UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

DANILLO BARROS CORDEIRO

AJUSTE DE MODELOS LINEARES GENERALIZADOS PARA DADOS POSITIVOS ASSIMÉTRICOS

CAMPINA GRANDE - PB

OUTUBRO 2016

DANILLO BARROS CORDEIRO

AJUSTE DE MODELOS LINEARES GENERALIZADOS PARA DADOS POSITIVOS ASSIMÉTRICOS

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: Prof. Dr. Ricardo Alves Olinda

CAMPINA GRANDE - PB
OUTUBRO 2016

C794a Cordeiro, Danillo Barros.
Ajuste de modelos lineares generalizados para dados positivos assimétricos [manuscrito] / Danillo Barros Cordeiro. - 2016.
70 p. : il. color.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2016.
"Orientação: Prof. Dr. Ricardo Alves Olinda, Departamento de Estatística".

1. Modelos lineares generalizados. 2. Dados positivos assimétricos. 3. Função pulmonar. I. Título.

21. ed. CDD 519.5

DANILLO BARROS CORDEIRO

AJUSTE DE MODELOS LINEARES GENERALIZADOS PARA DADOS POSITIVOS ASSIMÉTRICOS

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 05 de outubro de 2016.

BANCA EXAMINADORA

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

Prof. Dr. Kleber N. Nunes de Oliveira Barros
Universidade Estadual da Paraíba

Prof. Dr. Sílvio F. Alves Xavier Júnior
Universidade Estadual da Paraíba

Agradecimentos

A todos os professores do departamento de estatística que, direta ou indiretamente, me ajudaram a chegar onde cheguei. Ao longo desses anos, tive a oportunidade de aprender com os melhores e mais competentes professores que eu poderia ter.

Ao meu orientador Doutor Ricardo Alves Olinda pela sua dedicação, profissionalismo, humildade e sua disponibilidade de ajudar.

Ao Doutor Kleber Napoleão Nunes de Oliveira Barros pela amizade, conhecimento e por me ajudar sempre que podia, principalmente enquanto eu estava no meu intercâmbio.

Aos meus colegas que também contribuiram para meu aprendizado e por estarem unidos nos momentos difíceis ao longo do curso, enriquecendo minha vida acadêmica.

"A estatística é a gramática da ciência."
(Karl Pearson)

"Essencialmente, todos os modelos estão errados, mas alguns são úteis."
(George Box)

"A verdadeira ciência ensina sobretudo a duvidar e a ser ignorante."
(Miguel de Unamuno)

Resumo

O modelo de regressão clássico pressupõe que a variável resposta seja simétrica e homocedástica. Porém, em muitas situações esses pressupostos não são alcançados e precisa-se de uma abordagem mais flexível que alcance dados de natureza contínua com comportamento positivo assimétrico. Com efeito, o Modelo Linear Generalizado (MLG), por ser versátil, permite que a variável resposta se adeque a esse comportamento, sendo as distribuições normal inversa e gama adequadas para essa modelagem. Portanto, fez-se um estudo sobre os principais aspectos práticos e teóricos dos MLGs com o objetivo de ajustar modelos onde a variável independente é de natureza contínua com comportamento positivo assimétrico utilizando técnicas computacionais. Os dados utilizados são referentes à função pulmonar de adolescentes no munícipio de Campina Grande - PB. As variáveis dependentes se ajustaram bem aos modelos assimétricos propostos, constatando que as distribuições normal inversa e gama podem ser usadas para ajustar modelos com comportamento assimétrico na variável resposta.

Palavras-chave: Modelos Lineares Generalizados. Dados positivos assimétricos. Função Pulmonar.

Abstract

The linear regression model assumes that the response variable is symmetric and homoscedastic. However, in many situations these assumptions are not reach and it is necessary a more flexible approach that can handle with continuous data that has an asymmetric positive behavior. So, the Generalized Linear Model (GLM) allows response variables fit this behavior and the inverse gaussian and gama distributions are suitable for this modeling. Therefore, a study about the main theorical and pratical aspects of the GLMs was made to fit models where the dependent variable is continuous with asymmetric positive behavior using computational techniques. The data used are related to pulmonary function of teenagers of the city Campina Grande - PB. The dependent variables got good adjustments to the proposed asymmetric positive models, concluding that the inverse gaussian and gama distributions can be applied to fit models with such behavior.

Key-words: Generalized Linear Models. Asymmetric positive datasets. Pulmonary Function.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

O modelo normal linear, desenvolvido por Legendre e Gauss, segundo Stigler (1981), foi amplamente utilizado na descrição de fenômenos aleatórios, principalmente nas primeiras décadas do século XX. Gauss propôs a distribuição normal para descrever a variabilidade dos erros e, dessa forma, a suposição de normalidade dos erros e constância da variância são de suma importância para a adequação do modelo normal linear.

Para solucionar o problema de linearidade, homocedasticade da variância e normalidade dos dados foram desenvolvidas transformações com o objetivo de alcançar tais pressupostos. Uma transformação conceituada foi desenvolvida por Box e Cox (1964), que tem como objetivo transformar a variável ajustada a fim de se obter um melhor ajuste do modelo. No entanto, de acordo com Andrews (1971), a transformação de Box-Cox é muito sensível a valores atípicos e é recomendada quando a variável somente assume valores positivos.

Contudo, nem todas as situações estudadas se ajustavam a esses requisitos do modelo. Em face dessa adversidade, foram desenvolvidos modelos não lineares que melhor se adequassem aos dados, tais como: o modelo complemento log-log em ensaios de diluição ajustado à distribuição binomial (FISHER, 1922) e o modelo de testes de vida envolvendo a distribuição exponencial (FEIGL; ZELEN, 1965)

Os Modelos Lineares Generalizados (MLGs), apresentado por Nelder e Wedderburn (1972), vieram unificar todos os exemplos citados acima, que consistem em abordagens cuja variável resposta tem como característica comum pertencer à família exponencial de distribuições, possibilitando um aumento no número de modelos que podem ser ajustados. As aplicações dessa modelagem na solução de problemas estatísticos englobam diversos tipos de dados, sejam eles discretos ou contínuos.

A vantagem do uso dos MLGs na análise de dados positivos assimétricos deve-se ao fato de evitarmos transformações na variável dependente com a finalidade de encontrar normalidade e homocedasticidade dos erros, visto que as transformações podem não ser adequadas em determinadas situações e, portanto, os dados podem não se adequar aos pressupostos do modelo linear normal. Desse modo, as distribuições normal inversa e gama são úteis e bastante utilizadas na modelagem de dados com tal comportamento assimétrico.

Conforme Paula (2013), o avanço computacional proporcionou facilidade ao desenvolvimento dos MLGs, visto que processos iterativos como o método de Newton-Raphson e o método de Escore de Fisher, que necessitam de softwares adequados para sua fácil implementação, são essenciais na estimação dos parâmetros do modelo. Com efeito, *Softwares*

como o R (R Core Team) (2016) são bastante utilizados na estimação dos parâmetros e na análise dos modelos.

Diante disso, realizou-se um estudo dos principais aspectos teóricos e práticos acerca dos MLGs com o objetivo de fazer aplicações envolvendo dados reais de função pulmonar, fornecidos por Vânia (2016), cuja natureza é contínua com comportamento assimétrico positivo, utilizando o software R (R Core Team) (2016). Será ajustados modelos em que a variável resposta segue uma distribuição normal inversa e/ou gama para verificar a eficiência da aplicação dos MLG em dados com tal comportamento. Também será verificado a adequabilidade dos modelos propostos através de técnicas de diagnósticos, observando pontos atípicos e/ou influentes. Por fim, será comparado modelos via abordagem MLG e via modelos de regressão clássico para verificar se há vantagens e diferenças entre as abordagens.

2 Revisão de Literatura

Diversos trabalhos foram publicados desde o desenvolvimento dos modelos lineares generalizados, incluindo rotinas para o ajuste dos mesmos. Jorgensen (1987) propôs os modelos de dispersão que estende a variedade de distribuições que a variável resposta pode assumir. Tem-se ainda os modelos aditivos generalizados desenvolvido por Hastie e Tibshirani (1990) que supõe que o preditor linear pode ser formado por funções semiparamétricas e os modelos lineares generalizados mistos (BRESLOW; CLAYTON, 1993) que incluem efeitos aleatórios gaussianos no preditor linear.

A análise de diagnóstico é uma importante ferramenta para verificar a adequabilidade do modelo, observando se as suposições sob o modelo são sastifeitas, assim como a presença de pontos atípicos que possam influenciar nos resultados do ajuste. Nesse contexto, Atkinson (1981) sugere a construção, por simulação de Monte Carlo, de intervalos de confiança para os resíduos de modelos lineares, denominado como envelope, que compara os resíduos com os percentis de uma distribuição normal padrão.

Pregibon (1981) desenvolveu medidas de diagnóstico para detectar pontos atípicos em MLG, em especial o componente do desvio como resíduo. Williams (1984) verifica, através de simulação, a aproximação do componente do desvio padronizado proposta por Pregibon (1981) para os MLGs com a distribuição normal padrão. Ainda, Williams (1987) discute a construção de envelopes para os MLGs.

Os MLGs têm sido utilizados nas áreas de astronomia, agronomia, agricultura, saúde e pesca. Pela sua versatilidade, a variável resposta do modelo, quer seja de natureza contínua ou discreta, pode ter diversos comportamentos dependendo do tipo de dado que se deseja ajustar. Logo, os MLGs são bastante utilizados para dados de natureza discreta de contagem ou proporção e dados de natureza contínua com comportamentos simétricos e assimétricos.

Aplicações em dados de natureza discreta em que a variável resposta é de contagem foram feitas por Barros e Nascimento (2008), que utilizaram a distribuição de Poisson com função de ligação raiz quadrada para verificar o número de casos de AIDS na cidade de Recife-PE, encontrando variáveis explicativas como Sexo e Cor/Raça para explicar o aumento no número de infectados. Não obstante, Almeida et al. (2015) ajustaram um modelo de contagem binomial negativo inflado de zeros para avaliar a influência da iluminação artifical na floração do maracujazeiro, constatando que a iluminação artifical ao ínicio da manhã antecipa a abertura das flores.

Já para dados de proporção, Souza et al. (2015) compararam modelos com distribuição binomial com função de ligação *logit* e *probit* para verificar a atividade de formação de estrelas, sendo a probabilidade da presença dessa atividade a variável de interesse. Os resultados encontrados pelos autores evidenciam que ambas as ligações produzem estimativas similares e que um aumento na fração de moléculas de gás aumenta a probabilidade de formação das estrelas.

A utilização dos MLGs também se amplia para as variáveis contínuas com comportamento assimétrico. As distribuições normal inversa e gama, por apresentarem tal comportamento, são recomendadas no ajuste desses dados. Possamai (2009) fez um revisão dos modelos com variável resposta pertecendo à família exponencial e fazendo aplicações com as distribuições gama, normal inversa e outras distribuições, constatando a importância desses modelos em diversas aplicações.

Holanda, Vasconcellos e Silva (2012) compararam modelos com variável resposta gama, log-normal e normal inversa com o objetivo de identificar a influência da atividade sísmica sobre a captura de peixes no litoral do Rio de Janeiro, não encontrando modelos significativos com evidências de que a atividade sísmica influenciava na captura dos peixes estudados. Entretanto, os referidos autores ressaltam a importância dos MLGs visto que esse dispensa as condições de normalidade e homocedasticidade, condições nem sempre sastifeitas em dados de pesca.

Modelos com variável resposta normal inversa foram utilizados por Heller et al. (2006), que desenvolveu uma modelagem para dados de seguro através dos modelos aditivos generalizados para posição, escala e forma (Generalized Aditive Models for Location, Scale and Shape - GAMLSS), utilizando a distribuição normal inversa visto que essa comporta adequadamente a assimetria positiva dos dados. Santos et al. (2010) ajustou um MLG com função de ligação logarítmica para verificar quais fatores influenciam na taxa de mortalidade infantil do Brasil, chegando a conclusão de que a diminuição da mortaliadde infantil está relacionado com o aumento da cobertura de esgotamento sanitário, com a redução da taxa de analfabetismo e com a diminuição do índice de pobreza na população. Por fim, Dias (2014) ajustou um modelo GAMLSS na análise da quantidade de sangue recebido em tranfusão por crianças com problemas hepáticos, verificando que a distribuição normal inversa, por acomodar dados com forte assimetria, foi adequada para o ajuste do modelo com as variáveis explicativas Kasai (ocorrência de operação prévia) e PELD (nível de uma medida de gravidade do paciente em 4 níveis).

No tocante a modelos com variável resposta gama, Acorsi (2002) avaliou um indicador de bem-estar de peixes, que fundamenta-se na relação peso-comprimento do animal, das espécies *Trachydoras paraguayensis, Liposarcus anisiti, Raphiodon vulpinus, Serrasalmus marginatus e Leporinus friderice* utilizando a função de ligação potência, encontrando um modelo capaz de descrever o peso do peixe em função do seu comprimento

segundo sua espécie e sexo.

Assunção (2012) avaliou os fatores relacionados à variação da abundância da espécie *Octopus vulgaris*[1] através de função de ligação logarítmica, identificando as variáveis ano, embarcação, o trimeste e a classe de profundidade como variáveis significativas para o modelo proposto. Santos e Silva (2014) realizaram um estudo com o objetivo de verificar quais variáveis influenciam na taxa de neoplasia pulmonar no Brasil sendo utilizada a função de ligação logarítmica para o ajuste, os autores observaram que as variáveis região e sexo são significativas e que nas regiões sul e sudeste há maiores taxas.

Elliott et al. (2015) estimaram os *redshifts*[2] fotométricos das galáxias usando uma função de ligação logarítmica, verificando que a função de ligação foi adequada e concluindo que os MLGs são eficientes para calcular os *redshifts* das galáxias. Hess et al. (2015) ajustaram um modelo para a estimativa do crescimento em altura da *Pinus Taeda L.*[3] usando uma função de ligação identidade, chegando a conclusão de que o modelo gama se ajustou bem aos dados e as variáveis explicativas diâmetro e idade das árvores afetam significativamente o crescimento em altura.

[1] uma espécie de polvo

[2] *redshift* é a alteração da frequência da luz quando é observada em função da velocidade relativa entre a fonte emissora e o receptor observador. Ver Elliott et al. (2015)

[3] Uma espécie de pinheiro.

3 Fundamentação Teórica

Realizou-se um levantamento teórico acerca dos MLGs com o objetivo de fazer aplicações reais em dados contínuos com assimetria positiva utilizando as distribuições normal inversa e/ou gama. O estudo teórico dos MLGs inclui conceitos fundamentais como componente aleatório, função de ligação, método de estimação dos parâmetros, qualidade do ajuste, seleção de modelos, técnicas de diagnóstico entre outros. Ainda, foi feito uma revisão sobre a família exponencial de distribuições de probabilidades, tendo em vista a sua importancia teórica na teoria dos modelos lineares generalizados.

3.1 Família Exponencial Uniparamétrica

Conforme Demétrio (2001), uma variável aleatória Y pertence à família exponencial uniparamétrica quando sua função densidade de probabilidade (f.d.p.) ou função massa de probabilidade (f.m.p.) pode ser escrita na forma

$$f(y;\theta) = h(y)t(\theta)exp\left\{s(y)b(\theta)\right\},$$

em que $t(\theta)$, $b(\theta)$, $s(y)$ e $h(y)$ são funções conhecidas e cujos valores pertencem ao conjunto dos reais. McCullagh e Nelder (1989) propuseram uma outra notação para a família exponencial canônica adicionando um parâmetro ($\phi > 0$), conhecido como parâmetro de pertubação. A notação é expressa por

$$f(y;\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{3.1}$$

em que $b(\theta)$ e $c(y,\phi)$ são funções conhecidas e $a(\phi)$ pode ou não ser conhecido. Na Equação (3.1) tem-se a família exponencial na forma canônica sendo θ o parâmetro canônico. Segundo Demétrio (2001), quando existem outros parâmetros além de θ eles são observados como parâmetros de pertubação.

A esperança e variância de uma distribuição pertencente a família exponencial é dada por

$$\mu = E(Y) = b'(\theta) \quad e \quad \text{Var}(Y) = a(\phi)b''(\theta),$$

em que $b''(\theta) = \dfrac{d\mu}{d\theta}$ é uma função de μ e pode ser representada por $V(\mu)$, conhecido como função de variância. Essas expressões podem ser deduzidas através da função escore. Bolfarine e Sandoval (2001) define a função escore como

$$U(\theta) = \frac{\partial \ell n f(Y|\theta)}{\partial \theta}, \tag{3.2}$$

em que $\ell n f(y|\theta) = \ell(\theta|y)$ é o logarítmo da função de verossimilhança. A função de verossimilhança de θ, conforme Bolfarine e Sandoval (2001), é uma função que associa o

valor de $f(y_i|\theta)$ a cada um dos possíveis valores de θ, ou seja, $L(\theta;\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i|\theta)$, em que $\boldsymbol{y} = (y_1, ..., y_n)$ corresponde à amostra aleatória observada.

De acordo com Turkman e Silva (2000), sob certas condições de regularidade[4], tem-se que

$$E[U(\theta)] = 0 \quad e \quad E[U^2(\theta)] = E\left[-\frac{\partial^2 \ell n f(Y|\theta)}{\partial\theta^2}\right] = I(\theta),$$

em que $I(\theta)$ é a informação esperada de Fisher. Logo:

$$\begin{aligned} U(\theta) &= \frac{\partial \ell n\left\{exp\left\{\dfrac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}\right\}}{\partial\theta} \\ &= \frac{y - b'(\theta)}{a(\phi)}. \end{aligned}$$

Como $E[U(\theta)] = 0$, obtém-se

$$E[U(\theta)] = \frac{E(Y) - b'(\theta)}{a(\phi)} = 0.$$

Portanto

$$E(Y) = b'(\theta). \tag{3.3}$$

Em relação à variância, tem-se

$$\begin{aligned} U(\theta) &= \frac{y - b'(\theta)}{a(\phi)} \\ \frac{\partial U(\theta)}{\partial\theta} &= \frac{-b''(\theta)}{a(\phi)} \end{aligned}$$

e

$$E[U(\theta)] = \frac{1}{a(\phi)}[E(Y) - b'(\theta)] = 0 \rightarrow E(Y) = b'(\theta).$$

Então,

$$\begin{aligned} \mathrm{Var}[U(\theta)] &= -E\left(\frac{\partial U(\theta)}{\partial\theta}\right) = \frac{b''(\theta)}{a(\phi)} \\ \mathrm{Var}[U(\theta)] &= E[U^2(\theta)] = \frac{\mathrm{Var}(Y)}{[a(\phi)]^2} \\ \frac{b''(\theta)}{a(\phi)} &= \frac{\mathrm{Var}(Y)}{[a(\phi)]^2} \end{aligned}$$

Logo,

$$\mathrm{Var}(Y) = a(\phi)b''(\theta),$$

em que $b''(\theta) = V(\mu)$. Na Tabela 1 são apresentadas algumas distribuições importantes da família exponencial, exibindo-se as funções $b(\theta)$, $c(y,\phi)$ e a função de variância $V(\mu)$, além dos parâmetros canônico e de pertubação. $\Gamma(.)$ é a função gama dada por $\Gamma(\upsilon) = \int_0^\infty y^{\upsilon-1}e^{-y}dy, \upsilon > 0$.

[4] ver Bolfarine e Sandoval (2001) pág. 16.

Tabela 1 – Identificadores da família exponencial para algumas distribuições

Distribuição	θ	$b(\theta)$	$a(\phi)$	$c(y,\phi)$	$V(\mu)$
Binomial $B(n,\tau)$	$\ell n\left(\frac{\mu}{n-\mu}\right)$	$n\ell n(1+e^{\theta})$	1	$\ell n\begin{pmatrix} n \\ y \end{pmatrix}$	$\frac{\mu(n-\mu)}{n}$
Binomial Negativa $BN(m,\tau)$	$\ell n\left(\frac{\mu}{\mu+m}\right)$	$-m\ell n(1-e^{\theta})$	1	$\ell n\left[\frac{\Gamma(m+y)}{\Gamma(k)y!}\right]$	$\mu\left(\frac{\mu}{m}+1\right)$
Poisson $P(\lambda)$	$\ell n(\lambda)$	e^{θ}	1	$-\ell n(y!)$	λ
Normal $N(\mu,\sigma^2)$	μ	$\frac{\theta^2}{2}$	σ^2	$-\frac{1}{2}\left(\frac{y^2}{\sigma^2}+\ell n(2\pi\sigma^2)\right)$	1
Normal Inversa $NI(\mu,\sigma^2)$	$-\frac{1}{2\mu^2}$	$-(-2\theta)^{1/2}$	σ^2	$-\frac{1}{2}\left[\ell n(2\pi\sigma^2y^3)+\frac{1}{\sigma^2y}\right]$	μ^3
Gama $G(\mu,\upsilon)$	$-\frac{1}{\mu}$	$-\ell n(-\theta)$	$\frac{1}{\upsilon}$	$\upsilon\ell n(\upsilon y)-\ell n(y)-\ell n(\Gamma(\upsilon))$	μ^2

3.1.1 Função Geradora de Momentos e de Cumulantes

A função geradora de momentos (f.g.m) é uma ferramenta essencial pois, a partir dela, consegue-se obter todos os momentos de uma variável aleatória e, em muitos casos, definir sua distribuição de probabilidade.

Seja Y uma variável aleatória, Dantas (2013) descreve a função geradora de momentos por:

$$M_Y(t) = E[e^{tY}].$$

Em relação a família exponencial uniparamétrica, a função geradora de momentos, segundo a notação desenvolvida por McCullagh e Nelder (1989), é dada por

$$M_Y(t;\theta,\phi) = E[e^{tY}] = exp\left\{\frac{b(a(\phi)t+\theta)-b(\theta)}{a(\phi)}\right\}.$$

Prova: Será demonstrado para variáveis contínuas, no caso de variáveis discretas substitui-se a integral pelo somatório. Tem-se que

$$\begin{aligned}
\int_A f(y)dy &= 1 \\
&= \int_A exp\left\{\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)\right\}dy = 1 \\
&= \frac{1}{exp\left(\frac{b(\theta)}{a(\phi)}\right)}\int_A exp\left\{\frac{1}{a(\phi)}\theta y+c(y,\phi)\right\}dy = 1 \\
\longrightarrow &\int_A exp\left\{\frac{1}{a(\phi)}\theta y+c(y,\phi)\right\}dy = exp\left(\frac{b(\theta)}{a(\phi)}\right). \qquad (3.4)
\end{aligned}$$

Portanto,

$$\begin{aligned} M_Y(t;\theta,\phi) &= E[e^{tY}] = \int_A e^{ty} exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} dy \\ &= \int_A exp\left\{\frac{1}{a(\phi)}[a(\phi)t + b(\theta)]y + c(y,\phi)\right\} dy \\ &= \frac{1}{exp\left(\dfrac{b(\theta)}{a(\phi)}\right)} \int_A \exp\left\{\frac{1}{a(\phi)}[a(\phi)t + \theta]y + c(y,\phi)\right\} dy. \end{aligned}$$

Utilizando a Equação (3.4) obtém-se

$$M_Y(t;\theta,\phi) = \frac{1}{exp\left(\dfrac{b(\theta)}{a(\phi)}\right)} exp\left\{\frac{b(a(\phi)t + \theta)}{a(\phi)}\right\}.$$

Logo:

$$M_Y(t;\theta,\phi) = E[e^{tY}] = exp\left\{\frac{b(a(\phi)t + \theta) - b(\theta)}{a(\phi)}\right\}.$$

Ainda, tem-se a função geradora de cumulantes (f.g.c.), que, segundo Demétrio (2001), é importante na obtenção de propriedades assintóticas dos modelos lineares generalizados. A f.g.c. é expressa por

$$\varphi_Y(t) = \ell n M_Y(t).$$

Segundo a notação de McCullagh e Nelder (1989), tem-se

$$\varphi(t;\theta,\phi) = \frac{1}{a(\phi)} \left\{b[a(\phi)t + \theta] - b(\theta)\right\}. \tag{3.5}$$

Derivando-se (3.5), sucessivamente, em relação à t, tem-se

$$\begin{aligned} \varphi'(t;\theta,\phi) &= \frac{1}{a(\phi)} b'[a(\phi)t + \theta]a(\phi) = b'[a(\phi)t + \theta] \\ \varphi''(t;\theta,\phi) &= b''[a(\phi)t + \theta]a(\phi) \\ \varphi'''(t;\theta,\phi) &= b'''[a(\phi)t + \theta][a(\phi)]^2 \\ \vdots &\quad \vdots \\ \varphi^{(r)}(t;\theta,\phi) &= b^{(r)}[a(\phi)t + \theta][a(\phi)]^{(r-1)}. \end{aligned}$$

Para t=0, obtém-se a esperança e a variância a partir do primeiro e segundo cumulantes, respectivamente.

$$\begin{aligned} \varphi^{(1)} &= b'(\theta) = E(Y) \\ \varphi^{(2)} &= a(\phi)b''(\theta) = \mathrm{Var}(Y) \end{aligned}$$

3.1.2 Estatística Suficiente

Considere $Y_1, ..., Y_n$ uma amostra aleatória de uma distribuição que pertence a família exponencial com observações $y_1, ..., y_n$. Então, conforme Cordeiro e Demétrio (2013), a f.d.p. conjunta de $Y_1, ..., Y_n$ é expressa por:

$$\begin{aligned} f(y;\theta,\phi) &= \prod_{i=1}^{n} f(y_i;\theta,\phi) = \prod_{i=1}^{n} exp\left\{\frac{1}{a(\phi)}[y_i\theta - b(\theta)] + c(y_i;\phi)\right\} \\ &= \prod_{i=1}^{n} exp\left\{\frac{1}{a(\phi)}[y_i\theta - b(\theta)]\right\} exp\{c(y_i;\phi)\} \\ &= exp\left\{\frac{1}{a(\phi)}\left[\theta\sum_{i=1}^{n} y_i - nb(\theta)\right]\right\} exp\left\{\sum_{1=1}^{n} c(y_i;\phi)\right\}. \end{aligned} \tag{3.6}$$

Contudo, segundo Bolfarine e Sandoval (2001), pelo critério de fatoração de Neyman tem-se que T(Y) é uma estatística suficiente para θ se, somente se,

$$L(\theta;\boldsymbol{y}) = g_\theta(T(y_1, ..., y_n))h(y_1, ..., y_n),$$

em que $g_\theta(T(y_1, ..., y_n))$ depende de θ e de $y_1, ..., y_n$ somente através de T(.). Logo, verificando a Equação (3.6) demonstra-se que $T = \sum_{i=1}^{n} Y_i$ é uma estatística suficiente para θ, pois

$$f(y;\theta,\phi) = g(t,\theta)h(y_1, ..., y_n),$$

em que $g(t,\theta)$ depende de θ e dos y's apenas através de t e $h(y_1, ..., y_n)$ independe de θ. Ou seja, se uma f.d.p pertence à família exponencial com um parâmetro existe uma estatística suficiente. Pode-se demonstrar, ainda, que $T = \sum_{i=1}^{n} Y_i$ é uma estatística suficiente minimal. De acordo com Casella e Berger (2002), uma estatística é suficiente minimal se for suficiente e se for função de qualquer outra estatística suficiente para θ. Será visto que ligações canônicas possuem a vantagem de obter uma estatística suficiente minimal para o vetor de parâmetro $\boldsymbol{\beta}$.

3.2 Modelos Lineares Generalizados

Considere Y como a variável resposta ou dependente do modelo de interesse do experimento e considere $\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{ip}$ o vetor coluna de variáveis explicativas. McCullagh e Nelder (1989) definem os três elementos que compõe o modelo linear generalizado:

1. **Componente Aleatório**

Corresponde as variáveis aleatórias $Y_1, ..., Y_n$ condicionalmente independentes e que seguem uma mesma distribuição, a qual pertence à família exponencial, onde $E(Y_i|X = x_i) = b'(\theta_i) = \mu_i$ para *i=1,...,n* e um parâmetro de dispersão $\phi > 0$ conhecido e independentes das i's variáveis explicativas.

2. **Componente Sistemático ou Estrutural**

As variáveis independentes produzem um preditor linear $\boldsymbol{\eta}$ dado por

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

em que $\boldsymbol{X}=(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^T$ representa a matriz do modelo, $\boldsymbol{\beta}= (\beta_1, ..., \beta_p)^T$ o vetor de parâmetros desconhecidos e $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)$ o preditor linear.

3. **Função de ligação**

Uma função monótona e diferenciável cujo objetivo é relacionar a média ao preditor linear, estabelecendo uma relação entre o componente aleatório e o componente sistemático do modelo. Assim:

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

Na Tabela 2 são apresentadas algumas distribuições da família exponencial que a variável resposta Y pode seguir e o tipo de dado no qual a distribuição é aplicada.

Tabela 2 – Algumas distribuições para a variável resposta Y e a natureza dos dados em que ela é utilizada.

Distribuição	**Tipo de dados**
Binomial	Proporção
Binomial Negativa	Contagem
Poisson	Contagem
Normal	Contínuos
Normal Inversa	Contínuos Assimétricos
Gama	Contínuos Assimétricos

3.2.1 Ligação Canônica

De acordo com Paula (2013), quando a função de ligação escolhida é tal que $\eta_i = \theta_i$, dizemos que a função de ligação correspondente se chama função de ligação canônica, visto que o preditor linear coincide com o parâmetro canônico. Supondo ϕ conhecido, uma vantagem na utilização dessa função de ligação é que pode-se obter uma estatística suficiente minimal para o vetor de parâmetros $\boldsymbol{\beta}$. Considere $Y_1, ..., Y_n$ uma amostra aleatória de uma distribuição com f.d.p. definida em (3.1) , o logaritmo da função de verossimilhança de um MLG como função apenas de $\boldsymbol{\beta}$ e com observações dada por $\boldsymbol{y} = (y_1, ..., y_n)^T$ é definido por

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{n} \frac{1}{a(\phi)} \{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^{n} c(y_i, \phi). \tag{3.7}$$

Utilizando a ligação canônica obtém-se que $\theta_i = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$, logo:

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{n} \frac{1}{a(\phi)} \left\{ y_i \sum_{j=1}^{p} x_{ij}\beta_j - b\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)\right\} + \sum_{i=1}^{n} c(y_i, \phi).$$

Determinando a estatística $S_j = \dfrac{1}{a(\phi)} \sum\limits_{i=1}^{n} Y_i x_{ij}$, obtém-se:

$$\ell(\boldsymbol{\beta}) = \sum_{j=1}^{p} S_j \beta_j - \frac{1}{a(\phi)} \sum_{i=1}^{n} b(\sum_{j=1}^{p} x_{ij}\beta_j) + \sum_{i=1}^{n} c(y_i, \phi).$$

Através do teorema da fatoração, verifica-se que a estatística $\boldsymbol{S} = (S_1, S_2, ...S_p)^T$ é suficiente minimal para o vetor de parâmetros $\boldsymbol{\beta}$. Na Tabela 3 tem-se as principais ligações canônicas.

Tabela 3 – Ligações canônicas de algumas distribuições da família exponencial.

Distribuição	**Ligação**
Normal	Identidade: $\eta = \mu$
Poisson	Logarítmica: $\eta = \ell n(\mu)$
Binomial	Logística: $\eta = \ell n \dfrac{\mu}{1-\mu}$
Gama	Recíproca: $\eta = \dfrac{1}{\mu}$
Normal inversa	Recíproca do quadrado: $\eta = \dfrac{1}{\mu^2}$

Fonte: Cordeiro e Demétrio (2013)

Como exemplo, para um modelo com variável resposta seguindo uma distribuição normal inversa e com ligação canônica, o componente sistemático é dado por

$$\eta = \frac{1}{\mu^2} = \boldsymbol{X}\boldsymbol{\beta}$$

3.2.2 Estimação dos parâmetros

A estimação pontual e intervalar dos MLGs são baseadas na verossimilhança. De fato, segundo Turkman e Silva (2000), o método da máxima verossimilhança também é aplicado nos testes de hipóteses sobre os parâmetros do modelo e da qualidade do ajuste do MLG. Métodos numéricos, como o método de Escore de Fisher, serão aplicados visto que alguns parâmetros necessitam do uso desses métodos para serem estimados.

Estimação de β

Seja a função de verossimilhança $L(\boldsymbol{\beta}; \boldsymbol{y})$ dada por

$$L(\boldsymbol{\beta}; \boldsymbol{y}) = exp\left\{ \frac{1}{a(\phi)} \sum_{i=1}^{n} \{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^{n} c(y_i, \phi) \right\},$$

o estimador de máxima verossimilhança (EMV) do vetor de parâmetros $\boldsymbol{\beta}$ é definido pelo logaritmo da função de verossimilhança definida em 3.7. Logo:

$$\ell(\boldsymbol{\beta};\boldsymbol{y}) = \sum_{i=1}^{n} \frac{1}{a(\phi)} \{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^{n} c(y_i, \phi).$$

Na Sessão 3.1 observou-se que a família exponencial satisfaz as condições de regularidade para a função escore, ou seja, $U = U(\theta) = \frac{\partial \ell n f(Y|\boldsymbol{\theta})}{\partial \theta} = 0$, então

$$U_j = \sum_{i=1}^{n} \frac{\partial L(\boldsymbol{\beta};\boldsymbol{y})}{\partial \beta_j} = 0, \quad j = 1, ..., p.$$

Pela regra da cadeia obtém-se

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\beta};\boldsymbol{y})}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\beta};\boldsymbol{y})}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_i} \\ &= \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\beta};\boldsymbol{y})}{\partial \theta_i} \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_i} \end{aligned}$$

onde

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\beta};\boldsymbol{y})}{\partial \theta_i} &= \frac{y_i - b'(\theta_i)}{a_i(\phi)} = \frac{y_i - \mu_i}{a_i(\phi)}, \\ \frac{\partial \mu_i}{\partial \theta_i} &= b''(\theta_i) = V(\mu_i), \\ \frac{\partial \eta_i}{\partial \beta_j} &= x_{ij} \end{aligned}$$

Assim

$$U_j = \frac{\partial \ell(\boldsymbol{\beta};\boldsymbol{y})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i(\phi)} \frac{1}{V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij}$$

Como $U_j = 0$ são equações não lineares, deve-se utilizar um método iterativo para encontrar o estimador. Para Turkman e Silva (2000), o Método de Escore de Fisher é, em geral, mais simples comparado ao método de Newton-Raphson visto que é mais viável calcular a matriz de informação de Fisher $I(\boldsymbol{\beta})$. Inicia-se o processo especificando um valor inicial $\hat{\boldsymbol{\beta}}$ e obtendo sucessivas aproximações até que a convergência seja alcançada. A relação é dada por:

$$\hat{\boldsymbol{\beta}}^{m+1} = \hat{\boldsymbol{\beta}}^{(m)} + [I(\hat{\boldsymbol{\beta}}^{(m))}]^{-1} U(\hat{\boldsymbol{\beta}}^{(m)})$$

ou ainda

$$[I(\hat{\boldsymbol{\beta}}^{(m)})]\hat{\boldsymbol{\beta}}^{(k+1)} = [I(\hat{\boldsymbol{\beta}}^{(m)})]\hat{\boldsymbol{\beta}}^{(m)} + U(\hat{\boldsymbol{\beta}}^{(m)}) \tag{3.8}$$

sendo $\hat{\boldsymbol{\beta}}^{(k)}$ e $\hat{\boldsymbol{\beta}}^{(\boldsymbol{k+1})}$ os parâmetros estimados na m-ésima e na *(m+1)*-ésima iteração, $\boldsymbol{U}^{(m)}$ o vetor escore e $I(.)$ a matriz de informação de Fisher, que é dada por

$$I(\boldsymbol{\beta}) = E\left[\frac{-\partial \ell(\boldsymbol{\beta};\boldsymbol{y})}{\partial \beta_j \partial \beta_k}\right] = E\left[\frac{\partial \ell(\boldsymbol{\beta};\boldsymbol{y})}{\partial \beta_j} \frac{\partial \ell(\boldsymbol{\beta};\boldsymbol{y})}{\partial \beta_k}\right]$$

portanto, tem-se

$$\begin{aligned} I(\boldsymbol{\beta}) &= E\left[\left(\frac{(Y_i-\mu_i)x_{ij}}{a_i(\phi)V(\mu_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)\right)\left(\frac{(Y_i-\mu_i)x_{ik}}{a_i(\phi)V(\mu_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)\right)\right] \\ &= E\left[\frac{(Y_i-\mu_i)^2x_{ij}x_{ik}}{a_i^2(\phi)[V(\mu_i)]^2}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\right] \\ &= E\left[\frac{a_i(\phi)V(\mu_i)}{a_i^2(\phi)[V(\mu_i)]^2}x_{ij}x_{ik}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\right] \\ &= \frac{x_{ij}x_{ik}}{a_i(\phi)V(\mu_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2. \end{aligned}$$

Fazendo-se $a_i(\phi)=\dfrac{\phi}{\omega_i}$, sendo $\phi>0$ constante, ω_i peso a *priori* e $W_i=\dfrac{\omega_i}{V(\mu_i)}\left(\dfrac{\partial\mu_i}{\partial\eta_i}\right)^2$ o vetor da matriz de informação de Fisher fica dado por

$$I(\boldsymbol{\beta})=\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}, \tag{3.9}$$

em que $\boldsymbol{X}=(x_1,...,x_j,x_{j+1},...,x_k)^T$ é a matriz do modelo particionada e $\boldsymbol{W}=diag(W_1,...,W_n)$ é a matriz diagonal com elementos $W_1,...,W_n$. Rearrajando-se os termos de U_j obtém-se

$$U_j=\frac{\partial\ell(\boldsymbol{\beta};\boldsymbol{y})}{\partial\beta_j}=\sum_{i=1}^{n}\frac{y_i-\mu_i}{a_i(\phi)}\frac{1}{V(\mu_i)}\frac{\partial\mu_i}{\partial\eta_i}x_{ij}=\sum_{i=1}^{n}\frac{1}{\phi}x_{ij}W_i\frac{\partial\eta_i}{\partial\mu_i}(y_i-\mu_i)$$

e o vetor escore $\boldsymbol{U}$ fica expresso por

$$\boldsymbol{U}=\frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}\Lambda(\boldsymbol{y}-\boldsymbol{\mu}), \tag{3.10}$$

em que $\boldsymbol{\Lambda}=diag\left(\dfrac{\partial\eta_1}{\partial\mu_1},...,\dfrac{\partial\eta_n}{\partial\mu_n}\right)$ e $\boldsymbol{X}=(x_1,...,x_n)^T$ a matriz do modelo. Substituindo-se (3.9) e (3.10) na Equação (3.8) tem-se

$$\begin{aligned} \frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(m+1)} &= \frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(m)}+\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{\Lambda}^{(m)}(\boldsymbol{y}-\boldsymbol{\mu})^{(m)} \\ \boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(m+1)} &= \boldsymbol{X}^T\boldsymbol{W}^{(m)}[\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(m)}+\boldsymbol{\Lambda}^{(m)}(\boldsymbol{y}-\boldsymbol{\mu})^{(m)}] \\ \boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(m+1)} &= \boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{\psi}^{(m)} \end{aligned}$$

em que $\boldsymbol{\psi}^{(m)}=\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(m)}+\boldsymbol{\Lambda}^{(m)}(\boldsymbol{y}-\boldsymbol{\mu})^{(m)}=\boldsymbol{\eta}^{(m)}+\boldsymbol{\Lambda}^{(m)}(\boldsymbol{y}-\boldsymbol{\mu})^{(m)}$, conhecida como variável dependente ajustada. Por conseguinte, o estimador $\hat{\boldsymbol{\beta}}$ no *(m+1)*-ésimo passo é obtida por

$$\hat{\boldsymbol{\beta}}^{(m+1)} = (\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{\psi}^{(m)}. \tag{3.11}$$

Na Equação (3.11), chega-se ao método de mínimos quadrados ponderados semelhante ao do modelo linear normal, com a diferença de que para os MLGs o vetor de

estimadores $\boldsymbol{\beta}$ é obtido através de métodos iterativos. É interessante notar que a fórmula para calcular $\hat{\boldsymbol{\beta}}^{(\boldsymbol{m+1})}$ é independente desse parâmetro.

O valor inicial $\hat{\boldsymbol{\beta}}^{(\mathbf{0})}$ da iteração pode ser calculado a partir das observações y_i visto que elas podem ser consideradas como estimativas das médias μ_i, ou seja, $\eta_i = g(\hat{\mu}_i) = g(y_i)$.

Sumarizando, o cálculo das estimativas de $\boldsymbol{\beta}$, através do processo iterativo, é obtido através das seguintes etapas:

1) Obter as estimativas

$$\eta_i^{(m)} = \sum_{j=1}^{p} x_{ij} \hat{\beta}_i{}^{(m)}$$

em que

$$\mu_i^{(m)} = g^{-1}(\eta_i^{(m)});$$

2) Calcular $\boldsymbol{\psi}^{(m)}$ e os pesos $\boldsymbol{W}_i$ que são dados, respectivamente, por

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

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

3) calcular

$$\hat{\boldsymbol{\beta}}^{(m+1)} = (\boldsymbol{X}^T \boldsymbol{W}^{(m)} \boldsymbol{Z})^{-1} \boldsymbol{X}^T \boldsymbol{W}^{(m)} \boldsymbol{v}^{(m)}$$

e voltar ao passo 1 substituindo $\hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{\beta}}^{(m+1)}$ e repetindo o processo até que a convergência seja alcançada. Para verificar se a convergência foi obtida, utiliza-se algum critério de parada. Um critério bastante utilizado é dado por

$$\sum_{j=1}^{p} \left| \frac{\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}{}^{(m)} - \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}{}^{(m+1)}}{\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}{}^{(m)}} \right|^2 < \epsilon,$$

sendo $\epsilon > 0$ um valor razoavelmente pequeno. De modo geral, a convergência é alcançada rapidamente com poucas iterações. Quando, para algum valor de y_i, $g(y_i)$ não estiver definido pode-se modificar a observação y_i de tal forma que ela possa estar definida para $g(y_i)$.

Propriedades assintóticas e distribuição amostral de $\hat{\boldsymbol{\beta}}$

Resultados assintóticos são utilizados para assegurar a consistência e a normalidade assintótica para $\boldsymbol{\beta}$. Turkman e Silva (2000) ressaltam que esses resultados são garantidos quando se tem as condições de regularidade vistas na Seção 3.1 e para os MLGs essas condições são, em geral, validadas. Logo, algumas propriedades do estimador $\hat{\boldsymbol{\beta}}$:

1) $\hat{\boldsymbol{\beta}}$ é assintoticamente não-viesado

De acordo com Bolfarine e Sandoval (2001), um estimador é assintoticamente não-viesado quando $\lim_{n\longrightarrow\infty} B(\hat{\theta}) = 0$, em que $B(\hat{\theta}) = E[\hat{\theta}] - \theta$. O estimador de máxima verossimilhança de $\boldsymbol{\beta}$ é obtido, como ja definido anteriormente, pelo vetor escore dado por

$$U(\hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

Ainda

$$E[U(\boldsymbol{\beta})] = \mathbf{0} \quad e \quad cov[U(\boldsymbol{\beta})] = E[U(\boldsymbol{\beta})U(\boldsymbol{\beta})^T] = E\left[\frac{-\partial^2 L(\boldsymbol{\beta})}{\partial\beta^T\partial\beta}\right] = I(\boldsymbol{\beta}).$$

Desenvolvendo-se $U(\boldsymbol{\beta})$ em séries de Taylor em relação à $\hat{\boldsymbol{\beta}}$ até os termos de primeira-ordem obtém-se

$$U(\boldsymbol{\beta}) \approx U(\hat{\boldsymbol{\beta}}) + \frac{\partial U(\boldsymbol{\beta})}{\partial\beta}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}). \tag{3.12}$$

Substituindo-se a matriz de derivadas parciais de segunda ordem $\frac{\partial U(\boldsymbol{\beta})}{\partial\beta}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ por $-I(\boldsymbol{\beta})$, onde $I(\boldsymbol{\beta})$ é a matriz de informação de Fisher, na Equação 3.12 (supondo que seja aproximadamente válido para grandes amostras) implica em

$$U(\boldsymbol{\beta}) = U(\hat{\boldsymbol{\beta}}) - I(\boldsymbol{\beta})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

Assim,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx I^{-1}(\boldsymbol{\beta})U(\boldsymbol{\beta}),$$

contanto que $I(\boldsymbol{\beta})$ seja não-singular. Então, tem-se

$$E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = I^{-1}(\boldsymbol{\beta})E[U(\boldsymbol{\beta})] = 0 \Longrightarrow E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}. \tag{3.13}$$

Pois, $E[U(\boldsymbol{\beta})] = \mathbf{0}$ e portanto $\hat{\boldsymbol{\beta}}$ é um estimador assintoticamente não-viesado para $\boldsymbol{\beta}$.

2) Segundo Demétrio (2001), a matriz de variância e covariância de $\hat{\boldsymbol{\beta}}$ para amostras grandes, através da Equação 3.13 e considerando $U = U(\boldsymbol{\beta}), I(\boldsymbol{\beta}) = I$ é apresentada por

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T] = \boldsymbol{I}^{-1}E(\boldsymbol{I}^T)(\boldsymbol{I}^{-1})^T = \boldsymbol{I}^{-1}\boldsymbol{I}\boldsymbol{I}^{-1} = \boldsymbol{I}^{-1},$$

em que $\boldsymbol{I}^{-1} = (\boldsymbol{I}^{-1})^T$ visto que $\boldsymbol{I}^{-1}(\boldsymbol{\beta})$ é simétrica.

3) Conforme Turkman e Silva (2000), para amostra grandes, a distribuição de amostragem de $\hat{\boldsymbol{\beta}}$ é uma normal p-variada com vetor médio $\boldsymbol{\beta}$ e matriz de covariância $\boldsymbol{I}^{-1}(\boldsymbol{\beta})$, ou seja,

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \boldsymbol{I}^{-1}(\boldsymbol{\beta})).$$

Outra aproximação equivalente para $\hat{\boldsymbol{\beta}}$ é conhecida como estatística de Wald e é dada por

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T I(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_p.$$

Com efeito, no caso de modelos lineares com a variável dependente seguindo uma distribuição normal as equações descritas acima são resultados exatos.

Como $\boldsymbol{\beta}$ é desconhecido, faz-se necessário a substituição do vetor $\boldsymbol{\beta}$ pela sua estimativa $\hat{\boldsymbol{\beta}}$ para calcular $\boldsymbol{I}^{-1}(\boldsymbol{\beta})$. O mesmo ocorre quando o parâmetro de dispersão ϕ é desconhecido, ou seja, altera-se ϕ pela sua estimativa $\hat{\phi}$.

Estimação de ϕ

Segundo Demétrio (2001), o logarítmo da máxima verossimilhança também pode ser utilizado para estimar o parâmetro ϕ e é dado por

$$\ell(\boldsymbol{\beta}, \phi; \boldsymbol{y}) = \sum_{i=0}^{n} \phi^{-1} \{y_i\theta_i - b(\theta_i)\} + \sum_{i=0}^{n} c(y_i, \phi).$$

A função escore de ϕ é expressa por

$$U(\phi) = \frac{\partial \ell(\boldsymbol{\beta}, \phi; \boldsymbol{y})}{\partial \phi} = -\phi^{-2} \sum_{i=1}^{n} [y_i\beta_i - b(\beta_i)] + \sum_{i=1}^{n} \frac{\partial c(y_i, \phi)}{\partial \phi}.$$

Assim, o estimador $\hat{\phi}$ é calculado igualando-se $U_\phi = 0$. Percebe-se que U_ϕ é função de ϕ e de β (através de θ). Verifica-se que os vetores ϕ e $\boldsymbol{\beta}$ são ortogonais, ou seja,

$$I(\boldsymbol{\beta}, \phi) = E\left[-\frac{\partial^2 L(\beta, \phi)}{\partial \beta \partial \phi}\right] = 0.$$

Contudo, conforme Cordeiro e Demétrio (2013), quando não existe solução explícita, o método da máxima verossimilhança pode não ser a melhor solução para calcular a estimativa de ϕ. Com efeito, há um método mais simples conhecido como método de Pearson, que é baseado na distribuição de amostragem da estatística de Pearson generalizada. Assim, a estimativa de Pearson para ϕ é

$$\hat{\phi}_p = \frac{1}{n-p} \sum_{i=1}^{n} \frac{\omega_i (y_i - \mu_i)^2}{V(\mu_i)},$$

que é um estimador consistente para ϕ e tem uma distribuição aproximada de uma χ^2 com $n - p$ graus de liberdade. Destaca-se que além de estimar ϕ, a estatística de Pearson generalizada também pode ser usada para avaliar a qualidade do ajuste de um modelo, que será discutido mais adiante.

Outro método de estimação de ϕ, que é baseado na aproximação χ^2_{n-p}, é o método do desvio que é dado por

$$\hat{\phi}_d = \frac{D_p}{n-p},$$

onde D_p é o desvio escalonado, função que será abordada na seção 3.3.1. Esse estimador é aproximadamente não-viesado para ϕ.

3.2.3 Teste de Hipóteses

Conforme Paula (2013), os testes de hipóteses dos MLGs são baseados em três estatísticas, as quais são deduzidas das distribuições assintóticas do vetor de parâmetros $\boldsymbol{\beta}$. São elas:

a) Estatística de Razão de Verossimilhança;

b) Estatística de Wald;

c) Estatística Escore.

Segundo Cordeiro e Demétrio (2013), as estatísticas são assintoticamente equivalentes, entretanto, a razão de verossimilhança é o critério que define o teste uniformemente mais poderoso[5]. A escolha das estatísticas para testar as hipóteses vai depender de qual estatística é mais apropriada para a formulação do teste. A estatística de Wald, por exemplo é, geralmente, mais adequada em hipóteses referentes à um único coeficiente β_j. Já a razão de verossimilhança é geralmente mais apropriada para um subconjunto de coeficientes $\boldsymbol{\beta}$'s. A seguir, serão definidas as formas dos testes e as estatísticas utilizadas.

Hipótese simples

As hipóteses simples são usadas para testar um vetor especificado $\boldsymbol{\beta}_0$ para o vetor de parâmetros $\boldsymbol{\beta}$. Sob H_0 e ϕ conhecido, as estatísticas convergem assintoticamente para uma variável com distribuição χ^2_p e são dadas por

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \quad \textit{versus} \quad H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0,$$

em que $\boldsymbol{\beta}_0$ é um vetor p-dimensional e ϕ é conhecido.

Teste de Razão de Verossimilhanças (TRV)

A estatística de razão de verossimilhança, também conhecida como estatística de Wilks, de acordo com Turkman e Silva (2000), é expressa por

$$\begin{aligned} \xi_{RV} &= -2\ell n \frac{max_{H_0}\ell(\boldsymbol{\beta};\boldsymbol{y})}{max_{H_0\cup H_1}\ell(\boldsymbol{\beta};\boldsymbol{y})} \\ &= \left\{\ell(\hat{\boldsymbol{\beta}};\boldsymbol{y}) - \ell(\hat{\boldsymbol{\beta}}_{\mathbf{0}};\boldsymbol{y})\right\}, \end{aligned}$$

sendo $\ell(\hat{\boldsymbol{\beta}};\boldsymbol{y})$ e $\ell(\hat{\boldsymbol{\beta}}_0;\boldsymbol{y})$ os valores do logaritmo da função de verossimilhança em $\hat{\boldsymbol{\beta}}$ e $\boldsymbol{\beta}_{\mathbf{0}}$, respectivamente. Sob a hipótese nula, ξ_{RV} tem uma distribuição assintótica de uma χ^2 com p graus de liberdade.

[5] ver Bolfarine e Sandoval (2001) pág. 100

Teste de Wald

O Teste de Wald é baseado na distribuição normal assintótica de $\hat{\boldsymbol{\beta}}$, ou seja, $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \boldsymbol{I}^{-1}(\boldsymbol{\beta}))$ e, nesse caso, é obtido por

$$\begin{aligned} \xi_W &= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})^T \hat{V}\text{ar}^{-1}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \\ &= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})^T I(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \\ &= \frac{1}{\phi}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})^T (\boldsymbol{X^T \hat{W} X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \end{aligned}$$

em que $I(\hat{\boldsymbol{\beta}}) = \hat{V}\text{ar}^{-1}(\hat{\boldsymbol{\beta}})$ é a matriz de variância-covariância assintótica de $\hat{\boldsymbol{\beta}}$ coforme visto na seção 3.2.3 e na Equação (3.9) informação estimada em $\boldsymbol{\beta}$. Quando p=1, o teste de Wald coincide com teste t^2. De acordo com Paula (2013), esse teste é dependente da parametrização utilizada, especialmente quando $\eta(\boldsymbol{\beta})$ é não linear em $\boldsymbol{\beta}$, ou seja, diferentes formas de $\boldsymbol{\eta}(\boldsymbol{\beta})$ podem levar à diferentes valores de ξ_W.

Teste Escore

O teste escore é definido por

$$\begin{aligned} \xi_R &= U^T(\boldsymbol{\beta}_0)\hat{V}\text{ar}_0(\hat{\boldsymbol{\beta}})U(\boldsymbol{\beta}_0) \\ &= U^T(\boldsymbol{\beta}_0)I_0^{-1}U(\boldsymbol{\beta}_0) \\ &= U^T(\boldsymbol{\beta}_0)(\boldsymbol{X^T \hat{W}_0 X})^{-1}U(\boldsymbol{\beta}_0) \end{aligned}$$

em que $\hat{\boldsymbol{W}}_0$ é estimado sob H_0. Conforme Cordeiro e Demétrio (2013), uma vantagem da estatística escore é que não é exigido calcular o EMV de $\boldsymbol{\beta}$ sob a hipotése H_1, ainda que na prática essa estatística seja importante.

3.2.4 Regiões de confiança

Conforme Demétrio (2001), pode-se construir intervalos de confiança assintóticos para $\boldsymbol{\beta}$ utilizando as estatísticas de Wald ou de razão de verossimilhança supondo ϕ conhecido. Uma região de confiança assintótica para $\boldsymbol{\beta}$ baseada na razão de verossimilhança e na estatística de Wald é dada, respectivamente, por

$$2[\ell(\hat{\boldsymbol{\beta}}; \boldsymbol{y}) - \ell(\boldsymbol{\beta}; \boldsymbol{y})] \leq \chi^2_{q,1-\alpha}$$

e

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T (\boldsymbol{X^T \hat{W} X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq \phi\chi^2_{q,1-\alpha}$$

em que $\chi^2_{p,1-\alpha}$ é o percentil $(1-\alpha)$ de uma qui-quadrado com p graus de liberdade.

3.3 Seleção e Validação de Modelos

De acordo com Turkman e Silva (2000), o objetivo principal na análise e seleção de modelos é encontrar um modelo parcimonioso, ou seja, um modelo que ofereça o maior

número de informação em relação à variavel resposta com o menor número de covariáveis. Contudo, encontrar esse modelo, na prática, não é fácil, pois é preciso encontrar um equilíbrio entre um bom ajuste, isto é, os dados observados não sejam discrepantes das médias esperadas, e obter um modelo menos complexo. Os autores descrevem uma série de modelos que são referidos durante o processo de seleção e validação dos modelos. Considerando n observações, pode-se obter modelos com até n parâmetros, como o **modelo completo** ou saturado, que possui um parâmetro para cada observação (p=n), atribuindo toda a variação dos dados ao componente sistemático. Não é explicativo visto que reproduz os próprios dados. Já o **modelo nulo** é o mais simples, ou seja, contendo apenas um único parâmetro para todas as observações. Dificilmente este modelo consegue definir a estrutura inerente dos dados e atribui toda a variação ao componente aleatório.

Existem, ainda, os **modelos minimal** e **maximal** que não são tão extremos como os modelos nulo e completo. O modelo minimal contém apenas certos parâmetros que são necessários ao ajuste, como é o caso de efeitos de blocos. Já maximal contém o maior número de termos que podem ser considerados. Por fim, tem-se o **modelo corrente**, que é qualquer modelo com q parâmetros linearmente independentes situado entre o modelo máximal e o minimal. O modelo corrente é o modelo sob pesquisa, ou seja, o modelo parcimonioso que se deseja encontrar.

3.3.1 Qualidade do Ajuste *(Goodness of fit)*

Função Desvio

Conforme Paula (2013), a qualidade do ajuste de um MLG pode ser avaliada através da função desvio, que é uma medida da distância entre o modelo saturado e o modelo corrente, sendo maior ou igual a zero e é expressa por

$$D^*(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = \frac{D(\boldsymbol{y},\hat{\boldsymbol{\mu}})}{a(\phi)} = 2\left\{\ell(\boldsymbol{y};\boldsymbol{y}) - \ell(\hat{\boldsymbol{\mu}};\boldsymbol{y})\right\},$$

em que $\ell(\boldsymbol{y};\boldsymbol{y})$ é o logaritmo da função de verossimilhança do modelo saturado, e $\ell(\hat{\boldsymbol{\mu}};\boldsymbol{y})$ é o E.M.V do modelo corrente onde $p < n$. A função $D^*(\boldsymbol{y};\hat{\boldsymbol{\mu}})$ é conhecida como desvio escalonado e $D(\boldsymbol{y};\hat{\boldsymbol{\mu}})$ dar-se o nome de desvio. Outra forma de denotar a função desvio é obtida por

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}\left\{y_i(\tilde{\theta}_i - \hat{\theta}_i) + (b(\hat{\theta}_i) - b(\tilde{\theta}_i))\right\} = \sum_{i=1}^{n} d^2(\boldsymbol{y}_i,\hat{\boldsymbol{\mu}}_i),$$

em que $\hat{\theta}_i = \theta_i(\mu_i)$ e $\tilde{\theta}_i = \theta_i(\tilde{\mu_i})$ são os E.M.V de $\boldsymbol{\theta}$ do modelo corrente e saturado, respectivamente e $d^2(\boldsymbol{y}_i,\hat{\boldsymbol{\mu}}_i)$ é chamado de componente do desvio, que mede a diferença entre o logaritmo das verossimilhança observado e ajustado para cada observação. Um valor pequeno da função desvio sugere que com um menor número de parâmetros tem-se um ajuste tão bom quanto o ajuste com um modelo saturado. O desvio decresce para

zero conforme aumenta-se o número de parâmetros no modelo corrente, chegando a zero quando torna-se o modelo saturado.

Segundo Paula (2013), apesar de ser comum comparar os valores calculados da função desvio com o quantil da distribuição χ^2_{n-p}, em geral, $D(\boldsymbol{y};\hat{\boldsymbol{\mu}})$ não segue assintoticamente uma qui-quadrado com $n-p$ graus de liberdade. Na Tabela 4 apresenta-se as funções desvios das principais distribuições.

Tabela 4 – Funções Desvios para algumas distribuições da família exponencial.

Distribuição	**Desvio**
Poisson	$D_p = 2\sum_{i=1}^{n}\left[y_i \ell n\left(\frac{y_i}{\hat{\mu}_i}\right) + (\hat{\mu}_i - y_i)\right]$
Binomial	$D_p = 2\sum_{i=1}^{n}\left[y_i \ell n\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i)\ell n\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right)\right]$
Binomial Negativo	$D_p = 2\sum_{i=1}^{n}\left[y_i \ell n\left(\frac{y_i}{\hat{\mu}_i}\right) + (y_i + k)\ell n\left(\frac{\hat{\mu}_i + k}{y_i + k}\right)\right]$
Normal	$D_p = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$
Normal inversa	$D_p = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{y_i\hat{\mu}_i^2}$
Gama	$D_p = 2\sum_{i=1}^{n}\left[\ell n\left(\frac{\hat{\mu}_i}{y_i}\right) + \frac{y_i + \hat{\mu}_i}{\hat{\mu}_i}\right]$

Fonte: Cordeiro e Demétrio (2013)

Estatística de Pearson Generalizada

Outra medida interessante para verificar o ajuste de um modelo aos dados é a estatística de Pearson Generalizada, que é dada por

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

em que $V(\hat{\mu}_i)$ a função de variância estimada sob o modelo que está sendo ajustado. Conforme Turkman e Silva (2000), no caso da distribuição normal, a χ^2 coincide, assim como o desvio, com a soma dos quadrados dos resíduos. Já para o modelo Binomial e Poisson coincide com a estatística original de Pearson.

Apesar de a distribuição assintótica χ^2_{n-p} ser usada para verificar se o ajuste do modelo é adequeado, como aponta Demétrio (2001), essa aproximação, em muitos casos, pode ser pobre. A estatística de Pearson Generalizada é geralmente mais utilizada em comparação com a função desvio por ser de fácil interpretação. Porém, Pierce e Schafer (1986) sugerem que a função desvio fornece resultadores melhores do que a estatística de Pearson.

3.3.2 Seleção de Modelos

O objetivo da seleção de modelos é encontrar o melhor modelo, ou melhores modelos, visto que geralmente tem-se mais de um modelo que melhor se ajusta aos dados, para explicar a variabilidade da variável resposta. Conforme Cordeiro e Demétrio (2013), duas estatisticas que servem para comparar a qualidade do ajuste do modelo são os critérios de informação de Akaike e de Bayes.

Critério de informação de Akaike

O Critério de Informação de Akaike (AIC), desenvolvido por Akaike (1974), é baseado na função de verossimilhança e é dado por

$$AIC_p = D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) - 2\ell(\hat{\boldsymbol{\beta}}; \boldsymbol{y}) + 2p,$$

em que p é o número de parâmetros. Quanto menor o valor do AIC, possivelmente melhor será o modelo.

Critério de informação de Bayes

Schwarz et al. (1978) desenvolveu outra estatística usada na comparação e seleção de modelos, conhecida por Critério de Informação de Bayes (BIC) e dado por

$$BIC_p = D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) - 2\ell(\hat{\boldsymbol{\mu}}_i; \boldsymbol{y}) + p\ell n(n),$$

onde p é o numero de parâmetros. Assim como o AIC, o menor valor do BIC pode indicar um melhor ajuste do modelo.

Burnham e Anderson (2004) recomenda uma equação para comparar a perda de informação em se utilizar outros modelos com valores de BIC maiores que o melhor modelo. A equação é expressa por

$$\Delta_i = BIC_i - BIC_{min},$$

em que BIC_{min} é o BIC dado pelo melhor modelo, ou seja, o que tem menor BIC. Quanto maior o Δ_i, mais evidência tem de se rejeitar o modelo i. Kass e Raftery (1995) sugere a Tabela 5 para verificar a evidência contra o modelo i.

Tabela 5 – Regras de evidência para se rejeitar o modelo i.

$BIC_i - BIC_{min}$	**Evidência contra o modelo i**
0 - 2	apenas uma menção
2 - 6	substancial
6 - 10	forte
> 10	muito forte

Fonte: Kass e Raftery (1995)

3.4 Técnicas de Diagnóstico

Segundo Cordeiro e Demétrio (2013), os resíduos de um modelo são importantes para verificar a discrepância entre os valores observados y_i e ajustados $\hat{\mu}_i$ pelo modelo a fim de observar a qualidade do ajuste e, consequentemente, na escolha do modelo. Com efeito, A análise de resíduos avalia a escolha da distribuição, a função de ligação e o preditor linear. Além disso, ela é útil para identificar observações mal ajustadas ou mesmo atípicas que o modelo não consegue explicar.

Uma matriz de interesse na análise de resíduos é a matriz de projeção que, para os modelos lineares generalizados, tem a seguinte expressão:

$$\boldsymbol{H} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{T}\boldsymbol{W}^{\frac{1}{2}} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X}I^{-1}(\boldsymbol{\beta})\boldsymbol{X}^{T}\boldsymbol{W}^{\frac{1}{2}},$$

em que $\boldsymbol{H}$ é uma matriz idempotente. Essa matriz tem certas propriedades como $tr(\boldsymbol{H}) = \sum_{i=1}^{n} h_{ii} = p$ e $0 \leq h_{ii} \leq 1$. A matriz $\boldsymbol{H}$ depende das variáveis independentes, da função de ligação e da função de variância, o que torna mais difícil a interpretação dos pontos de alavanca, conceito que será visto adiante.

Diante disso, pode-se definir os resíduos R_i para os modelos lineares generalizados como

$$R_i = h_i(y_i, \hat{\mu}_i).$$

Conforme Cordeiro e Demétrio (2013), R_i é escolhida para estabilizar a variância e/ou induzir simetria na distribuição amostral de R_i. Como não se conhece a distribuição exata de R_i utiliza-se resultados assintóticos para definir os tipos de resíduos mais comuns nos MLGs.

Resíduo de Pearson

O resíduo de Pearson é expresso por:

$$R_i^P = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\frac{\hat{\phi}}{\omega_i}V(\hat{\mu}_i)}}.$$

Considerando que, assintoticamente, $\text{Var}(Y_i - \hat{\mu}_i) \approx \text{Var}(Y_i)(1 - h_{ii})$, o resíduo de Pearson padronizado é

$$R_i^{*P} = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\frac{\hat{\phi}}{\omega_i}V(\hat{\mu}_i)(1 - h_{ii})}}.$$

A desvantagem desse resíduo, de acordo com Turkman e Silva (2000), é que sua distribuição é, frequentemente, assimétrica para modelos não normais.

Desvio Residual

O desvio residual é baseado na função desvio e é dado por

$$R_i^D = \pm d^*(y_i, \hat{\mu}_i),$$

em que $d^*(y_i, \hat{\mu_i}) = \sqrt{\frac{2\omega_i}{\hat{\phi}}\left[y_i(\tilde{\theta}_i - \hat{\theta}_i) + (b(\hat{\theta}_i) - b(\tilde{\theta}_i))\right]}$. O desvio residual padronizado é obtido dividindo R_i^D por $\sqrt{(1-h_{ii})}$. Então,

$$R_i^{*^D} = \frac{R_i^D}{\sqrt{\hat{\phi}(1-h_{ii})}}.$$

Um valor grande para R_i^D sugere que a i-ésima observação pode estar mal ajustada em relação ao modelo. Williams (1984) verificou, através de simulações, que a distribuição do desvio residual estar mais próximo da normalidade do que a distribuição de Pearson.

3.4.1 Estatísticas para Diagnóstico

As estatísticas para diagnóstico são úteis para verificar pontos atípicos que ocorrem quando o modelo é ajustado, ou seja, quando certas observações não seguem o padrão das outras observações. Segundo McCullagh e Nelder (1989), essas observações podem ser classificadas por terem h_{ii} e/ou resíduos grandes, serem inconsistentes e/ou influentes.

Com efeito, as estatísticas mais usadas para verificar tais pontos são:

a) Pontos de alavanca;

b) Medida de inconsistência;

c) Medida de influência.

Pontos de alavanca

Conforme Paula (2013), os pontos de alavanca tem o objetivo de avaliar a influência de y_i sobre o próprio valor ajustado $\hat{y}_i$ e essa medida é dada por h_{ii}, ou seja, pelo i-ésimo elemento da diagonal principal da matriz de projeção $\boldsymbol{H}$. Como vimos que $tr(\boldsymbol{H}) = \sum_{i=1}^{n} h_{ii} = p$ e considerando-se que, em média, cada valor h_{ii} deve estar próximo de p/n, então um ponto de alavanca é dado por

$$h_{ii} > \frac{2p}{n}.$$

Uma maneira conveniente, porém informal, de visualizar observações consideradas pontos de alavanca consiste em utilizar um gráfico dos h_{ii} contra os valores ajustados com limite $h = 2p/n$.

Medida de influência

De acordo com Turkman e Silva (2000), um ponto influente é uma observação que afeta significativamente o modelo, ou seja, sua modificação ou exclusão produz alterações relevantes nas estimativas dos parâmetros ou de um determinado parâmetro.

Pode-se usar uma medida baseada na verossimilhança que verifica a influência da retirada da i-ésima observação em $\hat{\boldsymbol{\beta}}$ e é dado por

$$D_i = 2\left\{\ell(\hat{\boldsymbol{\beta}}; \boldsymbol{y}) - \ell(\hat{\boldsymbol{\beta}}_{(i)}; \boldsymbol{y})\right\},$$

sendo $\hat{\boldsymbol{\beta}_{(i)}}$ as estimativas de máxima verossimilhança do vetor $\boldsymbol{\beta}$ sem a observação y_i. Um valor de D alto então indica que a observação y_i é considerada influente. Dado que $\text{Cov}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$, obtém-se a generalização da medida de influência de Cook expresso por

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{p\hat{\boldsymbol{\phi}}}. \tag{3.14}$$

Para estimar $\hat{\boldsymbol{\beta}}_i$ utiliza-se métodos iterativos e seria trabalhoso fazer o processo para todas as observações. Nesse caso, uma aproximação utilizada para $\boldsymbol{\beta}_i$ é fazer apenas o 1º passo do processo iterativo, tendo $\hat{\boldsymbol{\beta}}$ como valor inicial. Portanto,

$$\hat{\boldsymbol{\beta}}^1_{(i)} = \boldsymbol{I}^{-1}_{(i)}(\hat{\boldsymbol{\beta}})\boldsymbol{X}^T_{(i)}\boldsymbol{W}_{(i)}(\hat{\boldsymbol{\beta}})\boldsymbol{v}(\hat{\boldsymbol{\beta}}).$$

Uma forma mais simples para $\hat{\beta^1}_{(i)}$ é dada por

$$\hat{\boldsymbol{\beta}}^{\mathbf{1}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{R_i^{*^P}}{\sqrt{1-h_{ii}}}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{x}_i. \tag{3.15}$$

Logo, substituindo a expressão acima na Equação 3.14 obtém-se

$$D_i \cong \frac{\hat{h}_{ii}}{(1-\hat{h}_{ii})}R_i^P$$

De acordo com Paula (2013), essa aproximação subestima o valor real da distância de cook generalizada, porém ela é suficiente para identificar os casos mais influentes.

Medida de consistência

Uma observação é dita consistente quando ela segue o comportamento sugerido pelas demais observações mesmo ela tendo um alto valor de leverage. Já uma observação inconsistente *(outlier)*, segundo McCullagh e Nelder (1989), é uma observação que possui um resíduo elevado. Essa inconsistência pode ser justificada a um valor extremo da variável dependente e/ou de uma ou mais covariáveis.

Williams (1987) propôs a utilização de um resíduo baseado no desvio reduzido quando se elimina do modelo a i-ésima observação e é denominado por resíduo de verossimilhança. Então, tem-se:

$$G_i = D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) - D^*(\boldsymbol{y}_{(i)}, \hat{\boldsymbol{\mu}}_{(i)}) = \phi^{-1}[d_i + \sum_{j\neq i} d_j - \sum_{j\neq i} d_{(i),j}],$$

sendo $\boldsymbol{y}_{(i)}$ o vetor sem o elemento y_i. Desenvolvendo em série de Taylor a função desvio e considerando a aproximação de $\boldsymbol{\beta}_{(i)}$ da Equação 3.15 chega-se à $D^*(\boldsymbol{y}_{(i)}, \hat{\boldsymbol{\mu}}_{\boldsymbol{(i)}}) = \phi h_{ii}(R_i^{*^P})^2$. Logo, G_i pode ser aproximado por $R_i^{2^G}$ que é expresso por

$$R_i^{2^G} = \phi^{-1}d_i + h_{ii}(R_i^{*^P})^2 = (1-h_{ii})(R_i^{*^D})^2 + h_{ii}(R_i^{*^P})^2.$$

Portanto, o resíduo de verossimilhança é dado por

$$R_i^{*^G} = \pm\sqrt{(1-h_{ii})(R_i^{*^D})^2 + h_{ii}(R_i^{*^P})^2}.$$

Uma observação poderá ser considerada inconsistente quando possui um valor alto para $R_i^{*^G}$. Para analisar as observações a respeito da consistência, Williams (1987) recomenda fazer um gráfico de $R_i^{*^G}$ versus i, h_{ii} ou $\hat{\eta}_i$ e/ou usar $maxR_i^{*^G}$ como estatística para testar se a observação é um *outlier*.

3.4.2 Análise Gráfica

De acordo com Turkman e Silva (2000), o uso de representações gráficas é uma ferramenta informal, porém bastante utilizada e útil, na análise de resíduos. Através delas podemos encontrar desvios tanto no componente aleatório como no componente sistemático. Os gráficos variam conforme os desvios que se pretende encontrar. Os mais recomendados são os seguintes:

a) Gráfico de R^{*^D} versus a ordem das observações, ou versus os valores ajustados;

b) Gráfico normal de probabilidade de R^{*^D} com envelope (sendo os residuos criados pela distribuição pertinente ao modelo);

c) Gráfico de $\hat{\psi}_i$ contra $\hat{\eta}_i$ para a verificação da função de ligação (uma tendência linear indica que a ligação está adequada);

4 Material e Métodos

4.1 Material

Os dados utilizados são referentes à função pulmonar e síndrome metabólica de adolescentes escolares do munícipio de Campina Grande - PB e foram disponibilizados por Vânia (2016). Foram avaliados 525 adolescentes escolares segundo covariáveis socioeconômicas, clínicas, bioquímicas e de estilo de vida. Na Tabela 6 encontram-se as variáveis que serão utilizadas para a análise dos modelos, sendo as variáveis Pressão inspiratória máxima média (Pimáx) e Pressão expiratória máxima média (Pemáx) as variáveis sob estudo. Será feito ajustes utilizando as distribuições normal inversa e gama e utilizando a distribuição normal com função de ligação identidade, conhecido como modelo de regressão clássico, para as duas variáveis para verificar se ambas as abordagens produzem os mesmos resultados ou se há vantagens em se usar os MLGs.

A Pressão Inspiratória, de acordo com Costa et al. (2003), mede a força dos músculos inspiratórios em conjunto enquanto a pressão expiratória indica a força dos músculos abdominais e intercostais. Ambos são utilizados na avaliação da função pulmonar, contribuindo no diagnóstico de algumas doenças como a Doença Pulmonar Obstrutiva Crônica (DPOC) (ver Zanchet, Viegas e Lima (2005)).

O modelo inicial proposto para variáveis dependentes utilizando abordagem dos MLGs incluirá todas as covariáveis. As análises foram feitas no *software* `R` e foram utilizados os pacotes *MASS* (VENABLES; RIPLEY, 2002), *car* (FOX; WEISBERG, 2011), *psych* (REVELLE, 2016), *effect* (FOX et al., 2003) e o pacote *stats*, que já faz parte do `R`. Os gráficos normal de probabilidade com envelope dos desvios padronizados foram feitos através de *scripts* disponibilizados por Paula (2013) em seu site <https://www.ime.usp.br/~giapaula/textoregressao.htm>.

Tabela 6 – Codificação das variáveis explicativas e as variáveis resposta Pimáx e Pemáx

Variável	**Codificação**
Pimáx	Pressão inspiratória máxima média
Pemáx	Pressão expiratória máxima média
SEXO	0 (Masculino) e 1 (Feminino)
TABAGIS	0 (Fumante) e 1 (Não Fumante)
SMCOD	0 (Com Síndrome Metabólica) e 1 (Sem Síndrome Metabólica)
TOTAFIS	Soma da quantidade de minutos de atividade física na semana
ESCMATER	Quantidade de anos que as mães estudaram
IMC	Índice de Massa Corporal
HRSEDCAL	Média do tempo gasto com computador, tv e videogame
NMEDPAS	Média de pressão arterial sistólica
NMEDPAD	Média de pressão arterial diastólica
MEDCABDO	Média da circunferência abdominal
HDL	Medida bioquímica do colesterol
TG	Medida bioquímica dos triglicerídeos
GLICEMIA	Medida bioquímica da glicemia
IDADE	em escala contínua

4.2 Métodos

4.2.1 Distribuição Normal Inversa

Função de densidade de probabilidade

Seja Y uma variável aleatória seguindo uma distribuição normal inversa, sua função de densidade de probabilidade (p.d.f) é dada por

$$f(y,\mu,\phi) = (2\pi\phi y^3)^{-1/2} exp\left\{-\frac{-(y-\mu)^2}{2\mu^2\phi y}\right\},$$

ou ainda, segundo a notação de McCullagh e Nelder (1989),

$$exp\left\{\frac{1}{\phi}\left(-\frac{y}{2\mu^2}+\frac{1}{\mu}\right)-\frac{1}{2}\left(\ell n\left(2\pi\phi y^3\right)+\frac{1}{\phi y}\right)\right\},$$

em que $y > 0$, $\mu > 0$ e $\phi > 0$, sendo μ o parâmetro da média e ϕ o parâmetro de dispersão. Na Figura 1 observa-se o comportamento da função de densidade de probabilidade (a) e a função de distribuição acumulada (b) para determinados valores de ϕ para $\mu = 2$ fixo.

Percebe-se que quando $\phi \longrightarrow 0$, a distribuição se torna simétrica em torno da média. Conforme Paula (2013), Y se aproxima assintoticamente de uma distribuição normal $N(\mu, \phi\mu^3)$ quando $\phi \longrightarrow 0$. Dessa forma, a normal inversa pode tanto ser usada para dados assimétricos quanto para dados simétricos que dependem da forma cúbica da média.

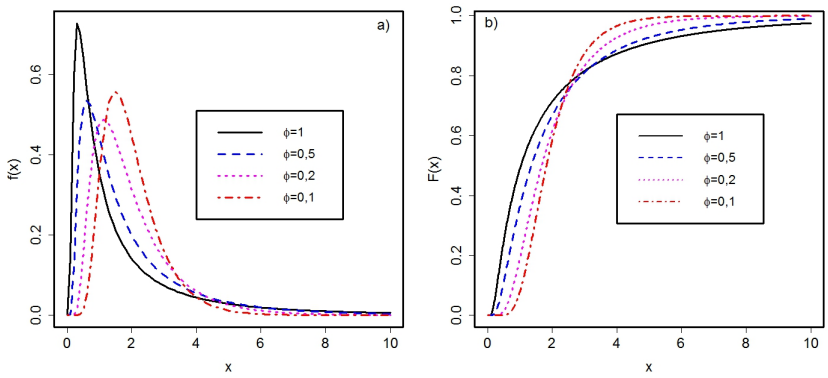

Figura 1 – Gráfico de densidade de probabilidade e função de distribuição acumulada da normal inversa para certos valores de ϕ e $\mu = 2$ fixo.

Função Geradora de Momentos e Cumulantes

A função geradora de momentos para a distribuição normal inversa é dada por

$$M(t,\theta,\phi) = exp\left\{\frac{1}{\phi}\left[\frac{1}{\mu} - \left(\frac{1}{\mu^2} - 2t\phi\right)^{1/2}\right]\right\}, t < \frac{1}{2\phi\mu^2}.$$

Ainda, a função geradora de cumulantes é expressa por

$$\varphi(t,\theta,\phi) = lnM(t,\theta,\phi) = \left\{\frac{1}{\phi}\left[\frac{1}{\mu} - \left(\frac{1}{\mu^2} - 2t\phi\right)^{1/2}\right]\right\}, t < \frac{1}{2\phi\mu^2}.$$

Derivando sucessivamente em relação a t, obtém-se o r-ésimo cumulante

$$\varphi^{(r)}(t,\theta,\phi) = (\phi)^{(r-1)}\left[\frac{1}{\mu} - \left(\frac{1}{\mu^2} - t\phi\right)^{1/2}\right]^{(r)}.$$

Determinando $t = 0$, pode-se obter a esperança e a variância a partir do primeiro e segundo cumulantes, respectivamente. Então, tem-se

$$\begin{aligned}
\varphi^{(1)}(t,\theta,\phi) &= (\phi)^{(1-1)}\left[\frac{1}{\mu} - \left(\frac{1}{\mu^2} - t\phi\right)^{1/2}\right]^{(1)} = b'(\theta) = \mu = E(Y) \\
\varphi^{(2)}(t,\theta,\phi) &= (\phi)^{(2-1)}\left[\frac{1}{\mu} - \left(\frac{1}{\mu^2} - t\phi\right)^{1/2}\right]^{(2)} = \phi\mu^3 = \phi b''(\theta) = \mathrm{Var}(Y).
\end{aligned}$$

Função de verossimilhança e Função Escore

Considere $Y_i \sim NI(\mu_i, \phi)$ com parâmetro de dispersão conhecido e suponha que $\boldsymbol{y} = (y_1, ..., y_n)^T$ são as observações a serem analisadas. Seja $g(\mu_i) = \eta_i$ dado por $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ em que η_i é a função de ligação, $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$ é o vetor de valores das covariáveis e $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ o vetor de parâmetros a serem estimados. Logo, o logarítmo da função de verossimilhança para a distribuição normal inversa é dada por

$$\ell(\boldsymbol{\beta};\boldsymbol{y}) = \sum_{i=1}^{n} lnf(y_i,\mu_i,\phi) = \frac{1}{\phi}\left(-\frac{y_i}{\mu_i^2} + \frac{1}{\mu_i}\right) - \frac{1}{2}\left(\ell n(2\pi\phi y_i^3) + \frac{1}{\phi y_i}\right).$$

A função escore é importante para estimar os β_j e, como foi visto, tem-se que $U_j = \dfrac{\partial \ell(\boldsymbol{\beta}; \boldsymbol{y})}{\partial \beta} = 0$ e a equação é expressa por

$$U_j = \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\beta}; \boldsymbol{y})}{\partial \theta_i} \frac{1}{\dfrac{\partial \mu_i}{\partial \theta_i}} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_i}$$

Assim,

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\beta}; \boldsymbol{y})}{\partial \theta_i} &= \frac{y_i - \mu_i}{\phi} \\ \frac{\partial \mu_i}{\partial \theta_i} &= b''(\theta) = \mu^3 \\ \frac{\partial \eta_i}{\partial \beta_i} &= x_{ij}. \end{aligned}$$

Logo,

$$U_j = \frac{(y_i - \mu_i)}{\phi} \frac{1}{\mu^3} \frac{d\mu_i}{d\eta_i} x_{ij}$$

em que $\dfrac{d\mu_i}{d\eta_i}$ será a derivada da função de ligação. As ligações mais usadas para a normal inversa, de acordo com Paula (2013), são identidade ($\mu_i = \eta_i$), logarítmica ($\ell n \mu_i = \eta_i$), recíproca ($\mu_i = \eta_i^{-1}$) e recíproca ao quadrado ($\mu_i = \eta_i^{-2}$), sendo esta última a ligação canônica.

4.2.2 Distribuição Gama

Função de densidade de probabilidade

Considere Y uma variável aleatória que tem distribuição gama, sua p.d.f é expressa por

$$f(y, \mu, \phi) = \frac{\left(\dfrac{\phi}{\mu}\right)^{\phi}}{\Gamma(\phi)} y^{\phi - 1} e^{\frac{-\phi y}{\mu}}$$

ou ainda, segundo a notação de McCullagh e Nelder (1989),

$$= \exp\left\{ \frac{1}{\phi} \left[-\frac{y}{\mu} - \ell n \mu \right] + \frac{1}{\phi} log\left(\frac{y}{\phi}\right) - \ell n y - \ell n \Gamma\left(\frac{1}{\phi}\right) \right\},$$

em que $y > 0$, $\mu > 0$, sendo μ o parâmetro da média e ϕ o parâmetro de pertubação. Na Figura 2 se percebe o comportamento da função de densidade (a) e da função acumulada (b) da distribuição para alguns valores de ϕ e para $\mu = 4$ fixo.

Observa-se que, assim como a normal inversa, quando $\phi \longrightarrow 0$ a distribuição gama se aproxima assintoticamente de uma distribuição normal $N(\mu, \phi\mu^2)$. Dessa forma, a distribuição gama pode ser adequada para o estudo de variáveis assimétricas e também simétricas que dependem da forma quadrática da média.

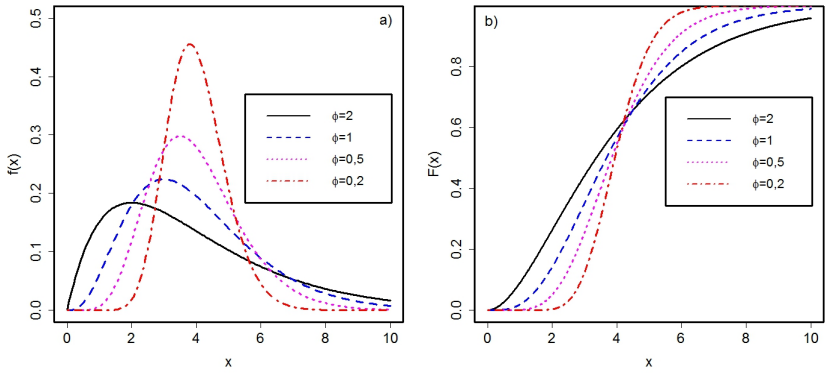

Figura 2 – Gráfico de densidade de probabilidade e função de distribuição acumulada da distribuição gama para certos valores de ϕ e $\mu = 4$ fixo.

Função geradora de Momentos e Cumulantes

A função geradora de momentos para a distribuição gama é expressa por

$$M(t, \theta, \phi) = (1 - t\phi\mu)^{-1/\phi}, t < \frac{1}{\phi\mu}.$$

E a função de cumulantes é dada por

$$\varphi(t, \theta, \phi) = \ell n[M(t, \theta, \phi)] = \ell n(1 - t\phi\mu)^{-1/\phi}.$$

Derivando sucessivamente em relação a t, obtém-se o r-ésimo cumulante

$$\varphi(t, \theta, \mu) = (-\phi\mu)^{(r-1)} \left[(1 - t\phi\mu)^{1/\phi}\right]^{(r)}.$$

Fazendo t=0, tem-se a esperança e a variância a partir do primeiro e do segundo cumulante, respectivamente. Então:

$$\begin{aligned} \varphi^{(1)}(t, \theta, \phi) &= \left[(1 - t\phi\mu)^{1/\phi}\right] = b'(\theta) = \mu = E(Y) \\ \varphi^{(2)}(t, \theta, \phi) &= (-\phi\mu)\left[(1 - t\phi\mu)^{1/\phi}\right]^{(2)} = \phi b''(\theta) = \phi\mu^2 = Var(Y) \end{aligned}$$

Função de Verossimilhança e Função escore

Considere $Y_i \sim G(\mu_i, \phi)$ com parâmetro de dispersão conhecido e suponha que $\boldsymbol{y} = (y_1, ..., y_n)^T$ são as observações a serem analisadas. Seja $g(\mu_i) = \eta_i$ dado por $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ em que η_i é a função de ligação, $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$ é o vetor de valores das covariáveis e $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ o vetor de parâmetros a serem estimados. Logo, o logarítmo da função de verossimilhança para a distribuição gama é dada por

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{n} \exp\left\{\frac{1}{\phi}\left[-\frac{y_i}{\mu_i} - \ell n\mu_i\right] + \frac{1}{\phi}\ell n\left(\frac{y_i}{\phi}\right) - \ell n y_i - \ell n\Gamma\left(\frac{1}{\phi}\right)\right\}$$

A função escore é expressa por

$$\boldsymbol{U}_j = \sum_{i=1}^{n} \frac{\partial\ell(\boldsymbol{\beta}; \boldsymbol{y})}{\partial\theta_i} \frac{1}{\frac{\partial\mu_i}{\partial\theta_i}} \frac{\partial\mu_i}{\partial\eta_i} \frac{\partial\eta_i}{\partial\beta_i}$$

Então

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\beta};\boldsymbol{y})}{\partial \theta_i} &= \frac{y_i-\mu_i}{a(\phi)}\\ \frac{\partial \mu_i}{\partial \theta_i} &= b''(\theta)=\mu^2\\ \frac{\partial \eta_i}{\partial \beta_i} &= x_{ij}.\end{aligned}$$

Portanto,

$$U_j = \frac{(y_i-\mu_i)}{\phi}\frac{1}{\mu^2}\frac{\partial \mu_i}{\partial \eta_i}x_{ij}$$

em que $\frac{d\mu_i}{d\eta_i}$ será a derivada da função de ligação. De acordo com Paula (2013), as principais funções de ligação para a distribuição gama são a identidade ($\mu_i = \eta_i$), a logarítmica ($\ell n\mu_i = \eta_i$) e a recíproca ($\mu_i^{-1} = \eta_i$), sendo esta última a ligação canônica dada por $\frac{d\theta}{d\mu_i} = \frac{d\mu_i}{d\eta_i} = \frac{1}{\mu}$

4.2.3 Modelos de regressão clássico

O modelo de regressão clássico, ou modelo normal linear, conforme Rawlings, Pantula e Dickey (2001), é dado por

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon_i.$$

O modelo de regressão clássico pode ser visto como um MLG em que a variável resposta segue uma distribuição normal com função de ligação identidade ($\eta = \mu$). De acordo com Rawlings, Pantula e Dickey (2001), existem suposições que os modelos de regressão normal devem sastifazer para que se tenha um modelo apropriado, são elas:

i ϵ_i e ϵ_j são independentes ($i \neq j$);

ii $\text{Var}(\epsilon_i) = \sigma^2$ (constante para todo ϵ_i);

iii $\epsilon_i \sim N(0, \sigma^2)$ (os resíduos seguem uma distribuição normal).

A violação de alguma dessas suposições podem levar à resultados não confiáveis. Segundo Pino (2014), a falta de normalidade dos resíduos pode introduzir viéses nos desvios padrão, afetando a validade dos intervalos de confiança e testes de hipóteses. Ainda, os desvios da normalidade frequentemente são seguidos de heteroscedasticidade das variâncias, o qual, em muitas situações, ocorre quando há uma correlação entre a variância e a média, implicando assimetria. Para resolver o problema da normalidade, foram propostas transformações nos dados a fim de que os dados transformados tenham distribuição normal ou aproximadamente normal. Transformções comuns na variável Y são $\ell n(Y)$, $1/Y$,$\sqrt{Y}$, entre outras.

Uma família de transformações bastante conhecida na literatura é a transformação Box-Cox, também conhecida como transformação potência, desenvolvida por Box e Cox (1964) e é expressa por

$$y(\lambda) = \begin{cases} \dfrac{y^{(\lambda)} - 1}{\lambda}, & se\ \lambda \neq 0 \\ \ell n(y), & se\ \lambda = 0 \end{cases}$$

O valor de λ escolhido deve ser aquele que maximiza o logarítmo da função de verossimilhança. Contudo, a transformação Box-Cox é recomendada quando a variável assume apenas valores positivos. Em muitas situações também é preciso transformar a variável independente X a fim de se obter relações lineares e/ou homogeneidade das variâncias. Transformações nas variáveis Y e X geralmente é suficiente para tornar o modelo apropriado (RAWLINGS; PANTULA; DICKEY, 2001).

A seleção de modelos na regressão clássica também pode ser feito considerando os critérios AIC e BIC. Ainda, técnicas de diagnósticos e análises gráficas também são utilizadas para avaliar pontos de alavanca, influentes e *outliers*, embora algumas das estatisticas utilizadas possam ser diferentes das aplicadas nos MLGs. Informações sobre tais técnicas podem ser verificadas em (RAWLINGS; PANTULA; DICKEY, 2001).

5 Resultados e discurssão

5.1 Análise Descritiva

Na Tabela 9 encontram-se as análises descritivas de todas as variáveis ajustadas aos modelos. Pode-se verificar se a Pimáx e Pemáx seguem uma distribuição normal utilizando o teste de Shapiro-Wilk e o teste de Kolmogorov–Smirnov[6]. Através da Tabela 7 verifica-se que ambas Pimáx e Pemáx não seguem normalidade, com valores p menores que 0,05 para as duas estatísticas. De fato, observando-se o histograma de ambas variáveis, na Figura 3, confirma-se uma assimetria positiva. Dessa forma, é razoável supor, considerando uma abordagem via MLGs, que modelos seguindo distribuições assimétricas como a normal inversa ou gama se ajustem bem aos dados. Considerando-se uma abordagem via modelos de regressão clássico, é provável que necessite de transformações a fim de alcançar a normalidade dos resíduos.

Tabela 7 – Valores p dos Teste de Shapiro-Wilk e Kolmogorov-Smirnov da Pressão inspiratória máxima média e Pressão inspiratória máxima média.

Variável	**Shapiro-Wilk**	**Kolmogorov-Smirnov**
Pimáx	<0,001	<0,001
Pemáx	<0,001	<0,001

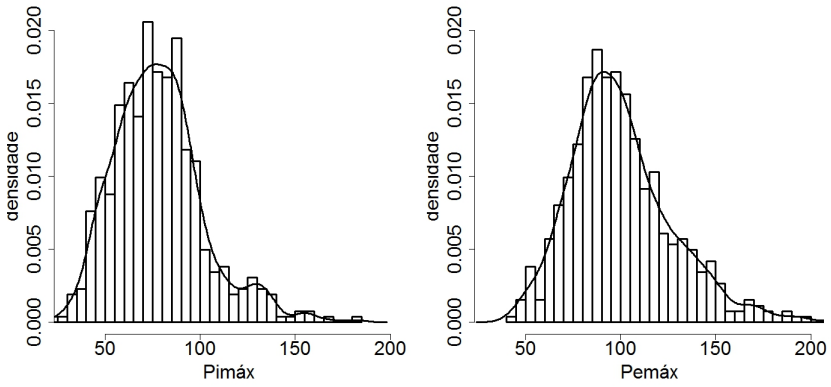

Figura 3 – Histograma das variáveis dependentes Pressão inspiratória máxima média e Pressão expiratória máxima média.

Freitas, Araujo e Alves (2012) encontraram uma diferença significativa entre a média da Pimáx de idosos não fumantes e fumantes. Já em relação a Pemáx, os autores não encontraram diferenças significativas na média dos dois grupos. Logo, tem-se por interesse verificar se esse resultados também são encontrados para adolescentes. Através

[6] para mais detalhes sobre essas estatísticas que testam a normalidade dos dados, ver Razali, Wah et al. (2011)

do teste não paramétrico de Mann-Whitney[7] verifica-se que a média da Pimáx não difere entre não fumantes e fumantes, com valor p > 0,05, conforme consta na Tabela 8. Já em relação a variável Pemáx, há diferença estatisticamente significativa entre a média dos fumantes e não fumantes, com valor p < 0,05. Aparentemente, verifica-se que os valores da Pimáx e Pemáx em relação à condição de fumante ou não do indívio difere entre idosos e adolescentes.

Tabela 8 – Pressão inspiratória máxima média e Pressão expiratória máxima média de fumantes e não fumantes e valores p da estatística de Mann-Whitney.

	Média fumante	**Média não fumante**	**Valor p**
Pimáx	91,600	78,050	0,090
Pemáx	123,060	100,240	0,042
Número de observações	515	10	

[7] Teste não paramétrico utilizado para comparar se dois grupos independentes pertecem ou não a mesma população. Ver (KRASKA-MILLER, 2013)

Tabela 9 – Análise descritiva das variáveis explicativas e das variáveis resposta Pressão inspiratória máxima média e Pressão expiratória máxima média.

Variável	**Média**	**D. Padrão**	**Mediana**	**Mínimo**	**Máximo**	**Amplitude**	**Simetria**	**Kurtose**
Pimáx	78,31	23,84	76	22	182	160	0,82	1,27
Pemáx	100,67	28,07	96,30	43	237,3	194,30	1,04	2,06
TOTAFIS	331,81	318,96	245	0	2830	2830	2,44	10,18
ESCMATER	8,81	3,60	9	0	17,00	17,00	-0,19	-0,34
IMC	21,57	3,90	20,70	14,70	40	25,30	1,45	2,94
HRSEDCAL	3,17	1,81	3	0	12	12	1,39	3,21
NMEDPAS	109,90	9,92	109	86,50	143,50	57	0,38	-0,05
NMEDPAD	66,87	7,02	66,50	48,50	93,50	45	0,29	0,18
MEDCABDO	71,45	8,77	69,60	56	116	60	1,59	3,59
HDL	41,93	9,56	41	20	142	122	2,65	22,52
TG	82,73	39,42	74	30	423	393	2,65	13,43
GLICEMIA	75,95	6,95	76	55	98	43	0,17	-0,11
IDADE	16,82	1,04	16,80	15	19,90	4,90	0,53	-0,10

5.2 Pressão inspiratória máxima média (Pimáx)

5.2.1 Ajuste utilizando a distribuição normal e função de ligação identidade (regressão clássica)

Será ajustado o modelo para a variável resposta Pimáx em que $Y_i \sim N(\mu, \sigma^2)$ com função de ligação identidade ($\eta = \mu$) para verificar se através do modelo de regressão normal obtém-se resultados apropriados que se adequem aos pressupostos do modelo.

Para a seleção dos modelos, pode-se utilizar os critérios AIC e BIC para verificar o melhor modelo que se adeque bem aos dados. Embora seja comum comparar valores de AIC e BIC para verificar qual o melhor modelo, os dois critérios tem filosofias diferentes do que seria o melhor modelo. Segundo Burnham e Anderson (2004), o critério AIC seleciona o melhor modelo que mais se adequa a realidade, sendo essa realidade complexa e desconhecida e nenhum dos modelos possíveis é realmente o melhor modelo. Logo, o AIC encontra o modelo que mais se aproxima dessa realidade, levando à modelos com maior número de parâmetros. Nesse contexto, o AIC é eficiente e assintoticamente escolhe o modelo que minimiza o erro quadrado médio (VRIEZE, 2012).

Já o critério BIC assume, para um tamanho amostral n grande, que existe um melhor modelo dentre o conjunto de modelos possíveis que são gerados pelos dados utiizados, conduzindo à modelos mais simples, com poucos parâmetros. Nessas condições, o BIC é assintoticamente consistente. Logo, será utilizado a metodologia BIC, ou seja, será verificado qual o melhor modelo dentre o conjunto de modelos para os dados da Pimáx. Através do critério de seleção BIC *stepwise* o melhor modelo encontrado, com menor valor de BIC (4667,619), é dado pelas as variáveis IMC, GLICEMIA e SEXO, todas significativas com valor $p < 0,05$. Utilizando o fator de inflação da variância[8] (Variance Inflaction Factor - VIF) verificou-se que não há problemas de multicolinearidade entre as covariáveis, com valores VIF próximos de 1. Logo o modelo proposto é dado por

$$\mu_i = \beta_0 + \beta_1 \text{IMC} + \beta_2 \text{SEXO} + \beta_3 \text{GLICEMIA} + \epsilon_i$$

Conforme foi visto, a variável Pimáx não segue normalidade e para alcançar tal pressuposto fez-se algumas transformações na variável Y a fim de verificar a normalidade dos resíduos. Pino (2014) sugere que as transformações mais comuns para dados positivos assimétricos são a inversa, a logarítmica e a raiz quadrada. Ainda, utilizou-se a transformação Box-Cox para verificar qual valor de λ pode ser usado para transformar Y. Na Figura 4 observa-se o gráfico da transformação de Box-Cox e verifica-se que o valor que maximiza o logarítmo da função de verossimilhança é aproxidamente $\hat{\lambda} = \frac{1}{3}$.

[8] Tem-se por objetivo analisar se há correlação entre as variáveis independentes (colinearidade), pois se houver uma forte correlação entre as variáveis pode aumentar a variância dos coeficientes de regressão. Valores acima de 10 indica problemas de multicolinearidade. Ver (MANSFIELD; HELMS, 1982)

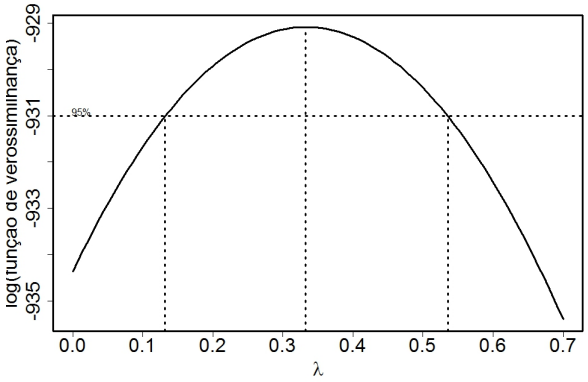

Figura 4 – Gráfico para a família de transformações Box-Cox do modelo ajustado aos dados da Pressão inspiratória máxima média.

Nota-se, na Tabela 10, que apenas a transformação raiz quadrada normalizou os resíduos em ambos os testes, com valores p maiores que 0,05 para os testes de Shapiro-Wilk e Kolmogorov-Smirnov. Na Figura 5 tem-se os gráfico de probabilidades dos resíduos ordinários[9] para as transformações feitas e confirma-se que as transformações logaritmica e inversa aparentam ser afastar da suposição de normalidade. Logo, as análises serão feitas a partir da transformação raiz quadrada. Na Tabela 11 encontra-se as estimativas dos parâmetros para o modelo com a variável dependente transformada.

Tabela 10 – Valores p dos testes de Shapiro-Wilk e Kolmogorov-Smirnov para os resíduos ordinários das transformações logarítmica, inversa, raiz quadrada e Box-Cox, respectivamente, do modelo normal linear ajustado aos dados da Pressão inspiratória máxima média.

transformação	**Shapiro-Wilk**	**Kolmogorov-Smirnov**
logarítmica	<0,001	<0,001
inversa	<0,001	<0,001
raiz quadrada	0,064	0,167
Box-Cox	0,047	0,139

Tabela 11 – Estimativas dos parâmetros referente ao modelo normal linear com transformação raiz quadrada ajustado aos dados da Pressão inspiratória máxima média.

Coeficientes	**Estimativas**	**Erro Padrão**	**Valor p**
Constante	6.693	0,591	<0,001
IMC	0,070	0,013	<0,001
GLICEMIA	0,019	0,007	< 0,009

[9] Também denominado apenas como resíduo, é dado por $\epsilon_i = y_i - \hat{y}_i$ e mede a discrepância entre a observação e seu valor ajustado, ver mais em Rawlings, Pantula e Dickey (2001) pág. 6

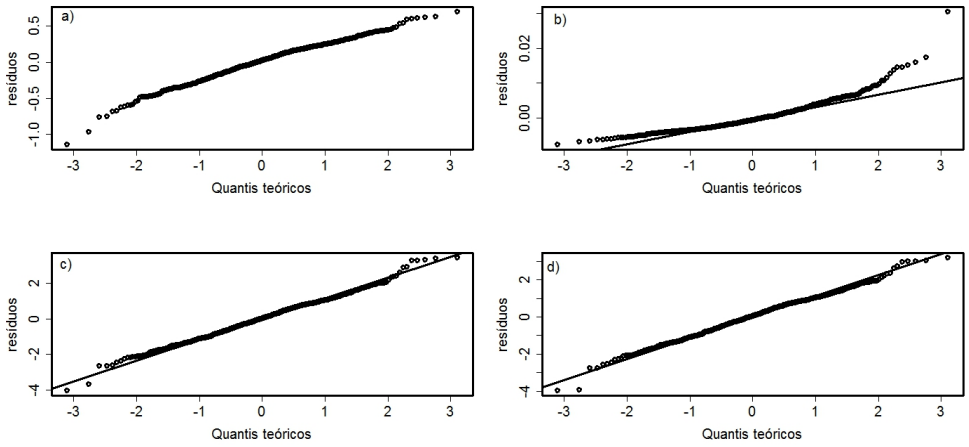

Figura 5 – Gráfico normal de probabilidade dos resíduos ordinários ajustado ao modelo normal linear com transformações logarítmica, inversa, raiz quadrada e Box-Cox, respectivamente, na Pressão inspiratória máxima média.

Como o pressuposto de normalidade foi alcançado, deseja-se verificar se os resíduos são homoscedasticos. Rawlings, Pantula e Dickey (2001) sugerem fazer um gráfico da variável independente *versus* os resíduos do modelo com o objetivo de verificar heteroscedasticidade, considerando um padrão aleatório em torno do 0 como indicativo de que a variância é constante.

Constata-se, através do gráfico entre os resíduos ordinários do modelo e as variáveis IMC e GLICEMIA (Figura 6), que os resíduos não parecem ter um padrão homogêneo em torno do zero para as variável IMC. Foi feita uma transformação inversa na variável IMC e observa-se que a mesma está mais aleatória em torno do 0 (Figura 7), indicando que houve uma melhora em relação à heteroscedasticidade. Quanto a variável GlICEMIA, não foi feito nenhuma transformação, pois a mesma já aparenta ter um comportamento aleatório.

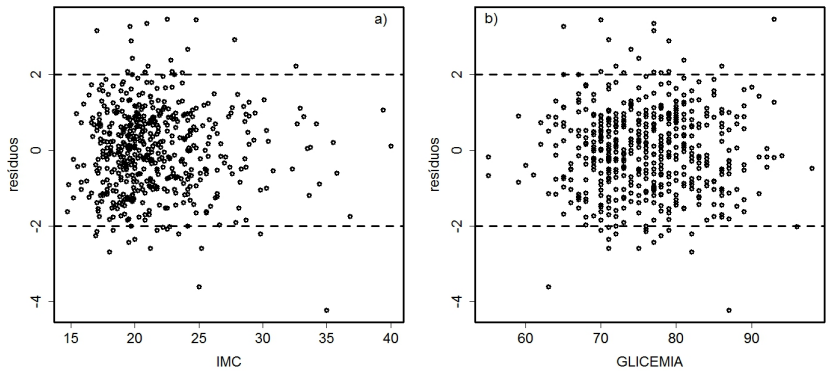

Figura 6 – Gráfico dos resíduos *versus* o Índice de Massa Corporal e Glicemia para a transformação raiz quadrada ajustado aos dados da Pressão inspiratória máxima média.

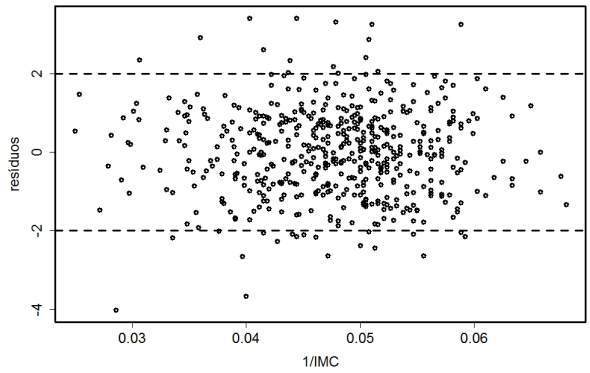

Figura 7 – Gráfico dos resíduos *versus* o Índice de Massa Corporal com transformação inversa para a transformação Box-Cox ajustado aos dados da Pressão inspiratória máxima média.

Na Figura 8 tem-se o gráfico dos resíduos studentizados[10] dá transformação e nota-se que, apesar de haver uma melhora nos resíduos, ainda há indicios de heteroscedasticidade, violando a suposição de homogeneidade das variâncias. Com isso, o modelo de regressão clássico não parece adequado para explica os dados da Pimáx com as variáveis explicativas SEXO, IMC e GLICEMIA.

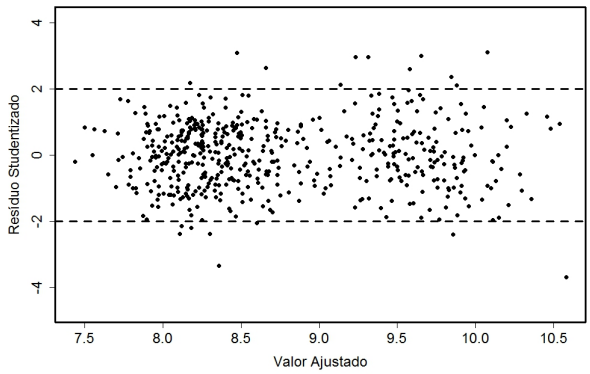

Figura 8 – Gráfico dos resíduos studentizados para a Pressão inspiratória máxima média com tranformação raiz quadrada e Box-Cox, respectivamente.

5.2.2 Ajuste utilizando as distribuições normal inversa e gama

Será suposto que $Y_i \sim G(\mu_i, \phi)$ e $Y_i \sim NI(\mu_i, \phi)$ a fim de verificar qual distribuição se ajusta melhor aos dados. Utiliza-se o critério de seleção BIC com o intuito de comparar os valores de BIC do melhor modelo de cada distribuição e função de ligação. Nas Tabelas 12 e 13 tem-se os modelos com menores valores de BIC das funções de ligação para a distribuição gama e normal inversa.

[10] resíduos studentizados são uma maneira de corrigir a heterogeneidade das variâncias padronizando os resíduos assim como observa-se nos desvios residuais padronizados para os MLGs (ver Rawlings, Pantula e Dickey (2001) pág. 342)

Tabela 12 – Critério de informação BIC para as funções de ligação da distribuição gama ajustado aos dados da Pressão inspiratória máxima média.

Função de ligação	**Modelo**	**BIC**
inversa $\eta = 1/\mu$	$\eta_i = \beta_0 + \beta_1 \text{IMC} + \beta_2 \text{SEXO} + \beta_3 \text{GLICEMIA}$	4628,024
logarítmica $\eta = \ell n(\mu)$	$\eta_i = \beta_0 + \beta_1 \text{IMC} + \beta_2 \text{SEXO}$	4622,657
identidade $\eta = \mu$	$\eta_i = \beta_0 + \beta_1 \text{IMC} + \beta_2 \text{SEXO}$	4618,407

Tabela 13 – Critério de informação BIC para as funções de ligação da distribuição normal inversa ajustado aos dados da Pressão inspiratória máxima média.

Função de ligação	**Modelo**	**BIC**
inversa $\eta = 1/\mu$	$\eta_i = \beta_0 + \beta_1 \text{IMC} + \beta_2 \text{SEXO}$	4647,803
inversa ao quadrado $\eta = 1/\mu^2$	$\eta_i = \beta_0 + \beta_1 \text{IMC} + \beta_2 \text{SEXO} + \beta_3 \text{GLICEMIA}$	4654,696
logarítmica $\eta = \ell n(\mu)$	$\eta_i = \beta_0 + \beta_1 \text{IMC} + \beta_2 \text{SEXO}$	4642,628
identidade $\eta = \mu$	$\eta_i = \beta_0 + \beta_1 \text{IMC} + \beta_2 \text{SEXO}$	4638,21

Com isso, nota-se que o modelo com menor BIC e, consequentemente, o melhor modelo foi o ajuste da distribuição gama com função de ligação identidade. Ainda, percebe-se que nenhum outro ajuste pode ser usado sem perca de informação em comparação à esse modelo. Logo, a distribuição gama com função de ligação identidade será utilizada com as variáveis IMC e SEXO, ambas significativas com valores $p < 0,05$ e não havendo problemas de multicolinearidade. Na Tabela 14 são apresentadas as estimativas dos parâmetros para o modelo gama.

Tabela 14 – Estimativas dos parâmetros referente ao modelo gama ajustado aos dados da Pressão inspiratória máxima média.

Coeficientes	**Estimativas**	**Erro Padrão**	**Valor p**
Constante	63,287	5,240	<0,001
IMC	1,480	0,235	<0,001
SEXOMulheres	-25.380	2,020	<0,001

Analisando as estimativas, percebe-se que a pressão inspiratória difere entre homens e mulheres. Através da Figura 9 nota-se que o sexo masculino apresenta uma média de Pimáx maior em relação ao sexo feminino.

A estimativa do parâmetro de dispersão foi de $\hat{\phi} = 0,06330482$. Com isso, pode-se averiguar o ajuste do modelo através do desvio escalonado ou, ainda, da estatística de Pearson. Não há um consenso entre qual das duas estatisticas é a melhor. Conforme Smyth (2003) a estatística de Pearson é a estatística baseada no teste escore e, portanto, a relação entre a estatística de Pearson e a função desvio é a relação entre a estatística de escore e a estatística da razão de verossimilhança. Sendo a estatística de Pearson mais utilizada por ser de fácil interpretação e por seu valor esperado depender apenas dos dois primeiros momentos da distribuição das observações y_i.

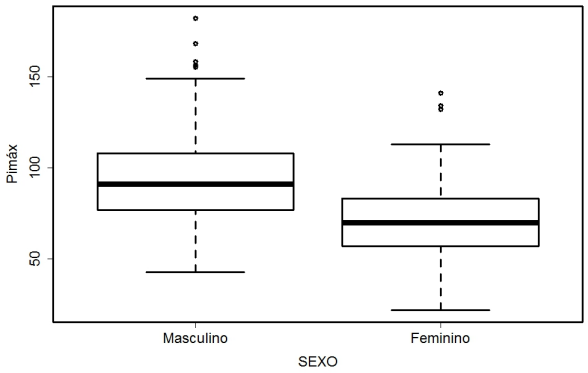

Figura 9 – Valores da Pressão inspiratória máxima média em relação ao Sexo do indivíduo.

Já para Pierce e Schafer (1986), apesar de a estatística de Pearson ter uma melhor aproximação de uma distribuição qui-quadrado e ser mais adequada para dados discretos, os desvios residuais são aproxidamente normais, o que contribui para uma melhor análise dos resíduos. Jorgensen (1987) afirma que quando o parâmetro de dispersão é pequeno, fica razoável comparar os valores observados da função desvio com os percentis da χ^2_{n-p}, ou seja, tem-se que

$$D^*(\boldsymbol{y};\hat{\boldsymbol{\mu}}) \sim \chi^2_{n-p}, \quad quando \quad \phi \longrightarrow 0.$$

Nesse caso, o desvio escalonado do modelo foi de $D^*(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = \frac{34{,}251}{0{,}0633482} = 541{,}0488$ com 522 graus de liberdade e, portanto, o Valor p encontrado foi de p= $0{,}273$, indicando a adequabilidade do modelo.

Por meio dos gráficos de diagnóstico da Figura 10, nota-se alguns pontos de alavanca e observações influentes e aberrantes. As observações [252] e [385] (Figura 10a) aparentam ser pontos de alavanca. Já as observações [466] e [42] apresentam-se como pontos atípicos e influentes (Figura 10b e 10c).

O elemento [42] é do sexo masculino, possui Pimáx de 43 e IMC de 35. Sendo assim, um valor de Pimáx muito pequeno para um valor muito alto de IMC. Já o elemento [466] é do sexo feminino, possui Pimáx de 22, o menor valor encontrado e IMC de 25. Por fim, o elemento [305] é do sexo masculino, tem Pimáx de 182, maior valor encontrado e IMC de 22,5. Percebe-se que as observações encotradas atípicas e influentes contrastam com o comportamento do modelo, pois há observações discrepantes tanto na Pimáx quanto no IMC. Na Figura 10d há indícios que a função de ligação está adequada, visto que, de acordo com McCullagh e Nelder (1989), um padrão adequado para o gráfico (d) é uma linha reta.

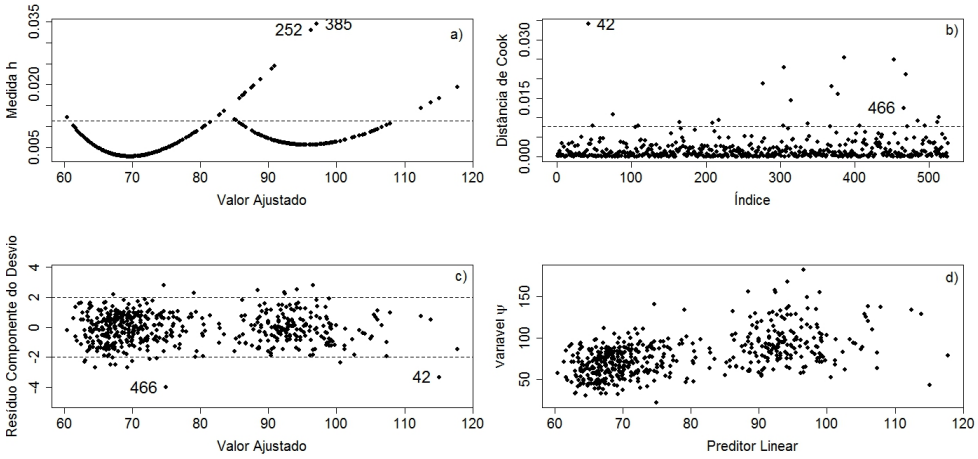

Figura 10 – Gráficos de diagnóstico referente ao modelo gama ajustado aos dados da Pressão inspiratória máxima média.

Por meio da Figura 10c e 10d notam-se indícios de heteroscedasticidade nos resíduos padronizados. Dias (2014) também verificou heteroscedasticade na análise de dados voltados à área da saúde, especificamente na quantidade de sangue recebido em tranfusão. Conforme McCullagh e Nelder (1989) pode-se ajustar modelos com distribuição gama, ou normal inversa, com resíduos heteroscedásticos em que a variância varia em função da média, considerando-se que o parâmetro de dispersão é constante para todas as observações. Essa propriedade se dá pelo fato de que os parâmetros dos modelos são estimados através do método de mínimos quadrados ponderados, conforme foi visto na seção 3.2.2, dando uma vantagem na aplicação dos MLGs. Assim, desconsidera-se a suposição de homoscedasticidade dos resíduos, diferentemente do que ocorre no modelo de regressão clássico.

Porém essa flexibilidade não se aplica em todas as situações. Segundo Dias (2014), para experimentos em que há diferentes grupos de observações, torna-se necessário que o coeficiente de variação (razão entre a média e o desvio padrão) seja constante entre os grupos. Caso não o seja, deve-se resolver o problema da heteroscedasticidade utilizando uma modelagem conjunta da média e do parâmetro de dispersão através dos modelos lineares generalizados duplos (*Double Generalized Linear Models* - DGLM), introduzido por Smyth (1989), ou através dos modelos aditivos generalizados para posição, escala e forma (*Generalized Additive Models for Location, Scale and Shape* - GAMLSS) desenvolvidos por Rigby e Stasinopoulos (2005).

Diante disso, como o nosso modelo não envolve comparações entre grupos de observações, pode-se considerar a heteroscedasticidade dos resíduos. Para analisar o impacto das observações atípicas e influentes nas estimativas dos parâmetros utiliza-se a variação percentual das estimativas que, conforme Possamai (2009), é definido por

$$\mathbf{VP} = \frac{\hat{\beta}_j - \hat{\beta}_{j(m)}}{\hat{\beta}_j} \times 100,$$

em que $\hat{\beta}_j$ é a estimativa de β_j com todas as observações e $\hat{\beta}_{(m)}$ é a estimativa de β sem a m-ésima observação. Ajustou-se o modelo retirando-se as observações [42] e [466] individualmente e também em conjunto. As estimativas dos parâmetros do modelo para as variáveis IMC (β_1) e SEXO (β_2) retirando-se as observações assim como os valores da **VP** encontram-se na Tabela 15.

Tabela 15 – Variação das estimativas do modelo gama ajustado aos dados da Pressão inspiratória máxima média ao excluir as observações [42] e [466], individualmente e em conjunto.

	Estimativas		**Valor p**		**VP(%)**	
Obs. retirada	β_1	β_2	β_1	β_2	β_1	β_2
[42]	1,550	-25,680	<0,001	<0,001	-4,730	-1,186
[466]	1,513	-25,261	<0,001	<0,001	-2,223	0,465
Todas	1,584	-25,562	<0,001	<0,001	-7,027	-0,721

Em relação às variáveis IMC e SEXO, não houveram diferenças significativas nos valores dos parâmetros nem nos valores p mesmo retirando-se todas as observações, sendo a maior variação percentual da IMC de 7,027% e do SEXO de -1,186%. Como a retirada das observações não alterou as estimativas em mais de 10%, optou-se por exclui-las e verificar se houve uma melhora no valor do BIC. De fato, o modelo com todas as observações tem BIC de 4618,407 e sem as observações [42] e [466] tem BIC de 4578,466, uma diferença de 39,941. Logo, percebe-se que houve uma melhora considerável no modelo após a retirada de tais observações.

Na Tabela 16 constata-se as estimativas dos parâmetros do modelo ajustado sem as observações [42] e [466]. O desvio escalonado para o novo modelo foi de $D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = \frac{32{,}483}{0{,}062} = 528,0187$ com 519 graus de liberdade e o valor p igual a $p = 0,394$, dando maiores indícios de um ajuste adequado.

Tabela 16 – Estimativas dos parâmetros referente ao modelo gama ajustado aos dados da Pressão inspiratória máxima média sem as observações [42] e [466].

Coeficientes	**Estimativas**	**Erro Padrão**	**Valor p**
Constante	61,402	5,220	<0,001
IMC	1,584	0,235	<0,001
SEXOMulheres	-25,562	2,000	<0,001

Na Figura 11, tem-se o gráfico normal de probabilidade para o componente do desvio padronizado e confirma-se que não há indicios de afastamentos da suposição de uma distribuição gama para a variável Pimáx. Portanto, conclui-se que o modelo final é dado por

$$\mu_i = \text{Pimax}_i = 61,402 + 1,584\text{IMC} - 25,562\text{SEXO}$$

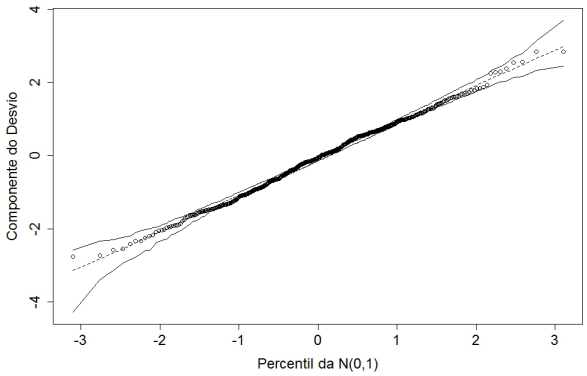

Figura 11 – Gráfico normal de probabilidade referente ao modelo gama ajustado aos dados da Pressão inspiratória máxima média.

De acordo com o modelo ajustado aos dados, pode-se observar que, para cada aumento em uma unidade na variável IMC a um aumento na Pimáx de 1,584 enquanto as outras variáveis permanecem constantes. Já para a variável SEXO, interpreta-se que há uma diminuição de 25,562 na Pimáx para o sexo feminino em relação ao sexo masculino. O modelo encontrado corrobora com a literatura da área. Simões et al. (2007) constataram que a Pimáx dos homens são maiores que os da mulheres com idade entre 40 a 89 anos. O autor também verificou que a idade é uma variável significativa para a Pimáx, observando que indivíduos mais velhos tendem a ter valores de Pimax menores em comparação com indivíduos mais novos.

A variável IDADE pode não ter sido significativa no modelo proposto para a população de adolescentes devido ao fato de que a amplitude das idades dos adolescentes foi pequena, sendo a menor idade encontrada de 15 anos e a maior de 19,9, impossibilitando de se verificar uma diferença significativa nos valores da Pimáx conforme a idade. Obando, López e Ávila (2012) encontraram um modelo bastante semelhante com as variáveis sexo, IMC e idade em pessoas saudáveis acima de 20 anos. Os referidos autores acreditam que a influência do IMC na Pimáx se deve ao fato de um aumento na massa muscular do indivíduo, resultando em melhores níveis na força respiratória. Apesar disso, pessoas com IMC acima de 35 tendem a ter mais peso e não significa, necessariamente, que o indivíduo tenha uma força maior nos músculos respiratórios, provavelmente por haver restrições pulmonares.

Sendo assim, percebe-se uma vantagem na utilizacão dos MLGs para o problema exposto pois, como foi visto, pode-se ter modelos com distribuições gama e normal inversa heteroscedásticos, flexibilizando as suposições que são impostas pelo modelo de regressão clássico. Outra vantagem dos MLGs é o fato de não precisar fazer transformações nas variáveis respostas e/ou explicativas, facilitando, em muitos casos, a interpretação dos resultados. Embora, segundo Cordeiro e Demétrio (2013), não é incomum nos MLGs casos em que os dados são primeiramente transformados para depois ajustar o modelo.

Já em relação ao modelo de regressão clássico, precisou-se fazer transformações tanto na variável resposta quanto na variável independente IMC para alcançar as suposições impostas, dificultando a interpretação dos resultados tendo em vista que precisa-se retransformar a variável resposta para sua escala original para entender a influência das covariáveis na mesma. Com isso, tão importante quanto suposições e estimativas significativas, o modelo precisa ter um sentido científico que tenha praticidade.

Contudo, ainda há outras formas de modelagem através dos modelos normal linear que poderiam ajustar os dados da Pimáx, como é o caso dos modelos lineares heteroscedásticos ((CARROLL; RUPPERT, 1982) e (LONG; ERVIN, 2000)) que poderiam obter modelos sem a necessidade da suposição de homogeneidade dos resíduos. Ou ainda, ter outro conjunto de variáveis que tenham uma correlação com a Pimáx.

5.3 Pressão expiratória máxima média (Pemáx)

5.3.1 Ajuste utilizando a distribuição normal com função de ligação identidade (regressão clássica)

Será suposto que $Y_i \sim N(\mu, \sigma^2)$ com função de ligação identidade ($\eta = \mu$) para analisar se o modelo de regressão clássico é apropriado para os dados Pemáx. Através do critério de seleção BIC, o modelo com menor BIC, com valor de 4841,751, é dado por

$$\mu_i = \beta_0 + \beta_1 \text{IMC} + \beta_2 \text{SEXO} + \epsilon_i.$$

Assim como a Pimáx, a Pemáx também não segue normalidade e as transformações logarítmica, inversa, raiz quadrada e Box-Cox são utilizadas para tentar normalizar os resíduos. O valor λ que maximiza o logarítmo da função de verossimilhança da transformação Box-Cox é aproximadamente $0,1$ como consta na Figura 12, sendo esse o valor usado para a transformação.

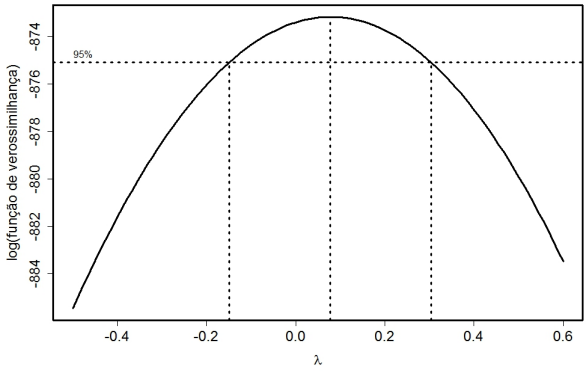

Figura 12 – Gráfico para a família de transformações Box-Cox da Pressão expiratória máxima média.

Na Tabela 17 tem-se os valores p dos testes de Shapiro-Wilk e Kolmogorov-Smirnov para os resíduos e nota-se que nenhum das transformações normalizaram os resíduos, podendo-se confirmar tais resultados através do gráfico normal de probabilidade (Figura 13).

Tabela 17 – Valores p dos testes de Shapiro-Wilk e Kolmogorov-Smirnov para os resíduos ordinários das transformações logarítmica, inversa, raiz quadrada e Box-Cox, respectivamente, do modelo normal linear ajustado aos dados da Pressão expiratória máxima média.

transformação	**Shapiro-Wilk**	**Kolmogorov-Smirnov**
logarítmica	0,005	<0,001
inversa	<0,001	<0,001
raiz quadrada	0,004	<0,001
Box-Cox	0,018	0,001

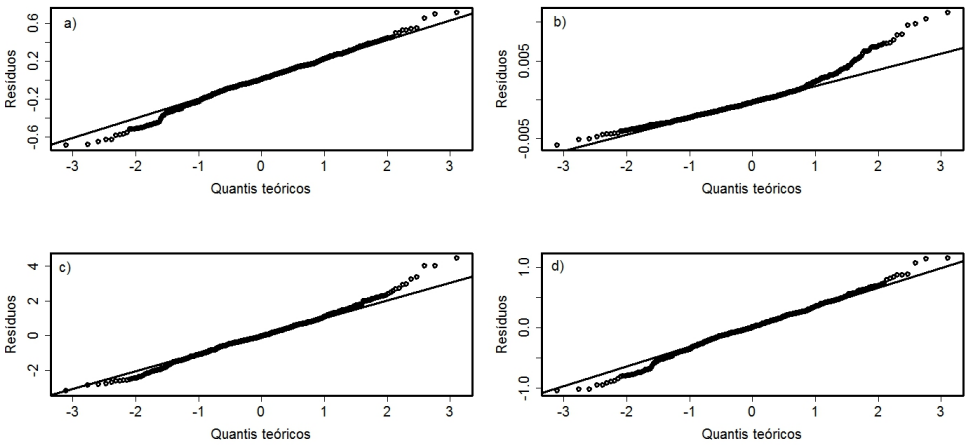

Figura 13 – Gráfico normal de probabilidade para o modelo normal linear com transformações logarítmica, inversa, raiz quadrada e Box-Cox, respectivamente, ajustado aos dados da Pressão expiratória máxima média.

A príncipio, o modelo de regressão normal linear não se ajusta bem ao comportamento assimétrico da Pemáx como pode-se verificar, mesmo após transformações na variável dependente. Outras abordagens através do modelo normal linear podem ser adotadas, como verificar outras variáveis que tenham correlação com a Pemáx e que podem normalizar os erros ou então utilizar modelos de regressão não paramétricos (FARAWAY, 2016), que consistem em outra alternativa para evitar a normalidade dos dados. A desvantagem de tais modelos, segundo Pino (2014), é abdicar de poderosas ferramentas estatísticas para análise dos dados.

5.3.2 Abordagem via MLGs

Para a variável Pemáx, denota-se que $Y \sim NI(\mu, \phi)$ e $Y \sim G(\mu, \phi)$ para verificar qual das distribuições produz um melhor ajuste. Através das Tabelas 18 e 19 constata-se

que a distribuição que obteve o menor valor de BIC foi a distribuição gama com função de ligação identidade, embora o modelo com distribuição gama e função de ligação logarítmica e a distribuição normal inversa com função de ligação identidade tiverem valores de BIC próximos ao valor do modelo com menor BIC, ou seja, não se tem fortes evidências contra esses modelos. Contudo, o modelo escolhido será a distribuição gama com função de ligação identidade, com as variáveis IMC, SEXO e TABAGIS, sendo essas significativas a 5%. Ainda, não houve problema de multicolinearidade nas variáveis independentes.

Tabela 18 – Critério de informação BIC para as funções de ligação da distribuição gama ajustado aos dados da Pressão expiratória máxima média.

Função de ligação	**Modelo**	**BIC**
inversa $\eta = 1/\mu$	$\eta_i = \beta_0 + \beta_1 \text{IMC} + \beta_2 \text{SEXO}$	4776,528
logarítmica $\eta = \ell n(\mu)$	$\eta_i = \beta_0 + \beta_1 \text{IMC} + \beta_2 \text{SEXO}$	4773,680
identidade $\eta = \mu$	$\eta_i = \beta_0 + \beta_1 \text{IMC} + \beta_2 \text{SEXO} + \beta_3 \text{TABAGIS}$	4771,682

Tabela 19 – Critério de informação BIC para as funções de ligação da distribuição normal inversa ajustado aos dados da Pressão expiratória máxima média.

Função de ligação	**Modelo**	**BIC**
inversa $\eta = 1/\mu$	$\eta_i = \beta_0 + \beta_1 \text{IMC} + \beta_2 \text{SEXO}$	4777,415
inversa ao quadrado $\eta = 1/\mu^2$	Não houve convergência	-
logarítmica $\eta = \ell n(\mu)$	$\eta_i = \beta_0 + \beta_1 \text{IMC} + \beta_2 \text{SEXO}$	4774,952
identidade $\eta = \mu$	$\eta_i = \beta_0 + \beta_1 \text{MEDCABDO} + \beta_2 \text{SEXO}$	4772,777

As estimativas dos parâmetros para o modelo se encontram na Tabela 20 e verifica-se que, assim como a Pimáx, a Pemáx dos homens são maiores que os das mulheres como observa-se na Figura 14. Na Figura 15 constata-se que a Pemáx dos fumantes são maiores que as do não fumantes conforme era esperado, pois através do teste não-paramétrico de Mann-Whitney, feito na análise descritiva, analisou-se que a média da Pemáx dos fumantes eram maior que as dos não fumantes.

Tabela 20 – Estimativas dos parâmetros referente ao modelo gama ajustado aos dados da Pressão expiratória máxima média.

Coeficientes	**Estimativas**	**Erro Padrão**	**Valor p**
Constante	112,382	10,558	<0,001
IMC	1,2833	0,2663	<0,001
SEXOMulheres	-30,410	2,333	<0,001
TABAGISnão fumantes	-19,5292	8,6620	0,025

A estimativa do parâmetro de dispersão foi de $\hat{\phi} = 0,05154474$. Com isso, analisa-se a adequabilidade do modelo através do desvio escalonado, cujo valor foi de $D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = \frac{26,994}{0,05154474} = 523,304$ com 521 graus de liberdade e obtendo um valor p de $p = 0,458$. De fato, tem-se evidências de que o modelo parece adequado.

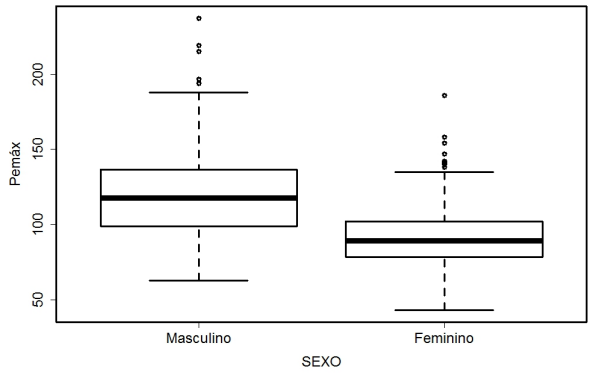

Figura 14 – Valores da Pressão expiratória máxima média em relação ao Sexo do indivíduo.

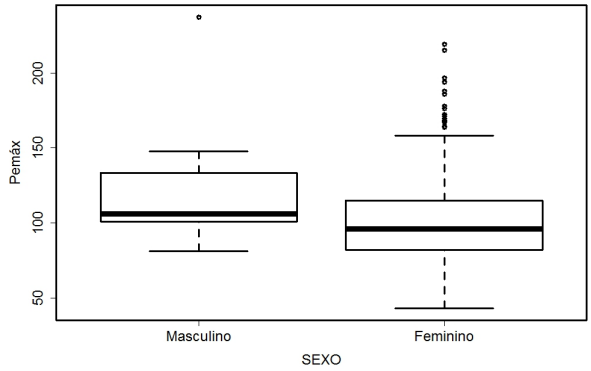

Figura 15 – Valores da Pressão expiratória máxima média em relação à condição de fumante do indivíduo.

Obervando os gráficos de diagnóstico da Figura 16, percebe-se alguns pontos de alavanca (Figura 16a), em especial as observações [25], [137], [167] e [347]. A observação [337]é um ponto influente (Figura 16b) e a observação [466] é um ponto aberrante. Já as observações [277] e [377] são pontos influentes e aberrantes. O elemento [277] é do sexo feminino, não fumante, Pemáx 185,7, o maior valor encontrado para as mulheres, e IMC de 24,8. O elemento [337] é do sexo masculino, fumante, Pemáx de 237,3, o maior valor encontrado, e IMC de 22,8. O elemento [377] é do sexo masculino, não fumante, Pemáx de 219, o segundo maior valor encontrado, e IMC de 19,6. Por fim, o elemento [466] é do sexo feminino, não fumante, com Pemáx de 46,3 e IMC de 25. As observações opõem-se ao modelo pelo fato de se ter valores elevados de PEMÁX para pessoas não fumantes, que, segundo o modelo proposto, deveriam ter Pemáx menores que os dos fumantes. Ainda, a observação [277] é do sexo feminino, o que causa ainda mais contraste tendo em vista que mulheres deveriam ter Pemáx inferiores aos dos homens. A função de ligação parece adequada conforme se ver na Figura 16d.

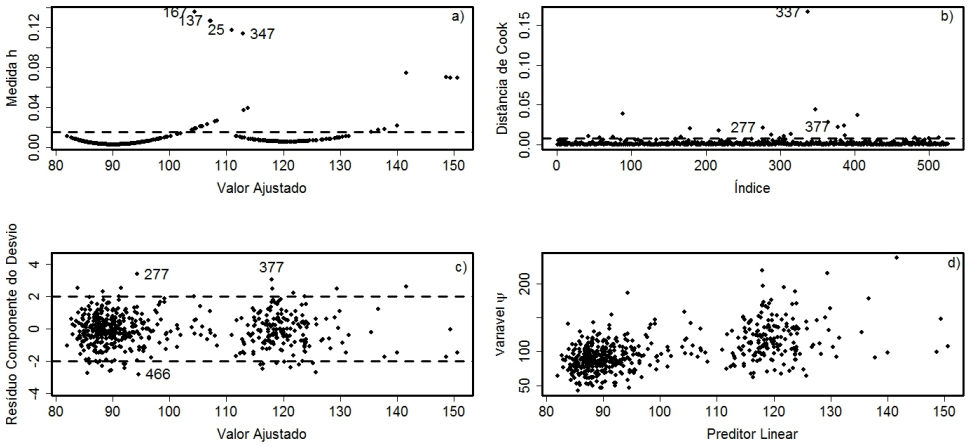

Figura 16 – Gráfico de diagnóstico ajustado ao modelo gama referente aos dados da Pressão expiratória máxima média.

Na Tabela 21 analisa-se as variações percentuais, as estimativas dos parâmetros e os valores p das estimativas após se tirar as observações [277], [337], [377] e [466] individualmente e em conjunto. Nota-se que as observações [277], [377] e [466], quando são retiradas, tem pouca influência nas estimativas dos parâmetros das variáveis, com *VPs* abaixo de 4%. Porém a observação [337] tem uma influência de 36,792% na variável TABAGIS e a mesma não é mais significativa a 5% com valor p de $p = 0,141$. Portanto, o modelo será analisado retirando-se todas as observações e retirando-se apenas as observações [277], [377] e [466].

Tabela 21 – Variação percentual e estimativa dos parâmetros do modelo gama ajustado aos dados da Pressão expiratória máxima média ao excluir as observações [277], [337], [377] e [466], individualmente e em conjunto.

	Estimativas			**Valor p**			**VP(%)**		
Obs. Esperada	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
[277]	1,233	-30,632	-19,713	<0,001	<0,001	<0,021	3,92	-0,730	-0,941
[337]	1,270	-29,966	-12,344	<0,001	<0,001	0,141	1,036	1,460	36,792
[377]	1,297	-29,814	-19,697	<0,001	<0,001	<0,022	-1,067	1,960	-0,859
[466]	1,311	-30,294	-19,433	<0,001	<0,001	0,025	-2,158	0,381	0,493
[Todas]	1,261	29,477	12,522	<0,001	<0,001	0,123	1,738	3,068	35,881

Na Tabela 22 tem-se as estimativas dos parâmetros excluindo-se as observações [277], [337], [377] e [466] e retirando-se a variável TABAGIS. Tal ajuste obteve um valor de BIC de 4696,023. Já o ajuste excluindo-se apenas as observações [277], [377] e [466] teve um valor de BIC de 4716,644 e as estimativas dos parâmetro do modelo encontra-se na Tabela 23. Diante disso, o modelo sem a variável TABAGIS parece representar melhor os dados, pois a diferença entre os valores de BIC encontrados foi maior que 20. Na Figura 17 obseva-se o gráfico normal de probabilidade para os dois ajustes e percebe-se ambos se encontram adequados.

Tabela 22 – Estimativas dos parâmetros referente ao modelo gama ajustado aos dados da Pressão expiratória máxima média sem as observações [277], [337], [377] e [466] e sem a variável TABAGIS.

Coeficientes	**Estimativas**	**Erro Padrão**	**Valor p**
Constante	92,570	5,749	<0,001
IMC	1,258	0,256	<0,001
SEXOMulheres	-29,352	2,241	<0,001

Tabela 23 – Estimativas dos parâmetros referente ao modelo gama ajustado aos dados da Pressão expiratória máxima média sem as observações [277], [377] e [466].

Coeficientes	**Estimativas**	**Erro Padrão**	**Valor p**
Constante	112,210	10,231	<0,001
IMC	1,275	0,256	<0,001
SEXOMulheres	-29,921	2,257	<0,001
TABAGISnão fumantes	-19,784	8,393	0,019

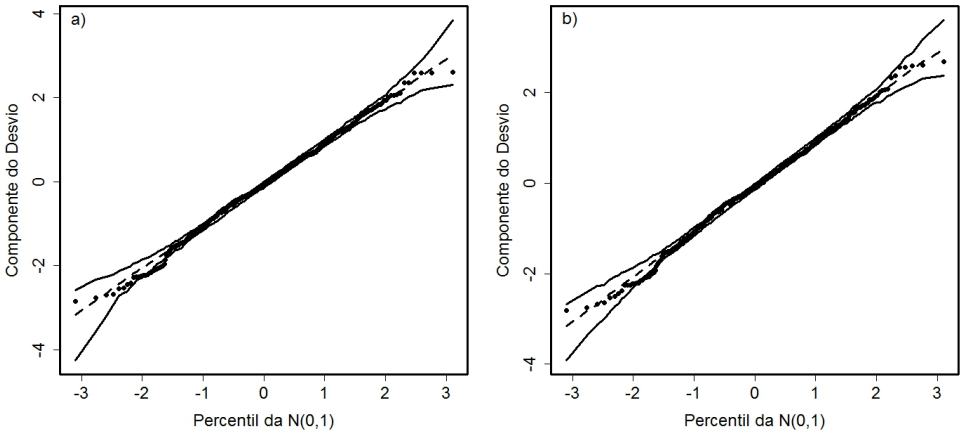

Figura 17 – Gráficos normais de probabilidade ajustado ao modelo gama sem as observações [277], [337], [377] e [466] e sem apenas as observações [277], [377] e [466], respectivamente, referente aos dados da Pressão expiratória máxima média.

Diante disso, os modelos sem a variável TABAGIS e com a variável TABAGIS ficam dados, respectivamente, por

$$\begin{aligned} \mu_i &= \text{Pemáx}_i = 92,570 + 1,258\text{IMC} - 29,352\text{SEXO} \\ \mu_i &= \text{Pemáx}_i = 112,210 + 1,275\text{IMC} - 29,921\text{SEXO} - 19,784\text{TABAGIS}. \end{aligned}$$

Algumas interpretações podem ser feitas para ambos os modelos. Em relação ao modelo sem a variável TABAGIS, nota-se que a Pemáx do sexo feminino tem uma diminuição de 29,352 em relação ao sexo masculino, enquanto as outras variáveis permanecem constantes. Já para o modelo com a variável TABAGIS, analisa-se que a Pemáx dos não fumantes é 19,784 menor do que as dos não fumantes, enquanto as outras variáveis

permanecem constantes.

Diferentemente do que ocorreu com o modelo normal, os ajustes da distribuição gama são adequados sem perder variáveis que são significativas para o modelo. Logo, chama-se atenção ao fato da facilidade na utilização dos MLGs para resolver tais problemas, abdicando das rígidas suposições de normalidade.

6 Conclusão

Os MLGs representam um instrumento adequado na modelagem de dados de natureza contínua com comportamento positivo assimétrico. Modelos com variável resposta seguindo distribuição normal inversa e gama heteroscedásticas podem ser aplicados em várias áreas de conhecimento visto que dados com natureza contínua assimétrica são comuns em situações reais. O *software* R constituiu-se de uma ferramenta de fundamental importância nas análises dos modelos visto que técnicas computacionais são necessárias para a estimação dos parâmetros e para verificar pontos atípicos.

As variáveis de estudo se adequaram bem aos gama propostos, contribuindo para o estudo da função pulmonar em adolescentes e concluindo-se que o sexo e o índice de massa corporal são variáveis que explicam os níveis de Pressão inspiratória máxima média e a Pressão expiratória máxima média.

Percebeu-se que a abordagem via MLGs conduziram à resultados mais adequados aos dados positivos assimétricos da Pimáx e da Pemáx em comparação aos modelos lineares, pois as suposições de normalidade dos erros e homogeneidade das variâncias não foram sastifeitas mesmo após serem feitas algumas transformações nas variáveis para alcançar essas propriedades. Portanto, uma desvantagem dos modelos lineares em relação aos MLGs é a necessidade de transformações nas variáveis, que em muitas situações reais não seguem uma distribuição normal e que podem dificultar a interpretação dos dados. Já uma vantagem dos MLGs em comparação com os modelos lineares é a flexibiização da variável resposta, que pode seguir outra distribuição da família exponencial além da normal e que há possibilidades de várias funções de ligação além da identidade, a qual é a função de ligação dos modelos lineares.

Como constatou-se que ambas Pimáx e Pemáx obtiveram resultados semelhantes, tem-se por interesse em trabalhos futuros verificar uma modelagem bivariada das duas variáveis a fim de analisar se há uma relação entre as mesmas. Ainda, será ajustado modelos GAMLSS a fim de constatar se há outras distribuições que se adequem melhor aos dados estudados e que possibilitem modelar a heteroscedasticidade encotrada.

Referências

ACORSI, C. *Estimação do fator de condição para peixes utilizando modelos lineares generalizados. 2002*. Dissertação (Mestrado) — Dissertação (Mestrado em Engenharia de Produção)–Universidade Federal de Santa Catarina, Florianópolis, 2002. AGOSTINHO, A, 2002. Citado na página 16.

AKAIKE, H. A new look at the statistical model identification. *IEEE transactions on automatic control*, Ieee, v. 19, n. 6, p. 716–723, 1974. Citado na página 34.

ALMEIDA, G. Q.; SILVA, J. de O.; CABRAL, L. T. S.; MATOS, G. R.; MENEGUCI, J. L. P. Influência da iluminação artificial no florescimento dos parentais de híbridos de maracujá (passiflora edulis). *Multi-Science Journal*, v. 1, n. 2, p. 117–123, 2015. Citado na página 15.

ANDREWS, D. A note on the selection of data transformations. *Biometrika*, Biometrika Trust, v. 58, n. 2, p. 249–254, 1971. Citado na página 13.

ASSUNÇÃO, R. *Análise da influência das variáveis pesqueiras e ambientais na abundância do polvo-comum, Octopus vulgaris (Cuvier, 1797), descarregado no estado de São Paulo entre 2003-2011*. Tese (Doutorado) — Instituto de Pesca, 2012. Citado na página 17.

ATKINSON, A. Two graphical displays for outlying and influential observations in regression. *Biometrika*, Biometrika Trust, v. 68, n. 1, p. 13–20, 1981. Citado na página 15.

BARROS, K. N. N. de O.; NASCIMENTO, P. S. Comparaçao entre modelos lineares generalizados utilizando dados sobre aids na cidade de recife. *8º Encontro Regional de Matemática Aplicada e Computacional*, 2008. Citado na página 15.

BOLFARINE, H.; SANDOVAL, M. C. *Introdução à inferência estatística*. [S.l.]: SBM, 2001. v. 2. Citado 5 vezes nas páginas 18, 19, 22, 28 e 30.

BOX, G. E.; COX, D. R. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 211–252, 1964. Citado 2 vezes nas páginas 13 e 45.

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, Taylor & Francis Group, v. 88, n. 421, p. 9–25, 1993. Citado na página 15.

BURNHAM, K. P.; ANDERSON, D. R. Multimodel inference understanding aic and bic in model selection. *Sociological methods & research*, Sage Publications, v. 33, n. 2, p. 261–304, 2004. Citado 2 vezes nas páginas 34 e 49.

CARROLL, R. J.; RUPPERT, D. A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 77, n. 380, p. 878–882, 1982. Citado na página 58.

CASELLA, G.; BERGER, R. L. *Statistical inference*. [S.l.]: Duxbury Pacific Grove, CA, 2002. v. 2. Citado na página 22.

CORDEIRO, G. M.; DEMÉTRIO, C. G. *Modelos lineares generalizados e extensões*. [S.l.: s.n.], 2013. Citado 9 vezes nas páginas 22, 24, 29, 30, 31, 33, 34, 35 e 57.

COSTA, D.; SAMPAIO, L. M. M.; LORENZZO, V. A. P. d.; JAMAMI, M.; DAMASO, A. R. Avaliação da força muscular respiratória e amplitudes torácicas e abdominais após a rfr em indivíduos obesos. *Revista Latino-Americana de Enfermagem*, SciELO Brasil, v. 11, n. 2, p. 156–160, 2003. Citado na página 39.

DANTAS, C. A. B. *Probabilidade: Um Curso Introdutório Vol. 10*. [S.l.]: Edusp, 2013. Citado na página 20.

DEMÉTRIO, C. G. B. *Modelos lineares generalizados em experimentação agronômica*. [S.l.]: USP/ESALQ, 2001. Citado 6 vezes nas páginas 18, 21, 28, 29, 31 e 33.

DIAS, M. F. *Modelos assimétricos inflaciconados de zeros*. Dissertação (Mestrado) — Universidade de São Paulo, 2014. Citado 2 vezes nas páginas 16 e 55.

ELLIOTT, J.; SOUZA, R. de; KRONE-MARTINS, A.; CAMERON, E.; ISHIDA, E.; HILBE, J.; COLLABORATION, C. et al. The overlooked potential of generalized linear models in astronomy-ii: Gamma regression and photometric redshifts. *Astronomy and Computing*, Elsevier, v. 10, p. 61–72, 2015. Citado na página 17.

FARAWAY, J. J. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. [S.l.]: CRC press, 2016. v. 124. Citado na página 59.

FEIGL, P.; ZELEN, M. Estimation of exponential survival probabilities with concomitant information. *Biometrics*, JSTOR, p. 826–838, 1965. Citado na página 13.

FISHER, R. On the mathematical foundations of theoretical statistics. The Royal Society, v. 222, p. 309–368, 1922. ISSN 02643952. Citado na página 13.

FOX, J. et al. Effect displays in r for generalised linear models. *Journal of statistical software*, v. 8, n. 15, p. 1–27, 2003. Citado na página 39.

FOX, J.; WEISBERG, S. *An R Companion to Applied Regression*. Second. Thousand Oaks CA: Sage, 2011. Disponível em: <http://socserv.socsci.mcmaster.ca/jfox/Books/Companion>. Citado na página 39.

FREITAS, E. R. F. S. de; ARAUJO, E. C. L. da S.; ALVES, K. da S. Influência do tabagismo na força muscular respiratória em idosos. *Fisioterapia e Pesquisa*, SciELO Brasil, v. 19, n. 4, p. 326–331, 2012. Citado na página 46.

HASTIE, T.; TIBSHIRANI, R. Generalized additive models. *Statistical science*, JSTOR, p. 297–310, 1990. Citado na página 15.

HELLER, G.; STASINOPOULOS, M.; RIGBY, B. et al. The zero-adjusted inverse gaussian distribution as a model for insurance claims. In: *Proceedings of the 21th International Workshop on Statistical Modelling*. [S.l.: s.n.], 2006. v. 226233. Citado na página 16.

HESS, A. F.; CIANORSCHI, L. D.; SILVESTRE, R.; SCARIOT, R.; RICKEN, P. Aplicação dos modelos lineares generalizados para estimativa do crescimento em altura. *Pesquisa Florestal Brasileira (Online)*, v. 35, p. 427, 2015. Citado na página 17.

HOLANDA, J. S.; VASCONCELLOS, M. C.; SILVA, A. C. *Avaliação do impacto da atividade sísmica sobre a captura de peixes através de monitoramento de desembarque pesqueiro: um estudo de caso no litoral do Rio de Janeiro, Brasil*. Dissertação (Mestrado) — Universidade Federal do Rio Grande, 2012. Citado na página 16.

JORGENSEN, B. Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 127–162, 1987. Citado 2 vezes nas páginas 15 e 54.

KASS, R. E.; RAFTERY, A. E. Bayes factors. *Journal of the american statistical association*, Taylor & Francis Group, v. 90, n. 430, p. 773–795, 1995. Citado na página 34.

KRASKA-MILLER, M. *Nonparametric statistics for social and behavioral sciences*. [S.l.]: CRC Press, 2013. Citado na página 47.

LONG, J. S.; ERVIN, L. H. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, Taylor & Francis Group, v. 54, n. 3, p. 217–224, 2000. Citado na página 58.

MANSFIELD, E. R.; HELMS, B. P. Detecting multicollinearity. *The American Statistician*, Taylor & Francis, v. 36, n. 3a, p. 158–160, 1982. Citado na página 49.

MCCULLAGH, P.; NELDER, J. A. *Generalized linear models*. [S.l.]: CRC press, 1989. v. 37. Citado 10 vezes nas páginas 18, 20, 21, 22, 36, 37, 40, 42, 54 e 55.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, [Royal Statistical Society, Wiley], v. 135, p. 370–384, 1972. Citado na página 13.

OBANDO, L. M. G.; LÓPEZ, A. L.; ÁVILA, C. L. Normal values of the maximal respiratory pressures in healthy people older than 20 years old in the city of manizales-colombia. *Colombia Médica*, Facultad de Salud, Universidad del Valle, Cali, Colombia, v. 43, n. 2, p. 119–125, 2012. Citado na página 57.

PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2013. Citado 12 vezes nas páginas 13, 23, 30, 31, 32, 33, 36, 37, 39, 40, 42 e 44.

PIERCE, D. A.; SCHAFER, D. W. Residuals in generalized linear models. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 81, n. 396, p. 977–986, 1986. Citado 2 vezes nas páginas 33 e 54.

PINO, F. A. A questão da não normalidade: uma revisão1. *Rev. de Economia Agrícola*, v. 61, p. 17–33, 2014. Citado 3 vezes nas páginas 44, 49 e 59.

POSSAMAI, A. A. *Modelos não lineares de famılia exponencial revisitados*. Dissertação (Mestrado) — Universidade de Sao Paulo, 2009. Citado 2 vezes nas páginas 16 e 55.

PREGIBON, D. Logistic regression diagnostics. *The Annals of Statistics*, JSTOR, p. 705–724, 1981. Citado na página 15.

R (R Core Team). *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2016. Disponível em: <https://www.R-project.org/>. Citado na página 14.

RAWLINGS, J. O.; PANTULA, S. G.; DICKEY, D. A. *Applied regression analysis: a research tool*. [S.l.]: Springer Science & Business Media, 2001. Citado 5 vezes nas páginas 44, 45, 50, 51 e 52.

RAZALI, N. M.; WAH, Y. B. et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, v. 2, n. 1, p. 21–33, 2011. Citado na página 46.

REVELLE, W. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois, 2016. R package version 1.6.6. Disponível em: <http://CRAN.R-project.org/package=psych>. Citado na página 39.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 54, n. 3, p. 507–554, 2005. Citado na página 55.

SANTOS, C. S. A. dos; SILVA, A. de O.; GOUVEIA, J. F.; CORDEIRO, G. M. Modelo para mortalidade infantil no brasil via modelos lineares generalizados. *19º Simpósio Nacional de Probabilidade e Estatística*, 2010. Citado na página 16.

SANTOS, T. R. d.; SILVA, R. W. C. Modelando a taxa de neoplasia pulmonar no brasil via modelos lineares generalizados. *Revista da Estatística da Universidade Federal de Ouro Preto*, v. 3, n. 1, 2014. Citado na página 17.

SCHWARZ, G. et al. Estimating the dimension of a model. *The annals of statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. Citado na página 34.

SIMÕES, R. P.; AUAD, M. A.; DIONÍSIO, J.; MAZZONETTO, M. Influência da idade e do sexo na força muscular respiratória. *Fisioterapia e pesquisa*, v. 14, n. 1, p. 36–41, 2007. Citado na página 57.

SMYTH, G. K. Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 47–60, 1989. Citado na página 55.

SMYTH, G. K. Pearson's goodness of fit statistic as a score test statistic. *Lecture Notes-Monograph Series*, JSTOR, p. 115–126, 2003. Citado na página 53.

SOUZA, R. D.; CAMERON, E.; KILLEDAR, M.; HILBE, J.; VILALTA, R.; MAIO, U.; BIFFI, V.; CIARDI, B.; RIGGS, J.; COLLABORATION, C. et al. The overlooked potential of generalized linear models in astronomy, i: Binomial regression. *Astronomy and Computing*, Elsevier, v. 12, p. 21–32, 2015. Citado na página 16.

STIGLER, S. M. Gauss and the invention of least squares. *The Annals of Statistics*, JSTOR, p. 465–474, 1981. Citado na página 13.

TURKMAN, M. A. A.; SILVA, G. L. Modelos lineares generalizados-da teoria à prática. In: *VIII Congresso Anual da Sociedade Portuguesa de Estatística, Lisboa*. [S.l.: s.n.], 2000. Citado 11 vezes nas páginas 19, 24, 25, 27, 28, 30, 31, 33, 35, 36 e 38.

VÂNIA, M. L. *Função Pulmonar e Síndrome Metabólica em adolescentes escolares do município de Campina Grande - Paraíba*. Dissertação (Mestrado) — Universidade Estadual da Paraíba, 2016. Citado 2 vezes nas páginas 14 e 39.

VENABLES, W. N.; RIPLEY, B. D. *Modern Applied Statistics with S*. Fourth. New York: Springer, 2002. ISBN 0-387-95457-0. Disponível em: <http://www.stats.ox.ac.uk/pub/MASS4>. Citado na página 39.

VRIEZE, S. I. Model selection and psychological theory: a discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic). *Psychological methods*, American Psychological Association, v. 17, n. 2, p. 228, 2012. Citado na página 49.

WILLIAMS, D. Residuals in generalized linear models. In: *International Biometrics Conference*. [S.l.: s.n.], 1984. p. 59–68. Citado 2 vezes nas páginas 15 e 36.

WILLIAMS, D. Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics*, JSTOR, p. 181–191, 1987. Citado 3 vezes nas páginas 15, 37 e 38.

ZANCHET, R. C.; VIEGAS, C. A. A.; LIMA, T. A eficácia da reabilitação pulmonar na capacidade de exercício, força da musculatura inspiratória e qualidade de vida de portadores de doença pulmonar obstrutiva crônica. *J Bras Pneumol*, SciELO Brasil, v. 31, n. 2, p. 118–24, 2005. Citado na página 39.