UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

Andreza Jardelino da Silva

Estudo Teórico sobre Modelos Lineares Generalizados com Aplicação a Dados Genéticos

Campina Grande - PB

Novembro de 2014

Andreza Jardelino da Silva

Estudo Teórico sobre Modelos Lineares Generalizados com Aplicação a Dados Genéticos

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Professor Gustavo Henrique Esteves

Campina Grande - PB

Novembro de 2014

S586e Silva, Andreza Jardelino da.
Estudo teórico sobre modelos lineares generalizados com aplicação a dados genéticos [manuscrito] / Andreza Jardelino da Silva. - 2014.
54 p. : il.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2014.
"Orientação: Prof. Dr. Gustavo Henrique Esteves, Departamento de Estatística".

1. Modelo linear generalizado. 2. Regressão logística. 3. Dados genéticos. I. Título.

21. ed. CDD 519.535 2

Andreza Jardelino da Silva

Estudo Teórico sobre Modelos Lineares Generalizados com Aplicação a Dados Genéticos

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Trabalho aprovado em 26 de Novembro de 2014.

BANCA EXAMINADORA

Prof. Dr. Gustavo Henrique Esteves
Universidade Estadual da Paraíba

Prof. Dr. João Gil de Luna
Universidade Estadual da Paraíba

Prof. Dr. Ricardo Alves de Olinda
Universidade Estadual da Paraíba

Ao meu marido, Ranieri, aos meus pais, Materna e Cleodon, que sempre acreditaram na minha capacidade e me apoiaram em todos os momentos, dedico com muito amor e carinho.

Agradecimentos

A DEUS por ter me concedido a vida e me dado forças nas horas difíceis, fazendo com que eu tenha superado todos os meus obstáculos.

Ao meu amado marido Ranieri da Silva Nascimento, por sempre estar ao meu lado dando-me força, coragem e companhia, fazendo com que eu sempre acreditasse nos meus sonhos e lutasse por meus objetivos.

Aos meus pais, Cleodon Vicente da Silva e Maria Materna Jardelino da Silva, a minha irmã Cícera Aline Jardelino da Silva, pelo amor, carinho e dedicação.

A minha avó Josefa da Silva, pelo seu amor e carinho.

Aos familiares do meu esposo que sempre estiveram ao meu lado dando-me carinho e atenção.

A todos os meus familiares que mesmo de longe, de forma especial, torceram por mim.

As minhas amigas, Otávia, Ana Patrícia, Eugênia, Tatiane, Luíza, Luzidark, Márcia, Ana Lúcia e aos meus amigos Tiago, Montini, por todos os momentos vividos juntos que jamais esquecerei.

Ao meu orientador Gustavo Henrique Esteves, pela paciência, respeito, amizade e dedicação que sempre teve para comigo.

Aos professores em especial Ana patrícia Bastos Peixoto, Tiago Almeida de Oliveira, Ricardo Alves de Olinda, Divanilda Maia Esteves, João Gil de Luna, pelos conhecimentos que enriqueceram minha vida acadêmica e pessoal, e também aos funcionários do Departamento de Estatística.

A todos que acreditaram em mim e fizeram/fazem parte da minha vida, o meu obrigada a todos.

"Por aqui, contudo, não olhamos para trás por muito tempo.
Seguimos em frente, abrindo novas portas e fazendo coisas novas...
E a curiosidade nos conduz a novos caminhos."
(WALT DISNEY)

Resumo

O modelo de regressão linear proporciona a mensuração do efeito de uma ou mais variáveis explicativas sobre uma variável resposta, porém quando uma ou mais suposições do modelo ordinário não são satisfatórias, principalmente a suposição dos erros seguirem uma distribuição normal, em alguns casos não será possível realizar a análise de forma confiável com apenas as transformações existentes. Para tal, precisa-se realizar a análise com base em uma modelagem específica e mais flexível, conhecida como Modelo Linear Generalizado (MLG) apresentada por Nelder e Wedderburn (1972). A ideia básica dessa técnica, consiste em ampliar as opções para a distribuição da variável resposta, permitindo que a mesma pertença a família exponencial de distribuições, bem como dar maior flexibilidade para a relação funcional entre a função de ligação (μ) da variável resposta univariada e a parte sistemática do modelo (η). O objetivo deste trabalho é estudar a teoria referente aos MLG's, especificamente o modelo de regressão logística, apresentando métodos e procedimentos computacionais para uma aplicação relacionada à pesquisa do câncer em esôfago e estômago, em dados oriundos de uma colaboração recente com pesquisadores do Instituto de Matemática e Estatística da USP e do Hospital Sírio-Libanês de São Paulo - SP. Todas as análises foram implementadas no *software* `R`.

Palavras-chaves: Modelos Lineares Generalizados; Regressão Logística; Câncer de Estômago e Esôfago.

Abstract

The linear regression model provides the measurement of the effect of one or more explanatory variables on a response variable, but when one or more of the usual model assumptions are not satisfactory, mainly the assumption of the errors follow a normal distribution, you can not perform analysis reliably with only existing transformations. To do this, one must perform the analysis based on a specific and more flexible modeling, known as Generalized Linear Model (GLM) by Nelder e Wedderburn (1972). The basic idea of this technique is to expand the options for the distribution of the response variable, allowing it to belong to the exponential family of distributions, as well as provide greater flexibility to the functional relationship between the link function (μ) of univariate response variable and the systematic part of the model (η). This paper comprises study on the theory related to GLM's, specifically the logistic regression model, presenting methods and computational procedures for research related to cancer in the stomach and esophagus, in data from a recent collaboration with researchers from the Institute of Mathematics and Application Statistics USP and Sírio-Libanês Hospital of São Paulo - SP. All analyzes were implemented in *software* `R`.

Key-words: Generalized Linear Model; Logistic Regression; Stomach Cancer and Esophagus.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Modelos de análise de regressão linear que envolvem as suposições usuais, especialmente aquela de normalidade dos erros associados ao modelo, são tradicionalmente estudados e aplicados em diversas áreas do conhecimento. Isso exige que a variável resposta associada à tal modelagem seja obrigatoriamente numérica, preferencialmente contínua. Formas comuns de se violar as suposições do modelo de regressão usual surgem quando temos variáveis resposta binárias ou associadas a processos de contagem, que são mais naturalmente modeladas por distribuições de Bernoulli, Binomial, Poisson, entre outras. Para estes casos, os modelos de regressão logística, de Poisson ou log-lineares podem se encaixar bem, tais modelos se enquadram em uma área da Estatística conhecida como Modelos Lineares Generalizados (MLG's).

A ideia básica envolvida em um MLG é estender as opções de distribuições de probabilidades dos erros associados ao modelo de regressão, e consequentemente da variável resposta, de forma a comportar aquelas que se enquadrem na família exponencial, além de flexibilizar a relação funcional entre a variável resposta e as variáveis explicativas do modelo (PAULA, 2013).

Segundo Cordeiro e Demétrio (2007), dada uma amostra aleatória de n observações independentes entre si, um modelo linear generalizado pode ser resumido da seguinte forma:

i) uma variável resposta Y, que é o componente aleatório do modelo, cuja distribuição deve se enquadrar na família exponencial;

ii) um conjunto de variáveis explicativas $X_1, X_2, \cdots, X_k$, que constituem o componente sistemático do modelo, e que entram de acordo com uma estrutura linear similar à regressão convencional;

iii) e finalmente, uma função adequada para a junção entre os componentes aleatório (variável resposta) e sistemático (variáveis explicativas) do modelo, conhecida como função de ligação.

Cabe salientar que boa parte das distribuições de probabilidade conhecidas atualmente pertencem à família exponencial. Especialmente no contexto dos modelos lineares generalizados, destacam-se as distribuições Normal, Normal Inversa e Gama para dados contínuos; Binomial para proporções; além das distribuições de Poisson e Binomial Negativa para dados decorrentes de processos de contagens.

Naturalmente, o processo de estimação dos parâmetros associados aos modelos lineares generalizados não é tão simples como o método dos mínimos quadrados, usado na estimação do modelo de regressão linear ordinário que é o mais usual. Na maioria dos casos de MLG's nem é possível se encontrar uma expressão fechada para os estimadores dos parâmetros, sendo necessário se usar algum método iterativo para tal finalidade, como o algoritmo de Newton-Raphson (CORDEIRO; DEMéTRIO, 2007; PAULA, 2013).

Desse modo, realizou-se um estudo teórico sobre a técnica dos MLG's, a qual é usada para modelar dados de natureza mais geral do que exclusivamente quantitativas. Foram estudados os conceitos fundamentais e os seus métodos estatísticos para obtenção dos estimadores dos parâmetros de interesse, tais como modelagem e função de ligação. Além do mais foram revisados alguns materiais para se chegar a um melhor entendimento, como por exemplo, técnicas computacionais (método iterativo de Newton-Raphson) e ainda inferência estatística para o Teorema da Informação de Fisher.

Com isso, foram usados para fins de aplicação dados reais associados a valores de expressão de diversos genes para diversas observações associadas a pacientes com problemas gástricos relacionados com o surgimento de câncer de esôfago e estômago, para se tentar estimar perfis de interações gênicas a partir do uso de modelos de regressão logísticos, que são um caso particular dos MLG's estudados anteriormente. Os dados nos são disponíveis através de uma colaboração que o orientador deste trabalho vem mantendo com pesquisadores do Instituto de Matemática e Estatística da USP e do Hospital Sírio-Libanês, ambos em São Paulo-SP.

2 Fundamentação Teórica

Este trabalho teve por finalidade estudar a teoria referente aos MLG's, sendo o objetivo principal em estudo o modelo de regressão logística. Inicialmente realizou-se uma profunda revisão acerca da família exponencial de distribuições de probabilidades, passando pelos conceitos de modelagem e funções de ligação, até culminar com os aspectos de estimação e inferências deste tipo de modelagem. Mais detalhadamente, pretendeu-se estudar a teoria dos MLG's, suas suposições, os métodos de estimação e os principais algoritmos iterativos, além de estudar também a construção de testes de hipóteses e intervalos de confiança.

Todo este estudo teórico foi finalizado com um estudo mais detalhado dos modelos de regressão logística, que são casos particulares dos MLG's, que foram usados com objetivo secundário de estimar relações funcionais na expressão de pares de genes de acordo com um modelo postulado biologicamente. Os dados utilizados são de tecidos gastro-esofágicos, e foram obtidos a partir de uma colaboração recente com pesquisadores do Instituto de Matemática e Estatística da USP e do Hospital Sírio-Libanês. Desta forma, esta seção está organizada de maneira a apresentar os principais aspectos teóricos e metodológicos acerca dos MLG's.

2.1 Marco Histórico

De acordo com Turkman e Silva (2000), a regressão clássica proposta por Legendre e Gauss no início do século XIX, foi a principal técnica de modelagem estatística, muito utilizada até meados do século XX, afim de descrever a maioria de alguns fenômenos aleatórios. Apesar das observações em estudo não apresentarem a suposição de normalidade, era sugerido algum tipo de transformação com o objetivo de alcançar a normalidade, sendo a mais conhecida proposta por Box e Cox (1964), onde transforma a observação y sendo esta positiva em:

$$z = \begin{cases} \frac{y^{\lambda}-1}{\lambda} \ se \ \ \lambda \neq 0 \\ \log(y) \ se \ \ \lambda = 0, \end{cases}$$

tal que, λ será uma constante não conhecida. Porém esta abordagem nem sempre produzia um efeito satisfatório.

Devido ao grande desenvolvimento computacional que ocorreu nos anos 70, o que possibilitou a utilização de métodos iterativos, tornou-se facilitado o surgimento de alguns modelos que não possuem características de normalidade.

A inovação veio em 1972 com a modelagem apresentada por Nelder e Wedderburn

(1972), os quais propuseram os modelos lineares generalizados (MLG's), em que se teria mais opções para a distribuição da variável resposta e que a mesma iria pertencer à família exponencial. Os MLG's permitem uma maior flexibilidade entre a relação da média da variável resposta univariada com o preditor linear η, sendo esta relação funcional obtida por meio de uma função monótona e diferenciável, denominada função de ligação $g(\cdot)$.

Nelder e Wedderburn (1972), propuseram um método iterativo para estimar os parâmetros, e ainda, introduziram o conceito de desvio, o qual é muito importante na avaliação da qualidade do ajuste dos modelos, contribuindo desta maneira, com o desenvolvimento dos resíduos e técnicas de diagnósticos. Vários estudos acerca dos MLG's são encontrados na literatura desde 1972. Wedderburn (1974) propôs os modelos de quase-verossimilhança, que estendem a ideia dos modelos lineares generalizados para situações mais gerais incluindo dados correlacionados. Na econometria houve a abordagem ao problema de se modelar as variâncias Harvey (1976). Modelos de dispersão por Jorgensen (1983), ampliam as opções para a distribuição da variável resposta. Alguns métodos gráficos para detectar heterocedasticidade (ver: Cook e Weisberg (1983) e Atkinson (1985)). Liang e Zeger (1986) estendem os modelos de quase-verossimilhança propondo as Equações de Estimação Generalizadas (EEG's) que permitem o estudo de variáveis aleatórias correlacionadas não gaussianas.

Além disso, Atkinson (1987) apresenta o *software* macro GLIM para se modelar a variância heterogênica. Carrol e Ruppert (1988) desenvolvem procedimentos de diagnósticos usando métodos de influência local para estimar o parâmetro da variância em vários modelos não lineares para a média, enquanto que os modelos não lineares de família exponencial Cordeiro e Paula (1989) admitem preditor não linear nos parâmetros, já Smyth (1989), descreve um método que permite modelar o parâmetro de dispersão em alguns modelos lineares generalizados. No mesmo ano McCullagh e Nelder (1989) publicam seu livro e se torna referência no assunto (MLG's).

Tem-se ainda os modelos aditivos generalizados propostos por Hastie e Tibshirani (1990) que supõem preditor linear formado também por funções semiparamétricas. Lee e Nelder (1996), Lee e Nelder (2001) estenderam o trabalho de Breslow e Clayton propondo modelos lineares generalizados hierárquicos em que o preditor linear pode ser formado por efeitos fixos e efeitos aleatórios não gaussianos. Barroso e Vasconcellos (2002) apresentam expressões para aperfeiçoamento do teste escore em modelos t-Student heterocedásticos e ainda recentemente Taylor e Verbyla (2004) propuseram uma modelagem conjunta de parâmetros de localização e escala em modelos t-Student. Têm-se ainda Paula (2013) com material introdutório aos modelos lineares generalizados, apresentando vários resultados relacionados com estimação, teste de hipóteses, métodos de diagnóstico e seleção de modelos na classe dos MLG's.

2.2 Definição

Segundo Fisher (1934), a família exponencial de distribuição garante aos parâmetros uma estatística suficiente. Desta maneira, sejam n variáveis aleatórias independentes $Y_1, \cdots, Y_n$ cada uma com função densidade ou função de probabilidade na família exponencial expressa por:

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

em que, de acordo com Paula (2013), $E(Y_i) = \mu_i = b'(\theta_i)$, $Var(Y_i) = \phi^{-1}b''(\theta_i) = \phi^{-1}V_i$, sendo $V_i = V(\mu_i) = d\mu_i/d\theta_i$ é a função de variância, a qual caracteriza a distribuição e $\phi^{-1} > 0$ é o parâmetro de dispersão (precisão) muitas vezes denominado σ^2, têm-se ainda θ que será o parâmetro de localização. Uma vez que se conhece a função de variância $V(\mu_i)$ que será a parcela correspondente à segunda derivada de $b(\theta_i)$, é possível determinar qual será a classe de distribuições correspondentes (vice-versa). Para algumas distribuições a variância dos dados muda conforme sua média.

De modo a exemplificar a função de variância, é interessante definir como $V(\mu) = \mu(1-\mu)$, com $0 < \mu < 1$, caracterizando a distribuição binomial com probabilidade de sucesso μ ou $1-\mu$. Quando se têm ϕ indo para infinito (grande), o Teorema Central do Limite (TCL) garante, que Y segue uma distribuição aproximadamente normal de média μ e variância $\phi^{-1}V(\mu)$.

Os MLG's estão definidos por três componentes:

i) Um componente aleatório independente Y_i, com $i = 1, \cdots, n$, e distribuição pertencente à família exponencial linear dado pela expressão (2.1), com médias μ_i e parâmetro de escala constante ϕ.

ii) Um componente sistemático, sendo

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

em que,

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} \tag{2.2}$$

em que η_i é o preditor linear, ou seja, a parte sistemática que pode ser usada para realizar possíveis previsões para a equação da reta estimada.

iii) Têm-se ainda que, $g(.)$ é uma função monótona e diferenciável denominada função de ligação, a qual permite fazer a ligação entre a média μ_i e o preditor linear (parte sistemática), definindo também a forma com que os efeitos sistemáticos de $x_1, x_2, ..., x_k$ são transmitidos para a média. Observe que a esperança de Y_i, será:

$$E(Y_i) = \mu_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}).$$

A função de ligação garante ainda a concavidade de $L(\beta; y)$, ou seja, a unicidade da estimativa de máxima verossimilhança, quando essa existe.

2.3 Casos particulares para algumas distribuições

As principais distribuições que pertencem à família exponencial, juntamente com os tipos de dados onde podem ser aplicadas em um MLG são dadas na Tabela 1 a seguir.

Tabela 1 – Distribuições de probabilidade e tipo de dados

Distribuição	Tipos dos Dados
Normal	Contínuos Simétricos
Poisson	Contagens
Binomial	Proporções
Gama	Contínuos Assimétricos
Normal Inversa	Contínuos Assimétricos

Sendo assim, na sequência serão apresentadas as funções θ, ϕ, $b(\theta)$ e $c(y, \phi)$, específicas para cada uma dessas distribuições listadas na Tabela 1.

2.3.1 Distribuição Normal

Se uma variável aleatória Y segue uma distribuição Normal, ou seja, $Y \sim N(\mu, \sigma^2)$, sua função densidade será escrita por:

$$\begin{aligned}
f(y;\mu,\sigma) &= \frac{1}{\sqrt{2\pi\sigma_2}} exp\left[\frac{-(y-\mu)^2}{2\sigma^2}\right] \\
&= exp\left[\frac{-1}{2\sigma^2}(y-\mu)^2 + \log(2\pi\sigma^2)^{-1/2}\right] \\
&= exp\left[\frac{-1}{2\sigma^2}(y-\mu)^2 - \frac{1}{2}\log(2\pi\sigma^2)\right] \\
&= exp\left[-\left(\frac{y^2-2y\mu+\mu^2}{2\sigma^2}\right) - \frac{1}{2}\log(2\pi\sigma^2)\right] \\
&= exp\left[-\left(\frac{y^2}{2\sigma^2} - \frac{2y\mu}{2\sigma^2} + \frac{\mu^2}{2\sigma^2}\right) - \frac{1}{2}\log(2\pi\sigma^2)\right] \\
&= exp\left[\frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right] \\
&= exp\left[\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}\{\log(2\pi\sigma^2) + \frac{y^2}{\sigma^2}\}\right]
\end{aligned}$$

em que $\theta = \mu$, $\phi = \sigma^{-2}$, $b(\theta) = \theta^2/2 = \mu^2/2$ e $c(y, \phi) = 1/2\{\log(2\pi\sigma^2) + y^2/\sigma^2\}$. Têm-se ainda que a função de variância $V(\mu) = 1$. De forma geral:

$$E(Y) = \frac{db(\theta)}{d\theta} = \mu; \qquad Var(Y) = \frac{d^2b(\theta)}{d\theta^2}\phi = \sigma^2$$

2.3.2 Distribuição Poisson

Sendo uma variável aleatória Y com distribuição de Poisson, isto é, $Y \sim Poisson(\mu)$, sua função densidade ficará escrita como:

$$\begin{aligned} f(y;\mu) &= \frac{\mu^y exp^{-\mu}}{y!} \\ &= \mu^y exp^{-\mu}(y!)^{-1} \\ &= exp\{-\mu + \log(\mu^y) + \log(y!)^{-1}\} \\ &= exp\{-\mu + y\log(\mu) - \log(y!)\} \\ &= exp\{y\log(\mu) - \mu - \log(y!)\} \end{aligned}$$

em que, $\theta = \log(\mu)$, $\phi = 1$, $b(\theta) = e^\theta = \mu$ e $c(y,\phi) = -\log(y!)$. Têm-se ainda que a função de variância $V(\mu) = \mu$. De forma geral:

$$E(Y) = \frac{db(\theta)}{d\theta} = \mu; \qquad Var(Y) = \frac{d^2b(\theta)}{d\theta^2} = \mu$$

2.3.3 Distribuição Binomial

Assumindo que $Y \sim Binomial(n,\mu)$, onde Y será a proporção de sucessos com n ensaios independentes de Bernoulli, sendo cada um com probabilidade de sucesso μ. Com isso, verifica-se que a função de probabilidade de Y pode ser escrita sob a forma:

$$\begin{aligned} f(y;\mu) &= \binom{n}{y}\mu^y(1-\mu)^{n-y} \\ &= \binom{n}{y}\mu^y(1-\mu)^n(1-\mu)^{-y} \\ &= \binom{n}{y}(1-\mu)^n\frac{\mu^y}{(1-\mu)^y} \\ &= \binom{n}{y}(1-\mu)^n\left(\frac{\mu}{1-\mu}\right)^y \\ &= \binom{n}{y}exp\left\{\log(1-\mu)^n + \log\left(\frac{\mu}{1-\mu}\right)^y\right\} \\ &= exp\left\{\log(1-\mu)^n + \log\left(\frac{\mu}{1-\mu}\right)^y + \log\binom{n}{y}\right\} \\ &= exp\left\{n\log(1-\mu) + y\log\left(\frac{\mu}{1-\mu}\right) + \log\binom{n}{y}\right\} \end{aligned}$$

em que, $\theta = \log\left(\frac{\mu}{1-\mu}\right)$, $\phi = 1$, $b(\theta) = n\log(1-\mu)$ e $c(y,\phi) = \log\binom{n}{y}$. Têm-se ainda que a função de variância $V(\mu) = \mu(1-\mu)$. De forma geral:

$$E(y) = \frac{db(\theta)}{d\theta} = \mu; \qquad Var(y) = \frac{d^2b(\theta)}{d\theta^2}\phi = \mu(1-\mu)$$

2.3.4 Distribuição Gama

Seja Y é uma variável aleatória, ou seja, $Y \sim Gama(\mu, \alpha)$, então, sua função densidade ficará:

$$\begin{aligned} f(y;\mu,\alpha) &= \frac{1}{y\Gamma(\alpha)}\left(\frac{\alpha y}{\mu}\right)^{\alpha} exp\left\{\frac{-\alpha y}{\mu}\right\} \\ &= exp\left\{\frac{-\alpha y}{\mu} + \log\left(\frac{1}{y\Gamma(\alpha)}\right) + \log\left(\frac{\alpha y}{\mu}\right)\right\} \\ &= exp\left\{\frac{-\alpha y}{\mu} + \log(1) - \log(y\Gamma(\alpha)) + \log(\alpha y)^{\alpha} - \log(\mu^2)\right\} \\ &= exp\left\{\frac{-\alpha y}{\mu} - \log(y\Gamma(\alpha)) + \log(\alpha y)^{\alpha} - \log(\mu^2)\right\} \\ &= exp\left\{\alpha\left(\frac{-y}{\mu} + \log(\mu)\right) - \log(y\Gamma(\alpha)) + \alpha\log(\alpha y) - \log(y)\right\} \end{aligned}$$

em que, $\theta = -1/\mu$, $\phi = 1/\alpha$, $b(\theta) = -\log(-\theta)$ e $c(y,\phi) = \alpha\log(\alpha y) - \log(y\Gamma(\alpha))$. Têm-se ainda que a função de variância $V(\mu) = \mu^2$. De forma geral:

$$E(y) = \frac{db(\theta)}{d\theta} = \mu; \qquad Var(y) = \frac{d^2b(\theta)}{d\theta^2}\phi = \frac{\mu^2}{\alpha}$$

2.3.5 Distribuição Normal inversa

Seja Y uma variável aleatória, isto é, $Y \sim NI(\mu, \phi)$, sua função densidade ficará expressa como:

$$\begin{aligned} f(y;\mu,\phi) &= \frac{1}{\sqrt{2\pi\phi y^3}} exp\left[\frac{-(y-\mu)^2}{2\phi\mu^2 y}\right] \\ &= exp\left[\frac{-1}{2\phi\mu^2 y}(y-\mu)^2 + \log(2\pi\phi y^3)^{-1/2}\right] \\ &= exp\left[\frac{-1}{2\phi\mu^2 y}(y^2 - 2y\mu + \mu^2) - \frac{1}{2}\log(2\pi\phi y^3)\right] \\ &= exp\left[-\frac{y}{2\phi\mu^2} + \frac{1}{\phi\mu} - \frac{1}{2\phi y} - \frac{1}{2}\log(2\pi\phi y^3)\right] \\ &= exp\left[\frac{\frac{-y}{2\mu^2} + \frac{1}{\mu}}{\phi} - \frac{1}{2y\phi} - \frac{1}{2}\log(2\pi\phi y^3)\right] \end{aligned}$$

em que $\theta = -1/2\mu^2$, $\phi = \sigma^2$, $b(\theta) = 1/\mu$ e $c(y,\phi) = (-1/2y\phi) - (1/2)\log(2\pi\phi y^3)$. Têm-se ainda que a função de variância $V(\mu) = \mu^3$. De forma geral:

$$E(y) = \frac{db(\theta)}{d\theta} = \mu; \qquad Var(y) = \frac{d^2b(\theta)}{d\theta^2}\phi = \mu^3\sigma^2$$

Na Tabela 2 a seguir é apresentado um resumo das distribuições mostradas anteriormente, com os principais resultados e funções referentes à família exponencial.

Tabela 2 – Distribuições de probabilidade pertencentes à família exponencial

Distribuição	$b(\theta)$	θ	ϕ	$V(\mu_i)$
Normal	$\theta^2/2$	μ	σ^{-2}	1
Poisson	e^{θ}	$\log \mu$	1	μ
Binomial	$n \log(1-\mu)$	$\log\{\mu/(1-\mu)\}$	1	$\mu(1-\mu)$
Gama	$-\log(-\theta)$	$-1/\mu$	$1/\alpha$	μ^2
Normal Inversa	$1/\mu$	$-1/2\mu^2$	σ^2	μ^3

2.4 Ligações canônicas

Quando o preditor linear é igual ao parâmetro canônico θ, têm-se uma ligação canônica sob a forma de identidade do tipo:

$$\theta_i = \eta_i$$

em que,

$$L(\boldsymbol{\beta}) = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}).$$

Segundo Myers, Montgomery e Vining (2002) a utilização da função de ligação canônica implica em algumas interessantes propriedades, mas isso não quer dizer que ela deva ser utilizada sempre. A sua escolha é conveniente porque não só simplifica as estimativas de máxima verossimilhança dos parâmetros do modelo, mas também o cálculo do intervalo de confiança para a média da resposta. Contudo, a conveniência não implica necessariamente em qualidade de ajuste do modelo. As ligações canônicas para os modelos mais usuais estão expostas na Tabela 3, a seguir.

Tabela 3 – Distribuições de probabilidade da família exponencial e sua ligação canônica.

Distribuição	η
Normal	μ
Poisson	$\log \mu$
Binomial	$\log\{\mu/(1-\mu)\}$
Gama	$1/\mu$
Normal Inversa	$1/\mu^2$

2.5 Função Desvio (*deviance*)

A qualidade do ajuste do MLG é avaliada por meio da função desvio, sendo esta a distância entre o logaritmo da função de verossimilhança do modelo saturado e do modelo sob investigação. Dessa forma, seja a função de verossimilhança dada por:

$$L(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^{n} L(\mu_i; y_i),$$

em que, $\mu_i = g^{-1}(\eta_i)$ e ainda $\boldsymbol{\eta}_i = \mathbf{x}_i^T\boldsymbol{\beta}$. Têm-se para o modelo saturado, onde $p = n$, que a função $L(\boldsymbol{\mu}; \mathbf{y})$ será escrita como:

$$L(\mathbf{y}; \mathbf{y}) = \sum_{i=1}^{n} L(y_i; y_i),$$

pois, neste caso a estimativa de máxima verossimilhança de μ_i é dada por $\tilde{\mu} = y_i$.

Quando o desvio não é igual a zero, ou seja, o modelo deixa de ser saturado com $p < n$, a estimativa de $L(\boldsymbol{\mu}; \mathbf{y})$ passa a ser escrita na forma $L(\hat{\boldsymbol{\mu}}; \mathbf{y})$, pois neste caso, μ_i será $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$, onde se têm que $\hat{\eta}_i = x_i^T\hat{\boldsymbol{\beta}}$. Daí a função desvio é escrita como sendo:

$$D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \phi D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\{L(\mathbf{y}; \mathbf{y}) - L(\hat{\boldsymbol{\mu}}; \mathbf{y})\}. \tag{2.3}$$

Ao se ter um valor significativo pequeno para a função desvio dada em (2.3), pode-se concluir que, para uma quantidade de parâmetros pequena, chega-se a um bom ajuste, tanto quanto o ajuste do modelo saturado. Reescrevendo a função desvio de (2.3) em uma outra forma segue-se que:

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}[y_i(\tilde{\theta}_i - \hat{\theta}_i) + (b(\hat{\theta}_i) - b(\tilde{\theta}_i))]$$

em que, $\hat{\theta}_i = \theta_i(\hat{\mu}_i)$ e $\tilde{\theta}_i = \theta(\tilde{\mu}_i)$, sendo as estimativas de máxima verossimilhança para θ_i dos modelos com p parâmetros ($p < n$) e saturado ($p = n$) respectivamente. A seguir são apresentadas as funções desvio para as principais distribuições de probabilidade de acordo com Paula (2013).

2.5.1 Distribuição Normal

Sendo $\theta_i = \mu_i$, têm-se que $\tilde{\theta}_i = y_i$ e $\hat{\theta}_i = \hat{\mu}_i$, e o desvio ficará:

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}[y_i(y_i - \hat{\mu}_i) + \hat{\mu}_i^2/2 - y_i^2/2] = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2,$$

que, naturalmente, coincide com a soma de quadrados de resíduos do modelo de regressão linear ordinário.

2.5.2 Distribuição Poisson

Aqui, $\theta_i = \log(\mu_i)$, têm-se que $\tilde{\theta}_i = \log(y_i)$ para a $y_i > 0$ e $\hat{\theta}_i = \log(\hat{\mu}_i)$, e o desvio ficará:

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}[y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)],$$

Se $y_i = 0$ o i-ésimo termo de $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ vale $2\hat{\mu}_i$, sendo que deste modo, para o modelo de Poisson têm-se:

$$d^2(y_i; \hat{\mu}_i) = \begin{cases} 2\{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\} & \text{se} \quad y_i > 0; \\ 2\hat{\mu}_i & \text{se} \quad y_i = 0. \end{cases}$$

2.5.3 Distribuição Binomial

Assumindo que $Y_i \sim Binomial(n_i, \mu_i)$, $i = 1, ..., k$, têm-se que $\tilde{\theta}_i = \log\{y_i/(n_i - y_i)\}$ para a $0 < y_i < n_i$ e $\hat{\theta}_i = \log\{\hat{\mu}_i/(1 - \hat{\mu}_i)$, e o desvio será dado por:

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}[y_i \log(y_i/n_i\hat{\mu}_i) + (n_i - y_i)\log\{(1 - y_i/n_i)/(1 - \hat{\mu}_i)\}].$$

Porém se $y_i = 0$ ou $y_i = n_i$, o i-ésimo termo de $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ será $-2n_i \log(1 - \hat{\mu}_i)$ ou $-2n_i \log \hat{\mu}_i$ respectivamente. Assim, para a binomial cada desvio individual ficará:

$$d^2(y_i; \hat{\mu}_i) = \begin{cases} y_i \log(y_i/n_i\hat{\mu}_i) + (n_i - y_i)\log\{(1 - y_i/n_i)/(1 - \hat{\mu}_i)\} & \text{se} \quad 0 < y_i < n_i; \\ -2n_i \log(1 - \hat{\mu}_i) & \text{se} \quad y_i = 0; \\ -2n_i \log \hat{\mu}_i & \text{se} \quad y_i = n_i. \end{cases}$$

2.5.4 Distribuição Gama

Sendo $\tilde{\theta}_i = -1/y_i$ e $\hat{\theta}_i = -1/\hat{\mu}_i$, o desvio para valores positivos ficará:

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}[-\log(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i].$$

Se algum componente de y_i for igual a zero se terá um desvio indeterminado. McCullagh e Nelder (1989) sugerem substituir $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ nesse caso por:

$$D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\phi C(\mathbf{y}) + 2\phi \sum_{i=1}^{n} \log \hat{\mu}_i + 2\phi \sum_{i=1}^{n} y_i/\hat{\mu}_i,$$

em que, $C(\mathbf{y})$ será uma função arbitrária, todavia limitada. Pode-se usar por exemplo, $C(\mathbf{y}) = \sum_{i=1}^{n} y_i/(1 + y_i)$.

2.5.5 Distribuição Normal inversa

Nesse caso, $\tilde{\theta}_i = -1/2y_i^2$ e $\hat{\theta}_i = -1/2\hat{\mu}_i^2$, o desvio será expresso por:

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2/(y_i\hat{\mu}_i^2).$$

2.6 Função escore e informação de Fisher

2.6.1 Escore e Fisher para $\boldsymbol{\beta}$

Muitos métodos podem ser utilizados para a obtenção das estimativas dos $\boldsymbol{\beta}$'s, inclusive o qui-quadrado mínimo, o Bayesiano e o método da máxima verossimilhança. Este último será de nosso interesse, pois possui algumas propriedades importantes, dentre elas a consistência e a eficiência assintótica.

Dessa forma, ainda de acordo com Paula (2013), para a estimação dos $\boldsymbol{\beta}$'s se faz necessário realizar a maximização da função do logaritmo da verossimilhança. Segue-se então a equação

$$L = \log l(\mathbf{y}; \boldsymbol{\theta}, \phi) = \sum_{i=1}^{n} \log f(y_i; \theta_i, \phi) = \sum_{i=1}^{n} \left\{ \frac{[y_i\theta_i - b(\theta_i)]}{a(\phi)} + c(y_i, \phi) \right\} \tag{2.4}$$

Fazendo a derivação de (2.4) em relação a $\boldsymbol{\beta}$, têm-se que:

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \phi\{y_i\theta_i - b(\theta_i)\},$$

como ϕ é constante, tem-se que

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \phi \left\{ y_i \frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j} - \frac{db(\theta_i)}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j} \right\}$$

Separando as derivadas anteriores para melhor entendimento, de acordo com

$$\frac{d\theta_i}{d\mu_i} = V^{-1}, \quad \frac{db(\theta_i)}{d\theta_i} = \mu_i \quad \text{e} \quad \frac{\partial \eta_i}{\partial \beta_j} = X_{ij},$$

de forma que têm-se

$$\begin{aligned} \frac{\partial L}{\partial \beta_j} &= \sum_{i=1}^{n} \phi\{y_i V_i^{-1}(d\mu_i/d\eta_i)x_{ij} - \mu_i V_i^{-1}(d\mu_i d\eta_i)x_{ij}\} \\ &= \sum_{i=1}^{n} \phi \left\{ \sqrt{\frac{w_i}{V_i}}(y_i - \mu_i)x_{ij} \right\}, \end{aligned}$$

em que

$$w_i = (d\mu_i/d\eta_i)^2/V_i.$$

Dessa maneira, a partir dos resultados acima define-se a função escore matricialmente da seguinte maneira

$$U(\boldsymbol{\beta}) = \frac{\partial L}{\partial \boldsymbol{\beta}} = \phi X^T W^{1/2} V^{-1/2}(\mathbf{y} - \boldsymbol{\mu}), \tag{2.5}$$

sendo X uma matriz $n \times p$ de posto completo onde suas linhas serão X_i^T com $i = 1, \cdots, n$, $W = diag\{w_1, \cdots, w_n\}$ sendo a matriz dos pesos, $V = diag\{V_1, \cdots, V_n\}$, $\mathbf{y} = (y_1, \cdots, y_n)^T$ e ainda $\boldsymbol{\mu} = (y_1, \cdots, \mu_n)^T$.

Garantindo-se um bom estimador de Máxima Verossimilhança obtêm-se a matriz da informação de Fisher. Com ele, será possível encontrar as estimativas de máxima verossimilhança através de um algoritmo iterativo, conhecido como Método Escore de Fisher, o qual se baseia no método de Newton-Raphson (NELDER; WEDDERBURN, 1972). Assim, temos que

$$\frac{\partial^2 L}{\partial \beta_j \partial \beta_l} = \phi \sum_{i=1}^{n} (y_i - \mu_i)\frac{d^2\theta_i}{d\mu_i^2}\left(\frac{d\mu_i}{d\eta_i}\right)^2 x_{ij}x_{il} + \phi \sum_{i=1}^{n} (y_i - \mu_i)\frac{d\theta_i}{d\mu_i}\frac{d^2\mu_i}{d\eta_i^2}x_{ij}x_{il}$$

$$-\phi \sum_{i=1}^{n} \frac{d\theta_i}{d\mu_i} \left(\frac{d\mu_i}{d\eta_i}\right)^2 x_{ij} x_{il}$$

logo:

$$\begin{aligned} E(\partial^2 L/\partial\beta_j \partial\beta_l) &= -\phi \sum_{i=1}^{n} \frac{d\theta_i}{d\mu_i} \left(\frac{d\mu_i}{d\eta_i}\right)^2 x_{ij} x_{il} \\ &= -\phi \sum_{i=1}^{n} \frac{(d\mu_i/d\eta_i^2)}{V_i} x_{ij} x_{il} \\ &= -\phi \sum_{i=1}^{n} w_i x_{ij} x_{il}. \end{aligned}$$

Dessa forma, escrevendo a informação de Fisher matricialmente, têm-se que:

$$K(\boldsymbol{\beta}) = E\left(-\frac{\partial^2 L}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right) = \phi X^T W X. \tag{2.6}$$

Se for possível a alternativa pela ligação canônica, onde $\eta = \theta$, as equações de (2.5) e (2.6), ficarão na seguinte forma respectivamente:

$$U(\boldsymbol{\beta}) = \phi X^T (y - \mu) \quad \text{e} \quad K(\boldsymbol{\beta}) = \phi X^T V X.$$

2.6.2 Escore e Fisher para ϕ

Para o parâmetro de dispersão ϕ, a função escore é expressa por:

$$\begin{aligned} U(\phi) &= \frac{\partial L}{\partial \phi} \\ &= \sum_{i=1}^{n} \{y_i \theta_i - b(\theta_i)\} + \sum_{i=1}^{n} c'(y_i, \phi), \end{aligned}$$

sendo $c'(y_i, \phi) = dc(y_i, \phi)/d\phi$.

Para obtenção da informação de Fisher para ϕ, se faz necessário realizar o cálculo dado por: $\partial^2 L/\partial\phi^2 = \sum_{i=1}^{n} c''(y_i, \phi)$, onde $c''(y_i, \phi) = d^2 c(y_i, \phi)/d\phi^2$.

Deste modo, a informação de Fisher para o parâmetro de dispersão ϕ, será:

$$K(\phi) = -\sum_{i=1}^{n} E\{c''(Y_i, \phi)\}.$$

2.7 Estimando os parâmetros

2.7.1 Estimando os $\boldsymbol{\beta}'s$

A partir dos resultados dados pelas expressões (2.5) e (2.6) da seção anterior, pode-se finalmente calcular as estimativas de máxima verossimilhança dos coeficientes de

$\boldsymbol{\beta}$ através do algoritmo iterativo de Newton-Raphson. Para isso é usual definir um valor inicial $\boldsymbol{\beta}^{(0)}$ para $\boldsymbol{\beta}$ e expandir a função escore $U(\boldsymbol{\beta})$ em torno dele

$$U(\boldsymbol{\beta}) \cong U(\boldsymbol{\beta}^{(0)}) + U'(\boldsymbol{\beta}^{(0)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}),$$

onde $U'(\boldsymbol{\beta})$ será a derivada primeira de $U(\boldsymbol{\beta})$ em relação a $\boldsymbol{\beta}$. Daí, fazendo repetições no processo anterior, chega-se ao modelo iterativo dado por

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \{-U'(\boldsymbol{\beta}^{(m)})\}^{-1} U(\boldsymbol{\beta}^{(m)}), \quad \text{com} \quad m = 0, 1, \cdots.$$

É possível que a matriz $-U'(\boldsymbol{\beta})$ não seja positiva definida, de modo que a substituição do valor esperado correspondente a matriz de $-U'(\boldsymbol{\beta})$ do método *scoring* de Fisher, poderá ser mais apropriada. Com isso, a equação do processo iterativo ficará:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + K^{-1}(\boldsymbol{\beta}^{(m)}) U(\boldsymbol{\beta}^{(m)}), \quad \text{com} \quad m = 0, 1, \cdots.$$

Substituindo as equações encontradas anteriormente, teremos o algoritmo dos Mínimos Quadrados Ponderados iterativos (MQPI) como se segue:

$$\boldsymbol{\beta}^{(m+1)} = (X^T W^{(m)} X)^{-1} X^T W^{(m)} z^{(m)}, \quad \text{com} \quad m = 0, 1, \cdots, \tag{2.7}$$

onde m refere-se à m-ésima iteração; $\hat{\boldsymbol{\beta}}^{(m)}$ é a estimativa do vetor de parâmetros na iteração m; X é a matriz dos valores das variáveis regressoras; W é a matriz de pesos, a qual muda de acordo com cada passo do processo iterativo, e ainda, os elementos da sua diagonal são definidos por

$$W_{ii}^{(m)} = \frac{1}{var(y_i)} \left(\left(\frac{\partial \mu_i}{\partial \eta_i} \right)^{(m)} \right)^2, \quad \text{com} \quad i = 1, 2, \cdots, n;$$

e $\boldsymbol{z}$ é o vetor das variáveis de ajuste na m-ésima iteração, que desempenha a função de uma variável dependente modificada, dada por

$$\mathbf{z} = \boldsymbol{\eta} + W^{-1/2} V^{-1/2} (\mathbf{y} - \boldsymbol{\mu}).$$

Em cada iteração, o vetor das estimativas dos parâmetros $\hat{\boldsymbol{\beta}}^{(m+1)}$ é calculado como uma função das estimativas anteriores $\hat{\boldsymbol{\beta}}^{(m)}$. A convergência em (2.7) ocorre em um número finito de passos independentemente dos valores iniciais dados, porém será mais rápida para valores iniciais mais próximos da solução.

2.7.2 Estimando o parâmetro de dispersão ϕ

Ainda de acordo com Paula (2013), igualando a função escore $U(\phi)$ a zero têm-se que:

$$\sum_{i=1}^{n} c'(y_i, \hat{\phi}) = \frac{1}{2} D(\mathbf{y}; \hat{\boldsymbol{\mu}}) - \sum_{i=1}^{n} \{ y_i \tilde{\theta}_i - b(\tilde{\theta}_i) \},$$

em que, $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ será o desvio do modelo de interesse. Se fazendo necessário trabalhar com a expressão acima de acordo com o modelo que estiver sendo trabalhado, para se chegar a um estimador para o parâmetro ϕ.

2.8 Testes de hipóteses

2.8.1 Hipótese simples

Segundo Paula (2013) supõe-se a seguinte situação para hipótese simples no caso dos MLG's:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^0 \quad contra \quad H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}^0,$$

sendo $\boldsymbol{\beta}^0$ um vetor conhecido de dimensão p e ϕ considerado conhecido.

As estatísticas ξ_{RV} e F que serão abordadas em seguida, apresentam a propriedade de que são invariantes com reparametrizações, onde será útil na construção dos intervalos de confiança para os parâmetros de interesse.

2.8.2 Teste da Razão de Verossimilhança (ξ_{RV})

Esse teste para o caso da hipótese simples é definido como:

$$\xi_{RV} = 2\{L(\hat{\boldsymbol{\beta}}) - L(\boldsymbol{\beta}^0)\}.$$

Por meio dessa estatística pode-se expressar os MLG's como sendo a diferença entre dois desvios, dado por:

$$\xi_{RV} = \phi\{D(\mathbf{y}; \hat{\boldsymbol{\mu}}^0) - D(\mathbf{y}; \hat{\boldsymbol{\mu}})\},$$

onde as funções desvios correspondentes aos modelos sob H_0 e H_1 de acordo com Paula (2013) serão denotados por $D(\mathbf{y}; \hat{\boldsymbol{\mu}}^0)$ e $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ respectivamente, em que $\hat{\boldsymbol{\mu}}^0 = g^{-1}(\hat{\boldsymbol{\eta}}^0)$, ou seja, é a estimativa de máxima verossimilhança sob H_0, e ainda que, $\hat{\eta}^0 = X\boldsymbol{\beta}^0$. Sob a hipótese nula $\xi_{RV} \sim \chi^2_q$ quando $n \to \infty$.

Para o caso normal linear, têm-se:

$$\xi_{RV} = \{\sum_{i=1}^{n}(y_i - \hat{\mu}_i^0)^2 - \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2\}/\sigma^2.$$

2.8.3 Teste F

Usando s desvios, o teste F para o caso de hipótese simples fica dado por:

$$F = \frac{\{D(\mathbf{y}; \hat{\boldsymbol{\mu}}^0) - D(\mathbf{y}; \hat{\boldsymbol{\mu}})\}/p}{D(\mathbf{y}; \hat{\boldsymbol{\mu}})/(n-p)},$$

em que $\phi \to \infty$ e sob H_0 têm-se $F \sim F_{[p,n-p]}$. Essa expressão é considerada para $n \to \infty$, no caso em que se tem no denominador da estatística F uma estimativa consistente para ϕ^{-1}. A vantagem de se utilizar a estatística F é que o teste não depende do parâmetro de dispersão ϕ^{-1}.

2.9 Bandas de Confiança para os MLG's

A banda de confiança assintótica de coeficiente $1-\alpha$ pode ser construída a partir de $\mu(\mathbf{z}) = g^{-1}(\mathbf{z^T}\boldsymbol{\beta})$, $\forall \mathbf{z} \in \mathbb{R}^p$, Paula (2013). Ainda de forma assintótica têm-se que $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N_p(\mathbf{0}, \phi^{-1}(\mathbf{X^T W X})^{-1})$. Assim, uma banda de confiança assintótica de coeficiente $1-\alpha$ para o preditor linear $\boldsymbol{z^T\beta}$, $\forall \mathbf{z} \in \mathbb{R}^p$ será:

$$\mathbf{z^T}\hat{\boldsymbol{\beta}} \pm \sqrt{\phi^{-1}c_\alpha}\{\mathbf{z^T}(\mathbf{X^T W X})^{-1}\mathbf{z}\}^{1/2}, \forall \mathbf{z} \in \mathbb{R}^p,$$

onde c_α será $Pr\{\chi^2_p \leq c_\alpha\} = 1-\alpha$. Realizando uma transformação do tipo $g^{-1}(\cdot)$ pode-se encontrar uma banda de confiança assintótica de coeficiente $1-\alpha$ para $\mu(\mathbf{z})$, de acordo com Paula (2013) dada por:

$$g^{-1}[\mathbf{z^T}\hat{\boldsymbol{\beta}} \pm \sqrt{\phi^{-1}c_\alpha}\{\mathbf{z^T}(\mathbf{X^T W X})^{-1}\mathbf{z}\}^{1/2}], \forall \mathbf{z} \in \mathbb{R}^p.$$

Vale salientar que $\mathbf{z}$ será um vetor do tipo $p \times 1$ que está variando livremente no $\mathbb{R}^p$, e que $\mathbf{X}$ por sua vez, será uma matriz fixa dos valores das variáveis explicativas. $\mathbf{W}$ e ϕ deverão ser estimadas por várias vezes.

2.10 Técnicas de diagnóstico

2.10.1 Pontos de alavanca

Segundo Cook e Weisberg (1983), o ponto de alavanca têm como objetivo avaliar a influência que y_i exerce sobre o seu valor ajustado $\hat{y}_i$. Uma vez que, para se obter a influência se faz necessário calcular a derivada de $\partial\hat{y}_i/\partial y_i$, foi que, Wei, Hu e Fung (1998) propuseram a matriz $(\partial\hat{\mathbf{y}}/\partial\mathbf{y}^T)_{(n\times n)}$ que é utilizada quando se têm uma resposta do tipo contínuo que pode ser aplicada em várias situações de estimação.

Assim, quando ϕ for conhecido, a matriz $(\partial\hat{y}/\partial y_T)$ é obtida sob a forma:

$$\widehat{GL} = \frac{\partial\hat{\mathbf{y}}}{\partial\mathbf{y}^T} = \{\mathbf{D}_{\boldsymbol{\beta}}(-\ddot{\mathbf{L}}_{\boldsymbol{\beta}_\beta})^{-1}\ddot{\mathbf{L}}_{\boldsymbol{\beta}_y}\}|_{\boldsymbol{\beta}},$$

em que, $\mathbf{D}_{\boldsymbol{\beta}} = \partial\boldsymbol{\mu}/\partial\boldsymbol{\beta}$, $\ddot{\mathbf{L}}_{\boldsymbol{\beta}_\beta} = \partial^2 L(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T$, e $\ddot{\mathbf{L}}_{\boldsymbol{\beta}_y} = \partial^2 L(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\mathbf{y}^T$. Daí, têm-se que:

$$\mathbf{D}_{\boldsymbol{\beta}} = \mathbf{NX} \quad e \quad \ddot{\mathbf{L}}_{\boldsymbol{\beta y}} = \phi\mathbf{X^T V^{-1} N},$$

tal que $\mathbf{N} = diag\{d\mu_1/d\eta_1, \cdots, d\mu_n/d\eta_n\}$.

Ao substituir $-\ddot{\mathbf{L}}_{\boldsymbol{\beta\beta}}$ por seu valor esperado dado por $\phi(\mathbf{X_T W X})$, se terá aproximadamente

$$\widehat{GL} = \hat{\mathbf{N}}\mathbf{X}(\mathbf{X^T}\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X^T}\hat{\mathbf{V}}^{-1}\hat{\mathbf{N}}.$$

Desse modo, $\widehat{GL}_{ii}$ poderá ser escrito sob a forma:

$$\widehat{GL}_{ii} = \hat{w}_i x_i^T (\mathbf{X^T \hat{W} X})^{-1} x_i,$$

tal que $w_i = (d\mu_1/d\eta_1)^2/V_i$. Quando se está com a ligação canônica, em que $-\ddot{\mathbf{L}}_{\boldsymbol{\beta\beta}} = \phi(\mathbf{X^T V X})$ obtêm-se

$$\widehat{GL} = \hat{V}\mathbf{X}(\mathbf{X^T \hat{V} X})^{-1}\mathbf{X^T}.$$

Outra proposta para os MLG's e muito utilizada ultimamente foi definida por Pregibon (1981) o qual afirmou que, por meio da solução de máxima verossimilhança dos $\hat{\boldsymbol{\beta}}'s$ juntamente com a solução dos mínimos quadrados da regressão normal linear ponderada é possível detectar os pontos de alavanca, embora só coincida com a expressão anterior nos casos em que a resposta for contínua e a ligação do tipo canônico.

Assim, obteve-se por meio da convergência do processo iterativo mostrado em (2.7) a seguinte expressão para $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X^T \hat{V} X})^{-1}\mathbf{X^T \hat{W} \hat{z}},$$

sendo $\hat{\mathbf{z}} = \hat{\eta} + \hat{\mathbf{W}}^{-\mathbf{1/2}}\hat{\mathbf{V}}^{-\mathbf{1/2}}(\mathbf{y} - \hat{\mu})$. Desta forma, segundo Paula (2013) pode-se interpretar $\hat{\beta}$ como a solução de mínimos quadrados da regressão linear de $\hat{\mathbf{W}}^{1/2}\hat{\boldsymbol{z}}$ contra as colunas de $\hat{\mathbf{W}}^{1/2}\boldsymbol{X}$. Com isso, o resultado da matriz de projeção $\hat{\mathbf{H}}$ por meio da solução de mínimos quadrados da regressão linear de $\hat{\mathbf{z}}$ contra $\mathbf{X}$ com os pesos de $\hat{\mathbf{W}}$ ficará dada por:

$$\hat{\mathbf{H}} = \hat{\mathbf{W}}^{1/2}\boldsymbol{X}(\mathbf{X^T \hat{V} X})^{-1}\mathbf{X}\hat{\boldsymbol{W}}^{1/2},$$

a qual sugere utilizar os elementos da diagonal de $\hat{\mathbf{H}}$ referenciada como $\hat{h}_{ii}$ para detectar a presença ou não desses pontos de alavanca para esse modelo de regressão normal linear ponderada.

Verifica-se que para grandes amostras $\widehat{GL}$ e $\hat{\mathbf{H}}$ são iguais, sendo assim, $\hat{h}_{ii} = \widehat{GL}_{ii}$, sendo essa igualdade válida para qualquer tamanho amostral quando utilizada a ligação canônica. Sabendo que no geral $\hat{h}_{ii}$ é dependente de $\hat{\mu}_{ii}$, Paula (2013) sugere detectar pontos de alavanca por meio do gráfico de $\hat{h}_{ii}$ contra os valores ajustados.

No caso da regressão logística que será nosso foco principal, Hosmer e Lemeshow (1989) afirmam que o uso da matriz $\hat{\mathbf{H}}$ têm interpretações diferentes daquelas no caso normal linear.

2.10.2 Resíduos

Ao se ajustar um modelo de regressão a um conjunto de observações após uma escolha minuciosa de um modelo, é importante verificar possíveis afastamentos dos pontos

observados com os pontos do modelo estimado, levando em consideração a parte aleatória e a parte sistemática do modelo.

Segundo Cordeiro e Neto (2006), os resíduos no contexto dos MLG's são comumente usados para explorar a adequação do modelo ajustado no que diz respeito a escolha da distribuição proposta para a variável resposta. Isso é explicado devido a preocupação da ocorrência de desvios sistemáticos, ocasionados por uma escolha inadequada da função de ligação e da função de variância, e ainda pela presença de pontos discrepantes, que podem ter sido ocasionados devido a pontos que estão nos extremos da amplitude da variável independente ou que foram erroneamente coletados, ou ainda por algum motivo não controlado influenciou na obtenção destes dados.

Assim, um potencial problema encontrado no resíduo *studentizado* para os MLG's será a sua aplicabilidade de forma análoga ao da regressão normal linear, uma vez que as propriedades não necessariamente continuam valendo. Desta maneira, é de extrema importância que se tenha outros tipos de resíduos, os quais suas propriedades sejam conhecidas, ou que pelo menos esteja o mais próximo possível das propriedades de

$$t_i^* = t_i \left(\frac{n-p-1}{n-p-t_i^2} \right)^{1/2},$$

em que t_i^* segue uma distribuição t_{n-p-1}. Sendo $t_i = \frac{r_i}{s(1-h_i)^{1/2}}$ com $i = 1, \cdots, n$ e ainda que $s^2 = \sum_{i=1}^{n} r_i^2/(n-p)$.

Apresentam-se a seguir, os tipos de resíduos mais comuns nos MLG's:

i) Resíduo ordinário: É dado a partir da solução de mínimos quadrados da regressão linear ponderada de $\hat{\mathbf{z}}$ contra X, definida por:

$$\mathbf{r}^* = \hat{\mathbf{W}}^{1/2}[\hat{\mathbf{z}} - \hat{\boldsymbol{\eta}}] = \hat{\mathbf{V}}^{-\mathbf{1/2}}(\mathbf{y} - \hat{\boldsymbol{\mu}}).$$

ii) Resíduo padronizado: Assumindo que $Var(\mathbf{z}) \cong \hat{\mathbf{W}}^{-1}\phi^{-1}$, têm por aproximação que $Var(\mathbf{r}^*) \cong \phi^{-1}(\mathbf{I}_n - \hat{\mathbf{H}})$. Assim o resíduo fica sob a forma:

$$t_{S_i} = \frac{\phi^{-1/2}(y_i - \hat{\mu}_i)}{\sqrt{\hat{V}_i(1-\hat{h}_{ii})}},$$

sabendo que, $\hat{h}_{ii}$ será o i-ésimo elemento da diagonal principal da matriz de projeção $\mathbf{H}$, por resultados têm-se que $\mathbf{r}^* = (\mathbf{I}_n - \hat{\mathbf{H}})\hat{\mathbf{W}}^{1/2}\hat{\mathbf{z}}$, ou seja, $\hat{\mathbf{H}}$ tem a função de projeção ortogonal local da mesma forma como na regressão normal linear, onde $\mathbf{W}$ será uma identidade.

Todavia, $\hat{\boldsymbol{\eta}}$ é desconhecido e não é fixo, assim como $\mathbf{z}$ não têm distribuição normal. Portanto, as propriedades de t_i^* não podem ser mais verificadas para t_{S_i}, pois, um estudo de Williams (1984) apresenta por meio de Monte Carlo, que t_{S_i} é geralmente assimétrico, até mesmo para amostras grandes.

Outros resíduos são sugeridos, uma vez que podem se aproximar melhor da normalidade. São eles:

iii) Resíduo de Anscombe: Anscombe (1953) mostra que por meio de uma transformação de $\psi(\cdot)$ é utilizada visando tornar a distribuição de Y a mais próxima possível da distribuição normal. Deste modo, Barndorff-Nielsen (1978) apresenta que para MLG's essa transformação é dada por:

$$\psi(\mu) = \int_0^{\mu} V_{(t)}^{-1/3} dt.$$

Portanto, visando a normalização e a estabilidade da variância, chega-se ao resíduo de Anscombe definido como:

$$t_{A_i} = \frac{\phi^{1/2}\{\psi(y_i) - \psi(\hat{\mu}_i)\}}{\hat{V}^{1/2}(\hat{\mu}_i)\psi'(\hat{\mu}_i)},$$

Na Tabela 4, têm-se para as principais distribuições do $\psi(\cdot)$ para o resíduo de Anscombe.

Tabela 4 – Resíduo de Ascombe para as principais distribuições.

	Distribuição				
	Normal	Poisson	Binomial	Gama	Normal Inversa
$\psi(\mu)$	μ	$\frac{3}{2}\mu^{2/3}$	$\int_0^{\mu} t^{-1/3}(1-t)^{-1/3}dt$	$3\mu^{1/3}$	$\log \mu$

iv) Desvio residual: Os resíduos mais utilizados nos MLG's surgem por meio dos componentes da função desvio. De acordo com a padronização de McCullagh (1987), Davison e Gigli (1989), têm-se que:

$$t_{D_i} = \frac{d^*(y_i;\hat{\mu}_i)}{\sqrt{1-\hat{h}_{ii}}} = \frac{\phi^{1/2} d(y_i;\hat{\mu}_i)}{\sqrt{1-\hat{h}_{ii}}}$$

onde $d(y_i;\hat{\mu}_i) = \pm\sqrt{2}\{y_i(\tilde{\theta}_i - \hat{\boldsymbol{\theta}}) + (b(\hat{\theta}_i) - b(\tilde{\theta}_i))\}^{1/2}$, sendo o sinal de $d(y_i;\hat{\mu}_i)$ análogo à $y_i - \hat{\mu}_i$. McCullagh (1987) mostrou ainda que, a distribuição de probabilidade que se segue apresenta uma aproximação $N(0,1)$

$$\frac{d^*(Y_i;\mu_i) + \rho_{3i}/6}{\sqrt{1 + (14\rho_{3i}^2 - 9\rho_{4i})/36}}$$

tal que, ρ_{3i} e ρ_{4i} serão respectivamente os coeficientes de assimetria e curtose de $\partial L(\eta_i)/\partial\eta_i$, e ainda que $d^*(Y_i;\mu_i)$, será o i-ésimo componente associado ao desvio $D^*(\mathbf{y};\boldsymbol{\mu})$ do parâmetro verdadeiro.

v) Desvio residual ponderado: Definido por Williams (1987) será uma ponderação média entre t_{S_i} e t_{D_i}, como se segue

$$t_{G_i} = sinal(y_i - \hat{\mu}_i)\{(1 - \hat{h}_{ii})t_{D_i}^2 + h_{ii}t_{S_i}^2\}^{1/2}.$$

Williams (1987) afirmou também por meio de simulações, que para alguns casos de MLG's, t_{G_i} apresenta esperança ligeiramente diferente de zero, uma variância acima de um, assimetria desconsideráveis e alguma curtose.

2.10.3 Influência

Cordeiro e Neto (2006) mostram que o $\hat{h}_{ii}$ reflete de forma parcial a influência de uma determinada observação. Assim, considerando as estimativas dos parâmetros e os valores ajustados entre outros, faz-se necessária a comparação entre as estimativas de $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\beta}}_i$, sendo $\hat{\boldsymbol{\beta}}_i$ obtida por meio da exclusão da i-ésima observação.

Admitindo ϕ conhecido, a distância entre as verossimilhanças é dada por

$$L_{G_i} = 2\{L(\hat{\boldsymbol{\beta}}) - L(\hat{\boldsymbol{\beta}}_{(i)})\},$$

em que L_{G_i} será uma medida de verificação da influência que $\hat{\boldsymbol{\beta}}$ sofre quando retirada a i-ésima observação.

Fazendo a utilização da segunda aproximação por meio de série de Taylor em torno de $\hat{\boldsymbol{\beta}}$, chega-se a seguinte expansão:

$$L_{D_i} \cong (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T\{-\ddot{\mathbf{L}}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}})\}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).$$

Substituindo $-\ddot{\mathbf{L}}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}})$ por seu valor esperado, e ainda $\boldsymbol{\beta}$ por $\hat{\boldsymbol{\beta}}$, têm-se que

$$L_{D_i} \cong \phi(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T(\mathbf{X}^\mathbf{T}\hat{\mathbf{W}}\mathbf{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}). \tag{2.8}$$

Tendo desta maneira, uma melhor aproximação para L_{D_i} quando $L(\boldsymbol{\beta})$ estiver quadraticamente aproximado em torno de $\hat{\boldsymbol{\beta}}$.

Pregibon (1981) introduz uma aproximação onde considera a primeira iteração do processo iterativo por meio do método escore de Fisher quando o mesmo for inicializado em $\boldsymbol{\beta}$ uma possível obtenção para $\boldsymbol{\beta}_{(i)}$, expressado por

$$\hat{\boldsymbol{\beta}}_{(i)}^1 = \hat{\boldsymbol{\beta}} + \{-\ddot{L}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}})\}^{-1}\mathbf{L}_{(i)}(\hat{\boldsymbol{\beta}}),$$

tal que, $\mathbf{L}_{(i)}(\boldsymbol{\beta})$ será o logaritmo da função de verossimilhança sem a presença da i-ésima observação. Fazendo a substituição de $-\ddot{L}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}})$ por $\mathbf{K}(\hat{\boldsymbol{\beta}})$ obtêm-se

$$\hat{\boldsymbol{\beta}}_{(i)}^1 = \hat{\boldsymbol{\beta}} - \frac{\hat{r}p_i\sqrt{\hat{w}\phi^{-1}}}{(1 - \hat{h}_{ii})}(\mathbf{X}^\mathbf{T}\hat{\mathbf{W}}\mathbf{X})^{-1}x_i. \tag{2.9}$$

Assim, substituindo a expressão (2.9) em (2.8), a equação pela distância generalizada de Cook será dada por

$$L_{D_i} \cong \left\{ \frac{\hat{h}_{ii}}{(1-\hat{h}_{ii})} \right\} t^2_{S_i}.$$

Lee (1987) propõe que os pontos que apresentarem $L_{D_i} \sim \frac{\chi^2_p(\alpha)}{p}$ serão considerados influentes. Entretanto, McCullagh e Nelder (1989) diz que, por meio da estatística modificada de Cook, a influência de uma observação é medida por $T_i = \left\{ \frac{n-p}{p} \frac{h_{ii}}{1-h_{ii}} \right\}^{1/2} |r^2_{D_{(i)}}|$, onde $r^2_{D_{(i)}}$ é definido pela variação no desvio residual ocasionada pela omissão da i-ésima observação.

2.10.4 Técnicas gráficas

As técnicas gráficas são recursos aos quais o pesquisador pode recorrer para uma melhor análise do diagnóstico do modelo. Segundo Paula (2013), os gráficos mais recomendados no MLG são descritos a seguir:

i) Gráficos de t_{Di} contra a ordem das observações, contra os valores ajustados e contra as variáveis explicativas, ou contra o tempo ou alguma ordem onde haja suspeita de correlação entre as observações;

ii) Gráfico normal de probabilidades para t_{Di} com envelope;

iii) Gráfico de $\hat{z}_i$ contra $\hat{\eta}_i$ para verificação da adequação da função de ligação (uma tendência linear indica adequação da ligação);

iv) Gráficos de L_{D_i}, C_i ou $|\ell_{max}|$ contra a ordem das observações.

Gráficos normais de probabilidades apresentam características importantes, tais como, a identificação da distribuição originária dos dados e a identificação de valores que se sobressaem no conjunto das observações. Os envelopes, para a classe dos MLG's com distribuições diferentes da normal, são feitos com os resíduos gerados por meio do modelo ajustado, ver Williams (1984).

2.11 O Modelo logístico

Estudos que apresentam conclusões dicotômicas são comuns em diversas áreas do conhecimento, principalmente em pesquisas clínicas. Para tal situação, o modelo de regressão logística é uma ferramenta importante para descrever a relação existente entre a variável resposta com as variáveis explanatórias, sendo essa técnica amplamente utilizada para estimar associações por meio da medida de razão de chances.

Essa técnica é um dos casos particulares que estão apresentados nos modelos lineares generalizados MLG's por Dobson (1990), Paula (2013), onde esses modelos são classificados para variáveis que se apresentam apenas em duas categorias, ou que estão categorizadas assumindo valores 1 para " *sucesso*" ou 0 para "*fracasso*", conhecidas usualmente como variáveis *dummy*, dicotômicas ou binárias.

Como essas variáveis assumem respostas do tipo sucesso ou fracasso estão classificadas na distribuição Bernoulli, e ao se ter n ensaios dessa distribuição chega-se à distribuição Binomial. Têm-se deste modo, a base fundamental para o modelo de regressão logística, ao qual está baseado na transformação *logit* para proporções.

2.11.1 Regressão logística simples

Considerando aqui um modelo logístico simples, tal que $\pi(x)$ é a probabilidade de "sucesso" da variável explicativa X, e a transformação logística de $\pi(.)$ sob a forma linear no modelo têm-se que:

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = g(x), \tag{2.10}$$

em que

$$g(x) = \beta_0 + \beta_1 x.$$

Aplicando exponencial em ambos os membros da equação acima, têm-se que

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

em que e^{β_0} será o risco basal em escala exponencial e e^{β_1} será o tamanho do efeito associado ao fator de risco em escala exponencial, e ainda que, β_0 e β_1 são parâmetros desconhecidos ao qual se pretende estimar.

Deste modo, se quiséssemos analisar a associação entre uma determinada doença e a ocorrência ou não de um fator particular, se teria independentes amostras com n_1 indivíduos com $(x = 1)$ indicando presença do fator, e n_2 indivíduos com $(x = 0)$ indicando ausência do fator, considerando ainda que $\pi(x)$ será a probabilidade da doença se desenvolver após fixar um período. Assim, a chance de desenvolvimento para um indivíduo com presença do fator ficará

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

por outro lado, a chance de desenvolvimento para um indivíduo com ausência do fator ficará

$$\frac{\pi(0)}{1-\pi(0)} = e^{\beta_0}.$$

Logo, a razão de chances que dependerá apenas do parâmetro β_1 será expresso por

$$\psi = \frac{\pi(1)\{1-\pi(0)\}}{\pi(0)\{1-\pi(1)\}} = e^{\beta_1}.$$

Supondo agora que se tenha dois estratos sendo representados por x_1 ($x_1 = 0$ para estrato 1 e $x_1 = 1$ para estrato 2) sendo amostrados no estrato 1 com n_{11} indivíduos na presença do fator e n_{21} indivíduos na ausência do fator, e para o estrato 2 têm-se respectivamente n_{12} e n_{22} descritos do mesmo modo que o outro estrato. Para o desenvolvimento da doença, a probabilidade será $\pi(x_1, x_2)$ com $x_2 = 1$ para a presença do fator e $x_2 = 0$ para ausência do fator. Logo, se têm quatro parâmetros para se estimar, sendo: $\pi(0,0)$, $\pi(0,1)$, $\pi(1,0)$, $\pi(1,1)$. Qualquer reparametrização terá que conter para um modelo saturado no máximo quatro parâmetros.

Considerando a seguinte reparametrização, segue-se que:

$$\log\left[\frac{\pi(x_1,x_2)}{1-\pi(x_1,x_2)}\right] = \beta_0 + \gamma x_1 + \beta x_2 + \delta x_1 x_2,$$

onde segundo Paula (2013), γ será o efeito do estrato, β o efeito do fator e δ a interação entre estrato e fator. Assim, de acordo com essa reparametrização, a razão de chances para cada estrato será dada por:

$$\psi_1 = \frac{\pi(0,1)\{1-\pi(0,0)\}}{\pi(0,0)\{1-\pi(0,1)\}} = e^{\beta}$$

e

$$\psi_2 = \frac{\pi(1,1)\{1-\pi(1,0)\}}{\pi(1,0)\{1-\pi(1,1)\}} = e^{\beta+\delta}$$

Com isso, realizando a hipótese sob homogeneidade das razões de chances dado por ($H_0 : \psi_1 = \psi_2$), será o mesmo que testar a hipótese da ausência de interação de ($H_0 : \delta = 0$). Desta maneira, ao se ter a ausência da interação do fator com o estrato, pode-se dizer que a associação entre o fator e a doença são considerados os mesmos nos estratos. Todavia, há possibilidades de ocorrer efeito de estrato, por este motivo, supondo que não se rejeita a hipótese nula ($H_0 : \delta = 0$), a expressão que mostra como o logaritmo da chance do desenvolvimento da doença será equivalente nos dois extratos diferenciando apenas da quantidade de γ ficará dada por

$$\log = \left\{\frac{\pi(x_1,x_2)}{1-\pi(x_1,x_2)}\right\} = \beta_0 + \gamma x_1 + \beta x_2.$$

Isso poderá ser interpretado como, mesmo não obtendo interação entre os dois estratos, ou seja, a ausência da razão de chances de forma constante, as probabilidades de desenvolvimento da doença podem estar em patamares diferentes, sendo essas probabilidades diferentes de um estrato para outro (um dos estratos apresenta-se maior que o outro) Paula (2013).

2.11.2 Regressão logística múltipla

Considerando neste momento uma forma geral da regressão logística, têm-se que:

$$\log\left\{\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right\} = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \tag{2.11}$$

em que, $\mathbf{x} = (1, x_2, \cdots, x_p)^T$ assumirá valores observados das variáveis independentes. De acordo com o que foi mostrado anteriormente por meio do processo iterativo de mínimos quadrados reponderados afim de se obter $\hat{\beta}$, que

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

em que, $\mathbf{V} = diag\{\pi_1(1-\pi_1), \cdots, \pi_n(1-\pi_n)\}$, $\mathbf{z} = (z_1, \cdots, z_n)^T$ apresenta-se como variável independente modificada, $z_1 = \eta_1 + (y_1 - \pi_1)/\pi_1(1-\pi_1)$, com $m = 0, 1, \cdots$ e $i = 1, \cdots, n$. Ao se ter dados de forma agrupada com k grupos, substitui-se n por k, daí, $\mathbf{V} = diag\{n_1\pi_1(1-\pi_1), \cdots, n_k\pi_k(1-\pi_k)\}$ e $z_1 = \eta_1 + (y_1 - n_1\pi_1)/\{n_1\pi_1(1-\pi_1)\}$. Assintoticamente, quando $n \to \infty$ na primeira situação e para a segunda situação $\frac{n_1}{n} \to a_1 > 0$, ambos terão, $\boldsymbol{\hat{\beta} - \beta \sim N_p(0, (X^T V X)^{-1})}$.

Quando se têm $(q-1)(q \leq p)$ das $(p-1)$ variáveis explicativas binárias, é possível se ter interpretações bastante interessantes para as razões de chances.

Por exemplo, ao se ter $q = 4$ onde $x_2(x_2 = 1$ para presença, e $x_2 = 0$ para ausência) e $x_3(x_3 = 1$ para presença, e $x_3 = 0$ para ausência) representando dois fatores, e ainda que, $x_4 = x_2x_3$ representa a interação entre os dois fatores, o modelo poderá ficar sob a seguinte forma:

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \sum_{j=5}^{p} x_j\beta_j.$$

Segundo Paula (2013), denotando ψ_{ij} como sendo a razão de chances entre o indivíduo sob condição $(x_2 = i, x_3 = j)$ em relação a um indivíduo sob condição $(x_2 = 0, x_3 = 0)$ com $i, j = 0, 1$, admitindo os mesmos valores observados para os dois indivíduos e para as demais $(p-4)$ variáveis explicativas, mostra-se que

$$\psi_{10} = \exp(\beta_2), \quad \psi_{01} = \exp(\beta_3), \quad \psi_{11} = \exp(\beta_2 + \beta_3 + \beta_4).$$

Assim, testar a hipótese $H_0 : \beta_4 = 0$ (ausência de interação), será o mesmo que testar a hipótese de efeito multiplicativo $H_0 : \psi_{11} = \psi_{10}\psi_{01}$. De forma particular, se x_3 representar dois estratos ($x_3 = 0$ estrato 1, e $x_3 = 1$ estrato 2), se terá a razão de chances no primeiro estrato entre a presença e ausência do fator dado por $\psi_{10} = \exp(\beta_2)$, ao passo que, no segundo estrato a razão de chances ficará $\psi_{11}/\psi_{01} = \exp(\beta_2 + \beta_4)$. Portanto, testar $H_0 : \beta_4 = 0$ é o mesmo que testar a hipótese de homogeneidade das razões de chances nos dois estratos, Paula (2013).

De acordo com Paula (2013), ainda é possível calcular intervalos de confiança assintóticos para ψ com coeficiente $(1-\alpha)$.

$$(\hat{\psi}_I, \hat{\psi}_S) = \exp\{\hat{\beta} \pm z_{(1-\alpha/2)}\sqrt{Var(\hat{\beta})}\}.$$

3 Resultados Obtidos

Com base no estudo desenvolvido por Esteves (2007), foi possível por meio de uma rede de interação gênica elaborar um grafo o qual representa uma hipótese biológica acerca da existência de câncer em tecidos gastro-esofágicos. Assim, o grafo esquematizado na Figura 1 apresenta um perfil de interação entre um grupo de genes onde, existe uma hipótese de que a expressão dos genes CCL20, CCL18 e IFNAR2, deveriam induzir também a expressão dos genes ADH1B, AKR1B10, ALDH3A2 e IL1R2 em tecidos gastro-esofágicos normais, o que não deveria acontecer em tecidos afetados pelo câncer.

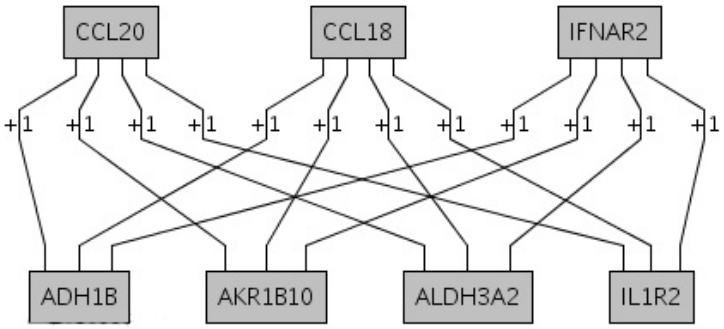

Figura 1 – Rede de interação gênica para os tecidos normais e tumorais. Grafo representando uma rede de interação gênica trazendo um modelo biológico que tenta explicar a existência ou não de tumores nos pacientes. Segundo esta rede, aumento de expressão nos genes CCL20, CCL18 e IFNAR2 deve ser acompanhado de aumento para os genes ADH1B, AKR1B10, ALDH3A2 e IL1R2 nos tecidos normais [Extraída de Esteves (2007)].

Neste trabalho foi usado um conjunto de dados (ainda não publicado) de expressão gênica obtido a partir de uma colaboração com um grupo de professores do Hospital Sírio-Libanês e do Instituto de Matemática e Estatística da USP, ambos de São Paulo-SP. A partir destes dados, foi possível obter os valores de expressão para os genes CCL20, CCL18, IFNAR2 e AKR1B10 para 57 observações de tecidos gastro-esofágicos normais, o que nos permitiu focar nosso estudo em um subgrafo obtido a partir da Figura 1, que está representado na Figura 2.

Os valores de expressão obtidos são registrados como níveis de intensidade de sinal luminoso obtidos a partir da técnica de *microarray* e, para o uso do modelo de regressão logística foi preciso discretizar estes dados. Para isso, cada um dos quatro genes representados na Figura 2 foi classificado como expresso (1) ou não expresso (0) se seu valor de expressão estava maior ou menor, respectivamente, do que sua média para as 57 observações disponíveis.

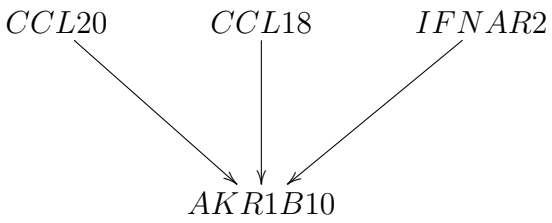

Figura 2 – Subgrafo construído a partir do grafo dado pela Figura 1.

Desta forma, o objetivo principal deste trabalho foi usar a regressão logística para tentar modelar cada aresta dada no grafo da Figura 2, tentando medir se a expressão de cada um dos genes CCL, CCL18 e IFNAR2 era capaz de explicar a expressão do gene AKR1B10. Desta forma, a expressão ou não deste último gene foi usada como variável resposta em três modelos logísticos com os outros três genes como variáveis explicativas.

Assim, a Tabela 5 apresenta as frequências conjuntas para as expressões dos genes AKR1B10 e CCL20, sendo que a partir destes valores é possível se calcular os parâmetros de interesse para a regressão logística entre estes dois genes.

Tabela 5 – Tabela resumo fixando AKR1B10 com o gene CCL20.

	CCL20	
AKR1B10	Não expresso	Expresso
Não expresso	10	11
Expresso	21	15

Aqui denota-se por $\pi(x)$ a probabilidade do gene AKR1B10 estar expresso, dada a expressão ou não do gene CCL20, que é denotada pela variável x definida por

$$x = \begin{cases} 1, & \text{CCL20 ser expresso} \\ 0, & \text{CCL20 não ser expresso.} \end{cases}$$

Aqui os parâmetros de interesse são $\pi(1)$ - a probabilidade do gene AKR1B10 ser expresso dado que o gene CCL20 também está expresso e $\pi(0)$ - probabilidade de AKR1B10 ser expresso dado que CCL20 não é. Daí, fazendo alguns cálculos simples têm-se que

$$\pi(1) = \frac{15}{26} \quad \text{e} \quad \pi(0) = \frac{21}{31}$$

Sabendo que,

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Fazendo as contas para $x = 0$,

$$\log\left(\frac{\pi(0)}{1-\pi(0)}\right) \Rightarrow \log\left(\frac{21/31}{10/31}\right) \Rightarrow \log\left(\frac{21}{10}\right) \cong 0,742 \cong \hat{\beta}_0,$$

e para $x = 1$,

$$\log\left(\frac{\pi(1)}{1-\pi(1)}\right) \Rightarrow \log\left(\frac{15/26}{11/26}\right) \Rightarrow \log\left(\frac{15}{11}\right) \cong 0,310 \cong \hat{\beta}_0 + \hat{\beta}_1,$$

onde

$$\begin{aligned} \hat{\beta}_0 + \hat{\beta}_1 &\cong 0,310 \\ \hat{\beta}_1 &\cong 0,310 - 0,742 \\ \hat{\beta}_1 &\cong -0,432 \end{aligned}$$

Por fim, a razão de chances é dada por

$$\begin{aligned} \hat{\psi} &= \frac{15/11}{21/10} \\ \hat{\psi} &= 0,65, \end{aligned}$$

cujo intervalo de 95% de confiança é $(0,22; 1,92)$.

Portanto, têm-se que

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) \cong 0,742 - 0,432x,$$

ou ainda que

$$\frac{\pi(x)}{1-\pi(x)} \cong e^{0,742-0,432x}.$$

Assim, nota-se por meio da Tabela 5 que a princípio o modelo de regressão apresentou o efeito basal de forma positiva, ou seja, a proporção desconsiderando a variável x – sendo x igual a zero – está possivelmente induzindo a expressão do gene AKR1B10, porém, ao se considerar x – sendo x igual a um – têm-se que o tamanho do efeito que será a expressão do gene CCL20 afeta na expressão do AKR1B10 de forma a reprimí-lo.

Os mesmos cálculos apresentados acima foram feitos também pelo *software* `R`, cujos resultados estão na Tabela 6. A vantagem do uso do `R` é o fato de que o mesmo já apresenta as estimativas de erro padrão e do teste z para a significância dos parâmetros, onde constata-se que a expressão do gene CCL20 não altera significativamente a expressão do gene AKR1B10, com nível de significância 0,4345.

Como conclusão desta primeira análise, a razão de chances do gene AKR1B10 ser expresso quando o CCL20 é expresso em comparação de quando não é expresso em pacientes normais será 0,65, ou seja, a chance do AKR1B10 ser expresso é 0,65 mais vezes quando o CCL20 for expresso comparativamente quando CCL20 for expresso. O

Tabela 6 – Ajuste do modelo logístico referente ao gene AKR1B10 com o gene CCL20.

Parâmetro	Estimativa	Erro Padrão	Valor z	Pr(>|z|)
β_0	0,7419	0,3842	1,93	0,0535
β_1	-0,4318	0,5524	-0,78	0,4345

desvio residual apresentou $74,41$ para 55 graus de liberdade. Porém, este resultado não é significativo como foi apontado na Tabela 6.

Esta mesma modelagem foi feita também para a aresta do grafo da Figura 2, que representa a interação entre os genes CCL18 e AKR1B10. Desta forma, a Tabela 7 apresenta as frequências conjuntas entre os valores de expressão para estes dois genes.

Tabela 7 – Tabela resumo fixando AKR1B10 com o gene CCL18.

	CCL18	
AKR1B10	Não expresso	Expresso
Não expresso	14	7
Expresso	11	25

Agora $\pi(x)$ é redefinida como a probabilidade do AKR1B10 estar expresso, dado que o CCL18 foi expresso ou não, onde os valores de x são dados por

$$x = \begin{cases} 1, & \text{CCL18 ser expresso} \\ 0, & \text{CCL18 não ser expresso} \end{cases}$$

Após atualização dos cálculos para o parâmetros de interesse, $\pi(1)$ e $\pi(0)$, têm-se que

$$\pi(1) = \frac{25}{32} \quad \text{e} \quad \pi(0) = \frac{11}{25}.$$

Fazendo as contas para $x = 0$,

$$\log\left(\frac{\pi(0)}{1-\pi(0)}\right) \Rightarrow \log\left(\frac{11/25}{14/25}\right) \Rightarrow \log\left(\frac{11}{14}\right) \cong -0,241 \cong \hat{\beta}_0,$$

e para $x = 1$,

$$\log\left(\frac{\pi(1)}{1-\pi(1)}\right) \Rightarrow \log\left(\frac{25/32}{7/32}\right) \Rightarrow \log\left(\frac{25}{7}\right) \cong 1,273 \cong \hat{\beta}_0 + \hat{\beta}_1,$$

onde

$$\begin{aligned} \hat{\beta}_0 + \hat{\beta}_1 &\cong 1,273 \\ \hat{\beta}_1 &\cong 1,273 + 0,241 \\ \hat{\beta}_1 &\cong 1,514 \end{aligned}$$

E enfim, a razão de chances fica dada por

$$\hat{\psi} = \frac{25/7}{11/14}$$
$$\hat{\psi} = 4,55$$

com intervalo de 95% de confiança dado por $(1,44; 14,38)$.

Portanto, pode-se concluir que

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) \cong -0,241 + 1,514x,$$

ou ainda que

$$\frac{\pi(x)}{1-\pi(x)} \cong e^{-0,241+1,514x}$$

Deste modo, na Tabela 7 o modelo de regressão indica que o efeito basal desconsiderando a variável x – sendo x igual a zero – está possivelmente reprimindo a expressão do gene AKR1B10. Entretanto, considerando x igual a um, o tamanho do efeito que será a expressão do gene CCL18 afeta na expressão do AKR1B10 de forma a induzí-lo.

Como no caso anterior, a Tabela 8 proporciona a saída do `R`, onde pode-se constatar agora que a expressão do gene CCL18 altera significativamente a expressão do gene AKR1B10 (nível descritivo de $0,01$).

Tabela 8 – Ajuste do modelo logístico referente ao gene AKR1B10 com o gene CCL18.

Parâmetro	Estimativa	Erro Padrão	Valor z	Pr(>|z|)
β_0	-0,2412	0,4029	-0,60	0,5495
β_1	1,5141	0,5875	2,58	0,0100

Logo, a razão de chances do AKR1B10 ser expresso quando o CCL18 é expresso em relação a chance de AKR1B10 ser expresso quando CCL18 não é em pacientes normais é $4,55$, ou seja, a chance do AKR1B10 ser expresso é 4,55 vezes maior quando o CCL18 for expresso comparativamente quando CCL18 não for expresso. O desvio residual apresentou $67,917$ para 55 graus de liberdade, e este resultado é significativo como foi mostrado na Tabela 8.

Finalmente, a Tabela 9 apresenta as frequências conjuntas para os valores de expressão dos genes AKR1B10 e IFNAR2, com intuito de estimar os parâmetros da regressão logística para a associação entre estes dois genes.

O valor $\pi(x)$ agora representa a probabilidade do gene AKR1B10 estar expresso, dada a expressão do gene IFNAR2, sendo que x é definido por

$$x = \begin{cases} 1, & \text{IFNAR2 ser expresso} \\ 0, & \text{IFNAR2 não ser expresso} \end{cases}$$

Tabela 9 – Tabela resumo fixando AKR1B10 com o gene IFNAR2.

	IFNAR2	
AKR1B10	Não expresso	Expresso
Não expresso	7	14
Expresso	16	20

Assim, atualizando $\pi(1)$ e $\pi(0)$, e refazendo as contas têm-se que

$$\pi(1) = \frac{20}{34} \quad \text{e} \quad \pi(0) = \frac{16}{23}.$$

Fazendo para $x = 0$,

$$\log\left(\frac{\pi(0)}{1-\pi(0)}\right) \Rightarrow \log\left(\frac{16/23}{7/23}\right) \Rightarrow \log\left(\frac{16}{7}\right) \cong 0,827 \cong \hat{\beta}_0,$$

e para $x = 1$,

$$\log\left(\frac{\pi(1)}{1-\pi(1)}\right) \Rightarrow \log\left(\frac{20/34}{14/34}\right) \Rightarrow \log\left(\frac{20}{14}\right) \cong 0,357 \cong \hat{\beta}_0 + \hat{\beta}_1,$$

onde

$$\begin{aligned} \hat{\beta}_0 + \hat{\beta}_1 &\cong 0,357 \\ \hat{\beta}_1 &\cong 0,357 - 0,827 \\ \hat{\beta}_1 &\cong -0,47 \end{aligned}$$

Finalmente a razão de chances é dada por

$$\begin{aligned} \hat{\psi} &= \frac{20/14}{16/7} \\ \hat{\psi} &= 0,63 \end{aligned}$$

apresentando um intervalo de 95% de confiança de $(0,2; 1,92)$.

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) \cong 0,827 - 0,47x,$$

ou ainda que

$$\frac{\pi(x)}{1-\pi(x)} \cong e^{0,827-0,47x}.$$

Assim, têm-se que na Tabela 9 os resultados são parecidos com os da Tabela 7, porém com intensidade menor na questão da indução.

A Tabela 10 apresenta os resultados dos cálculos feitos pelo `R`, onde da mesma forma que no caso do gene CCL20, o resultado não foi significativo, com nível descritivo do teste de 0,411.

Tabela 10 – Ajuste do modelo logístico referente ao gene AKR1B10 com o gene IFNAR2.

Parâmetro	Estimativa	Erro Padrão	Valor z	Pr(>|z|)
β_0	0,8267	0,4532	1,82	0,0681
β_1	-0,4700	0,5717	-0,82	0,4110

Logo, a razão de chances do AKR1B10 ser expresso quando o IFNAR2 é expresso por AKR1B10 ser expresso quando IFNAR2 não é expresso em pacientes normais será $0,63$, ou seja, a chance do AKR1B10 ser expresso é $0,63$ mais vezes quando o IFNAR2 for expresso comparativamente quando IFNAR2 não for expresso. O desvio residual apresentou $74,337$ para 55 graus de liberdade, e o resultado não é significativo.

Até agora foi possível ajustar um modelo de regressão simples aos dados de forma mais fácil por assim dizer, onde se calculou o tamanho da interação entre eles. Mas ao se ter um modelo com mais variáveis, torna-se mais difícil achar uma expressão analítica. Para isso, se usa recursos mais usados na técnica dos MLG's, onde por meio de métodos iterativos se chega ao modelo desejado, que é o que é implementado no programa `R`. Assim, foi ajustado também um último modelo, considerando a regressão logística múltipla, onde ajustou-se a expressão do gene AKR1B10 em função dos valores de expressão dos outros três estudados simultaneamente.

Os resultados deste ajuste estão na Tabela 11, onde nota-se que os resultados anteriores se mantém, mesmo no modelo de regressão logística múltipla. Nesta tabela os parâmetros β_1, β_2 e β_3 estão associados aos fatores de expressão dos genes CCL20, IFNAR2 e CCL18, respectivamente. O desvio residual apresentou $67,408$ para 53 graus de liberdade, onde apenas o CCL18 foi significativo para o modelo.

Tabela 11 – Resumo das estimativas para a análise múltipla.

Parâmetro	Estimativa	Erro Padrão	Valor z	Pr(>|z|)
β_0	0,0547	0,5789	0,09	0,9248
β_1	-0,1696	0,6524	-0,26	0,7949
β_2	-0,3284	0,6704	-0,49	0,6243
β_3	1,4830	0,5918	2,51	0,0122

Assim, acredita-se que o melhor ajuste nesse caso seria o modelo que apresentasse apenas o efeito da covariável CCL18. Porém, por características apresentadas pelo pesquisador, o modelo geral com todas as covariáveis faz mais sentido biologicamente, assim para uma validação do ajuste realizou-se as técnicas de diagnóstico levando em consideração o modelo proposto com CCL20, IFNAR2 e CCL18.

Dessa maneira, têm-se que o desvio no modelo geral foi de $D(y;\hat{\mu}) = 67,408$ para 53 graus de liberdade, indicando um bom ajuste. A Figura 3 apresenta alguns gráficos da análise de diagnóstico do modelo ajustado.

Na Figura 3.*a* têm-se o gráfico de $\hat{h}_{ii}$ contra os valores ajustados, nota-se que o ponto 9 se encontra no limite, porém não ultrapassa o mesmo, sendo assim, por meio desse gráfico nenhum dos pontos podem ser considerados como alavanca. No gráfico de resíduos t_{Di} , Figura 3.*b*, todos os pontos se encontram dentro do intervalo [-2,2]. Em 3.*c* tem-se que o gráfico corrobora 3.*b*, por meio dos valores ajustados. Quanto ao gráfico de influência L_{Di} na Figura 3.*d*, observa-se que todos os pontos se encontram próximos, indicando que não se tem de fato pontos aberrantes. Isto nos leva a crer que o modelo está bem ajustado.

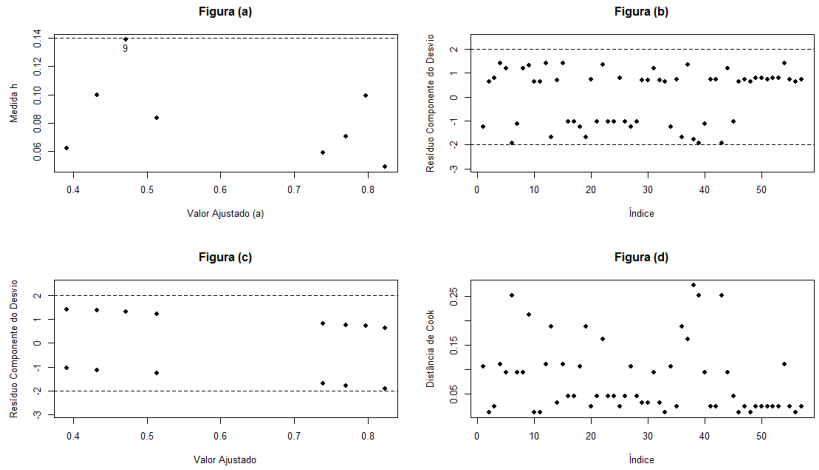

Figura 3 – Gráficos de diagnóstico referentes ao modelo logístico ajustado aos dados com a presença de todas as covariáveis.

Na Figura 4 apresenta-se o gráfico normal de probabilidades para o resíduo t_{Di} nos dois casos e não notamos nenhum indício de que a distribuição utilizada seja inadequada, observando ainda que, tanto para o caso múltiplo quanto para o simples, as bandas de confiança se apresentaram de forma bem parecidas.

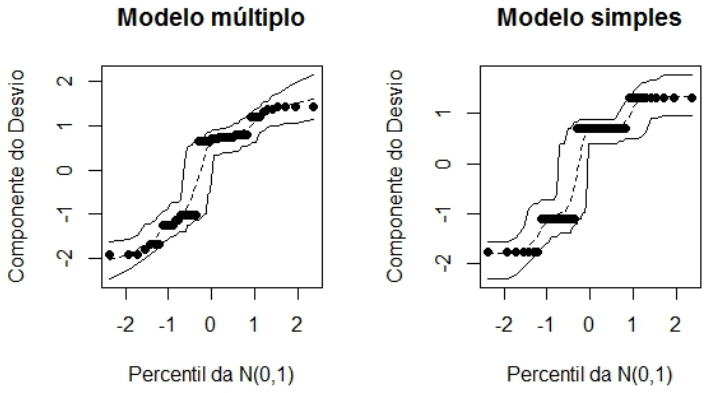

Figura 4 – Bandas de confiança referente aos modelos logísticos múltiplo (modelo completo) e simples (modelo apenas com o CCL18) respectivamente.

4 Conclusão

A necessidade de conhecer os efeitos das variáveis independentes sobre a variável dependente, faz com que a importância de se entender e utilizar modelos de regressão se torne uma técnica de bastante interesse, principalmente na área da saúde.

Neste trabalho foi apresentado um estudo teórico dos modelos lineares generalizados com ênfase na regressão logística. Percebe-se que a aplicação dos MLG's para dados referentes a pesquisa do estudo do câncer de esôfago e estômago se constitui de uma técnica bastante confiável, onde os algoritmos implementados no *software* `R` 3.0.1 se mostraram consistentes no ajuste de modelos lineares generalizados logísticos, com função de ligação *logit*.

Foram ajustados modelos para caso simples e múltiplo, onde para este último modelo realizou-se técnicas de diagnóstico. Dessa forma, foi possível concluir estatisticamente que, apenas o CCL18 contribui para o gene AKR1B10 está expresso, diferentemente da hipótese anterior, a qual apresentava como resultado que, quando os genes CCL18, CCL20 e IFNAR2 estavam expressos, o gene AKR1B10 também estaria expresso.

Entretanto, mesmo sabendo que apenas o CCL18 foi significativo no modelo, foi considerado aqui todas as covariáveis (CCL18, CCL20 e IFNAR2) devido a proposta inicial do pesquisador, sendo assim, com base nesse resultado e nas análises técnicas para a validação desse modelo, pode-se dizer que o ajuste está adequado aos dados.

Referências

ANSCOMBE, F. J. Contribution to the discussion of h. hotelling's paper. J. R. Statist. Soc. B,15, p. 229–230, 1953. Citado na página 30.

ATKINSON, A. C. Plots, Transformation and regression. Oxford: Clarendon Press, 1985. Citado na página 15.

ATKINSON, A. C. Two graphical display for outlying and influential observations in regression. Biometrika, p. 68, 13 – 20, 1987. Citado na página 15.

BARNDORFF-NIELSEN, O. E. Information and exponencial families in statistical theory. John Wiley & Sons, New York, 1978. Citado na página 30.

BARROSO, L. P.; VASCONCELLOS, K. L. P. Second-order asymtotic for score tests in heterocedastic t regression models. Comm. Statist.-Theory Methods, 2002. Citado na página 15.

BOX, G. E. P.; COX, D. R. An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B*, p. 26, 211–252, 1964. Citado na página 14.

CARROL, R. J.; RUPPERT, D. On robust test for heteroscedasticity. Ann. Statist, p. 48,133–169, 1988. Citado na página 15.

COOK, R. D.; WEISBERG, S. Diagnostics for heteroscedasticity in regression. Biometrika, p. 70, 1–10, 1983. Citado 2 vezes nas páginas 15 e 27.

CORDEIRO, G.; NETO, E. L. Modelos Paramétricos. Recife: Universidade Federal Rural de Pernambuco, Departamento de Estatística e Informática, 2006. Citado 2 vezes nas páginas 29 e 31.

CORDEIRO, G. M.; DEMéTRIO, C. G. B. *Modelos Lineares Generalizados e Extensões*. [S.l.], 2007. Citado 2 vezes nas páginas 12 e 13.

CORDEIRO, G. M.; PAULA, G. A. Improved likelihood ratio statistics for exponential family nonlinear models. Biometrika, p. 76, 93–100, 1989. Citado na página 15.

DAVISON, A. C.; GIGLI, A. Deviance residuals and normal scores plots. Biometrika 76, p. 211–221, 1989. Citado na página 30.

DOBSON, A. J. An introduction to generalized linear models. Chapman & Hall, London, 1990. Citado na página 33.

ESTEVES, G. H. *Métodos estatísticos para a análise de dados de cDNA microarray em um ambiente computacional integrado*. Tese (Tese (Doutorado)) — Universidade de São Paulo, 2007. Disponível em: <http://www.teses.usp.br/teses/disponiveis/95/95131/tde-03062007-210232/>. Citado 2 vezes nas páginas 8 e 37.

FISHER, R. Two new properties of mathematical likelihood. *Philosophical Transactions of the Royal Society A*, v. 144, p. 285,307, 1934. Citado na página 16.

HARVEY, A. C. Estimating regression models with multiplicative heteroscedasticity. Econometrika, p. 41,461–465, 1976. Citado na página 15.

HASTIE, T.; TIBSHIRANI, R. Generalized Additive Models. Chapman and Hall, London, 1990. Citado na página 15.

HOSMER, D. W.; LEMESHOW, S. Applied Logistic Regression. John Wiley, New York, 1989. Citado na página 28.

JORGENSEN, B. Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. Biometrika, p. 70, 19–28, 1983. Citado na página 15.

LEE, A. Diagnostic displays for assessing leverage and influence in generalized linear models. Austral. J. Statist., p. 29,233–243, 1987. Citado na página 32.

LEE, Y.; NELDER, J. A. Hierarchical generalized linear models. *Journal of the Royal Statistical Society B*, 1996. Citado na página 15.

LEE, Y.; NELDER, J. A. Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. Biomerika, 2001. Citado na página 15.

LIANG, K. Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. Biometrika, p. 73, 13–22, 1986. Citado na página 15.

MCCULLAGH, P. Tensor Methods in Statistics. Chapman and Hall, London, 1987. Citado na página 30.

MCCULLAGH, P.; NELDER, J. A. *Generalized Linear Models*. 2. ed. London: Chapman and Hall, 1989. Citado 3 vezes nas páginas 15, 22 e 32.

MYERS, R.; MONTGOMERY, D. C.; VINING, G. G. Generalized Linear Models: With Applications in Engineering and the Sciences. John Wiley, New York, 2002. Citado na página 20.

NELDER, J.; WEDDERBURN, R. Generalized Linear Models. *Journal of the Royal Statistical Society A*, v. 135, p. 370,384, 1972. Citado 4 vezes nas páginas 6, 7, 15 e 23.

PAULA, G. A. de. *Modelos de Regressão com apoio computacional*. São Paulo, 2013. Disponível em: <http://www.ime.usp.br/~giapaula/texto_2013.pdf>. Citado 15 vezes nas páginas 12, 13, 15, 16, 21, 23, 25, 26, 27, 28, 32, 33, 34, 35 e 36.

PREGIBON, D. Logistic regression diagnostics. Annals of Statistics 9, p. 705–724, 1981. Citado 2 vezes nas páginas 28 e 31.

SMYTH, G. K. Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society B*, p. 51, 47–60, 1989. Citado na página 15.

TAYLOR, J.; VERBYLA, A. P. Joint modelling of location andscaleparameterof the t distribution. Statist. Model, 2004. Citado na página 15.

TURKMAN, M. A. A.; SILVA, G. *Modelos Lineares Generalizados da Teoria à Prática*. Spe. Lisboa: [s.n.], 2000. Citado na página 14.

WEDDERBURN, R. W. M. Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. Biometrika, p. 61, 439–447, 1974. Citado na página 15.

WEI, B.; HU, Y.; FUNG, W. Generalized leverage and its applications. *Scandinavian Journal of Statistics*, v. 25, p. 25–37, 1998. Citado na página 27.

WILLIAMS, D. A. Residuals in generalized linear models. In: Proceedings of the 12th. International Biometrics Conference, Tokyo, 1984. Citado 2 vezes nas páginas 29 e 32.

WILLIAMS, D. A. Generalized linear model diagnostic using the deviance and single case deletion. Applied Statistics 36, 1987. Citado na página 31.

Apêndices

APÊNDICE A – Apêndice A

Script para análise no *software* R.

```
# Reconhecendo o banco de dados:
require(MASS)
require(epicalc)
#load("dados.RData")
names(dados)
str(dados)

########## Ajustando os dados##########
ajuste< −glm(AKR1B10 ∼ CCL20 + IFNAR2 + CCL18,family="binomial",data=dados)
stepAIC(ajuste)
anova(ajuste,test="Chisq")# testa hipótese sobre os efeitos
confint.default(ajuste)
summary.glm(ajuste)

logistic.display(ajuste)
exp(coef(ajuste))


        attach(dados)
fit.model = ajuste

########## Diagnóstico ##########
# Plot
X <- model.matrix(fit.model)
n <- nrow(X)
p <- ncol(X)
w <- fit.model $ weights
W <- diag(w)
H <- solve(t(X)%*%W%*%X)
H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
h <- diag(H)
ts <- resid(fit.model,type="pearson")/sqrt(1 − h)
td <- resid(fit.model,type="deviance")/sqrt(1 − h)
di <- (h/(1 − h))*(ts^2)
```

```
a <- max(td)
b <- min(td)
par(mfrow=c(2,2))

plot(fitted(fit.model),h,xlab="Valor Ajustado (a)",
ylab="Medida h",main="Figura (a)", pch=16)
cut <- 2 * p/n
abline(cut,0,lty=2)
identify(fitted(fit.model), h, n=1)

plot(di,xlab="Índice", ylab="Distância de Cook",main="Figura (d)",pch=16)

plot(td,xlab="Índice", ylab="Resíduo Componente do Desvio",
ylim=c(b - 1, a + 1),main="Figura (b)", pch=16)
abline(2,0,lty=2)
abline(-2,0,lty=2)

plot(fitted(fit.model), td ,xlab="Valor Ajustado",
ylab="Resíduo Componente do Desvio",ylim=c(b - 1, a + 1) ,main="Figura (c)",pch=16)
abline(2,0,lty=2)
abline(-2,0,lty=2)


      ####### Construindo o envelope para o modelo proposto ######
par(mfrow=c(1, 2))
X <- model.matrix(fit.model)
n <- nrow(X)
p <- ncol(X)
w <- fit.model$ weights
W <- diag(w)
H <- solve(t(X)% * % W % * % X)
H <- sqrt(W)% * % X % * % H % * % t(X) % * % sqrt(W)
h <- diag(H)
td <- resid(fit.model,type="deviance")/sqrt(1 - h)
e <- matrix(0,n,100)

for(i in 1 : 100){
dif <- runif(n) - fitted(fit.model)
dif[dif >= 0 ] <- 0
dif[dif< 0] <- 1
```

```
nresp <- dif
fit <- glm(nresp ~ X, family=binomial)
w <- fit$ weights
W <- diag(w)
H <- solve(t(X)% * % W % * % X)
H <- sqrt(W)% * % X % * % H % * % t(X) % * % sqrt(W)
h <- diag(H)
e[,i] <- sort(resid(fit,type="deviance")/sqrt(1 − h))}

e1 <- numeric(n)
e2 <- numeric(n)

for(i in 1 : n){
eo <- sort(e[i,])
e1[i] <- (eo[2] + eo[3])/2
e2[i] <- (eo[97] + eo[98])/2 }

med <- apply(e,1,mean)
faixa <- range(td,e1,e2)
par(pty="s")
qqnorm(td,xlab="Percentil da N(0, 1)",
ylab="Componente do Desvio", ylim=faixa, pch=16, main="Modelo múltiplo")
par(new=T)
qqnorm(e1,axes=F,xlab="",ylab="",type="l",ylim=faixa,lty=1, main="")
par(new=T)
qqnorm(e2,axes=F,xlab="",ylab=, type="l",ylim=faixa,lty=1, main="")
par(new=T)
qqnorm(med,axes=F,xlab="", ylab="", type="l",ylim=faixa,lty=2, main="")

########## Fazendo o envelope para o ajuste do CCL18 ##########
modelo<-glm(AKR1B10 ~ CCL18,family="binomial",data=dados)
fit.model = modelo
X <- model.matrix(fit.model)
n <- nrow(X)
p <- ncol(X)
w <- fit.model$ weights
W <- diag(w)
H <- solve(t(X)% * % W % * % X)
H <- sqrt(W)% *% X % * % H % * % t(X)% * % sqrt(W)
```

```
h <- diag(H)
td <- resid(fit.model,type="deviance")/sqrt(1 − h)
e <- matrix(0,n,100)

for(i in 1 : 100){
dif <- runif(n) - fitted(fit.model)
dif[dif >= 0 ] <- 0
dif[dif<0] <- 1
nresp <- dif
fit <- glm(nresp ∼ X, family=binomial)
w <- fit$ weights
W <- diag(w)
H <- solve(t(X)% * % W % * % X)
H <- sqrt(W)% *% X % * % H % * % t(X)% * % sqrt(W)
h <- diag(H)
e[,i] <- sort(resid(fit,type="deviance")/sqrt(1 − h))}

e1 <- numeric(n)
e2 <- numeric(n)

for(i in 1 : n){
eo <- sort(e[i,])
e1[i] <- (eo[2] + eo[3])/2
e2[i] <- (eo[97] + eo[98])/2 }

med <- apply(e,1,mean)
faixa <- range(td,e1,e2)
par(pty="s")
qqnorm(td,xlab="Percentil da N(0, 1)",
ylab="Componente do Desvio", ylim=faixa, pch=16, main="Modelo simples")
par(new=T)
qqnorm(e1,axes=F,xlab="",ylab="",type="l",ylim=faixa,lty=1, main="")
par(new=T)
qqnorm(e2,axes=F,xlab="",ylab=, type="l",ylim=faixa,lty=1, main="")
par(new=T)
qqnorm(med,axes=F,xlab="", ylab="", type="l",ylim=faixa,lty=2, main="")
```