



UNIVERSIDADE ESTADUAL DA PARAÍBA
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA

ALEXSANDRO VENCESLAU DE SOUZA

**Modelos Lineares Generalizados Aplicados a
Dados de Pressão em pré-adolescentes do
município de Campina Grande-PB**

Campina Grande - PB

Novembro 2018

ALEXSANDRO VENCESLAU DE SOUZA

Modelos Lineares Generalizados Aplicados a Dados de Pressão em pré-adolescentes do município de Campina Grande-PB

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 de Olinda

Campina Grande - PB

Novembro 2018

É expressamente proibido a comercialização deste documento, tanto na forma impressa como eletrônica. Sua reprodução total ou parcial é permitida exclusivamente para fins acadêmicos e científicos, desde que na reprodução figure a identificação do autor, título, instituição e ano do trabalho.

S729m Souza, Alexsandro Venceslau de.
Modelos lineares generalizados aplicados a dados de pressão em pré-adolescentes do município de Campina Grande-PB [manuscrito] / Alexsandro Venceslau de Souza. - 2018.
40 p.
Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2018.
"Orientação : Prof. Dr. Ricardo Alves de Olinda, Departamento de Estatística - CCT."
1. Modelos Lineares Generalizados. 2. Modelo Gama. 3. Pressão arterial. I. Título
21. ed. CDD 519.5

ALEXSANDRO VENCESLAU DE SOUZA

Modelos Lineares Generalizados Aplicados a Dados de Pressão em pré-adolescentes do município de Campina Grande-PB

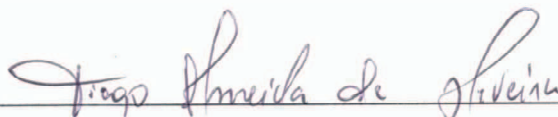
Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de Bacharel em Estatística.

Trabalho aprovado em 27 de Novembro de 2018.

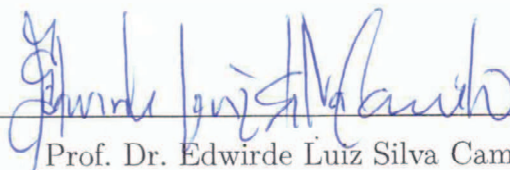
BANCA EXAMINADORA



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



Prof. Dr. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba



Prof. Dr. Edwirde Luiz Silva Camêlo
Universidade Estadual da Paraíba

Agradecimentos

Primeiramente a Deus por sempre me fortalecer e sempre estar à frente de minhas vitórias, e por sua infinita bondade, a todos os professores do curso de estatística que ao longo dos anos me ajudaram a chegar onde cheguei, durante esse período tive a oportunidade aprender com os melhores.

Ao meu orientador, Doutor Ricardo Alves de Olinda, por ter me ajudado com muita inteligência e paciência nessa reta final de curso, ao professor Tiago Almeida de Oliveira pela sua dedicação e paciência para conosco, por sempre ter acreditado no meu potencial, a professora Ana Patricia Bastos Peixoto, por nunca me deixar abater diante dos desafios. E ao professor Juarez Fernandes de Oliveira, por fortalecer a minha experiência acadêmica e pessoal.

Aos amigos que me acompanharam durante os anos de minha graduação, em especial a Ednaldo, a Thuenne, a Renata e Rafaela, companheiros de batalha em busca do conhecimento.

A meus pais, Antônio Venceslau de Souza e Luzinete Vitória Vieira, a meus irmãos, Adriana, Adriano, Alexandra e Raimunda, a minha querida esposa, Joselita Monteiro de Oliveira Venceslau, que sempre esteve ao meu lado.

“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

A classe dos Modelos Lineares Generalizados (MLG) é utilizada quando se quer ajustar modelos probabilísticos à dados que não seguem normalidade e não podem ser ajustados usando apenas a regressão linear. Os modelos apresentados nesse trabalho foram ajustados com auxílio do software R (R core Team, 2017), e o modelo final foi avaliado através do gráfico de diagnóstico residual e do gráfico de envelope simulado. Este estudo teve como objetivo avaliar o comportamento da pressão arterial de um grupo de indivíduos da cidade de Campina Grande-pb. Foram avaliados 576 indivíduos, a pressão arterial sistólica é também conhecida como a pressão máxima ou alta, e se trata da pressão do sangue no momento da sístole cardíaca, em outras palavras, no momento da contração do coração, ocasionando o impulso do sangue para as artérias. Com relação a pressão arterial diastólica, também conhecida como mínima ou baixa, se opõe a pressão arterial sistólica e é influenciada pela resistência imposta pelos vasos contra a passagem do sangue.

Palavras-chave: Saúde Pública, Pressão Arterial, Modelo Gama

Abstract

The Generalized Linear Models (MLG) class is used when you want to fit probabilistic models to data that do not follow normality and can not be adjusted using only linear regression. The models presented in this study were adjusted using the software R (R core Team, 2017), and the final model was evaluated through the residual diagnostic chart and the simulated envelope graph. The objective of the study was to evaluate the blood pressure behavior of group of individuals from the city of Campina Grande-pb. We evaluated 576 individuals, systolic blood pressure is also known as maximum or high pressure, and it deals with blood pressure at the time of cardiac systole in other words, at the time of the contraction of the heart, causing the blood impulse to the arteries. With regard to diastolic blood pressure, also known as minimal or low, it is opposed to systolic blood pressure and is influenced by the resistance imposed by the vessels against the passage of blood.

Key-words: Public health, Blood Pressure, Range Model.

Lista de ilustrações

Figura 1 – Gráfico boxplot das variáveis independentes e dependente.	30
Figura 2 – Diagrama de correlação entre PAS e TG, GLI, HDL, CA.	32
Figura 3 – Diagrama de correlação entre PAD e TG, GLI, HDL, CA.	33
Figura 4 – Gráfico para diagnóstico residual do modelo com a variável PAS e PAD	35
Figura 5 – Gráficos de envelope simulado com as variáveis PAS e PAD	36

Lista de tabelas

Tabela 1 – Funções Geradoras de Momentos para algumas Distribuições.	13
Tabela 2 – Algumas distribuições para a variável resposta Y e a natureza dos dados em que ela é utilizado.	16
Tabela 3 – Ligações Canônicas de algumas distribuições da família exponencial. . .	17
Tabela 4 – Análise descritiva das variáveis independentes, GLI, TG, HDL, CA e variáveis dependentes, PAS, PAD.	29
Tabela 5 – Análise descritiva das variáveis independentes, GLI, TG, HDL, CA e variáveis dependentes, PAS, PAD.	29
Tabela 6 – Categorias de Pressão Sanguínea Sistólica e Diastólica.	31
Tabela 7 – Valores para o teste de normalidade de Anderson-Darling.	31
Tabela 8 – Correlação de Spearman entre Pressão Arterial Sistólica e as variáveis independentes.	32
Tabela 9 – Correlação de Spearman entre Pressão Arterial Diastólica e as variáveis independentes.	32
Tabela 10 – Modelos ajustados e desvios residuais.	34

Sumário

1	INTRODUÇÃO	11
2	FUNDAMENTAÇÃO TEÓRICA	12
2.1	Família exponencial uniparamétrica	12
2.2	Família exponencial multiparamétrica	12
2.2.1	Função Geradora de Momentos	13
2.2.2	Estatística Suficiente	13
2.3	Modelos Lineares Generalizados	14
2.3.1	Seleção do modelo	14
2.3.2	Componente Aleatório	15
2.3.3	Componente Sistemático	15
2.3.4	Função de Ligação	16
2.3.5	Ligação Canônica	16
2.4	Estimação dos parâmetros	17
2.4.1	Método da Máximo Verossimilhança	17
2.5	Teste de hipóteses	18
2.5.1	Região de Confiança	19
2.6	Seleção e Validação de Modelos	20
2.6.1	Modelo completo ou saturado	20
2.6.2	Modelo nulo	20
2.6.3	Modelo maximal	21
2.6.4	Modelo minimal	21
2.6.5	Modelo corrente	21
3	APLICAÇÕES	22
3.1	Material	22
3.2	Métodos	22
3.2.1	Distribuição Gama	22
3.2.2	Máxima Verossimilhança	22
3.3	Estimação dos parâmetros	23
3.3.1	Estimação de β	23
3.3.2	Estimação de ϕ	23
3.4	Teste de Normalidade de Anderson-Darling¹	24
3.5	Tipos de Resíduos	25
3.6	Estatísticas para diagnósticos	26

4	RESULTADOS	29
4.1	Análise Descritiva	29
4.2	Pressão Arterial	30
4.2.1	Teste de Normalidade de Anderson-Darling.	31
4.2.2	Correlações de Spearman entre PAS, PAD e as variáveis GLI, TG, HDL, CA.	31
4.3	Modelos ajustados e desvios residuais.	33
4.3.1	Gráficos para diagnósticos dos modelos com as variáveis PAS e PAD.	34
4.3.2	Gráficos de envelope simulado com as variáveis PAS e PAD.	35
5	CONCLUSÃO	38
	REFERÊNCIAS	39

1 Introdução

Os Modelos Lineares Generalizados (MLG) foram formulados por Nelder e Wedderburn (1972), que vieram assim unificar vários modelos estatísticos, incluindo a regressão linear, regressão logística e regressão de poisson, sobre um só marco teórico. Isto lhes permitiu formular um algoritmo geral para a estimativa de máxima verossimilhança em todos os modelos, que pertencem a família exponencial de distribuições.

De acordo com Turkman e Silva (2000), em muitos estudos estatísticos, quer sejam de natureza experimental ou observacional, nos deparamos com problemas em que o objetivo principal é o de estudar a relação entre variáveis, ou seja, analisar a influência que uma ou mais variáveis (explicativas), medidas em indivíduos ou objetos, tem sobre uma variável de interesse a que damos o nome de variável resposta. O modo como, em geral, o estatístico aborda tal problema é através do estudo de um modelo de regressão que relacione essa variável de interesse com as variáveis ditas explicativas.

Segundo Turkman e Silva (2000), a importância dos Modelos Lineares Generalizados não é apenas de índole prática. Do ponto de vista teórico a sua importância advém, essencialmente, do fato de a metodologia destes modelos constituir uma abordagem unificada de muitos procedimentos estatísticos correntemente usados nas aplicações e promover o papel central da verossimilhança na teoria da inferência.

Para avaliar os dados, selecionarmos as variáveis independentes e ajustarmos os modelos. A precisão de cada modelo ajustado foi avaliado usando a deviance residual. O modelo com menor deviance foi avaliado usando o gráfico para diagnóstico de pontos e o gráfico de envelope. Os envelopes, no caso de MLG's com distribuições diferentes da normal, são construídos com os resíduos sendo gerados a partir do modelo ajustado.

A pressão arterial é a pressão que o sangue exerce sobre a parede das arteriais: existe duas pressões, a máxima que é a sistólica e a mínima que é a diastólica. A pressão arterial sistólica quando o seu coração bate, ele contrai e bombeia sangue pelas arteriais para o resto do seu corpo. A pressão arterial diastólica quando o coração estar em repouso, entre uma batida e outro.

Dentre os fatos, devemos verificar a relação entre as variáveis independentes e as variáveis dependentes, avaliar o comportamento da pressão arterial e da frequência cardíaca de indivíduos hipertensos. Será ajustados modelos em que as variáveis dependentes segue uma distribuição gama, para verificar a eficiência da aplicação dos MLG's em dados com tal comportamento. Avaliar a precisão de cada modelo ajustado, usando a desvio residual, as análises de diagnóstico, as correlações de spearman e o teste de normalidade.

2 Fundamentação Teórica

Os Modelos Lineares Generalizados têm sido utilizados nas áreas de astronomia, agronomia, agricultura, saúde pública, ciências sociais, etc. Pela sua versatilidade resposta de modelo, quer seja de natureza contínua ou discreta, pode ter diversos comportamentos dependendo do tipo de dado que se deseja ajustar. Logo, os Modelos Lineares Generalizados são bastante utilizados para dados de natureza discreta de contagem ou proporção e dados de natureza contínua com comportamento simétrico e assimétricos.

2.1 Família exponencial uniparamétrica

De acordo com Cordeiro e Demétrio (2013), a família exponencial uniparamétrica é caracterizada por uma função de probabilidade ou uma função densidade, da forma:

$$f(x; \theta) = h(x) \exp[\eta(\theta)t(x) - b(\theta)], \quad (2.1)$$

em que as funções $\eta(\theta)$, $b(\theta)$, $t(x)$ e $h(x)$ assumem valores no subconjuntos dos reais. As funções $\eta(\theta)$, $b(\theta)$ e $t(x)$ não são únicas. Por exemplo, $\eta(\theta)$ pode ser multiplicada por uma constante k e $t(x)$ pode ser dividida pela mesma constante.

Várias distribuições importantes podem ser escritas na forma (2.1), tais como: Poisson, binomial, Rayleigh, normal, gama e normal inversa (as três últimas com a suposição de que um dos parâmetros é conhecido). A família exponencial na forma canônica é definida a partir de (2.1), considerando que as funções $\eta(\theta)$ e $t(x)$ são iguais á função identidade, de forma que

$$f(x; \theta) = h(x) \exp[\theta x - b(\theta)]. \quad (2.2)$$

Na parametrização (2.2), θ é chamado de parâmetro canônico. O logaritmo da função de verossimilhança correspondente a uma única observação no modelo (2.2).

2.2 Família exponencial multiparamétrica

Segundo Cordeiro e Demétrio (2013), a família exponencial multiparamétrica de dimensão k é caracterizada por uma função (de probabilidade ou densidade) da forma:

$$f(x; \boldsymbol{\theta}) = h(x) \exp \left[\sum_{i=1}^k \eta_i(\boldsymbol{\theta}) t_i(x) - b(\boldsymbol{\theta}) \right], \quad (2.3)$$

em que $\boldsymbol{\theta}$ é um vetor de parâmetros, usualmente, de dimensão k , e as funções $\eta_i(\boldsymbol{\theta})$, $b(\boldsymbol{\theta})$, $t_i(x)$ e $h(x)$ assumem valores em subconjuntos dos reais. Pelo teorema da fatoração, o

vetor $\mathbf{T} = [T_1(X), \dots, T_k(X)]^T$ é suficiente para o vetor de parâmetros $\boldsymbol{\theta}$. Quando $\eta_i(\boldsymbol{\theta}) = \theta_i$, $i = 1, \dots, k$, obtém-se de (2.3) a família exponencial na forma canônica com parâmetros canônicos $\theta_1, \dots, \theta_k$ e estatísticas canônicas $T_1(X), \dots, T_k(X)$. Tem-se

$$f(x; \boldsymbol{\theta}) = h(x) \exp \left[\sum_{i=1}^k \theta_i t_i(x) - b(\boldsymbol{\theta}) \right]. \quad (2.4)$$

Segundo Gelfand e Dalal (1990), estudaram a família exponencial biparamétrica $f(x; \boldsymbol{\theta}, \tau) = h(x) \exp[\boldsymbol{\theta}x + \tau t(x) - b(\boldsymbol{\theta}, \tau)]$ que é um caso especial de (2.3) com $k = 2$. Essa família tem despertado interesse, recentemente, como o componente aleatório dos modelos lineares generalizados superdispersos (Dey et al., 1997).

2.2.1 Função Geradora de Momentos

Segundo Cordeiro e Demétrio (2013), a função geradora de momentos (f.g.m.) da família exponencial é dada por:

$$M(t; \theta, \phi) = E(e^{ty}) = \exp\{\phi^{-1}[b(\phi t + \theta) - b(\theta)]\}. \quad (2.5)$$

Tabela 1 – Funções Geradoras de Momentos para algumas Distribuições.

Distribuições	Função geradora de momentos $M(t; \theta, \phi)$
Normal: $N(\mu, \sigma^2)$	$\exp\left(t\mu + \frac{\sigma^2 t^2}{2}\right)$
Poisson: $P(\mu)$	$\exp[\mu(e^t - 1)]$
Binomial: $B(m, \pi)$	$\left(\frac{m-\mu}{m} + \frac{\mu}{m} e^t\right)^m$
Bin. Negativa: $BN(\mu, k)$	$[1 + \mu/k(1 - e^t)]^{-k}$
Gama: $G(\mu, \nu)$	$(1 - \frac{t\mu}{\nu})^{-\nu}, \quad t < \frac{\nu}{\mu}$
Normal Inversa: $IG(\mu, \sigma^2)$	$\exp\left\{\frac{1}{\sigma^2}\left[\frac{1}{\mu} - \left(\frac{1}{\mu^2} - 2t\sigma^2\right)^{\frac{1}{2}}\right]\right\}, \quad t < \frac{1}{2\sigma^2\mu^2}$

Fonte: Cordeiro e Demétrio (2013)

essa relação não caracteriza a distribuição na família exponencial não-linear

$$\pi(y; \theta, \phi) = \exp\{\phi^{-1}[t(y)\theta - b(\theta)] + c(y, \phi)\}.$$

2.2.2 Estatística Suficiente

Conforme Cordeiro e Demétrio (2013), uma estatística $T = T(Y)$ é suficiente para um parâmetro θ (que pode ser um vetor) quando contém toda informação sobre esse parâmetro contida na amostra Y . Se T é suficiente para θ , então, a distribuição condicional de Y dada a estatística $T(Y)$ é independente de θ , isto é,

$$P(\mathbf{Y} = \mathbf{y} | T = t, \theta) = P(\mathbf{Y} = \mathbf{y} | T = t). \quad (2.6)$$

O critério da fatoração é uma forma conveniente de caracterizar uma estatística suficiente. Uma condição necessária e suficiente para T ser suficiente para um parâmetro θ é que a função densidade de probabilidade (f.d.p.) $f_{\mathbf{Y}}(\mathbf{y}, \theta)$ possa ser decomposta como:

$$f_{\mathbf{Y}}(\mathbf{y}, \theta) = h(\mathbf{y})g(t, \theta), \quad (2.7)$$

em que $t = T(\mathbf{y})$ e $h(\mathbf{y})$ não dependem de θ .

Seja Y_1, \dots, Y_n uma amostra aleatória (a.a.) de uma distribuição que pertence à família exponencial. A distribuição conjunta de Y_1, \dots, Y_n é dada por

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

Pelo teorema da fatoração de Neyman-Fisher e supondo ϕ conhecido, tem-se que $T = \sum_{i=1}^n y_i$ é uma estatística suficiente para θ , pois

$$f(\mathbf{y}; \theta, \phi) = g(t, \theta)h(y_1, \dots, y_n),$$

sendo que $g(t, \theta)$ depende de θ e dos y 's apenas através de t e $h(y_1, \dots, y_n)$ independe de θ . Isso mostra, que se uma distribuição pertence à família exponencial uniparamétrica, então, existe uma estatística suficiente. Na realidade, usando-se o Teorema de Lehmann-Scheffé (Mendenhall et al., 1981) mostra-se que $T = \sum_{i=1}^n y_i$ é uma estatística suficiente minimal.

2.3 Modelos Lineares Generalizados

2.3.1 Seleção do modelo

Conforme Cordeiro e Demétrio (2013), é difícil propor uma estratégia geral para o processo de escolha de um MLG a ser ajustado às observações que se dispõe. Em geral, o algoritmo de ajuste deve ser aplicado não a um MLG isolado, mas a vários modelos de um conjunto bem amplo que deve ser, realmente, relevante para o tipo de observações que se pretende analisar. Se o processo é aplicado a um único modelo, não levando em conta possíveis modelos alternativos, existe o risco de não se obter um dos modelos mais adequados aos dados. Um processo simples de seleção é de natureza sequencial, adicionando (ou eliminando) covariáveis (uma de cada vez) a partir de um modelo original até se obterem modelos adequados

Segundo Turkman e Silva (2000), existe três etapas essenciais que devemos dá prioridade ao tentar modelar dados através de um MLG:

1. **Formulação do modelo:** Escolher uma distribuição para a variável resposta. Para isso há necessidade de examinar cuidadosamente os dados; por exemplo, a distribuição gama e normal inversa são apropriadas para modelar dados de natureza contínua e que mostram assimetrias.
2. **Ajuste dos modelos:** Os modelos passam pela estimação dos parâmetros, isto é, pela estimação dos coeficientes β 's associados às covariáveis, e do parâmetro de dispersão ϕ caso ele esteja presente.
3. **Seleção e validação dos modelos:** tem por objetivo encontrar submodelos com um número moderado de parâmetros que ainda sejam adequados aos dados, detectar discrepâncias entre os dados e os valores preditos, averiguar a existência de outliers ou/e observações influentes, etc.

Esses modelos envolvem uma variável resposta univariada, variáveis explanatórias e uma amostra aleatória de n observações independentes, seja Y uma variável resposta ou dependente do modelo de interesse do experimento e considerando $x_{i1}, x_{i2}, \dots, x_{ip}$ o vetor coluna de variáveis independentes. McCullagh e Nelder (1989) definem os três elementos que compõe o modelo linear generalizado, São eles: Componente Aleatório, Componente Sistemático, e a Função de Ligação.

2.3.2 Componente Aleatório

Segundo Cordeiro e Demétrio (2013), o componente aleatório de um modelo linear generalizado é definido a partir da família exponencial uniparamétrica na forma canônica (2.2) com a introdução de um parâmetro $\phi > 0$ de perturbação. Nelder e Wedderburn (1972) ao fazerem isso, conseguiram incorporar distribuições biparamétricas no componente aleatório do modelo.

$$f(y; \theta, \phi) = \exp\{\phi^{-1}[y\theta - b(\theta)] + c(y, \phi)\}, \quad (2.8)$$

em que $b(\cdot)$ e $c(\cdot)$ são funções conhecidas. Quando é conhecido, a família de distribuições (2.8) é idêntica à família exponencial na forma canônica (2.2).

2.3.3 Componente Sistemático

Conforme Cordeiro e Demétrio (2013), as variáveis independentes produzem um preditor linear η dado por

$$\eta = \mathbf{X}\boldsymbol{\beta}, \quad (2.9)$$

em que $\mathbf{X} = (x_1, x_2, \dots, x_n)^T$ representa a matriz do modelo, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ o vetor de parâmetros desconhecidos e $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_n)$ o preditor linear.

2.3.4 Função de Ligação

Segundo Cordeiro e Demétrio (2013), uma função monótona e diferenciável cujo objetivo é relacionar a média ao preditor linear, estabelecendo uma relação entre o componente aleatório e o sistemático do modelo, logo:

$$\eta_i = g(\mu_i), \tag{2.10}$$

sendo $g(\cdot)$ uma função monótona e diferenciável. 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 utilizado.

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

Fonte : Cordeiro e Demétrio (2013)

2.3.5 Ligação Canônica

Segundo 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, o logaritmo da função de verossimilhança de um MLG com respostas independentes pode ser expresso na forma

$$L(\beta) = \sum_{i=1}^n \phi\{y_i \theta_i - b(\theta_i)\} + \sum_{i=1}^n c(y_i, \phi).$$

Um caso particular importante ocorre quando o parâmetro canônico (θ) coincide com o preditor linear, isto é, quando

$$\theta_i = \eta_i = \sum_{j=1}^p x_{ij} \beta_j.$$

Nesse caso, $L(\beta)$ fica dado por

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

Definindo a estatística $S_j = \phi \sum_{i=1}^n y_i x_{ij}$, $L(\beta)$ fica então reexpresso na forma

$$L(\beta) = \sum_{j=1}^p s_j \beta_j - \phi \sum_{i=1}^n b(\sum_{j=1}^p x_{ij} \beta_j) + \sum_{i=1}^n c(y_i, \phi).$$

Logo, pelo teorema da fatorização a estatística $S = (S_1, \dots, S_p)^T$ é suficiente minimal para o vetor $\beta = (\beta_1, \dots, \beta_p)^T$. As ligações que correspondem a tais estatísticas são chamadas de ligações canônicas e desempenham um papel importante na teoria dos MLG's. As ligações canônicas mais comuns são dadas abaixo.

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

Distribuição	Ligação
Normal	Identidade: $\eta = \mu$
Poisson	Logarítmica: $\eta = \ln(\mu)$
Gama	Recíproca: $\eta = \frac{1}{\mu}$
Normal Inversa	Recíproca do quadrado: $\eta = \frac{1}{\mu^2}$

Fonte :Cordeiro e Demétrio (2013)

De acordo Cordeiro e Demétrio (2013), deve ser lembrado, porém, que embora as funções de ligação canônicas conduzam a propriedades estatísticas desejáveis para o modelo, principalmente, no caso de amostras pequenas, não há nenhuma razão a priori para que os efeitos sistemáticos do modelo devam ser aditivos na escala dada por tais funções. Para o modelo clássico de regressão, a função de ligação canônica é a identidade, pois o preditor linear é igual à média. Essa função de ligação é adequada no sentido em que ambos, η e μ podem assumir valores na reta real.

2.4 Estimação dos parâmetros

Segundo Turkman e Silva (2000), num modelo linear generalizado o parâmetro β é o parâmetro de interesse, o qual é estimado pelo método da máxima verossimilhança. O parâmetro de dispersão ϕ , quando existe, é considerado um parâmetro perturbador, sendo a sua estimação feita pelo método dos momentos.

2.4.1 Método da Máximo Verossimilhança

Conforme Bolfarine e Sandoval (2001), o conceito de função de verossimilhança, enunciado a seguir, é central na teoria da verossimilhança.

Definição: Sejam X_1, \dots, X_n uma amostra aleatória de tamanho n da variável aleatória X com função de densidade (ou de probabilidade) $f(x|\theta)$, com $\theta \in \Theta$, onde Θ é o espaço paramétrico. A função de verossimilhança de θ correspondente à amostra aleatória observada é dada por

$$L(\theta; x) = \prod_{i=1}^n f(x_i|\theta).$$

Definição: O estimador de máxima verossimilhança de θ é o valor $\hat{\theta} \in \theta$ que maximiza a função de verossimilhança $L(\theta; x)$. O logaritmo natural da função de verossimilhança

de θ é denotado por

$$l(\theta; x) = \log L(\theta; x).$$

Não é difícil verificar que o valor de θ que maximiza a função de verossimilhança $L(\theta; x)$, também maximiza $l(\theta; x)$ dada pela equação acima.

2.5 Teste de hipóteses

Buse (1982) apresenta de uma forma bastante didática a interpretação geométrica dos Testes da Razão de Verossimilhanças, Escore e Wald para o caso de hipóteses simples. Apresentamos a seguir as generalizações para os MLG's. Vamos supor, inicialmente, a seguinte situação de hipóteses simples:

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

em que β^0 é um vetor p-dimensional conhecido e ϕ é também assumido conhecido.

- **Teste da Razão de Verossimilhanças**

De acordo com Paula (2013), o teste da Razão de Verossimilhanças, no caso de hipóteses simples, é usualmente definido por

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

Essa estatística pode também ser expressa, para os MLGs, como a diferença entre duas funções desvio

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

em que $\hat{\mu}^0 = g^{-1}(\hat{\eta}^0)$, $\hat{\eta}^0 = X\beta^0$. Em particular, para o caso normal linear, obtemos

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

- **Teste de Wald**

Segundo Paula (2013), o teste de Wald é definido, nesse caso, por

$$\xi_W = [\hat{\beta} - \beta^0]^T \hat{Var}^{-1}(\hat{\beta}) [\hat{\beta} - \beta^0],$$

em que $\hat{Var}(\hat{\beta})$ denota a matriz de variância-covariância assintótica de $\hat{\beta}$ estimada em $\hat{\beta}$. Para os MLGs, $\hat{Var}(\hat{\beta}) = K^{-1}(\hat{\beta})$. Assim, a estatística de Wald fica reexpressa na forma

$$\xi_W = \phi[\hat{\beta} - \beta^0]^T (X^T \hat{W} X) [\hat{\beta} - \beta^0].$$

Em particular, para o caso de $p = 1$, o teste de Wald é equivalente ao teste t^2 usual

$$\xi_W = (\hat{\beta} - \beta^0)^2 \hat{V}ar(\hat{\beta}).$$

Um problema com a estatística de Wald, especialmente quando $n(\beta)$ é não linear em β , é a dependência de ξ_W com a parametrização utilizada. Isto é, duas formas diferentes e equivalentes para $n(\beta)$, podem levar a diferentes valores de ξ_W .

- **Teste de Escore**

conforme Paula (2013), o teste de Escore, também conhecido como teste de Rao, é definido quando $U_\beta(\hat{\beta}) = 0$ por:

$$\xi_{SR} = U_\beta(\beta^0)^T \hat{V}ar_0(\hat{\beta}) U_\beta(\beta^0),$$

em que $\hat{V}ar_0(\hat{\beta})$ denota que a variância assintótica de $\hat{\beta}$ está sendo estimada sob H_0 . Para os MLGs temos que,

$$\xi_{SR} = \phi^{-1} U_\beta(\beta^0)^T (X^T \hat{W}_0 X)^{-1} U_\beta(\beta^0).$$

em que \hat{W}_0 é estimado sob H_0 , embora tenha a forma do modelo em H_1 . A estatística de escore pode ser muito conveniente em situações em que a hipótese alternativa é bem mais complicada do que a hipótese nula. Nesses casos, somente seria necessário estimarmos os parâmetros sob H_1 quando o modelo em H_0 fosse rejeitado.

2.5.1 Região de Confiança

De acordo com Cordeiro e Demétrio (2013), as regiões de confiança assintóticas para β_1 podem ser construídas usando-se qualquer uma das três estatísticas de teste. A partir da estatística da razão de verossimilhanças, uma região de confiança para β_1 , com um coeficiente de confiança de $100(1 - \alpha)\%$, inclui todos os valores de β_1 tais que:

$$2[l(\hat{\beta}_1, \hat{\beta}_2) - l(\beta_1, \tilde{\beta}_2)] < \chi_q^2, 1 - \alpha.$$

em que $\tilde{\beta}_2$ é a EMV de β_2 para cada valor de β_1 que é testado ser pertencente, ou não, ao intervalo, e $\chi_q^2, 1 - \alpha$ é o percentil da distribuição χ^2 com q graus de liberdade, correspondente a um nível de significância igual a $100\phi\%$.

Usando-se a estatística de Wald, uma região de confiança para β_1 , com um coeficiente de confiança de $100(1 - \alpha)\%$, inclui todos os valores de β_1 tais que:

$$(\hat{\beta}_1 - \beta_1)^T \hat{V}ar(\hat{\beta}_1)^{-1} (\hat{\beta}_1 - \beta_1) < \chi_q^2, 1 - \alpha.$$

Alternativamente, regiões de confiança para os parâmetros lineares β_1, \dots, β_p de um MLG podem ser construídos através da função desvio. Deseja-se uma região de confiança aproximada para um conjunto particular de parâmetros β_1, \dots, β_q de interesse.

2.6 Seleção e Validação de Modelos

De acordo com Turkman e Silva (2000), no estudo feito até aqui admitiu-se que o modelo proposto, em termos da combinação a distribuição da variável resposta e função de ligação era um modelo adequado. No entanto, quando se trabalha com muitas covariáveis, tem interesse de saber qual o modelo mais parcimonioso, ou seja, com o menor número de variáveis explicativas, que ofereça uma boa interpretação do problema posto e que ainda se ajuste bem aos dados.

O problema da seleção do modelo corresponde à procura do melhor modelo, no sentido de ser um modelo que atinge um bom equilíbrio. Sabendo que no processo de seleção há uma série de modelos em consideração, convém descrever vários que são comumente referidos durante o processo.

2.6.1 Modelo completo ou saturado

Segundo Turkman e Silva (2000), consideremos o modelo linear generalizado sem a estrutura linear $\eta = Z\beta$, isto é com n parâmetros μ_1, \dots, μ_n , linearmente independentes, sendo a matriz do modelo a matriz identidade $n \times n$. Este modelo atribui toda a variação dos dados à componente sistemática.

Como as estimativas de máxima verossimilhança dos μ_i são as próprias observações, isto é, $\hat{\mu}_i = y_i$, o modelo ajusta-se exatamente, reproduzindo os próprios dados. Não oferece qualquer simplificação e, como tal, não tem interesse na interpretação do problema, já que não faz sobressair características importantes transmitidas pelos dados. Além disso tem pouca hipótese de ser um modelo adequado em réplicas do estudo.

Conforme Cordeiro e Demétrio (2013), o modelo saturado ou completo que tem n parâmetros especificados pelas médias μ_1, \dots, μ_n linearmente independentes, ou seja, correspondendo a uma matriz modelo igual à matriz identidade de ordem n .

2.6.2 Modelo nulo

Conforme Turkman e Silva (2000), o modelo mais simples que se pode imaginar é o modelo com um único parâmetro. Corresponde a assumir que todas as variáveis Y_i têm o mesmo valor médio μ . É um modelo, de interpretação sem dúvida simples, mas que raramente captura a estrutura inerente aos dados. A matriz do modelo é, neste caso,

um vetor coluna unitário. Contrariamente ao modelo anterior, este modelo atribui toda a variação nos dados à componente aleatória.

De acordo com Cordeiro e Demétrio (2013), na prática, o modelo nulo é muito simples e o saturado é não-informativo, pois não sumariza os dados, mas, simplesmente, os repete. Existem dois outros modelos, não tão extremos, quanto os modelos nulo e saturado.

2.6.3 Modelo maximal

De acordo com Turkman e Silva (2000), o modelo maximal é o modelo que contém o maior número de parâmetros, e portanto, o mais complexo, que estamos preparados a considerar.

2.6.4 Modelo minimal

Segundo Turkman e Silva (2000), contrariamente ao modelo maximal, o modelo minimal é o modelo mais simples, com o menor número de parâmetros, que ainda se ajusta adequadamente aos dados. Este modelo embora adaptando-se aos dados e podendo até ser adequado para réplicas do estudo, pode esconder características ainda importantes dos dados.

2.6.5 Modelo corrente

De acordo com Turkman e Silva (2000), em geral trabalha-se com modelos encaixados, ou seja, passasse do modelo maximal para o modelo minimal por exclusão de termos da desvio. O modelo corrente, é qualquer modelo com que parâmetros linearmente independentes situado entre o modelo maximal e o modelo minimal, e que está a ser sujeito a investigação.

3 Aplicações

3.1 Material

Os dados utilizados são referentes à função pulmonar e síndrome metabólica de adolescentes escolares no município de Campina Grande-PB e foram disponibilizados por Vânia (2016). A hipertensão é conceituada como uma síndrome caracterizada por valores pressóricos permanentemente elevados, associados a alterações metabólicas e hormonais, e a fenômenos tróficos como hipertrofia cardíaca e vascular (1, 2). A hipertensão arterial é uma das formas de doença cardiovascular das mais prevalentes, constituindo-se em um problema de saúde pública no Brasil e no mundo, sendo a sua prevalência no Brasil. A distribuição gama é associada a dados contínuos assimétricos.

3.2 Métodos

3.2.1 Distribuição Gama

Função de Densidade de Probabilidade

Segundo Paula (2013), a distribuição gama é associada a dados contínuos assimétricos. A distribuição gama de Y , $G(\mu, \phi)$, com parâmetros positivos μ e ϕ , é dada por:

$$f(y; \mu, \phi) = \frac{\left(\frac{\phi}{\mu}\right)^\phi}{\Gamma(\phi)} y^{\phi-1} \exp\left(-\frac{\phi y}{\mu}\right), \quad y > 0 \quad (3.1)$$

sendo a média μ e o coeficiente de variação igual a $\sqrt{\phi}$, Clarice e Demetrio (1989). Essa distribuição pode ser escrita na forma

$$\exp[\phi\{-y/\mu\} - \log \mu] - \log \Gamma(\phi) + \log(\phi y) - \log y],$$

em que $y > 0$, $\phi > 0$ e $\Gamma(\phi) = \int_0^\infty t^{\phi-1} e^{-t} dt$ é a função gama. A ligação usada, neste caso, foi a logarítmica ($\log \mu_i = \eta_i$).

A ligação logarítmica tem um atrativo especial de possibilitar o desenvolvimento de experimentos ortogonais como são bem conhecidos em modelos de regressão normal linear.

3.2.2 Máxima Verossimilhança

Conforme Bolfarine e Sandoval (2001), se a distribuição da variável aleatória X pertence à família exponencial unidimensional de distribuições, então o estimador de máxima verossimilhança de θ baseado na amostra $X = (X_1, \dots, X_n)$ é solução da equação

$E[T(X)] = T(X)$, desde que a solução pertença ao espaço paramétrico correspondente ao parâmetro θ . Esse resultado pode ser estendido para o caso k-paramétrico em que os estimadores de máxima verossimilhança de $\theta_1, \dots, \theta_k$ seguem como soluções das equações $E[T_j(X)] = T_j(X)$, $j = 1, \dots, k$.

3.3 Estimação dos parâmetros

3.3.1 Estimação de β

Conforme Paula (2013), O processo iterativo de Newton-Raphson para a obtenção da estimativa de máxima verossimilhança de β é definido expandindo a função escore U_β em torno de um valor inicial $\beta^{(0)}$, tal que:

$$U_\beta \cong U_\beta^{(0)} + U_\beta^{\prime(0)}(\beta - \beta^{(0)})$$

Em que U_β' denota a primeira derivada de U_β com respeito a β^T , sendo $U_{\beta'}(0)$ e $U_\beta^{(0)}$, respectivamente, essas quantidades avaliadas em $\beta^{(0)}$. Assim, repetindo o procedimento acima, chegamos ao processo iterativo.

$$\beta^{(m+1)} = \beta^{(m)} + \{(-U_\beta')^{-1}\}^{(m)} U_\beta^{(m)}$$

Sendo $m = 0, 1, \dots$. Como a matriz $-U_\beta'$ pode não ser positiva definida, a aplicação do método escore de Fisher substituindo a matriz $-U_\beta'$ pelo correspondente valor esperado $K_{\beta\beta}$ pode ser mais conveniente. Isso resulta no seguinte processo iterativo:

$$\beta^{(m+1)} = \beta^{(m)} + \{K_{\beta\beta}^{-1}\}^{(m)} U_\beta^{(m)}$$

Sendo $m = 0, \dots$. Se trabalharmos um pouco o lado direito da expressão acima, chegaremos a um processo iterativo de mínimos quadrados ponderados.

3.3.2 Estimação de ϕ

Conforme Paula (2013), Igualando a função escore U_ϕ a zero chegamos à seguinte solução:

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

em que $D(y; \hat{\mu})$ denota o desvio do modelo sob investigação. Verificamos que a estimativa de máxima verossimilhança para ϕ nos casos normal e normal inversa, igualando U_ϕ a zero, é dada por:

$$\hat{\phi} = \frac{n}{D(y; \hat{\mu})}.$$

Para o caso gama, a estimativa de máxima verossimilhança de ϕ sai da equação

$$2n\{\log\hat{\phi} - \psi(\hat{\phi})\} = D(y; \hat{\mu}).$$

A equação acima pode ser resolvida diretamente pelo R através do comando `require(MASS)`, Venables e Ripley (1999). Como ilustração, vamos supor que os resultados do ajuste sejam guardados em `fit.model`. Então, para encontrarmos a estimativa de máxima verossimilhança de ϕ com o respectivo erro padrão aproximado devemos usar os comandos, `require(MASS)`, `gamma.shape(fit.model)`.

Um outro estimador consistente para ϕ (de momentos) que não envolve processo iterativo é baseado na estatística de Pearson, sendo dado por:

$$\hat{\phi} = \frac{(n-p)}{\sum_{i=1}^n \left\{ \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu})} \right\}}$$

A suposição aqui é que $\hat{\beta}$ tem sido consistentemente estimado. O R solta a estimativa $\hat{\phi} = (n-p)/D(y; \hat{\mu})$ que não é consistente para ϕ .

3.4 Teste de Normalidade de Anderson-Darling¹

A estatística de Anderson-Darling (AD) mede o quão bem os dados seguem uma distribuição normal. Em geral, quanto melhor distribuição se ajusta aos dados, menor a estatística (AD).

A estatística (AD) é usada para calcular o valor de p para o teste de qualidade do ajuste, que ajuda a determinar o que melhor se adapta a distribuição de seus dados. Por exemplo, a estatística (AD) é calculada para cada distribuição, quando você realizar Identificação de distribuição individual. Os valores de p calculados a partir da estatística ajudam a determinar qual modelo de distribuição deverá ser usado para uma análise de capacidade ou uma análise de confiabilidade. A estatística (AD) também é utilizada para testar se uma amostra de dados é proveniente de uma população com uma distribuição especificada. Por exemplo, você pode precisar testar se os seus dados atendem à suposição de normalidade para um teste t. As hipóteses para o teste Anderson-Darling são:

H0: os dados seguem uma distribuição normal.

H1: os dados não seguem uma distribuição normal.

Se o valor-p do teste de Anderson-Darling for menor do que o nível de significância escolhido (normalmente 0,05 ou 0,10), conclui que os dados não seguem a distribuição normal, para o teste de Anderson-Darling porque ele não existe matematicamente para determinados casos.

Se você estiver comparando o ajuste de diversas distribuições, a distribuição com o maior valor-p, normalmente, tem o ajuste mais próximo aos dados. Se as distribuições tiverem valores-p similares, escolha uma das distribuições com base no conhecimento prático.

Alguns comandos geram uma estatística Anderson-Darling ajustada ou (AD). A estatística Anderson-Darling não ajustada usa a função da etapa não-paramétrica baseada no método de Kaplan-Meier de cálculo pontos do gráfico, enquanto a estatística de Anderson-Darling ajustada usa outros métodos para calcular os pontos do gráfico.

1 : <https://support.minitab.com/pt.../anderson-darling-and-distribution-fit/> (Como a estatística de Anderson-Darling é usada para avaliar o ajuste)

3.5 Tipos de Resíduos

Segundo Cordeiro e Demétrio (2013), vale destacar que os resíduos têm papel fundamental na verificação do ajuste de um modelo. Vários tipos de resíduos foram propostos na literatura (Cook e Weisberg, 1982; Atkinson, 1985).

a) Resíduos Ordinários

Os resíduos do processo de ajuste por mínimos quadrados são dados por:

$$r_i = y_i - \hat{\mu}_i.$$

Enquanto os erros ϵ_i 's são independentes e têm a mesma variância, o mesmo não ocorre com os resíduos obtidos a partir do ajuste do modelo através de mínimos quadrados. Tem-se,

$$Var(r) = Var[(I - H)Y] = (I - H)\sigma^2.$$

Assim, o uso dos resíduos ordinários pode não ser adequado devido à heterogeneidade de variâncias. Foram, então, propostas diferentes padronizações para minimizar esse problema.

b) Resíduos Estudentizados Internamente

Considerando-se $s^2 = QMRes$ como a estimativa de σ^2 , tem-se que um estimador não tendencioso para $Var(r_i)$ é dado por

$$\hat{Var}(r_i) = (1 - h_{ii})\sigma^2 = (1 - h_{ii})QMRes$$

e como $E(r_i) = E(Y_i - \hat{\mu}_i) = 0$, então, o resíduo estudentizado internamente é dado por

$$rsi_i = \frac{r_i}{s\sqrt{(1 - h_{ii})}} = \frac{Y_i - \hat{\mu}_i}{\sqrt{(1 - h_{ii})QMRes}}.$$

Esses resíduos são mais sensíveis do que os anteriores por considerarem variâncias distintas. Entretanto, um valor discrepante pode alterar profundamente a variância residual dependendo do modo como se afasta do grupo maior das observações.

Além disso, o numerador e o denominador dessa expressão são variáveis dependentes, isto é, $Cov(r, QMRes) \neq 0$.

c) Resíduos Estudentizados Externamente

Para garantir a independência do numerador e denominador na padronização dos resíduos, define-se o resíduo estudentizado externamente, como

$$rse_{(i)} = \frac{ri}{s_{(i)}\sqrt{(1 - h_{ii})}}$$

sendo $s_{(i)}^2$ o quadrado médio residual livre da influência da observação i , ou seja, a estimativa de σ^2 , omitindo-se a observação i . Prova-se que

$$rse_{(i)} = rsi_i \sqrt{\frac{n - p - 1}{(n - p - rsi_i^2)}}.$$

A vantagem de usar o resíduo $rse_{(i)}$ é que, sob normalidade, tem distribuição t de Student com $(n-p-1)$ graus de liberdade. Embora não seja recomendada a prática de testes de significância na análise de resíduos, sugere-se que a i -ésima observação seja merecedora de atenção especial se $|rse_{(i)}|$ for maior do que o $100[1 - \alpha/(2n)]$ -ésimo percentil da distribuição t com $(n-p-1)$ graus de liberdade, sendo que o nível de significância α é dividido por n por ser esse o número de pontos sob análise.

3.6 Estatísticas para diagnósticos

Segundo Cordeiro e Demétrio (2013), discrepâncias isoladas (pontos atípicos) podem ser caracterizadas por terem h_{ii} e/ou resíduos grandes, serem inconsistentes e/ou serem influentes (McCullagh e Nelder, 1989). Uma observação inconsistente é aquela que destoa da tendência geral das demais.

Quando uma observação está distante das outras em termos das variáveis explanatórias ela pode ser, ou não, influente. Uma observação influente é aquela cuja omissão do conjunto de dados resulta em mudanças substanciais em certos aspectos do modelo. Essa observação pode ser um "outlier" (observação aberrante), ou não. Uma observação pode ser influente de diversas maneiras, isto é,

- no ajuste geral do modelo;
- no conjunto de estimativas dos parâmetros;
- na estimativa de um determinado parâmetro;

- na escolha de uma transformação da variável resposta ou de uma variável explanatória.

As estatísticas mais utilizadas para verificar de pontos atípicos são:

- Medida de "leverage": h_{ii} ;
- Medida de inconsistência: $rse_{(i)}$;
- Medida de influência sobre o parâmetro β_j : $DFBeta_{S(i)}$ para β_j ;
- Medidas de influência geral: $DFFit_{S(i)}$, $D_{(i)}$ ou $C_{(i)}$.

A seguir são descritas as estatísticas citadas.

a) Medida de "leverage"

A distância de uma observação em relação às demais é medida por h_{ii} (medida de "leverage"). No caso particular da regressão linear simples, usando-se a variável centrada $x_i = X_i - \bar{X}$,

e, portanto, o que mostra que medida que X se afasta de \bar{X} o valor de h_{ii} aumenta e que seu valor mínimo, é $1/n$. Esse valor mínimo ocorre para todos os modelos que incluem uma constante. No caso em que o modelo de regressão passa pela origem, o valor mínimo de h_{ii} é 0 para uma observação $X_i = 0$. O valor máximo de h_{ii} é 1, ocorrendo quando o modelo ajustado é irrelevante para a predição em X_i e o resíduo é igual a 0.

b) Medida de inconsistência

Ponto inconsistente: ponto com $rse_{(i)}$ grande, isto é, tal que $|rse_{(i)}| \geq t_{\gamma/(2n);n-p-1}$, com nível de significância igual a $100\gamma\%$;

c) Medida de influência

Essas estatísticas são importantes quando o coeficiente de regressão tem um significado prático. $DFBeta_{(i)}$ mede a alteração no vetor estimado $\hat{\beta}$ ao se retirar o i -ésimo ponto da análise, isto é,

$$DFBeta_{(i)} = \hat{\beta} - \hat{\beta}_{(i)} = \frac{r_i}{(1 - h_{ii})} (X^T X)^{-1} x_i.$$

ou ainda, considerando que $\hat{\beta} = (X^T X)^{-1} X^T Y = CY$ em que $C = (X^T X)^{-1} X^T$,

$$DFBeta_{(i)} = \frac{r_i}{(1 - h_{ii})} (X^T X)^{-1} x_i c_i^T, i = 1, \dots, n,$$

sendo c_i^T a i -ésima linha de C . Então,

$$DFBeta_{(i)} = \frac{r_i}{(1 - h_{ii})} (X^T X)^{-1} x_i c_{ji}, i = 1, \dots, n, j = 0, \dots, p - 1.$$

d) Medidas de influência geral

A estatística DFFit e sua versão estudentizada DFFitS medem a alteração provocada no valor ajustado pela retirada da observação i . São dadas por

$$DFFit_{(i)} = x_i^T (\hat{\beta} - \hat{\beta}_{(i)}) = \hat{y} - \hat{y}_{(i)}$$

sendo o quociente $\frac{h_{ii}}{1-h_{ii}}$, chamado potencial de influência, uma medida da distância do ponto x_i em relação às demais observações. Nota-se que DFFitS pode ser grande porque h_{ii} é grande ou porque o resíduo estudentizado externamente é grande.

4 Resultados

Para ajustar os modelos foram avaliados 576 indivíduos, dos quais foram coletadas a Pressão Arterial Diastólica (PAD), Pressão Arterial Sistólica (PAS), o nível de Glicose (GLI), o nível de Colesterol (HDL), a medida da Circunferência Abdominal (CA) e o nível de Triglicérides (TG). Os dados utilizados são referentes à função pulmonar e síndrome metabólica de adolescentes escolares no município de Campina Grande-PB e foram disponibilizados por Vânia (2016).

4.1 Análise Descritiva

Os resultados aqui obtidos, foram calculados usando o software R (R core Team, 2017). Na Tabela 4 e 5 foi realizada uma análise descritiva das variáveis independentes GLI, TG, HDL, CA e das variáveis dependentes PAS e PAD, quanto às variáveis assimétricas devemos destacar TG com assimetria igual a 2,59 e a variável CA com assimetria igual a 1,55. Já na Tabela 7 temos o teste de normalidade de Anderson-Darling, já nas Tabelas 8 e 9 temos as correlações das variáveis independentes GLI, TG, HDL, CA com as variáveis dependentes Pressão Arterial Sistólica (PAS) e Pressão Arterial Diastólica (PAD), usando o método de spearman.

Tabela 4 – Análise descritiva das variáveis independentes, GLI, TG, HDL, CA e variáveis dependentes, PAS, PAD.

Variável	Média	Variância	Mediana	Mínimo	Máximo
PAS	109,93	106,92	109,00	86,50	143,50
PAD	66,95	49,60	66,50	48,50	93,50
GLI	76,006	50,57	76,00	54,00	98,00
TG	82,46	1558,65	74,00	30,00	423,00
HDL	41,78	73,14	41,00	16,00	81,00
CA	71,43	78,25	69,55	56,00	116,00

Tabela 5 – Análise descritiva das variáveis independentes, GLI, TG, HDL, CA e variáveis dependentes, PAS, PAD.

Variável	Assimetria	Curtose
PAS	0,42	-0,031
PAD	0,31	015
GLI	0,17	-0046
TG	2,59	12,57
HDL	0,65	1,22
CA	1,55	3,37

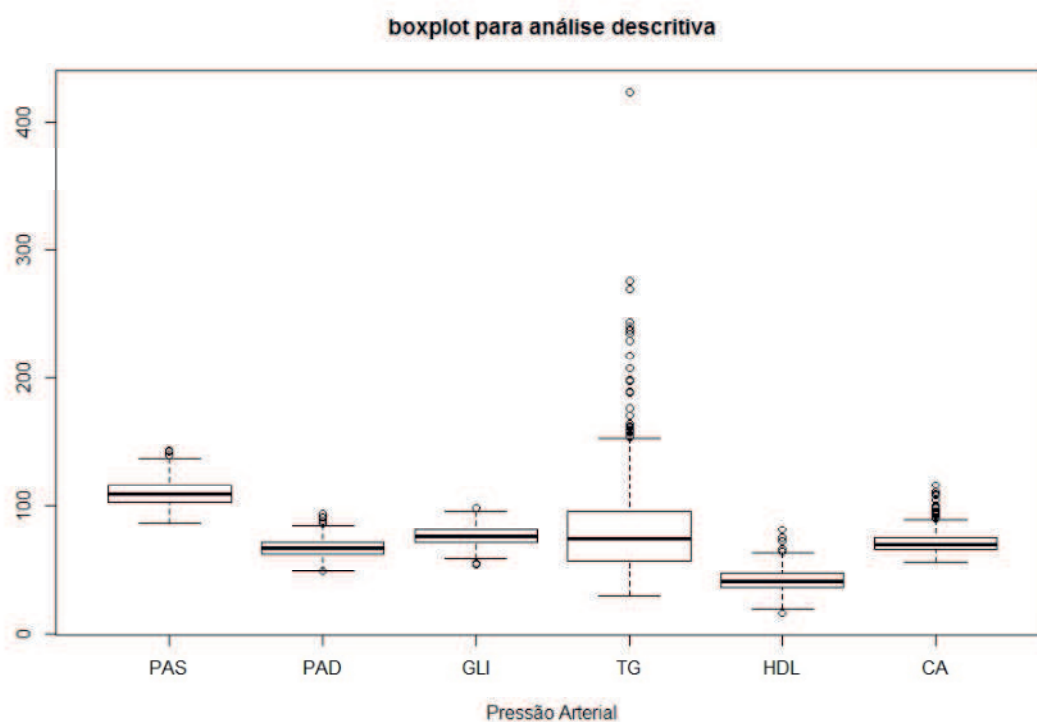


Figura 1 – Gráfico boxplot das variáveis independentes e dependente.

4.2 Pressão Arterial

A pressão arterial é a pressão que o sangue exerce na parede das artérias. Quando o seu coração bate, ele contrai e bombeia sangue pelas artérias para o resto do seu corpo, chamamos de pressão arterial sistólica, já quando o coração está em repouso, entre uma batida e outra chamamos de pressão arterial diastólica. Ela é medida em milímetros de mercúrio. Com essa medida, são determinadas duas pressões:

- i) **A máxima:** Quando o coração se contrai, temos uma pressão máxima (sistólica).
- ii) **A mínima:** Quando ele se dilata, temos uma pressão mínima (diastólica).

De acordo com a American Heart Association¹ (Associação Americana do Coração) e a American Stroke Association¹ (Associação Americana do AVC), a hipertensão é uma doença grave que precisa de tratamento e orientação personalizados. Algumas condições específicas de cada indivíduo podem fazer com que uma pequena variação nos números já mereça atenção médica. Portanto, siga as orientações do cardiologista com disciplina.

Pode-se observar por meio da Tabela 6 as diversas categorias de Pressão Sanguínea Sistólica e Diastólica, variando de normal à crise hipertensiva.

Tabela 6 – Categorias de Pressão Sanguínea Sistólica e Diastólica.

Categoria	<i>PS*</i> (Nº mais Alto)		<i>PD**</i> (Nº mais Baixo)
Normal	Menor que 120	e	Menor que 80
Elevada	120-129	e	Menor que 80
Hipertensão Estágio 1	130-139	ou	80-89
Hipertensão Estágio 2	140 ou maior	ou	90 ou maior
Crise hipertensiva	180 ou maior	e/ou	Maior que 120

*Pressão Sistólica=PS

**Pressão Diastólica=PD

1 :<https://drauziovarella.uol.com.br/.../saiba-como-interpretar-sua-medida-de-pressao-art...> (Saiba como interpretar os números da sua pressão arterial | Portal ...)

4.2.1 Teste de Normalidade de Anderson-Darling.

Para conhecer o teste de normalidade, conforme apresentado na Tabela 7, o resultado encontrado mostra que as variáveis não seguem um comportamento normal ao nível de 5% de significância, veja que os p-valores são todos menos que 0,05.

Tabela 7 – Valores para o teste de normalidade de Anderson-Darling.

Variável	Estatística	valor-p
PAS	2,1366	0,001
PAD	1,1452	0,005
GLI	1,465	0,001
TG	22,352	0,001
HDL	2,7985	0,001
CA	15,805	0,001

A análise dos dados para o teste de normalidade apresentou estatística de Anderson-Darling de 22,352 para o nível de Triglicérides (TG) e de 15,805 para medida da Circunferência Abdominal (CA) e valor-p menores que 0,05, rejeitando-se a hipótese de normalidade dos resíduos. Nestes resultados, a hipótese nula afirma que os dados não seguem uma distribuição normal. Como os valores de p são menores que o nível de significância de 0,05, então rejeitamos a hipótese nula, assim concluímos que os dados não seguem uma distribuição normal.

4.2.2 Correlações de Spearman entre PAS, PAD e as variáveis GLI, TG, HDL, CA.

Podemos notar pela análise das Tabelas abaixo, a relação linear entre as variáveis, os coeficientes apresentados a seguir nos auxiliam na quantificação do grau de relacionamento entre as variáveis de interesse. Na Tabela 8 temos a correlação das variáveis independentes

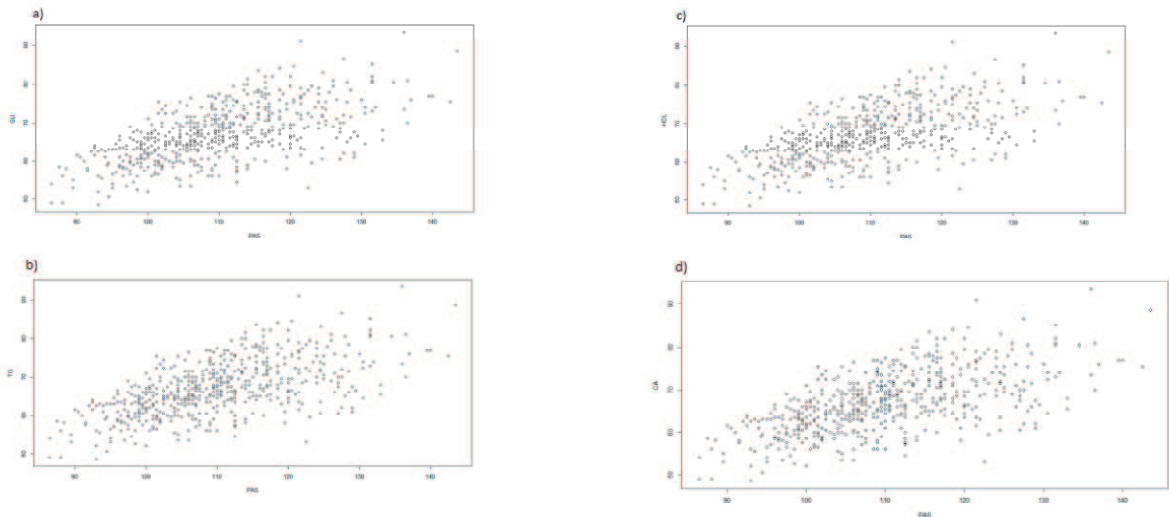


Figura 2 – Diagrama de correlação entre PAS e TG, GLI, HDL, CA.

GLI, TG, HDL, CA com a variável dependente Pressão Arterial Sistólica (PAS), usando método de spearman destacamos a variável medida da Circunferência Abdominal (CA) com correlação 0,412.

Tabela 8 – Correlação de Spearman entre Pressão Arterial Sistólica e as variáveis independentes.

Variável	Correlação	valor-p
GLI	0,120	0,004
TG	0,143	0,001
HDL	-0,159	0,001
CA	0,412	0,001

Nas Figuras a), b), c) e d), há uma dispersão dos dados muito forte, portanto quanto maior a correlação entre as variáveis, maior será a proximidade dos pontos, ou seja, estarão menos dispersos. Cada ponto no diagrama de dispersão corresponde às medidas de pressão sistólica.

Tabela 9 – Correlação de Spearman entre Pressão Arterial Diastólica e as variáveis independentes.

Variável	Correlação	valor-p
GLI	0,051	0,223
TG	0,222	0,001
HDL	-0,066	0,112
CA	0,208	0,001

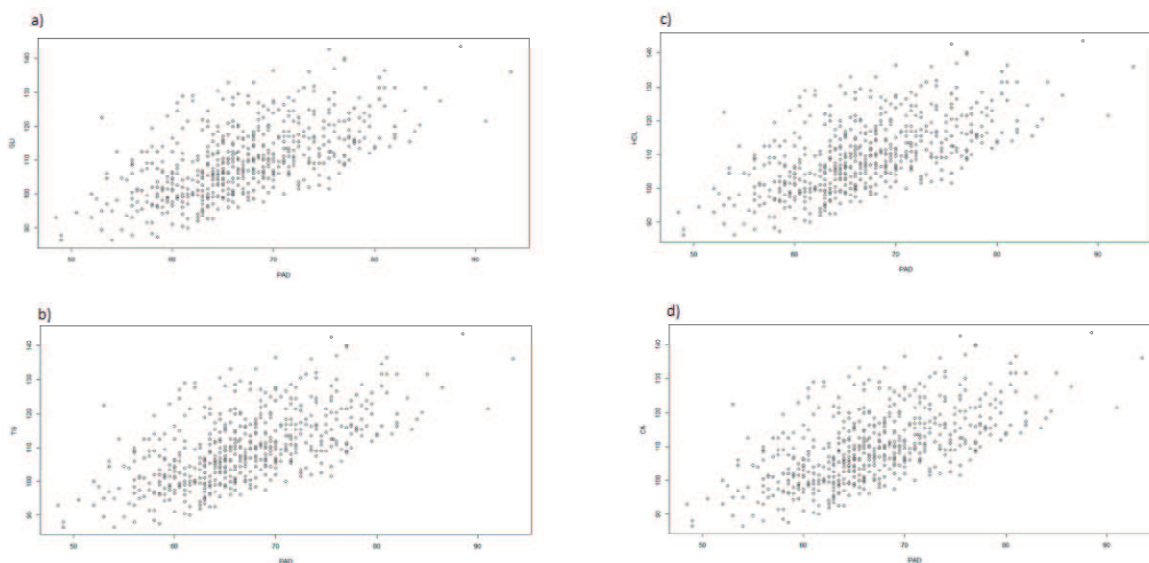


Figura 3 – Diagrama de correlação entre PAD e TG, GLI, HDL, CA.

Para a variável dependente Pressão Arterial Diastólica (PAD) a correlação é vista na Tabela 9. Observe que o nível de Triglicerídeos (TG) e a medida da Circunferência Abdominal (CA) tem relação positiva, embora não seja forte, com a variável Pressão Arterial Diastólica (PAD). A correlação com as outras variáveis podemos desprezar dado que têm valores quase zero.

Através da correlação de Spearman, o nível de Triglicerídeos (TG) e a medida da Circunferência Abdominal (CA), apresentaram uma correlação positiva, porém com significância estatística ($p = 0,001$), pra duas variáveis.

Observando as Figuras a), b), c) e d), observa-se que há uma dispersão dos dados muito fraca, portanto quanto menor a correlação entre as variáveis, mais dispersos estarão os pontos. Cada ponto no diagrama de dispersão corresponde às medidas de pressão diastólica. Podemos notar que, conforme aumenta o nível da Pressão Arterial Sistólica (PAS), o nível da Pressão Arterial Diastólica (PAD) tende a aumentar, nota-se também uma tendência linear.

4.3 Modelos ajustados e desvios residuais.

Foram ajustados dois modelos, o primeiro modelo com variável dependente Pressão Arterial Diastólica (PAD), o segundo modelo com variável dependente Pressão Arterial Sistólica (PAS). Ambos, inicialmente, em função das variáveis independentes, o nível de Glicose (GLI), o nível de Triglicerídeos (TG), o nível de Colesterol (HDL) e a medida da Circunferência Abdominal (CA).

Após o ajuste do modelo saturado foi observado a significância das variáveis. A variável não significativa foi retirada até encontramos um modelo mais que nos explique o comportamento da Pressão Arterial Diastólica (PAD) e da Pressão Arterial Sistólica (PAS) e de melhor interpretação e que melhor se ajuste aos dados em análise, os modelos ajustados estão descritos na Tabela 10.

Tabela 10 – Modelos ajustados e desvios residuais.

Modelo com variável dependente PAD				
Coefficientes	Estimativa	Erro Padrão	Pr(> z)	Desvio Residual
Intercepto	3,994	0,034	0,001 ***	5,797
TG	0,001	0,001	0,001 ***	
CA	0,002	0,0005	0,001 ***	
Modelo com variável dependente PAS				
Coefficientes	Estimativa	Erro Padrão	Pr(> z)	Desvio Residual
Intercepto	4,304	0,045	0,001 ***	4,129
GLI	0,001	0,001	0,013 *	
CA	0,004	0,005	0,001 ***	

Para o modelo com variável dependente Pressão Arterial Diastólica (PAD) as variáveis significativas foram o nível de Triglicérides (TG) e a medida da Circunferência Abdominal (CA). O modelo fica da forma $\log(\hat{\mu}_{PAD}) = 3,994 + 0,001TG + 0,002CA$ com desvio de 5,797. O segundo modelo é da forma $\log(\hat{\mu}_{PAS}) = 4,304 + 0,001GLI + 0,004CA$ com desvio de 4,129.

Pela análise dos dados, é possível perceber que o grupo de indivíduos hipertensos apresenta uma Pressão Arterial Sistólica (PAS) significativamente maior que a dos normotensos, comprovando estatisticamente o diagnóstico de hipertensão arterial. E ambos os casos com o erro padrão do coeficiente (TG) e (GLI) é menor do que aquele em (CA). Portanto, seu modelo foi capaz de estimar o coeficiente de (CA) com maior precisão.

4.3.1 Gráficos para diagnósticos dos modelos com as variáveis PAS e PAD.

A análise de diagnóstico é uma importante ferramenta para verificar a adequabilidade do modelo, observando se as suposições sob os modelos são satisfeitas, assim como a presença de pontos atípicos que possam influenciar nos resultados do ajuste. Desta forma Pregibon (1981), desenvolveu algumas medidas de diagnóstico para detectar pontos atípicos em MLG's, em especial o componente do desvio como resíduo. Possamai (2009) fez uma revisão dos modelos com a variável resposta pertencendo a família exponencial e fazendo aplicações com a distribuição gama e entre outras distribuições. segundo Paula (2013), as técnicas gráficas são basicamente iguais às que foram descritas para o modelo linear clássico, sendo os gráficos mais recomendadas para os MLG's.

Inicialmente, realizamos a análise de resíduos para detectar possíveis pontos extremos e avaliar a adequação da distribuição proposta para a variável resposta. Assim como

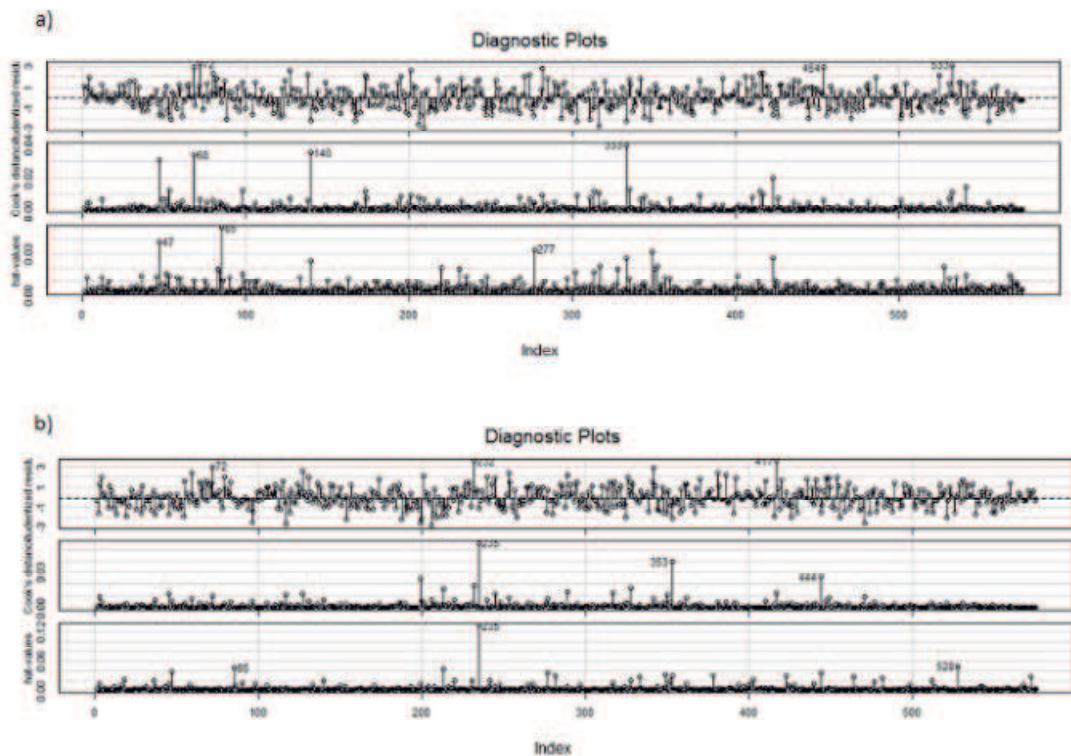


Figura 4 – Gráfico para diagnóstico residual do modelo com a variável PAS e PAD

no modelo clássico de regressão, as técnicas usadas para análise de resíduos e diagnóstico para os modelos lineares generalizados são semelhantes, com algumas adaptações, devido à estruturas MLG's.

Ao ajustarmos um modelo a um conjunto de dados, uma etapa muito importante é a verificação de possíveis afastamentos das suposições do modelo, levando-se em consideração a parte aleatória e sistemática do modelo, da mesma forma que verificamos a presença de observações com alguma influência fora de padrão nos resultados do ajuste.

Nas Figuras 4a e 4b estão os gráficos para diagnósticos dos modelos. Conforme podemos ver, na Figura 1a temos o diagnóstico de pontos para o modelo com variável dependente Pressão Arterial Diastólica (PAD), observe que apesar de termos pontos de alavanca não foi detectados outliers. Analisando as Figuras 4a e 4b, observe que não existe pontos influentes que possam alterar significativamente o modelo ajustado. Apesar dos resíduos estudentizados das observações não se mostrarem grandes, não possuem valores alto suficiente para distorcer o modelo logístico ajustado. Assim podemos concluir que não há pontos de alavanca e também não foi detectado nem um outliers.

4.3.2 Gráficos de envelope simulado com as variáveis PAS e PAD.

De acordo com Paula (2013), os gráficos normais de probabilidades com envelope destacam-se em dois aspectos: a identificação da distribuição originária dos dados e a identificação de valores que se destacam no conjunto de observações.

Os envelopes, no caso dos MLG's com distribuições diferentes da normal, são construídos com os resíduos sendo gerados a partir do modelo ajustado. O gráfico de resíduos simulados permite checar a adequação do modelo ainda que os resíduos não tenham uma aproximação adequada com a distribuição Normal.

Nesse tipo de gráfico, o padrão esperado, para um modelo bem ajustado, corresponde aos pontos (resíduos) dispersos aleatoriamente entre os limites do envelope. A presença de pontos externos ao envelope ou de pontos internos apresentando padrões sistemáticos podem indicar problemas de ajuste.

Pelo gráfico de envelope simulado, nas Figuras 5a e 5b, não há evidências de que o modelo esteja mal ajustado. Isso quer dizer que ao nível de 5% de significância o modelo é válido. Como podemos observar na Figura 5a e na Figura 5b as dispersões dos pontos, assim dizemos que os modelos foram bons, os envelopes deu menos de 5% dos pontos fora. Repare que, no geral, os resíduos estão dispersos no interior do envelope.

Alguns pontos encontram-se acima do limite superior. O modelo se ajustou muito bem aos dados, pois todos os pontos (resíduos) estão dentro ou sobre as bandas de confiança.

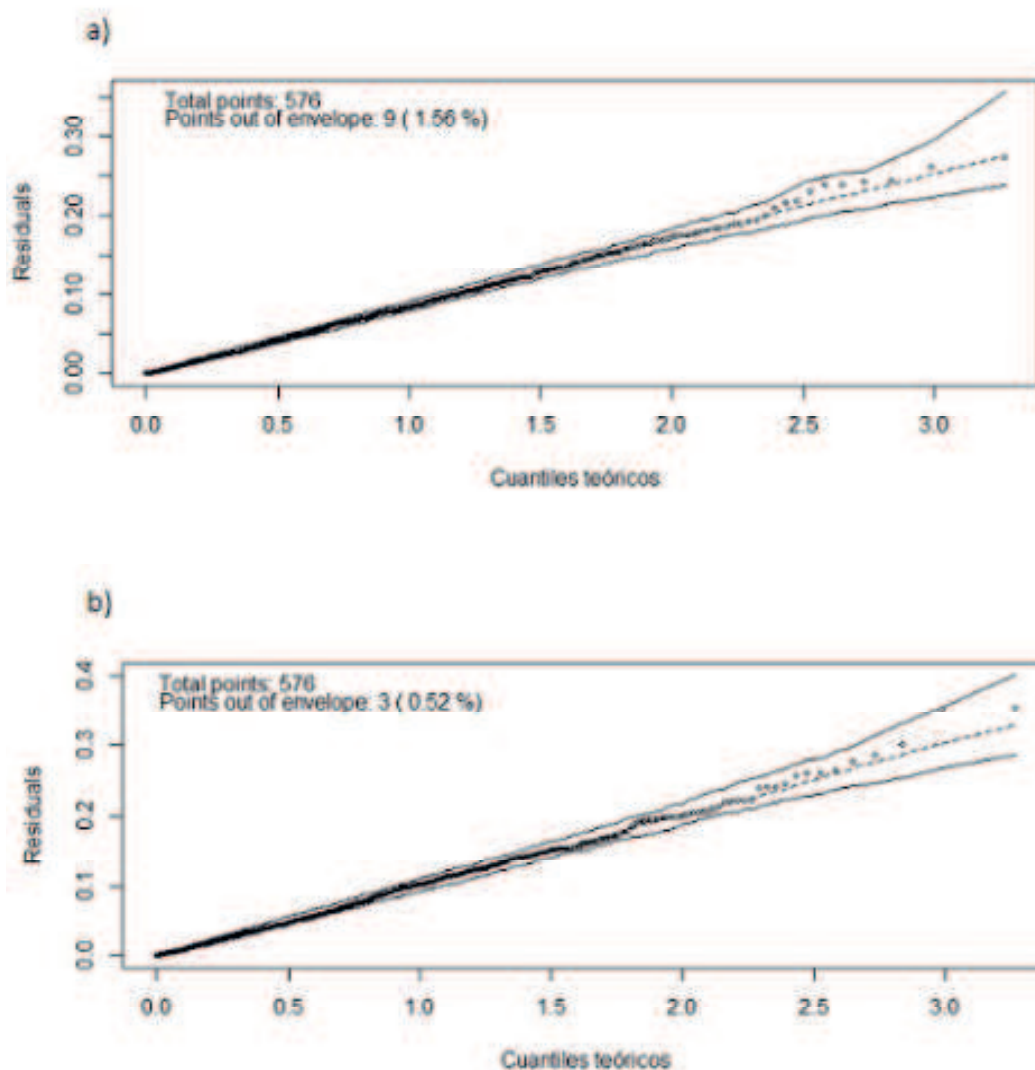


Figura 5 – Gráficos de envelope simulado com as variáveis PAS e PAD

5 Conclusão

Este estudo teve por objetivo avaliar o comportamento da pressão arterial e da frequência cardíaca de indivíduos hipertensos. A hipertensão é conceituada como uma síndrome caracterizada por valores pressóricos permanentemente elevados, associados a alterações metabólicas e hormonais, e a fenômenos tróficos como hipertrofia cardíaca e vascular (1, 2).

A hipertensão arterial é uma das formas de doença cardiovascular das mais prevalentes, constituindo-se em um problema de saúde pública no Brasil e no mundo, sendo a sua prevalência no Brasil. A distribuição gama é associada a dados contínuos assimétricos. Este estudo teve como objetivo avaliar o comportamento da pressão arterial de indivíduos.

A função de relação entre a variável Pressão Arterial Sistólica (PAS) com as variáveis nível de Triglicérides (TG) e a medida da Circunferência Abdominal (CA) e a função de relação entre Pressão Arterial Diastólica (PAD) e as variáveis o nível de Glicose (GLI) e a medida da Circunferência Abdominal (CA) por meio de um modelo linear generalizado com família gama e função de ligação log apresenta alta significância com duas variáveis independentes e desvio pequeno.

A análise residual indica que o modelo se ajusta muito bem aos dados, não apresenta resíduos discrepantes nem outliers. Assim observa-se que, os modelos lineares generalizados podem modelar a relação entre variáveis que um modelo linear normal não é capaz de fazer. Neste caso específico, usamos um modelo com família gama.

Podemos notar que os modelos lineares generalizados podem modelar a relação entre variáveis que um modelo linear normal não é capaz de fazer. Neste caso específico, usamos um modelo com família gama, mas os MLG's podem relacionar variáveis dependentes de diversas funções de distribuições, desde que pertençam a família exponencial, com as variáveis independentes.

Referências

- BOLFARINE, H.; SANDOVAL, M. C. *Introdução à inferência estatística*. [S.l.]: SBM, 2001. v. 2. Citado 2 vezes nas páginas 17 e 22.
- BUSE, A. The likelihood ratio, wald, and lagrange multiplier tests: An expository note. *The American Statistician*, Taylor & Francis, v. 36, n. 3a, p. 153–157, 1982. Citado na página 18.
- CLARICE, G. M. C.; DEMETRIO, B. Ccsa/ufpe esalq/usp. *Statistical modelling*, Springer Verlag, v. 57, p. 95, 1989. Citado na página 22.
- CORDEIRO, G. M.; DEMÉTRIO, C. G. B. *Modelos lineares generalizados e extensões*. [S.l.: s.n.], 2013. Citado 11 vezes nas páginas 12, 13, 14, 15, 16, 17, 19, 20, 21, 25 e 26.
- GELFAND, A. E.; DALAL, S. R. A note on overdispersed exponential families. *Biometrika*, Oxford University Press, v. 77, n. 1, p. 55–64, 1990. Citado na página 13.
- MCCULLAGH, P.; NELDER, J. A. *Generalized linear models*. [S.l.]: CRC press, 1989. v. 37. Citado na página 15.
- 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 2 vezes nas páginas 11 e 15.
- PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2004. Nenhuma citação no texto.
- PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2013. Citado 7 vezes nas páginas 16, 18, 19, 22, 23, 34 e 35.
- POSSAMAI, A. A. *Modelos não lineares de família exponencial revisitados*. Dissertação (Mestrado) — Universidade de Sao Paulo, 2009. Citado na página 34.
- PREGIBON, D. Logistic regression diagnostics. *The Annals of Statistics*, JSTOR, p. 705–724, 1981. Citado na página 34.
- 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 5 vezes nas páginas 11, 14, 17, 20 e 21.
- 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 22 e 29.
- VENABLES, W.; RIPLEY, B. Tree-based methods. In: *Modern applied statistics with S-Plus*. [S.l.]: Springer, 1999. p. 303–327. Citado na página 24.

