioterapia ou por meio de medicamentos.

Um experimento duplo-cego aleatorizado descrito por Bombardier *et al.* (1986), com 303 pacientes de 14 clínicas dos Estados Unidos da América e Canadá, realizado no ano de 1986, buscava avaliar a dor em pacientes pré-diagnósticos com atrite reumatoide. Foram avaliadas as seguintes características dos pacientes: sexo, idade, além de um score de autoavaliação de artrite reumatoide no inicio dos tratamentos. Estas características foram avaliadas em dois momentos, após um mês de tratamento e após três meses.

Os pacientes foram divididos em dois grupos, com um recebendo como tratamento Auranofin e outro grupo recebendo placebo. As variáveis respostas obtidas foram cate-

gorizadas em 3 scores (ruim, regular, bom) também autoavaliativo, em dois momentos diferentes. Apesar de o banco de dados inicialmente conter 303 pacientes, o seu tamanho foi reduzido para 287 pacientes após serem removidas observações faltantes.

A equação do modelo marginal considerada para esse conjunto de dados é:

$$\boldsymbol{log}\left(\frac{\boldsymbol{P(Y_{im} \leq j|X_{im})}}{\boldsymbol{P(Y_{im} > j|X_{im})}}\right) = \boldsymbol{\alpha_j + X_{im}\beta}, \qquad \boldsymbol{j = 1, 2, 3.}$$

tal que,

$$\boldsymbol{Y} = \begin{cases} 1, \text{ para o score ruim;} \\ 2, \text{ para o score regular;} \\ 3, \text{ para o score bom.} \end{cases}$$

$\boldsymbol{\beta_{0j}}$: proporção global para categoria j;

$\boldsymbol{\beta_1}$: efeito do gênero, sendo $\boldsymbol{X_1}$=0 do sexo masculino e $\boldsymbol{X_1}$=1 do sexo feminino;

$\boldsymbol{\beta_2}$: efeito da idade do paciente;

$\boldsymbol{\beta_3}$: efeito do tratamento recebido, com 0 sendo o placebo e 1 Auranofin.

$\boldsymbol{\beta_4}$: efeito do score inicial atribuído pelo paciente, com 1 sendo o score ruim 2 sendo o score regular, e 3 o score bom.

Para estimar o modelo marginal 3.6, foi utilizado o pacote *multgee* do software R Core Team (2019). O código utilizado para a estimação do modelo pode ser visto no Apêndice A. A estrutura de correlação considerada para a matriz de correlação R foi a estrutura uniforme, escolhida por meio dos parâmetros intrínsecos estimados de L.

Os parâmetros referentes a idade e ao gênero não resultaram significativos. O modelo ajustado sem esses dois preditores teve como resultado do ajuste os valores apresentados na tabela 3.5.

Tabela 3.5: Parâmetros estimados para o modelo marginal

	Estimado	exp(Estimado)	Erro padrão	Valor z	p-valor
$\boldsymbol{Intercepto_1}$	-0.3752	0.6872	0.2019	-1.8578	0.0632
$\boldsymbol{Intercepto_2}$	1.6294	5.1006	0.2299	7.0872	<0.0001
$\boldsymbol{Tratamento}$	-0.5166	0.5966	0.1866	-2.7676	0.0056
$\boldsymbol{Score_2}$	-0.6569	0.5184	0.2175	-3.0203	0.0025
$\boldsymbol{Score_3}$	-2.2743	0.1029	0.3063	-7.4238	<0.0001

Da tabela 3.5 pode-se concluir que, mantendo o Score inicial constante, estima-se que o Score de dor aumente em 59% nos pacientes que receberam placebo com relação aos que receberam Auranofin.

E, estima-se que o Score de dor aumente em 51.84% quando passamos do paciente que deu um Score ruim inicial para um paciente que deu score regular e em 10.29% quando passamos do paciente que deu um Score ruim inicial para um paciente que deu score bom, mantendo o tratamento constante.

3.2.4 Aplicação a dados de Retinopatia Diabética

O modelo marginal, para esse conjunto de dados é:

$$\boldsymbol{log}\left(\frac{\boldsymbol{P(Y_{im} \leq j|X_{im})}}{\boldsymbol{P(Y_{im} > j|X_{im})}}\right) = \boldsymbol{\alpha_j + X_{im}\beta}, \qquad \boldsymbol{j = 1, 2, 3.}$$

em que,

$$\boldsymbol{Y} = \begin{cases} 1, \text{ para a ausência de retinopatia naquele olho} \\ 2, \text{ para a retinopatia moderada naquele olho, e} \\ 3, \text{ para a retinopatia severa naquele olho.} \end{cases}$$

$\boldsymbol{\beta_{0j}}$:: proporção global para categoria j;

$\boldsymbol{\beta_1}$: efeito da idade do paciente;

$\boldsymbol{\beta_2}$: efeito da duração em anos da diabete no paciente;

$\boldsymbol{\beta_3}$: efeito do percentual de hemoglobina glicada no paciente;

$\boldsymbol{\beta_4}$: efeito de proteína na urina;

$\boldsymbol{\beta_5}$: efeito do uso de insulina.

Para estimar o modelo marginal 3.6, foi utilizado o pacote *multgee* do software R Core Team (2019). O código utilizado para a estimação do modelo pode ser visto no Apêndice A. A estrutura de correlação considerada para a matriz de correlação R foi a estrutura uniforme, escolhida por meio dos parâmetros intrínsecos estimados de L.

Tabela 3.6: Parâmetros estimados para o modelo marginal

	Estimado	exp(Estimado)	Erro padrão	Valor z	p-valor
$\boldsymbol{Intercepto_1}$	3.07542	21.6590	0.27809	11.0590	<0.0001
$\boldsymbol{Intercepto_2}$	6.35744	576.7706	0.31287	20.3195	<0.0001
Idade	0.00577	1.0058	0.00235	2.4546	0.0141
Diabetes	-0.12697	0.8807	0.00726	-17.4987	<0.0001
Hemoglobina	-0.11169	0.8943	0.01796	-6.2192	<0.0001
Proteína	-0.59119	0.5537	0.04328	-13.6595	<0.0001
Insulina	-0.69652	0.4983	0.12875	-5.4099	<0.0001

Da tabela 3.6 pode-se concluir que, mantendo as demais variáveis preditoras constantes, estima-se que a cada aumento de 1 ano de idade a chance de o nível de retinopatia ir de ausente para severa aumenta 0.58%.

Mantendo as demais variáveis constantes, estima-se que a cada aumento de 1 ano de duração da diabetes a chance de o nível de retinopatia ir de ausente para severa diminua em 12%.

Mantendo as demais variáveis constantes, estima-se que a cada aumento de 1 porcento da e hemoglobina glicada a chance de o nível de retinopatia ir de ausente para severa diminua em 11%.

Mantendo as demais variáveis constantes, estima-se que a cada aumento de 1 porcento de proteína na urina a chance de o nível de retinopatia ir de ausente para severa diminua em 45%.

Mantendo as demais variáveis constantes, estima-se que o paciente ao tomar insulina, a chance de o nível de retinopatia ir de ausente para severa diminua em 51%.

3.3 Modelo Linear Generalizado Misto

Os modelos Marginal, Condicional e os MLG's apresentados no capítulo anterior são considerados de efeito fixos, pois, não consideram em sua estrutura os efeitos aleatórios que não podem ser observados. Os efeitos aleatórios são efeitos únicos de cada individuo.

Os modelos lineares mistos foram propostos por Laird e Ware (1982), assim como as EEG's, também tinham como interesse inicial a análise de dados longitudinais. O modelo misto recebe esse nome por conter efeitos aleatórios na sua composição, além dos efeito fixos. A equação desse modelo, para o indivíduo i, é da forma:

$$\boldsymbol{y_i} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{\epsilon_i}, \tag{3.7}$$

sendo:

$\boldsymbol{y_i}$: o vetor de variável resposta para individuo i, com i=1,...,n;

$\mathbf{X}$: a matriz do modelo composta pelas p variáveis explicativas; e

$\boldsymbol{\beta}$: o vetor de (p+1) parâmetros.

$\mathbf{Z}$: a matriz dos q efeitos aleatórios;

$\boldsymbol{u}$: o vetor de parâmetros com os efeitos individuais; e

$\boldsymbol{\epsilon}$ é o vetor do erro aleatório.

Os vetores $\boldsymbol{\epsilon}$ e $\boldsymbol{u}$ são normalmente distribuídos com média zero e, respectivas, variâncias desconhecidas.

A generalização dos modelos lineares mistos para respostas não-normais, foi proposta em Breslow e Clayton (1993). A equação do modelo linear generalizado misto, MLGM, pode ser escrita similar a (2.3), porém nesse caso a função de ligação é da forma:

$$\boldsymbol{g(\mu_i) = \eta_i = X\beta + Zu,}$$

com $\boldsymbol{u}$ seguindo a distribuição normal multivariada com vetor de valores esperado zero e matriz $\boldsymbol{\Sigma}$ de covariância desconhecidas.

Para o caso em que as variáveis respostas categórica nominal Hartzel *et al.* (2001) propôs a equação do modelo logístico da seguinte forma:

$$\boldsymbol{log\left(\frac{\pi_{imj}}{\pi_{imJ}}\right) = \alpha_j + X\beta + Zu,} \qquad \text{j = 1,..., J-1.} \tag{3.8}$$

A equação 3.8 utilizada para descrever do modelo logístico para variáveis respostas categóricas ordinais, conhecido como Modelo Logístico Cumulativo, é apresentada a seguir:

$$\boldsymbol{log\left(\frac{\pi_{ir1} + \cdots + \pi_{imj}}{\pi_{im,j+1} + \cdots + \pi_{im,J-1}}\right) = \alpha_j + X\beta + Zu,} \qquad \text{j = 1,..., J-1,} \tag{3.9}$$

de forma alternativa a equação 3.9 pode ser escrita como:

$$\boldsymbol{log\left(\frac{\pi_{imj}}{\pi_{im,j+1}}\right) = \alpha_j + X\beta + Zu,} \qquad \text{j = 1,..., J-1.} \tag{3.10}$$

As equações 3.9 e 3.10, apesar de não oferecem os mesmos valores dos preditos, oferecem conclusões semelhantes. (Hartzel *et al.*, 2001)

3.3.1 Estimação dos Parâmetros

De forma semelhante ao MLG univariado, o método de estimação do MLG misto se baseia no método de máxima verossimilhança. A função de verossimilhança é da forma:

$$\boldsymbol{L(\beta, \Sigma) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\prod_{m=1}^{M} f(\bar{y}_{im}|u_i;\beta)\right] g(u_i;\Sigma) du_i} \tag{3.11}$$

Para o MLGM também é necessário a aplicação de métodos numéricos para o cálculo da integral da equação em 3.11, que pode ser aproximado por diversos métodos. Entretanto, neste trabalho só será apresentado o método adaptativo da quadratura de Gauss-Hermite, por ser o utilizado durante a estimação dos modelos na aplicação. Os demais métodos podem ser vistos em Hartzel *et al.* (2001).

3.3.2 Método adaptativo da quadratura de Gauss-Hermite

A quadratura de Gauss é um dos métodos mais populares de integração numérica. O método adaptativo da quadratura de Gauss-Hermite é mais indicado do que o método de Gauss-Hermite, mas pode ser problemático computacionalmente para o caso em que o conjunto de dados é muito grande.

Hartzel *et al.* (2001) mostra que a contribuição para a função de máxima verossimilhança para cada individuo é dada por:

$$\boldsymbol{L(\beta,\Sigma) \approx |\hat{Q}|^{1/2} 2^{m/2} \sum_{l} w_l \left[\prod_{m=1}^{M} f(\bar{y}_{im}|z_i;\beta)\right] g(z_i;\Sigma) exp(z_l^t z_l),} \qquad (3.12)$$

com $\boldsymbol{z_i}=\hat{\omega}_i + \sqrt{2}\hat{Q}^{1/2} z_l$, o nó para o indivíduo i; $\hat{\boldsymbol{Q}}^{\boldsymbol{1/2}}$ o resultado da decomposição de Cholesky da estimativa da curvatura $\hat{\boldsymbol{Q}}$.

A maximização da função de máxima verossimilhança 3.12, pode ser feita utilizando o PROC NLMIXED do software *Statistical Analysis System* (SAS®) que utiliza o métodos Quasi-Newton. Mais detalhes sobre o método Quasi-Newton podem ser consultados em Gill e Murray (1972).

3.3.3 Aplicação a dados de Artrite Reumatoide

Foi considerado para essa aplicação o mesmo conjunto de dados sobre pacientes com artrite reumatoide utilizado na seção 3.2. O modelo proposto para a estimação é da forma,

$$\boldsymbol{log\left(\frac{\pi_{imj}}{\pi_{imJ}}\right) = \alpha_j + X\beta + Zu.}$$

De acordo com Kuss e McLerran (2007), os valores estimados para um modelo fixo podem ser utilizados para valores inicias dos parâmetros do MLG misto. O modelo linear generalizado considerado para a obtenção dos valores inciais foi:

$$\log\left[\frac{\boldsymbol{P(Y=1|x)}}{\boldsymbol{P(Y=3|x)}}\right] = \boldsymbol{\beta_0} + \boldsymbol{\beta_{1i}x_1} + \boldsymbol{\beta_2x_2} + \boldsymbol{\beta_3x_3} + \boldsymbol{\beta_4x_4},$$

$$\log\left[\frac{\boldsymbol{P(Y=2|x)}}{\boldsymbol{P(Y=3|x)}}\right] = \boldsymbol{\beta_0} + \boldsymbol{\beta_{1i}x_1} + \boldsymbol{\beta_2x_2} + \boldsymbol{\beta_3x_3} + \boldsymbol{\beta_4x_4},$$

tal que,

$$\boldsymbol{Y} = \begin{cases} 1, \text{ para o score ruim;} \\ 2, \text{ para o score regular;} \\ 3, \text{ para o score bom.} \end{cases}$$

$\boldsymbol{\beta_0}$: proporção global;

$\boldsymbol{\beta_1}$: efeito do gênero, sendo $\boldsymbol{X_1}$=0 do sexo masculino e $\boldsymbol{X_1}$=1 do sexo feminino;

$\boldsymbol{\beta_2}$: efeito da idade do paciente;

$\boldsymbol{\beta_3}$: efeito do tratamento recebido, com 0 sendo o placebo e 1 Auranofin.

$\boldsymbol{\beta_4}$: efeito do score inicial atribuído pelo paciente, com 1 sendo o score ruim 2 sendo o score regular, e 3 o score bom.

A estimação do MLG foi feita utilizando a função *vglm* do pacote *VGAM* do software R Core Team (2019). Obtendo os valores iniciais disposto na tabela 3.7.

Tabela 3.7: Parâmetros iniciais para o modelo misto completo

	Intercepto	Sexo	Idade	Tratamento	Score_1	Score_2
Categoria 1	0.6728	0.0330	0.0033	-0.8762	-0.9661	2.5212
Categoria 2	-0.4679	0.1535	0.0179	-0.0767	0.1233	-1.7415

Para o ajuste do Modelo Linear Generalizado Misto foi utilizado o PROC NLMIXED do software *Statistical Analysis System* (SAS®), utilizando os valores iniciais de 3.7. O código utilizado para a estimação do modelo pode ser visto no Apêndice A.

A partir dos valores iniciais da tabela 3.7, foi ajustado o MLGM utilizando equação 3.8, do qual os parâmetros referentes a idade e sexo do paciente não resultaram significativos para o modelo. Assim sendo, foram ajustados novos modelos com a retirada uma-a-uma de cada um dos parâmetros não significativos. O modelo com os parâmetros significativos e de menor AIC foi o que teve parâmetros iniciais dado em 3.8.

Tabela 3.8: Parâmetros iniciais para o modelo misto final

	Intercepto	Tratamento	Score	
Categoria 1	0.8532	-0.8781	-0.9784	-2.5286
Categoria 2	0.5216	-0.1002	0.0772	-1.7833

As estimativas do modelo com todas os parâmetros significativos e de menor AIC podem ser vistas na tabela 3.9.

Tabela 3.9: Ajuste para o modelo misto final

	Estimado	Erro padrão	exp(Est.)	Valor t	p-valor
$\boldsymbol{Intercepto_1}$	-2.4771	0.0839	0.5711	-4.34	<0.0001
$\boldsymbol{Tratamento_1}$	-1.2144	0.2969	0.4222	-2.88	0.0043
$\boldsymbol{Score_{11}}$	4.0623	58.1078	0.6963	5.83	<0.0001
$\boldsymbol{Score_{12}}$	2.6904	14.7376	0.6204	4.34	<0.0001
$\boldsymbol{Intercepto_2}$	-1.4854	0.2264	0.3846	-3.86	0.0001
$\boldsymbol{Tratamento_2}$	-0.3479	0.7061	0.3219	-1.08	0.2808
$\boldsymbol{Score_{21}}$	2.9343	18.8083	0.5336	5.50	<0.0001
$\boldsymbol{Score_{22}}$	2.6913	14.7508	0.4628	5.81	<0.0001
$\boldsymbol{\hat{\sigma}_1}$	5.4917	-	1.8386	2.99	0.0031
$\boldsymbol{\hat{\sigma}_2}$	2.7028	-	0.9940	2.72	0.0069
$\boldsymbol{cov_{12}}$	3.8524	-	1.2049	- 3.20	0.0015

Obteve-se valores significantes para todos os parâmetros com exceção do efeito do placebo, o que já era esperado para o experimento. Da tabela 3.11 pode-se concluir que, mantendo o Score inicial constante, estima-se que o tratamento diminua em 71.31% a chance do Score de dor final ir de regular para ruim e diminua em 29.39% a chance do Score de dor final ir de bom para ruim.

Mantendo o tratamento constante, estima-se que o Score de dor inicial ruim aumente em 571% a chance do Score de dor final ir de regular para ruim e 178% a chance do Score de dor final ir de bom para ruim.

Mantendo o tratamento constante, estima-se que o Score de dor inicial bom aumente em 147% a chance do Score de dor final ir de regular para ruim e 147% a chance do Score de dor final ir de bom para ruim.

3.3.4 Aplicação a dados de Retinopatia Diabética

Foi considerado para essa aplicação o mesmo conjunto de dados sobre pacientes diabéticos utilizados na seção **??**. O modelo proposto para a estimação é da forma,

$$\boldsymbol{log\left(\frac{\pi_{imj}}{\pi_{imJ}}\right) = \alpha_j + X\beta + Zu.}$$

Novamente seguindo a recomendação de Kuss e McLerran (2007) de estimar os valores iniciais dos parâmetros, o modelo linear generalizado, para esse conjunto de dados é:

$$\log\left[\frac{\boldsymbol{P(Y=1|x)}}{\boldsymbol{P(Y=3|x)}}\right] = \boldsymbol{\beta_0 + \beta_{1i}x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5},$$

$$\log\left[\frac{\boldsymbol{P(Y=2|x)}}{\boldsymbol{P(Y=3|x)}}\right] = \boldsymbol{\beta_0 + \beta_{1i}x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5},$$

em que,

$$\boldsymbol{Y} = \begin{cases} 1, \text{ para a ausência de retinopatia naquele olho} \\ 2, \text{ para a retinopatia moderada naquele olho, e} \\ 3, \text{ para a retinopatia severa naquele olho.} \end{cases}$$

$\boldsymbol{\beta_0}$: proporção global;

$\boldsymbol{\beta_1}$: efeito da idade do paciente;

$\boldsymbol{\beta_2}$: efeito da duração em anos da diabete no paciente;

$\boldsymbol{\beta_3}$: efeito do percentual de hemoglobina glicada no paciente;

$\boldsymbol{\beta_4}$: efeito de proteína na urina;

$\boldsymbol{\beta_5}$: efeito do uso de insulina.

Da mesma forma que na aplicação com os dados sobre pacientes com artrite reumatoide, foi feita a estimação no software R Core Team (2019) para a obtenção dos valores a serem utilizados como valores iniciais do modelo misto. Esses valores podem ser vistos na tabela 3.10.

Tabela 3.10: Parâmetros iniciais para o modelo misto completo

	Intercepto	Idade	Diabetes	Hemoglobina	Proteína	Insulina
Categoria 2	-1.0451	-0.0010	0.1331	0.1207	6.7735	0.6710
Categoria 3	-2.9500	-0.0167	0.2167	0.1495	9.5828	1.0774

A partir dos valores iniciais da tabela 3.10, foi ajustado o MLGM utilizando equação 3.8. Para o ajuste do Modelo Linear Generalizado Misto foi utilizado o PROC NLMIXED do software *Statistical Analysis System* (SAS®). O código utilizado para a estimação do modelo pode ser visto no Apêndice A. Todos as estimativas dos parâmetros para este modelo resultaram em significativos, conforme pode ser visto na tabela 3.11.

Tabela 3.11: Ajuste para o modelo misto final

	Estimado	Erro padrão	exp(Est.)	Valor t	p-valor
Intercepto$_{\mathbf{2}}$	-7.8037	0.0004	0.7504	-10.40	<0.0001
Idade$_{\mathbf{2}}$	-0.0082	0.9918	0.0059	-1.37	0.1703
Diabetes$_{\mathbf{2}}$	0.2968	1.3455	0.0193	15.33	<0.0001
Hemoglobina$_{\mathbf{2}}$	0.2985	1.3478	0.0460	6.49	<0.0001
Proteína$_{\mathbf{2}}$	1.1328	3.1043	0.1352	8.38	<0.0001
Insulina$_{\mathbf{2}}$	1.6101	5.0033	0.3010	5.35	<0.0001
Intercepto$_{\mathbf{3}}$	-25.1035	0.0000	2.3200	-10.82	<0.0001
Idade$_{\mathbf{3}}$	-0.0390	0.9617	0.0169	-2.31	0.0209
Diabetes$_{\mathbf{3}}$	0.6791	1.9721	0.0460	14.75	<0.0001
Hemoglobina$_{\mathbf{3}}$	0.4695	1.5992	0.1226	3.83	0.0001
Proteína3	3.3870	29.5771	0.2938	11.53	<0.0001
Insulina$_{\mathbf{3}}$	3.2048	24.6506	0.9818	3.26	0.0011
$\hat{\boldsymbol{\sigma}}_{\mathbf{2}}$	11.1593	-	1.1760	9.49	<0.0001
$\hat{\boldsymbol{\sigma}}_{\mathbf{3}}$	76.4755	-	8.2138	9.31	<0.0001
cov$_{\mathbf{23}}$	26.2049	-	2.4588	2 10.66	<0.0001

Da tabela 3.11 pode-se concluir que, mantendo as demais variáveis preditoras constantes, estima-se que a cada aumento de 1 ano de idade, a chance de o nível de retinopatia ir de moderada para ausente diminua em 1.82% e de ir de severa para ausente diminua em 3.63%

Mantendo as demais variáveis preditoras constantes, estima-se que a cada aumento de 1 ano de duração da diabetes, a chance de o nível de retinopatia ir de para moderada ausente aumente em 34.55% e de ir de severa para ausente diminua em 97.21%.

Mantendo as demais variáveis preditoras constantes, estima-se que a cada aumento de 1 porcento da hemoglobina glicada, a chance de o nível de retinopatia ir de moderada para ausente diminua em 34.78% e de ir de severa para ausente diminua em 59.92%.

Mantendo as demais variáveis preditoras constantes, estima-se que a cada aumento de 1 porcento de proteína na urina, a chance de o nível de retinopatia ir de moderada para ausente aumente em 210.43% e ir de severa para ausente aumente em 2900%.

Mantendo as demais variáveis preditoras constantes, estima-se que o paciente ao tomar insulina a chance de o nível de retinopatia ir de moderada para ausente aumente em 500% e de ir de severa para ausente aumente em 2400%.

3.4 Considerações

Para os dados de pacientes de artrite reumatoide utilizando os modelos Marginal e Linear Generalizado Misto não houve diferenciação das significâncias dos parâmetros. Dentre as variáveis gênero, idade, tratamento e score inicial as variáveis preditoras tratamento e escore inicial foram consideradas significantes em ambos os casos.

Os dados de pacientes diabéticos possuíam como variáveis respostas os valores para ambos os olhos do paciente. Analisando essas observações de forma univariada, no Capítulo 2 as significância das estimativas dos parâmetros do olho esquerdo e do olho direito foram similares, com as variáveis preditoras Hemoglobina e Insulina não significantes para a categoria referente a retinopatia moderada. Analisando com a abordagem bivariada tivemos como resultado para o Modelo Marginal que todas as variáveis foram significativas para o modelo, e para o Modelo Linear Generalizado Misto que a estimativa do parâmetro idade para a categoria moderada resultou em não significante.

Capítulo 4

Conclusões

Este trabalho de conclusão de curso, ou TCC, apresentou características, fundamentos teóricos e algumas possibilidades de aplicação dos (modelos lineares usuais e) dos modelos lineares generalizados binomial univariado e multinomial univariado e bivariado.

Os modelos lineares generalizados multinomiais univariado e bivariado são muito adequados para uso em aplicações ou em problemas de regressão, em que se considera uma ou duas variáveis resposta categóricas que dependam de uma ou mais variáveis explicativas. Entretanto, após longa e árdua busca, verificamos que a literatura disponível sobre regressão logística multinomial e aplicações relacionadas, não é tão ampla para o caso univariado, mas é escassa para o caso bivariado ou multivariado.

Para o caso univariado e bivariado há inúmeras possibilidades de aplicações. A vantagem que os modelos bivariados possuem sobre o modelo unviariado é a possibilidade de considerar a correlação entre as variáveis respostas.

No capítulo 2 foram expostos os fundamentos dos modelos de regressão logística binomial e multinomial univariados. Quando os Modelos Lineares Generalizados são feitos utilizando como função de ligação logito obtêm-se como vantagem a possibilidade de interpretação dos parâmetros por meio da razão de chances. Foi apresentada uma aplicação do modelo de regressão logístico binomial, para analisar a eficácia de doses de um herbicida sobre o controle de plantas. Três aplicações do modelo de regressão logístico multinomial foram apresentadas. A primeira analisa a eficácia da concentração de raios gama sobre o controle de moscas domésticas. A segunda analisa o grau de severidade da retinopatia como função das variáveis idade, uso de insulina, níveis de hemoglobina e de proteína na urina, e anos da duração do diabetes. Esta análise foi feita para cada um dos olhos do paciente.

No capítulo 3 foram expostos os fundamentos de três modelos alternativos de regressão logística multinomial bivariada. O primeiro é o chamado Modelo Condicional de Efeitos Fixos, o segundo o chamado Modelo Marginal, e o terceiro o Modelo Linear Generalizado Misto. Duas aplicações do modelo de regressão logístico multinomial foram apresentadas. A primeira analisa o grau de severidade da artrite reumatoide como função das variáveis sexo, idade, aplicação de um tratamento e medida de um escore inicial. E, a segunda é referente aos dados de retinopatia diabética apresentada no capítulo 2, porém com a análise sendo feita considerando os dois olhos de forma conjunta

O Modelo Condicional de Efeitos Fixos nos permite a partir de duas variáveis condicionar a segunda variável a primeira e obter melhores estimativas dos parâmetros da segunda variável, pois considera em sua estrutura a correlação inerente as duas variáveis.

O Modelo Marginal possui a a vantagem ao Modelo Linear Generalizado por considerar as correlações entre as variáveis respostas. E também possuir um método de estimação consistente.

O Modelo Linear Generalizado Misto a partir da sua estrutura, que é composta por efeitos aleatórios individuais nos permite calcular os efeitos aleatórios referente a cada indivíduo. Ele é útil no caso bivariado justamente por essa componente aleatória individual que é acrescida no modelo.

Para a construção dos modelos apresentados no Capítulo 3 houve escassez no encontro de metodologia que contemplasse esse caso. Apesar disso não pode ser dito que referencial teórico seja inexistente. Também vale ressaltar que, para a construção deste trabalho houveram dificuldades em relação a obtenção de pacotes ou rotinas da estimação dos modelos, principalmente do Modelo Linear Generalizado Misto. Em sua maioria as rotinas disponíveis para resposta categórica são feitas considerando que a variável resposta tenha distribuição binomial. Os softwares que foram utilizados foi o Software R Core Team (2019) para os Modelos Lineares generalizados binomial e multinomial e o Software SAS para o Modelo Linear Generalizado Misto e o MATLAB para o Modelo Condicional de Efeitos Fixos. Os códigos para a estimação contam no Apêndice A.

Apêndice A

Códigos de programação

A seguir são apresentados os códigos de programação utilizados para a estimação dos parâmetros das metodologias apresentadas neste trabalho.

```
############################## MODELO BINOMIAL ##############################
#Código para a estimação do Modelo Linear Generalizado com resposta #
#binomial e função de ligação logito #
library(tidyverse)

picloram= read.table("~/picloram.txt", quote="\"")
picloram = picloram %>%
  mutate(V2 = factor(V2), V1=factor(V1))%>%
  rename("Quantidade" = "V2", "Bloco"="V1", "Y"="V3")

mb = glm(factor(Y) ~ factor(Bloco) + factor(Quantidade), data=picloram,
family=binomial(link='logit'))
exp(summary(mb)$coef)

############################## MODELO MULTINOMIAL ##############################
#Código para estimação do Modelo Linear Generalizado com resposta #
#multinomial e função de ligação logito#
#Dados sobre radiação e moscas#
library(VGAM)
```

```
x=c(rep(80,500),rep(100,500),rep(120,500),rep(140,500),rep(160,500),
    rep(180,500),rep(200,500))
y1=c(rep(1,62),rep(0,438),rep(1,94),rep(0,406),rep(1,179),rep(0,321),
     rep(1,335),rep(0,165),rep(1,432),rep(0,68),rep(1,487),rep(0,13),
     rep(1,498),rep(0,2))
y2=c(rep(0,62),rep(1,5),rep(0,433),rep(0,94),rep(1,24),rep(0,382),
     rep(0,179),rep(1,60),rep(0,261),rep(0,335),rep(1,80),rep(0,85),
     rep(0,432),rep(1,46),rep(0,22),rep(0,487),rep(1,11),rep(0,2),
     rep(0,498),rep(1,2))
y3=c(rep(0,62),rep(0,5),rep(1,433),rep(0,94),rep(0,24),rep(1,382),
     rep(0,179),rep(0,60),rep(1,261),rep(0,335),rep(0,80),rep(1,85),
     rep(0,432),rep(0,46),rep(1,22),rep(0,487),rep(0,11),rep(1,2),
     rep(0,498),rep(0,2))
y=c(rep("y1",62),rep("y2",5),rep("y3",433),rep("y1",94),rep("y2",24),
    rep("y3",382),rep("y1",179),rep("y2",60),rep("y3",261),rep("y1",335),
    rep("y2",80),rep("y3",85),rep("y1",432),rep("y2",46),rep("y3",22),
    rep("y1",487),rep("y2",11),rep("y3",2),rep("y1",498),rep("y2",2))
mosca=data.frame(x,y)
mm=vglm(y ~ factor(x), family = multinomial,data = mosca)
summary(mm)
exp(coefficients(mm))

############################## MODELO MULTINOMIAL ##############################
#Código para estimação do Modelo Linear Generalizado com resposta #
#multinomial e função de ligação logito#
#Dados sobre retinopatia diabética#

library(gss)
data(wesdr1)
retin_m=wesdr1
retin_m=retin_m %>% mutate(ret1=as.numeric(ret1),ret2=as.numeric(ret2),
    ID=seq(1,2049))
retin_m1=retin_m %>% filter(ret1 %in% c(1)) %>%
```

```
  mutate(direito=0)
retin_m2=retin_m %>% filter(ret1 %in% c(2,3,4,5)) %>%
  mutate(direito=1)
retin_m3=retin_m %>% filter(ret1 %in% c(6)) %>%
  mutate(direito=2)
retin_m5=retin_m %>% filter(ret2 %in% c(1)) %>%
  mutate(esquerdo=0)
retin_m6=retin_m %>% filter(ret2 %in% c(2,3,4,5)) %>%
  mutate(esquerdo=1)
retin_m7=retin_m %>% filter(ret2 %in% c(6)) %>%
  mutate(esquerdo=2)
retino_l=full_join(retin_m1,retin_m2)
retino_l=full_join(retino_l,retin_m3)
retino_l=retino_l%>% arrange(retino_l$ID,decreasing = FALSE)
retino_r=full_join(retin_m5,retin_m6)
retino_r=full_join(retino_r,retin_m7)
retino_r=retino_r%>% arrange(retino_l$ID,decreasing = FALSE)
retino_m=right_join(retino_r,retino_l)

retino=retino_m %>% select(-c(ret1,ret2)) %>%
  mutate(esquerdo=factor(esquerdo), direito=factor(direito)) %>%
  rename(idade=age,diabetes=dur,glicose=gly,insulina=insl)

#Modelo Multinomial para o olho esquerdo#
m_retino_e=retino_m%>% select(-direito)
mmre=vglm(formula = esquerdo ~ idade + diabetes + glicose + upro +
    insulina, family = multinomial, data = m_retino_e)

#Modelo Multinomial para o olho direito#
m_retino_d=retino_m%>% select(-esquerdo)
mmrd=vglm(formula = direito ~ idade + diabetes + glicose + upro +
    insulina, family = multinomial, data = m_retino_d)

#################### MODELO MARGINAL ####################
```

```
library(multGEE)
artrite_marg1=ordLORgee(formula = new_y ~ age + factor(sex_f)+
                factor(droga) + factor(baseline),
              link = "logit", id = id, repeated = time,
              data = artrite3,LORstr = "uniform")
summary(artrite_marg1)

artrite_marg2=ordLORgee(formula = new_y ~  factor(sex_f)+
            factor(droga) + factor(baseline),
                        link = "logit", id = id, repeated = time,
                        data = artrite3,LORstr = "uniform")
summary(artrite_marg2)

artrite_marg3=ordLORgee(formula = new_y ~   factor(droga) +
            factor(baseline),
                        link = "logit", id = id, repeated = time,
                        data = artrite3,LORstr = "uniform")
summary(artrite_marg3)

retinopatia_marg1=ordLORgee(formula = olho ~ idade + diabetes+ glicose +
u  pro + insulina,
                            link = "logit", id = ID, repeated = Ind_Olho,
                            data = retinopatia_teste,LORstr = "uniform")
summary(retinopatia_marg1)

#################### MODELO LINEAR GENERALIZADO MISTO ####################
# Obtenção dos parâmetros iniciais para a inicialização da função no SAS #
library(VGAM)
artrite_mm=vglm(new_y~sex+age+trt+baseline,
                family=multinomial, data=artrite3)
summary(artrite_mm)

artrite_mm1=vglm(new_y~age+trt+baseline,
                family=multinomial, data=artrite3)
```

```
summary(artrite_mm1)

artrite_mm2=vglm(new_y~trt+baseline,
                 family=multinomial, data=artrite3)
summary(artrite_mm2)

artrite_mm3=vglm(new_y~trt+baseline,
                 family=multinomial, data=artrite3)
summary(artrite_mm3)

artrite_mm4=vglm(new_y~trt,
                 family=multinomial, data=artrite3)
summary(artrite_mm4)

retinopatia_mm=vglm(olho~idade+diabetes+glicose+upro+insulina,
                    family=multinomial, data=retinopatia)
summary(retinopatia_mm)
##########################################################################

%%%%%%%%%%%%%%%%% CÓDIGOS EM SAS %%%%%%%%%%%%%%%%%
/* Código para a estimação do Modelo Linear Generalizado Misto */
/* com resposta Multinomial.Baseado na implementação proposta */
/* por Kuss, O. e McLerran, D. (2007).*/

/* Estimação para os dados de artrite*/
PROC NLMIXED DATA=artrite_teste;
/*Chutes iniciais para os parâmetros */
PARMS theta2=4.4457949 b_sex2=-0.1757379 b_age2=0.0008608
    b_trt2=-0.9106069 b_baseline2=-1.2332807 b_time2=0.0763111
    theta3=2.4362319 b_sex3=-0.2848431 b_age3=0.0140255
    b_trt3=-0.1366899 b_baseline3=-0.7458759 b_time3=-0.0468914
    logsu2=.5 logsu3=.5 z23=1;
```

```
eta1 = 0; /* Recebe valor 0 por ser a categoria de referência */
eta2 = theta2+b_sex2*sex+b_age2*age+b_trt2*trt+
        b_baseline2*baseline b_time2*time + u2;
eta3 = theta3+b_sex3*sex+b_age3*age+b_trt3*trt+
        b_baseline3*baseline+b_time3*time + u3;

ARRAY exp_eta {3};
exp_eta1 = 1;
exp_eta2 = exp(eta2);
exp_eta3 = exp(eta3);
bot = exp_eta1+exp_eta2+exp_eta3;

p_y = exp_eta{new_y}/bot;
ll=log(p_y);

su2 = exp(logsu2);
su3 = exp(logsu3);
rho23 = (exp(2*Z23)-1)/(exp(2*Z23)+1);
cov23 = rho23*su2*su3;
MODEL new_y~GENERAL(ll);
RANDOM u2 u3~NORMAL([0,0],[su2*su2,cov23,su3*su3]) SUBJECT=id;
ESTIMATE 'Var2' exp(2*logsu2);
ESTIMATE 'Var3' exp(2*logsu3);
ESTIMATE 'cov23' su2*su3*(exp(2*Z23)-1) / (exp(2*Z23)+1);
RUN;

PROC NLMIXED DATA=artrite_teste;
PARMS theta2=4.1318368894  b_age2=0.0007083063  b_trt2=-0.9093005359
b_baseline2=-1.2287184485 b_time2=0.0766336108
theta3=1.9351178808  b_age3=0.0138513827 b_trt3=-0.1405408210
b_baseline3=-0.7381415186 b_time3=-0.0466479879
logsu2=.5 logsu3=.5 z23=1;
eta1 = 0;
```

```
eta2 = theta2+b_age2*age+b_trt2*trt+b_baseline2*baseline+b_time2*time+u2;
eta3 = theta3+b_age3*age+b_trt3*trt+b_baseline3*baseline+b_time3*time+u3;

ARRAY exp_eta {3};
exp_eta1 = 1;
exp_eta2 = exp(eta2);
exp_eta3 = exp(eta3);
bot = exp_eta1 + exp_eta2 + exp_eta3;

p_y = exp_eta{new_y} / bot;
ll=log(p_y);

su2 = exp(logsu2);
su3 = exp(logsu3);
rho23 = (exp(2*Z23) - 1) / (exp(2*Z23) + 1);
cov23 = rho23*su2*su3;
MODEL new_y ~ GENERAL(ll);
RANDOM u2 u3 ~ NORMAL([0,0],[su2*su2,cov23,su3*su3]) SUBJECT=id;
ESTIMATE 'Var2' exp(2*logsu2);
ESTIMATE 'Var3' exp(2*logsu3);
ESTIMATE 'cov23' su2*su3*(exp(2*Z23)-1) / (exp(2*Z23)+1);
RUN;

PROC NLMIXED DATA=artrite_teste;
/*Chutes iniciais para os parâmetros */
PARMS theta2=4.15070694   b_trt2=-0.90835484
b_baseline2=-1.22377326 b_time2=0.07669054
theta3=2.68185131  b_trt3=-0.15077677
b_baseline3=-0.75091159 b_time3=-0.04498265
logsu2=.5 logsu3=.5 z23=1;
eta1 = 0;
eta2 = theta2  +b_trt2*trt + b_baseline2*baseline + b_time2*time + u2;
eta3 = theta3  +b_trt3*trt + b_baseline3*baseline + b_time3*time + u3;
```

```
ARRAY exp_eta {3};
exp_eta1 = 1;
exp_eta2 = exp(eta2);
exp_eta3 = exp(eta3);
bot = exp_eta1 + exp_eta2 + exp_eta3;

p_y = exp_eta{new_y} / bot;
ll=log(p_y);

su2 = exp(logsu2);
su3 = exp(logsu3);
rho23 = (exp(2*Z23) - 1) / (exp(2*Z23) + 1);
cov23 = rho23*su2*su3;
MODEL new_y ~ GENERAL(ll);
RANDOM u2 u3 ~ NORMAL([0,0],[su2*su2,cov23,su3*su3]) SUBJECT=id;
ESTIMATE 'Var2' exp(2*logsu2);
ESTIMATE 'Var3' exp(2*logsu3);
ESTIMATE 'cov23' su2*su3*(exp(2*Z23)-1) / (exp(2*Z23)+1);
RUN;

PROC NLMIXED DATA=artrite_teste;
/*Chutes iniciais para os parâmetros */
PARMS theta2=4.3036   b_trt2=-0.9082 b_baseline2=-1.2228
theta3=2.5945 b_trt3=-0.1508 b_baseline3=-0.7514
logsu2=.5 logsu3=.5 z23=1;
eta1 = 0; /* Recebe valor 0 por ser a categoria de referência */
eta2 = theta2 +b_trt2*trt + b_baseline2*baseline  + u2;
eta3 = theta3 +b_trt3*trt + b_baseline3*baseline  + u3;


ARRAY exp_eta {3};
```

```
exp_eta1 = 1;
exp_eta2 = exp(eta2);
exp_eta3 = exp(eta3);
bot = exp_eta1 + exp_eta2 + exp_eta3;

p_y = exp_eta{new_y} / bot;
ll=log(p_y);

su2 = exp(logsu2);
su3 = exp(logsu3);
rho23 = (exp(2*Z23) - 1) / (exp(2*Z23) + 1);
cov23 = rho23*su2*su3;
MODEL new_y ~ GENERAL(ll);
RANDOM u2 u3 ~ NORMAL([0,0],[su2*su2,cov23,su3*su3]) SUBJECT=id;
ESTIMATE 'Var2' exp(2*logsu2);
ESTIMATE 'Var3' exp(2*logsu3);
ESTIMATE 'cov23' su2*su3*(exp(2*Z23)-1) / (exp(2*Z23)+1);
RUN;

logsu2=.5 logsu3=.5 z23=1;
eta1 = 0; /* Recebe valor 0 por ser a categoria de referência */
eta2 = theta2 +b_trt2*trt +  u2;
eta3 = theta3 +b_trt3*trt +  u3;


ARRAY exp_eta {3};
exp_eta1 = 1;
exp_eta2 = exp(eta2);
exp_eta3 = exp(eta3);
bot = exp_eta1 + exp_eta2 + exp_eta3;

p_y = exp_eta{new_y} / bot;
ll=log(p_y);
```

```
su2 = exp(logsu2);
su3 = exp(logsu3);
rho23 = (exp(2*Z23) - 1) / (exp(2*Z23) + 1);
cov23 = rho23*su2*su3;
MODEL new_y ~ GENERAL(ll);
RANDOM u2 u3 ~ NORMAL([0,0],[su2*su2,cov23,su3*su3]) SUBJECT=id;
ESTIMATE 'Var2' exp(2*logsu2);
ESTIMATE 'Var3' exp(2*logsu3);
ESTIMATE 'cov23' su2*su3*(exp(2*Z23)-1) / (exp(2*Z23)+1);
RUN;

/* Estimação para os dados de retinopatia*/
PROC NLMIXED DATA=retinopatia_teste;
/*Chutes iniciais para os parâmetros */
PARMS theta2=-1.04507 b_idade2=-0.00103679 b_diabetes2=0.133094
    b_glicose2=0.1206676  b_upro2=6.773548  b_insulina2=0.6709761
    theta3=-2.95 b_idade3=-0.016722571 b_diabetes3=0.2166763
    b_glicose3=0.1495580 b_upro3=9.582781 b_insulina3=1.077383
    logsu2=.5 logsu3=.5 z23=1;

eta1 = 0; /* Recebe valor 0 por ser a categoria de referência */
eta2 = theta2+b_idade2*idade+b_diabetes2*diabetes+b_glicose2*glicose+
    b_upro2*upro+ b_insulina2*insulina+u2;
eta3 = theta3+b_idade3*idade+b_diabetes3*diabetes+b_glicose3*glicose+
    b_upro3*upro+ b_insulina3*insulina+u3;

ARRAY exp_eta {3};
exp_eta1 = 1;
exp_eta2 = exp(eta2);
exp_eta3 = exp(eta3);
bot = exp_eta1 + exp_eta2 + exp_eta3;
```

```
p_y = exp_eta{olho} / bot;
ll=log(p_y);

su2 = exp(logsu2);
su3 = exp(logsu3);
rho23 = (exp(2*Z23) - 1) / (exp(2*Z23) + 1);
cov23 = rho23*su2*su3;
MODEL olho~GENERAL(ll);
RANDOM u2 u3~NORMAL([0,0],[su2*su2,cov23,su3*su3]) SUBJECT=ID;
ESTIMATE 'Var2' exp(2*logsu2);
ESTIMATE 'Var3' exp(2*logsu3);
ESTIMATE 'cov23' su2*su3*(exp(2*Z23)-1) / (exp(2*Z23)+1);
RUN;
```

```
%%%%%%%%%%%%%%%%%% CÓDIGOS EM MATLAB %%%%%%%%%%%%%%%%%%
% Código para a estimação do Modelo Condicional com efeitos fixos, %
% com 2 resposta Multinomial e 3 categorias. Implementação proposta %
% por Uddin, M. N. e Begum, M. (2018).%
%Ajuste Multinomial
% 2 respostas e 3 categorias
bhat=logreg3MLE2(xretinopatia,yretinopatia);


function bhat = logreg3MLE2(X,Y)
%Finds MLE for explantory variables X and responses Y.
d=size(X,2);
Bhatg = ones(1,8*d);
%optimatization - derivative free
    %bhat=fminsearch(f,Bhatg);
%optimization with supplied gradient
options = optimset('GradObj','off','TolFun',1e-22,'Display','off');
    bhat=fminunc(@(B)loglike3_2(X,Y,B),Bhatg,options);
end
```

```
function L = loglike3_2(X,Y,B)
%Returns  the negative of loglike3 (since maximizing).
r=length(B); s=r/8; n=size(Y,1);
b1=B(1:s); b2=B(s+1:2*s);
b01=B(2*s+1:3*s); b11=B(3*s+1:4*s); b21=B(4*s+1:5*s);
b02=B(5*s+1:6*s); b12=B(6*s+1:7*s);b22=B(7*s+1:8*s);
sm=zeros(1,n);
for i=1:n
    x=X(i,:); z1z2=h(Y(i,1)); z3z4=h(Y(i,2)); z1=z1z2(1); z2=z1z2(2);
    z3=z3z4(1); z4=z3z4(2);
     sm(i)=eta0(x)+z1*eta1(x)+z2*eta2(x)+z3*eta3(x)+z4*eta4(x)+z1*z3*eta5(x)
     +z1*z4*eta6(x)+z2*z3*eta7(x)+z2*z4*eta8(x);
end
L=-sum(sm)/56;
%if nargout > 1 % gradient required
%   grads = -grad_lrMLE3(X,Y,B);
%end
    function a = evals(x,b,c)
         a=log(1+exp(x*b')+exp(x*c'));
    end
    function a = eta0(x)
         a=evals(x,b02,b01)+evals(x,b1,b2);
    end
    function a = eta1(x)
        a=x*b1'+evals(x,b02,b01)-evals(x,b11,b12);
    end
    function a = eta2(x)
        a=b2*x'-evals(x,b21,b22)+evals(x,b02,b01);
    end
    function a = eta3(x)
        a=b01*x';
    end
    function a = eta4(x)
```

```
        a=b02*x';
    end
    function a = eta5(x)
        a=(b11-b01)*x';
    end
    function a = eta6(x)
        a=(b12-b02)*x';
    end
    function a = eta7(x)
        a=(b21-b01)*x';
    end
    function a = eta8(x)
        a=(b22-b2)*x' + evals(x,b21,b22) - evals(x,b02,b01);
    end
    function  z=h(yp)
        if yp==0, z=[0 0]; elseif yp==1, z=[1 0]; else, z=[0 1];
        end
    end
end
```

Referências Bibliográficas

Abreu, T. P., Borille, G. M. R., Alves, C. J. P., Caetano, M. e Correia, A. R. (2016). Análise dos fatores que influenciam na escolha aeroportuária dos passageiros da tma-sp através da análise logística multinomial.

Bombardier, C., Ware, J., Russell, I. J., Larson, M., Chalmers, A., Read, J. L., Arnold, W., Bennett, R., Caldwell, J., Hench, P. K. *et al.* (1986). Auranofin therapy and quality of life in patients with rheumatoid arthritis. results of a multicenter trial. *The American journal of medicine*, **81**(4), 565–578.

Box, G. P. e Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, **26**(2), 211–252.

Breslow, N. E. e Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, **88**(421), 9–25.

Cordeiro, G. M. e Demétrio, C. G. B. (2008). Modelos lineares generalizados e extensões. *Piracicaba: USP*.

Cuyabano, B. C., Pinheiro, H. P. e Pinheiro, A. (2010). Models applied to dna sequences with multinomial correlated responses.

Figueira, C. V. (2006). *Modelos de regressão logística*. Ph.D. thesis, Universidade Federal do Rio Grande do Sul.

Fitzmaurice, G. M., Laird, N. M. e Ware, J. H. (2012). *Applied longitudinal analysis*, volume 998. John Wiley & Sons.

Galdino, M. V. (2015). *Modelos lineares generalizados mistos e equações de estimação generalizadas para dados binário aplicados em anestesiologia veterinária*. Master's thesis, Universidade Estadual Paulista (UNESP).

Gill, P. E. e Murray, W. (1972). Quasi-newton methods for unconstrained optimization. *IMA Journal of Applied Mathematics*, **9**(1), 91–108.

Hartzel, J., Agresti, A. e Caffo, B. (2001). Multinomial logit random effects models. *Statistical Modelling*, **1**(2), 81–102.

Itepan, N. M. (1995). *Aumento do periodo de aceitabilidade de pupas de Musa domestica L., 1758 (Diptera: Muscidae), irradiadas com raios gama, como hospedeiras de parasitoides (Hymenoptera: Pteromalidae)..* Master's thesis, Centro de Energia Nuclear na Agricultura da Universidade de São Paulo.

Klein, R., Klein, B. e Moss, S. E. (1989). The wisconsin epidemiological study of diabetic retinopathy: a review. *Diabetes/metabolism reviews*, **5**(7), 559–570.

Kuss, O. e McLerran, D. (2007). A note on the estimation of the multinomial logistic model with correlated responses in sas. *Computer methods and programs in biomedicine*, **87**(3), 262–269.

Laird, N. M. e Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Liang, K.-Y. e Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.

Nelder, J. A. e Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, **135**(3), 370–384.

Nicolau, J. (2014). Determinantes do voto no primeiro turno das eleições presidenciais brasileiras de 2010: uma análise exploratória. *Opinião Pública*, **20**(3), 311–325.

Paula, G. A. (2013). *Modelos de regressão: com apoio computacional*. IME-USP São Paulo.

Pino, F. A. (2014). A questão da não normalidade: Uma revisão. *Revista de economia agrícola*, **61**(2), 17–33.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sun, B. (2014). *Bivariate multinomial models*. Ph.D. thesis, Memorial University of Newfoundland.

Touloumis, A. (2014). R package multgee: A generalized estimating equations solver for multinomial responses. *arXiv preprint arXiv:1410.5232*.

Touloumis, A., Agresti, A. e Kateri, M. (2013). Gee for multinomial responses using a local odds ratios parameterization. *Biometrics*, **69**(3), 633–640.

Turner, D. L., Ralphs, M. H. e Evans, J. O. (1992). Logistic analysis for monitoring and assessing herbicide efficacy. *Weed Technology*, **6**(2), 424–430.

Tutz, G. (2011). *Regression for categorical data*, volume 34. Cambridge University Press.

Uddin, M. N. e Begum, M. (2018). A generalized linear model for multivariate correlated binary response data on mobility index. *Journal of Statistical Research*, **52**(1), 61–73.

Vieira, A. C. (2019). *Análise de Dados de Sinistralidade Rodoviária nas Zonas em Obras com Recurso à Regressão Logística Multinomial*. Ph.D. thesis, Universidade da Beira Interior.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, **61**(3), 439–447.

Zhu, Y. (2014). Correlated multinomial data. *Wiley StatsRef: Statistics Reference Online*, pages 1–6.