UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

JAIME PHASQUINEL LOPES CAVALCANTE

VARIABILIDADE E ADERÊNCIA EM MODELOS DE APRENDIZADO DE MÁQUINA COM DISTRIBUIÇÃO BETA

Recife
2022

JAIME PHASQUINEL LOPES CAVALCANTE

VARIABILIDADE E ADERÊNCIA EM MODELOS DE APRENDIZADO DE MÁQUINA COM DISTRIBUIÇÃO BETA

Dissertação apresentada ao Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial à obtenção do título de mestre em Estatística. Área de Concentração: Probabilidade e Estatística

Orientadora: Prof.ª Dra. Patrícia Leone Espinheira Ospina

Co-Orientador: Prof.º Dr. Juvêncio Santos Nobre

Recife

2022

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

C376v Cavalcante, Jaime Phasquinel Lopes
Variabilidade e aderência em modelos de aprendizado de máquina com distribuição beta / Jaime Phasquinel Lopes Cavalcante. – 2022.
69 f.: il., fig., tab.

Orientadora: Patrícia Leone Espinheira Ospina.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2022.
.
Inclui referências.

1. Probabilidade. 2. Regressão beta. I. Ospina, Patrícia Leone Espinheira (orientadora). II. Título.

519.2 CDD (23. ed.) UFPE - CCEN 2022-27

JAIME PHASQUINEL LOPES CAVALCANTE

"VARIABILIDADE E ADERÊNCIA EM MODELOS DE APRENDIZADO DE MÁQUINA COM DISTRIBUIÇÃO BETA"

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovada em: 7 de fevereiro de 2022.

BANCA EXAMINADORA

Profª. Drª Patrícia Leone Espinheira Ospina
DE/UFPE

Prof. Dr. Francisco Cribari Neto
DE/UFPE

Prof.Dr. Raydonal Ospina Martínez
DE/UFPE

Prof.Dr. Rafael Izbicki
DE/ UFSCar

AGRADECIMENTOS

A minha família, pelo total apoio diante desta trajetória.

À Professora Patrícia Ospina, por todo o aprendizado e acolhimento que recebi ao longo da elaboração deste trabalho.

À todos os professores do Programa de Pós Graduação em Estatística da UFPE, pela transmissão de conhecimentos.

Às colegas da turma de mestrado - Jaciele, Noemir, Penelope e Suelem - pelas horas de estudo, esforço coletivo, momentos de tensão e alegrias.

Aos irmãos que fiz em Recife - Érica, Raquel e Thiago - por todo o suporte e histórias compartilhadas.

Aos professores Luciana Moura Reinaldo e Juvêncio Santos Nobre, por desde a graduação me servirem de exemplo para nunca desistir dos meus objetivos.

Aos membros da banca examinadora, pela participação e contribuições ao trabalho.

Ao universo, pelo ensinamento de que nada é por acaso e que tudo tem um motivo.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro.

RESUMO

Proposto por Ferrari e Cribari-Neto (2004), o modelo de regressão beta tem sido objeto de estudo de diversos autores devido a sua relevância para a modelagem de fenômenos cuja variável resposta esteja definida no intervalo unitário (0,1). No tocante ao diagnóstico dos modelos de regressão beta, Espinheira *et al.* (2008) apresentaram a definição de resíduos baseados no processo iterativo Scoring de Fisher, sendo esta amplamente utilizada para a generalização e proposição de novos resíduos para as extensões dos modelos de regressão beta. Com o foco na distribuição de probabilidade e observando que a mesma forma uma família exponencial bidimensional, utilizamos o *Teorema da Função Integrável* - demonstrado por Barndorff-Nielsen (1978) e Lehmann (1986) - para propor uma nova classe de resíduos e critérios do tipo pseudo-R2 baseados nas estatísticas suficientes e completas com a finalidade de avaliar a variabilidade e aderência, além de realizar diagnósticos em modelos de aprendizado de máquina (*machine learning*) com distribuição beta. Além disso, para o modelo de regressão beta, propomos um novo resíduo baseado no processo iterativo Scoring de Fisher. Quanto à qualidade preditiva, utilizamos a estatística PRESS e o coeficiente de predição P2, introduzido por Espinheira *et al.* (2019) para a classe de modelos de regressão beta lineares e não-lineares. O desempenho das propostas é avaliado por meio de três aplicações, associadas a um conjunto de dados reais, relativas ao estudo do risco à doenças cardíacas.

Palavras-chave: regressão beta; distribuição beta; variabilidade; aderência; diagnóstico; predição.

ABSTRACT

Proposed by Ferrari and Cribari-Neto (2004), the beta regression model has been the object of study by several authors due to its relevance for the modeling of phenomena whose response variable is defined in the unit interval (0.1). With a diagnostic focus on beta regression models, Espinheira et al. (2008) presented the definition of residuals based on Fisher's Scoring iterative process, which is widely used for the generalization and proposition of new residuals for the extensions of the models of beta regression. Aiming at the probability distribution we verify that the same form a two-dimensional exponential family, we use the Integrable Function Theorem - demonstrated by Barndorff-Nielsen (1978) and Lehmann (1986) - to propose a new class of residues and criteria R2 type based on sufficient and complete statistics in order to assess variability and adherence, in addition to performing diagnostic in machine learning models with beta distribution. Furthermore, for the beta regression model, we propose a new residual based on Fisher's Scoring iterative process. As for the predictive quality, we used the PRESS statistic and the prediction coefficient P2, introduced by Espinheira et al. (2019) for the class of linear and non-linear beta regression models. The performance of the proposals is evaluated through three applications, associated with a set of real data, related to the study of the risk of cardiovascular diseases.

Keywords: beta distribution; beta regression; variability; adherence; diagnostic; prediction.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nas mais variadas situações práticas, sejam observacionais ou experimentais, observamos o interesse de investigar a relação entre uma variável aleatória, denominada variável resposta, e um conjunto de p variáveis explicativas $(x_1, x_2, \ldots, x_p)$, por meio de um modelo de regressão. Durante muito tempo, diversas propostas adotaram modelos lineares normais na tentativa de descrever a ocorrência da maioria dos fenômenos aleatórios. Mesmo quando não era razoável admitir a suposição de normalidade, algum tipo de transformação era sugerida com a finalidade de obtenção da normalidade desejada (PAULA,2013). Para além dessa problemática, a modelagem de fenômenos que envolvem variáveis relacionadas com taxas, razões e proporções utilizando o modelo normal linear apresenta-se inadequada, uma vez que os dados restritos ao intervalo (0,1), em geral, apresentam assimetria e possuem um padrão especifico de heteroce-dasticidade. Nesse contexto, o modelo de regressão beta proposto por Ferrari e Cribari-Neto (2004) assume que a variável resposta possui distribuição beta e que a média dessa variável está relacionada a um preditor linear através de uma função de ligação. Semelhante ao Modelo Linear Generalizado (MLG), o preditor linear envolve covariáveis e parâmetros de regressão desconhecidos. Além disso, os autores utilizam uma reparametrização da distribuição beta que considera um parâmetro de precisão ϕ supostamente constante. O modelo de regressão beta, tem sido objeto de estudo de diversos autores, tanto em aspectos inferenciais quanto em diagnóstico. Smithson e Verkulien (2006) definiram um modelo de regressão beta em que tanto a modelagem da variável resposta quanto do parâmetro de precisão ocorrem em função das covariáveis.

Direcionando ao diagnóstico dos modelos de regressão beta, Espinheira *et al.* (2008) apresentaram a definição de resíduos baseados no processo iterativo Scoring de Fisher, sendo esta amplamente utilizada para a generalização e proposição de novos resíduos para as extensões dos modelos de regressão beta. Adicionalmente, no contexto da seleção de modelos com foco no poder de predição, baseando-se em Allen (1974), Espinheira *et al.* (2019) introduziram a estatística PRESS (Predictive Residual Sum of Squares) para a classe de modelos de regressão beta lineares e não-lineares. Sequencialmente, baseando-se em Quan (1988) e Madiavilla *et al.* (2008) e na estatística PRESS, Espinheira *et al.* (2019) propõem um coeficiente de predição, denominado P^2, de modo que o mesmo pode ser utilizado para selecionar modelos com maior qualidade preditiva, considerando inclusive distribuições distintas para a variável resposta.

Outro aspecto importante é a adoção de modelos que sejam capazes de controlar, ao mesmo tempo, a variabilidade e o viés, questões que podem ser associadas caso sejam

consideradas estatísticas de adequação e predição. Desta forma, o objetivo dessa dissertação é propor critérios do tipo pseudo-R^2, com a finalidade da avaliação da variabilidade e predição dos modelos. Tomando os resíduos baseados no processo iterativo Scoring Fisher, propomos o resíduo *Variance*. Adicionalmente, observando que distribuição beta forma uma família exponencial bidimensional, utilizamos o *Teorema da Função Integrável* - demonstrado por Barndorff-Nielsen (1978) e Lehmann (1986) - para apresentar uma nova classe de resíduos para o modelo de regressão beta. Fazendo uso de tais resíduos definimos novos critérios do tipo pseudo-R^2 para avaliação da variabilidade e do viés dos modelos. Para a ilustração das medidas propostas estudamos a qualidade das mesmas em três aplicações ligadas ao estudo do risco à doenças cardiovasculares.

1.1 ORGANIZAÇÃO DA DISSERTAÇÃO

Esta dissertação encontra-se dividia em seis capítulos. No segundo capítulo, apresentamos o modelo de regressão beta com dispersão fixa e dispersão variável, além de definir um novo resíduo baseado no processo iterativo Scoring de Fisher. No terceiro capítulo definimos os novos resíduos para os modelos de regressão beta fazendo uso do *Teorema da Função Integrável*. No quarto capítulo abordamos os critérios de seleção de modelos para avaliação da qualidade do ajuste dos modelos em termos da variabilidade e da predição. Adicionalmente, baseando-se nos resíduos obtidos das estatísticas suficientes e completas da distribuição beta, apresentamos os novos critérios de seleção de modelos desenvolvidos. No quinto capítulo, apresentamos as aplicações da teoria desenvolvida em três casos relacionados com o estudo do risco à doenças cardiovasculares. Por fim, no sexto capítulo, apresentamos as conclusões desta dissertação.

1.2 SUPORTE COMPUTACIONAL

As operações computacionais deste trabalho foram realizadas por meio da linguagem de programação matricial Ox, sendo gratuita para fins acadêmicos e disponível em <http://www.doornik.com>. Maiores detalhes podem ser encontrados em Doornik (2009). Todos os resultados gráficos apresentados nesta dissertação foram obtidos através do ambiente de programação, análise de dados e gráficos R, em sua versão 4.0.3, disponível em <http://www.R-project.org>. O sistema de tipografia LaTeXfoi utilizado para a digitação deste trabalho. Detalhes sobre o sistema de tipografia podem ser encontrados em Lamport (1994).

2 MODELO DE REGRESSÃO BETA

2.1 DISTRIBUIÇÃO BETA

A família beta, apresentada em uma carta de Sir. Isaac Newton a Henry Oldenberg no ano de 1676, tem sido amplamente utilizada na estatística ao longo dos anos devido à necessidade da modelagem de dados que se apresentam como proporções, frações ou taxas, isto é, variáveis cujos valores estão limitados ao intervalo (0,1). Dessa forma, a distribuição beta pode assumir uma variedade surpreendentemente grande de formas, podendo ser ajustada a praticamente qualquer dado que represente um fenômeno em quase todos os campos de aplicação. Assume-se que uma variável aleatória y possui distribuição beta caso sua função de densidade seja definida por

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \qquad 0 < y < 1,$$

em que $p,q > 0$ são os parâmetros de forma e $\Gamma(.)$ é a função gama, sendo definida como $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx$. Além disso, a média e variância da variável aleatória são denotadas, respectivamente, por

$$\mathbb{E}(y) = \frac{p}{p+q} \qquad \text{e} \qquad \text{Var}(y) = \frac{pq}{(p+q)^2(p+q+1)}.$$

Tendo em vista que a função de densidade apresentada possui somente parâmetros de forma e que modelos de regressão usualmente são definidos para a modelagem da média, de modo que esteja relacionada com um parâmetro de precisão (ou dispersão), Ferrari e Cribari-Neto (2004) apresentaram uma reparametrização para a distribuição beta, em que $\mu = p/(p+q)$ e $\phi = p+q$, ou seja, $p = \mu\phi$ e $q = (1-\mu)\phi$. Então, a função de densidade reparametrizada pode ser escrita como

$$f(y;\mu\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \qquad 0 < y < 1, \tag{2.1}$$

em que $0 < \mu < 1$ e $\phi > 0$. Desta forma, a média e variância são expressas, respectivamente, como

$$\mathbb{E}(y) = \mu \qquad \text{e} \qquad \text{Var}(y) = \frac{\mu(1-\mu)}{1+\phi},$$

com $\mu(1-\mu)$ representando a função de variância. A respeito dos parâmetros μ e ϕ, observa-se que uma vez definida a média da variável y, quando o valor de ϕ aumenta, a variância de y diminui. Logo, ϕ pode ser tomado como o parâmetro de precisão do modelo e seu inverso

(ϕ^{-1}) é visto como o parâmetro de dispersão. Quando ϕ assume baixos valores e $\mu = 0,5$ a distribuição apresenta simetria e considerável achatamento. Para valores de μ nas extremidades do intervalo (0,1), a distribuição assume formas assimétricas. Porém, à medida que o valor de ϕ cresce, mesmo com μ tendo valores localizado nas extremidades, a distribuição tende a formas simétricas. Finalmente, a distribuição beta também assume formas afiladas quando o parâmetro ϕ aumenta.

2.2 MODELO DE REGRESSÃO BETA COM DISPERSÃO FIXA

Sejam $y_1, ..., y_n$ variáveis aleatórias independentes, nas quais cada y_t, $t = 1, ..., n$, tem função de densidade (2.1) com média μ_t e precisão ϕ desconhecida. O modelo de regressão beta com ϕ constante pode ser escrito através da seguinte relação funcional

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_t,$$

com $\eta_t = \mathbf{x}_t^\top \beta$ sendo o preditor linear, em que $\beta = (\beta_1, ..., \beta_k)^\top$ é um vetor de parâmetros desconhecido, de tal forma que $\beta \in \mathbb{R}^k$ e $\mathbf{x}_t = (x_{t1}, ..., x_{tk})^\top$ são valores fixos e conhecidos das k covariáveis ($k < n$). A função $g(.)$ é a função de ligação que é estritamente monótona e duplamente diferenciável, possuindo domínio no intervalo (0,1), imagem nos reais e relaciona o valor esperado μ_t com o preditor linear.

Como exemplos de função de ligação, menciona-se: a logito, $g(\mu) = \log\{\mu/(1-\mu)\}$, probito, $g(\mu) = \Phi^{-1}(\mu)$, com $\Phi(.)$ sendo a função de distribuição acumulada da normal padrão, a log-log, $g(\mu) = -\log\{-\log(\mu)\}$, entre outras. Ressalta-se que a escolha adequada da função de ligação depende do tipo da variável resposta e da condução do estudo. Neste modelo, verifica-se que $\mu_t = g^{-1}(\eta_t)$ e $\text{Var}(y_t) = V(g^{-1}(\eta_t))/(1+\phi)$. Então, sendo a variância da variável resposta dependente de μ_t, até nos casos em que o parâmetro de precisão é constante para todas as observações, observa-se que o modelo de regressão beta proposto por Ferrari e Cribari-Neto (2004) é naturalmente heterocedástico.

O processo de estimação pode ser realizado por meio do método da máxima verossimilhança. Desta forma, tomando a expressão (2.1) e assumindo uma amostra com n observações independentes, o logaritmo da função de verossimilhança é escrito como

$$\ell(\beta, \phi) = \sum_{t=1}^{n} \ell_t(\mu_t, \phi),$$

em que

$$\begin{aligned}\ell(\mu_t,\phi) =& \log\Gamma(\phi)-\log\Gamma(\mu_t\phi)-\log\Gamma((1-\mu)\phi)+(\mu_t\phi-1)\log y_t\\ &+\{(1-\mu_t)\phi-1\}\log(1-y_t).\end{aligned}$$

A função escore, obtida por meio da diferenciação do logaritmo da função de verossimilhança com relação aos parâmetros desconhecidos, é definida por $\left(U_\beta(\beta,\phi)^\top, U_\phi(\beta,\phi)\right)^\top$, um vetor de dimensão $(k+1)\times 1$. Os componentes do vetor escore de β são dados por

$$\frac{\partial\ell(\beta,\phi)}{\partial\beta_i}=\sum_{t=1}^{n}\frac{\partial\ell_t(\mu_t,\phi)}{\partial\mu_t}\frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_t}\frac{\partial\eta_t}{\partial\beta_i}, \qquad i=1,\ldots,k, \tag{2.2}$$

com

$$\frac{\partial\ell_t(\mu_t,\phi)}{\partial\mu_t}=\phi\left[\log\left(\frac{y_t}{1-y_t}\right)-\{\psi(\mu_t\phi)-\psi((1-\mu_t)\phi)\}\right], \tag{2.3}$$

sendo $\psi(.)$ a função digama, isto é, $\psi(z)=\mathrm{d}\log\Gamma(z)/\mathrm{d}z$. Além disso,$\mathrm{d}\mu_t/\mathrm{d}\eta_t=1/g'(\mu_t)$ e $\partial\eta_t/\partial\beta_i=x_{ti}$.

A Equação (2.2) pode ser reescrita como

$$\frac{\partial\ell(\beta,\phi)}{\partial\beta_i}=\sum_{t=1}^{n}\phi\,(y_t^*-\mu_t^*)\,\frac{1}{g'(\mu_t)}x_{ti},$$

de modo que

$$y_t^*=\log\left(\frac{y_t}{1-y_t}\right) \qquad \text{e} \qquad \mu_t^*=\psi(\mu_t\phi)-\psi((1-\mu_t)\phi). \tag{2.4}$$

Além disso, sob certas condições de regularidade, observa-se que

$$\mathbb{E}\left(\frac{\partial\ell_t(\mu_t,\phi)}{\partial\mu_t}\right)=0\Rightarrow\mathbb{E}(\phi(y_t^*-\mu_t^*))=0\Rightarrow\mathbb{E}(y_t^*)=\mu_t^*. \tag{2.5}$$

Matricialmente, a função escore para β, um vetor de dimensão k, é definida como

$$U_\beta(\beta,\phi)=\phi X^\top T(y^*-\mu^*),$$

em que X é a matriz de covariadas de ordem $n\times k$, de forma que a t-ésima linha de $x_t^\top$, $T=\mathrm{diag}\{1/g'(\mu_1),\ldots,1/g'(\mu_n)\}$, $y^*=(y_1^*,\ldots,y_n^*)^\top$ $\mu^*=(\mu_1^*,\ldots,\mu_n^*)^\top$.

Para o parâmetro de precisão, ϕ, a função escore é dada por

$$\frac{\partial\ell_t(\beta,\phi)}{\partial\phi}=\sum_{t=1}^{n}\frac{\partial\ell_t(\mu_t,\phi)}{\partial\phi},$$

com

$$\begin{aligned}\frac{\partial\ell_t(\mu_t,\phi)}{\partial\phi}=&\mu_t\left[\log\left(\frac{y_t}{1-y_t}\right)-\psi(\mu_t\phi)+\psi((1-\mu_t)\phi)\right]\\ &+\log(1-y_t)-\psi((1-\mu_t)\phi)+\psi(\phi),\end{aligned} \tag{2.6}$$

resultando no seguinte escalar

$$U_\phi(\beta,\phi)=\sum_{t=1}^{n}\mu_t(y_t^*-\mu_t^*)+\log(1-y_t)-\psi((1-\mu_t)\phi)+\psi(\phi).$$

O processo de obtenção da matriz de informação de Fisher para β e ϕ requer as derivadas de segunda ordem do logaritmo da função de verossimilhança em relação aos parâmetros. Tomando a Equação (2.2), as derivadas de segunda ordem de $\ell(\beta,\phi)$ relacionadas a β_i e β_j, com $i,j=1,\ldots,k$, são

$$\begin{aligned}\frac{\partial^2\ell(\beta,\phi)}{\partial\beta_i\partial\beta_j}&=\sum_{t=1}^{n}\frac{\partial}{\partial\mu_t}\left(\frac{\partial\ell_t(\mu_t,\phi)}{\partial\mu_t}\frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_t}\frac{\partial\eta_t}{\partial\beta_i}\right)\frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_t}\frac{\partial\eta_t}{\partial\beta_j}\\&=\sum_{t=1}^{n}\left\{\frac{\partial^2\ell_t(\mu_t,\phi)}{\partial\mu_t^2}\frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_t}+\frac{\partial\ell_t(\mu_t,\phi)}{\partial\mu_t}\frac{\partial}{\partial\mu_t}\left(\frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_t}\right)\right\}\frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_t}x_{ti}x_{tj}.\end{aligned}$$

Admitindo (2.5), segue que

$$\mathbb{E}\left(\frac{\partial^2\ell(\beta,\phi)}{\partial\beta_i\partial\beta_j}\right)=\sum_{t=1}^{n}\mathbb{E}\left(\frac{\partial^2\ell_t(\mu_t,\phi)}{\partial\mu_t^2}\right)\left(\frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_t}\right)^2x_{ti}x_{tj}.$$

De (2.3), obtém-se que

$$\frac{\partial^2\ell_t(\mu_t,\phi)}{\partial\mu_t^2}=-\phi^2\left\{\psi'(\mu_t\phi)+\psi'((1-\mu_t)\phi)\right\},$$

em que $\psi'(.)$ é a função trigama, ou seja, $\psi'(z)=\mathrm{d}^2\log\Gamma(z)/\mathrm{d}z^2$. Então,

$$\mathbb{E}\left(\frac{\partial^2\ell(\beta,\phi)}{\partial\beta_i\partial\beta_j}\right)=-\phi\sum_{t=1}^{n}w_tx_{ti}x_{tj},$$

com

$$\begin{aligned}w_t&=\phi\left[\psi'(\mu_t\phi)+\psi'((1-\mu_t)\phi)\right]\frac{1}{\{g'(\mu_t)\}^2}\qquad\text{e}\\ \upsilon&=\psi'(\mu_t\phi)+\psi'((1-\mu_t)\phi).\end{aligned}$$

Em forma matricial,

$$\mathbb{E}\left(\frac{\partial^2\ell(\beta,\phi)}{\partial\beta\partial\beta^\top}\right)=-\phi X^\top WX,$$

onde W=diag$\{w_1,\ldots,w_n\}$.

Com relação a ϕ, a derivada de (2.2) é expressa por

$$\frac{\partial^2\ell(\beta,\phi)}{\partial\beta_i\partial\phi}=\sum_{t=1}^{n}\left[(y_t^*-\mu_t^*)-\phi\frac{\partial\mu_t^*}{\partial\phi}\right]\frac{1}{g'(\mu_t)}x_{ti},$$

com $\partial\mu_t^*/\partial\phi=\mu_t\psi'(\mu_t\phi)-\psi'((1-\mu_t)\phi)(1-\mu_t)$. Além disso, com base em (2.5), segue-se que

$$\mathbb{E}\left(\frac{\partial^2\ell(\beta,\phi)}{\partial\beta_i\partial\phi}\right)=-\sum_{t=1}^{n}c_t\frac{1}{g'(\mu_t)}x_{ti},$$

em que $c_t = \phi\{\mu_t\psi'(\mu_t\phi) - (1-\mu_t)\psi'((1-\mu_t)\phi)\}$. Dessa maneira, a forma matricial é escrita como

$$\mathbb{E}\left(\frac{\partial^2\ell(\beta,\phi)}{\partial\beta\partial\phi}\right) = -X^\top Tc,$$

com $c = \{c_1,\ldots,c_n\}$. Da expressão (2.6), a segunda derivada de $\ell(\beta,\phi)$ com relação a ϕ é dada por

$$\frac{\partial^2\ell(\beta,\phi)}{\partial\phi^2} = -\sum_{t=1}^{n}\left[\mu_t^2\psi'(\mu_t\phi) + (1-\mu_t)^2\psi'((1-\mu_t)\phi) - \psi'(\phi)\right] = -\sum_{t=1}^{n} d_t,$$

com $d_t = \mu_t^2\psi'(\mu_t\phi) + (1-\mu_t)^2\psi'((1-\mu_t)\phi) - \psi'(\phi)$. Logo,

$$\mathbb{E}\left(\frac{\partial^2\ell(\beta,\phi)}{\partial\phi^2}\right) = -\sum_{t=1}^{n} d_t,$$

podendo ser escrito na seguinte forma matricial

$$\mathbb{E}\left(\frac{\partial^2\ell(\beta,\phi)}{\partial\phi^2}\right) = -\mathrm{tr}(D),$$

em que $D = \mathrm{diag}\{d_1,\ldots,d_n\}$ e tr(D) indica traço da matriz D. Assim, a matriz de informação de Fisher é dada por

$$K = K(\beta,\gamma) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{pmatrix} = \begin{pmatrix} \phi X^\top WX & X^\top Tc \\ X^\top Tc & \mathrm{tr}(D) \end{pmatrix}.$$

Em grandes amostras e sob certas condições de regularidade (SEN; SINGER,1993),

$$\begin{pmatrix} \hat\beta \\ \hat\phi \end{pmatrix} \sim N_{k+1}\left(\begin{pmatrix} \beta \\ \phi \end{pmatrix}, K^{-1}\right),$$

aproximadamente, em que $\hat\beta$ e $\hat\phi$ são os estimadores de máxima verossimilhança de β e ϕ, respectivamente. Além disso, diferente do que se observa em modelos lineares generalizadas, no modelo de regressão beta, tais parâmetros não são ortogonais e, assim, $\hat\beta$ e $\hat\phi$ não possuem independência assintótica. A obtenção destes estimadores se dá através do seguinte sistema:

$$\begin{cases} U_\beta(\beta,\phi) = 0 \\ U_\phi(\beta,\phi) = 0 \end{cases}.$$

No entanto, eles não podem ser expressos por meio de uma forma fechada, devendo ser obtidos numericamente através do uso de algoritmos de otimização não linear, tais como Newton-Raphson, Escore de Fisher, BHHH, BFGS, entre outros. Tal tipo de otimização requer a especificação de valores iniciais para serem utilizados no esquema iterativo. Uma sugestão para o

ponto inicial para β, de acordo com Ferrari e Cribari-Neto (2004), é a utilização da estimativa de mínimos quadrados ordinários desse vetor de parâmetros, que é obtida de uma regressão linear em que a variável resposta é a variável transformada através da função de ligação $g(.)$. Disso, o estimador de mínimos quadrados ordinários é $(X^\top X)^{-1}X^\top z$, em que $z = (g(y_1), \ldots, g(y_n))^\top$.

Quanto ao valor do parâmetro de precisão, a indicação do valor inicial é baseada em $\mathrm{Var}(y_t) = \mu_t(1-\mu_t)/(1+\phi)$, podendo ser reescrito como $\phi = [\mu_t(1-\mu_t)/\mathrm{Var}(y_t)] - 1$. A sugestão do valor inicial é

$$\phi^{(0)} = \frac{1}{n}\sum_{t=1}^{n} \frac{\breve{\mu}_t(1-\breve{\mu}_t)}{\breve{\sigma}_t^2} - 1,$$

de modo que $\breve{\mu}_t$ é resultante da aplicação de $g^{-1}(.)$ para o t-ésimo valor ajustado da regressão linear de $g(y_1), \ldots, g(y_n)$ em X, ou seja,

$$\breve{\mu}_t = g^{-1}(x_t^\top (X^\top X)^{-1}X^\top z) \qquad \text{e} \qquad \breve{\sigma}_t^2 = \frac{\breve{e}^\top \breve{e}}{(n-k)g'(\breve{\mu}_t)^2}.$$

Deve-se observar que $\breve{e} = z - X(X^\top X)^{-1}X^\top z$ é o vetor de resíduos de mínimos quadrados ordinários de uma regressão linear com a variável resposta transformada. O processo para a obtenção de $\breve{\sigma}_t^2$ utiliza a expansão até a primeira ordem da função $g(y_t)$ em série de Taylor em torno do ponto μ_t e aplica-se a variância, tal que

$$\mathrm{Var}\left[g(y_t)\right] \approx \mathrm{Var}\left[g(y_t) + (y_t - \mu_t)g'(\mu_t)\right] = \mathrm{Var}g(y_t)\left[g'(\mu_t)\right]^2,$$

então,

$$\mathrm{Var}(y_t) \approx \frac{\mathrm{Var}\left[g(y_t)\right]}{\left[g'(\mu_t)\right]^2} \Rightarrow \breve{\sigma}^2 = \breve{\mathrm{Var}}(y_t) \approx \frac{\widehat{\mathrm{Var}}\left[g(y_t)\right]}{\left[g'(\breve{\mu}_t)\right]^2},$$

em que $\widehat{\mathrm{Var}}\left[g(y_t)\right] = \breve{e}^\top \breve{e}/(n-k)$.

2.3 RESÍDUOS

A definição de resíduo para o modelo de regressão beta é realizada de maneira semelhante ao processo apresentado para os Modelos Lineares Generalizados (MLG's). Todavia, é verificado que as propriedades dos resíduos não são as mesmas, tornando-se importante a definição de resíduos com propriedades conhecidas. Desta forma, propostos por Espinheira *et al.*(2008), serão apresentados alguns dos resíduos já existentes na literatura: os resíduos ponderado padronizado 1 e ponderado padronizado 2. A fim de facilitar a nomenclatura, os mesmos serão denominados por resíduo ponderado e resíduo ponderado padronizado, respectivamente.

2.3.1 Resíduo Ponderado

A proposta de Espinheira *et al.* (2008) para o resíduo da regressão beta, quando é assumido que a dispersão é constante, é baseada no processo iterativo Scoring de Fisher para β, sendo definido como

$$\beta^{(m+1)} = \beta^{(m)} + \left(K_{\beta\beta}^{(m)}\right)^{-1} U_{\beta}^{(m)}(\beta, \gamma),$$

de modo que m representa os passos necessários até a convergência do processo. Então, o processo iterativo de interesse é dado por

$$\beta^{(m+1)} = \beta^{(m)} + \left(\phi^{(m)} X^{\top} W^{(m)} X\right)^{-1} \phi^{(m)} X^{\top} T^{(m)} (y^* - \mu^{*(m)})$$

podendo ser expresso como

$$\beta^{(m+1)} = \beta^{(m)} + \left(X^{\top} \phi^{(m)} W^{(m)} X\right)^{-1} X^{\top} \phi^{(m)} T^{(m)} (y^* - \mu^{*(m)}),$$

com $\Phi = \text{diag}\{\phi, \ldots, \phi\}$.

Assumindo um processo iterativo de mínimos quadrados reponderados, o processo descrito acima tem a seguinte forma:

$$\beta^{(m+1)} = \left(X^{\top} \phi^{(m)} W^{(m)} X\right)^{-1} X^{\top} \phi^{(m)} W^{(m)} z^{(m)},$$

em que $z^{(m)} = \eta^{(m)} + W^{-1(m)} T^{(m)} \left(y^* - \mu^{*(m)}\right)$, com $\eta = (\eta_1, \ldots, \eta_n)^{\top} = X\beta$. Após convergência, tem-se que

$$\hat{\beta} = \left(X^{\top} \hat{\Phi} \hat{W} X\right)^{-1} X^{\top} \hat{\Phi} \hat{W} z, \tag{2.7}$$

com

$$z = \hat{\eta} + \hat{W}^{-1} \hat{T} (y^* - \hat{\mu}^*),$$

que é o estimador de mínimos quadrados considerando a regressão linear de $\Phi^{1/2} W^{1/2} z$ em $\Phi^{1/2} W^{1/2} X$. Assim, o resíduo de mínimos quadrados dessa regressão é definido por

$$\begin{aligned} r^{\beta} =& \hat{\Phi}^{1/2} \hat{W}^{1/2} z - \hat{\Phi}^{1/2} \hat{W}^{1/2} \hat{\eta} = \hat{\Phi}^{1/2} \hat{W}^{1/2} (z - \hat{\eta}) \\ =& \hat{\Phi}^{1/2} \hat{W}^{1/2} (\hat{\eta} + \hat{W}^{-1} \hat{T} (y^* - \hat{\mu}^*) - \hat{\eta}) \\ =& \hat{\Phi}^{1/2} \hat{W}^{-1/2} \hat{T} (y^* - \hat{\mu}^*). \end{aligned} \tag{2.8}$$

Logo, o resíduo da t-ésima observação é dado por

$$r_t^{\beta} = \hat{\Phi}^{1/2} \frac{1}{\hat{w}_t^{1/2}} \frac{1}{g'((\hat{\mu}_t))} (y_t^* - \hat{\mu}_t^*),$$

com $w_t = \phi_t \upsilon_t \left[1/g'(\mu_t)\right]^2$. Logo, este resíduo é definido como

$$r_t^{\beta} = \frac{(y_t^* - \hat{\mu}_t^*)}{\sqrt{\hat{\upsilon}_t}}, \tag{2.9}$$

sendo

$$\upsilon_t = \psi'(\mu_t \phi) + \psi'((1-\mu_t)\phi). \tag{2.10}$$

2.3.2 Resíduo Ponderado Padronizado

Proposto por Espinheira *et al.* (2008), o resíduo ponderado padronizado é baseado na variância de z e apresenta-se como uma padronização de (2.9). A equação (2.7) pode ser reescrita como

$$\left(X^\top \hat{\Phi} \hat{W} X\right) \hat{\beta} = X^\top \hat{\Phi} \hat{W} z. \tag{2.11}$$

Dado que $\mathrm{Cov}(\hat{\beta}) \approx \phi^{-1}(X^\top W X)^{-1} = (X^\top \Phi W X)^{-1}$ e assumindo que $\hat{W} \approx W$ e $\hat{\Phi} \approx \Phi$ na equação (2.11), tem-se que

$$\begin{aligned}
\left(X^\top \Phi W X\right) \mathrm{Cov}(\hat{\beta}) \left(X^\top \Phi W X\right)^\top &\approx \left(X^\top \Phi W\right) \mathrm{Cov}(z) \left(X^\top \Phi W\right)^\top \\
\left(X^\top \Phi W X\right) \left(X^\top \Phi W X\right)^{-1} \left(X^\top \Phi W X\right)^\top &\approx \left(X^\top \Phi W\right) \mathrm{Cov}(z) \left(W \Phi X\right)^\top \\
\mathrm{Cov}(z) &\approx W^{-1} \Phi^{-1}.
\end{aligned}$$

Com isso, $\widehat{\mathrm{Cov}}(z) \approx \hat{W}^{-1}\hat{\Phi}^{-1}$. Tomando (2.7), o resíduo em (2.8) também pode ser expresso por

$$r^{\beta} = \hat{\Phi}^{1/2} \hat{W}^{1/2} \left(I - X \left(X^\top \hat{\Phi} \hat{W} X\right)^{-1} X^\top \hat{\Phi} \hat{W}\right) z.$$

Considerando novamente $\hat{W} \approx W$ e $\hat{\Phi} \approx \Phi$, segue que

$$\begin{aligned}
\mathrm{Cov}(r^{\beta}) &\approx \left(\Phi^{1/2} W^{1/2} - \Phi^{1/2} W^{1/2} X \left(X^\top \Phi W X\right)^{-1} X^\top \Phi W\right) \mathrm{Cov}(z) \\
&\times \left(\Phi^{1/2} W^{1/2} - \Phi^{1/2} W^{1/2} X \left(X^\top \Phi W X\right)^{-1} X^\top \Phi W\right)^\top \\
&\approx \left(I - \Phi^{1/2} W^{1/2} X \left(X^\top \Phi^{1/2} W X\right)^{-1} X^\top \Phi^{1/2} W^{1/2}\right) \\
&\approx (I - H^*).
\end{aligned}$$

Assim, $\widehat{\mathrm{Cov}}(r^{\beta}) \approx I - \hat{H}^*$, em que $\hat{H}^* = \hat{\Phi}^{1/2} \hat{W}^{1/2} X \left(X^\top \hat{\Phi}^{1/2} \hat{W} X\right)^{-1} X^\top \hat{\Phi}^{1/2} \hat{W}^{1/2}$, com H^* sendo a matriz de projeção de $\Phi^{1/2} W^{1/2} z$ contra $\Phi^{1/2} W^{1/2} X$, além de ser uma matriz simétrica e idempotente. Por fim, a padronização do resíduo é definida por

$$r_{pt}^{\beta} = \frac{r_t^{\beta}}{\sqrt{\widehat{\mathrm{Cov}}(r_t^{\beta})}} = \frac{y_t^* - \mu_t^*}{\sqrt{\hat{\upsilon}_t \left(1 - \hat{h}_{ttt}^*\right)}},$$

de modo que h_{tt}^* é o t-ésimo elemento da diagonal principal de H^* e υ_t encontra-se definido em (2.10).

2.4 MODELO DE REGRESSÃO BETA COM DISPERSÃO VARIÁVEL

O modelo de regressão beta proposto por Ferrari e Cribari-Neto (2004) assume que o parâmetro de dispersão é constante ao longo das observações. Analogamente ao que ocorre em modelos lineares generalizados, o conceito de heterocedasticidade e dispersão variável é distinto do utilizado em modelos lineares normais, de modo em am ambos as variâncias não são contantes. Quanto aos modelos normais, o parâmetro de dispersão é a própria variância.

Ao se considerar o modelo de regressão beta com dispersão variável, a modelagem da média e dispersão é realizada simultaneamente (SMITHSON; VERKULIEN, 2006). Para tal, além da definição de uma modelagem para o parâmetro da média, também é apresentada uma estrutura de regressão para o parâmetro de precisão, em que a estimação conjunta de tais parâmetros pode ser feita fazendo uso do método de máxima verossimilhança.

Sendo $y_1,...,y_n$ variáveis aleatórias independentes, nas quais cada y_t, $t = 1,...,n$, possui a densidade (2.1), a média e o parâmetro de precisão satisfazem as seguintes relações funcionais

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_{1t} \qquad \text{e} \qquad h(\phi_t) = \sum_{j=1}^{q} z_{tj}\gamma_j = \eta_{2t},$$

em que $\beta = (\beta_1,....,\beta_k)^\top$ e $\gamma = (\gamma_1,....,\gamma_q)^\top$, são vetores de parâmetros desconhecidos, tais que ($\beta \in \mathbb{R}^k$ e $\gamma \in \mathbb{R}^q$), $x_{t1},...,x_{tk}$ e $z_{t1},...,z_{tq}$ são valores fixos e conhecidos das k e q covariáveis ($k+q<n$). Além disso, as funções $g(.)$ e $h(.)$ são as funções de ligação, sendo estas estritamente monótonas e duplamente diferenciáveis. Quanto ao logaritmo da função de verossimilhança, segue que

$$\ell(\beta,\gamma) = \sum_{t=1}^{n} \ell_t(\mu_t,\phi_t),$$

em que

$$\begin{aligned}\ell_t(\mu_t,\phi_t) =& \log\Gamma(\phi_t) - \log\Gamma(\mu_t\phi_t) - \log\Gamma((1-\mu)\phi_t) + (\mu_t\phi_t - 1)\log y_t \\ &+ \{(1-\mu_t)\phi_t - 1\}\log(1-y_t).\end{aligned}$$

O vetor escore é definido por $\left(U_\beta(\beta,\gamma)^\top, U_\gamma(\beta,\gamma)\right)^\top$. Então, os componentes do

vetor escore de β são definidos por

$$\frac{\partial \ell(\beta,\gamma)}{\partial \beta_i} = \sum_{t=1}^{n} \frac{\partial \ell_t(\mu_t,\phi_t)}{\partial \mu_t} \frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_{1t}} \frac{\partial \eta_{1t}}{\partial \beta_i}, \qquad i = 1,\ldots,k, \tag{2.12}$$

com

$$\frac{\partial \ell_t(\mu_t,\phi_t)}{\partial \mu_t} = \phi_t \left[\log\left(\frac{y_t}{1-y_t}\right) - \{\psi(\mu_t\phi_t) - \psi((1-\mu_t)\phi_t)\} \right], \tag{2.13}$$

sendo $\psi(.)$ a função digama.Assim, a Equação (2.12) reduz-se a

$$\frac{\partial \ell(\beta,\gamma)}{\partial \beta_i} = \sum_{t=1}^{n} \phi_t \left(y_t^* - \mu_t^*\right) \frac{1}{g'(\mu_t)} x_{ti}, \tag{2.14}$$

de modo que

$$y_t^* = \log\left(\frac{y_t}{1-y_t}\right) \qquad \text{e} \qquad \mu_t^* = \psi(\mu_t\phi_t) - \psi((1-\mu_t)\phi_t). \tag{2.15}$$

A forma matricial da função escore para $\beta = (\beta_1,\ldots,\beta_k)$ é definida como

$$U_\beta(\beta,\phi) = X^\top \Phi T (y^* - \mu^*),$$

em que X é a matriz de covariadas de ordem $n \times k$, de forma que a t-ésima linha de $x_t^\top$, Φ =diag$\{\phi_1,\ldots,\phi_n\}$, T = diag$\{1/g'(\mu_1),\ldots,1/g'(\mu_n)\}$, $y^* = (y_1^*,\ldots,y_n^*)^\top$ $\mu^* = (\mu_1^*,\ldots,\mu_n^*)^\top$. A função escore para γ_j, $j = 1,...,n$, é definida por

$$\frac{\partial \ell_t(\beta,\gamma)}{\partial \gamma_j} = \sum_{t=1}^{n} \frac{\partial \ell_t(\mu_t,\phi_t)}{\partial \phi_t} \frac{\mathrm{d}\phi_t}{\mathrm{d}\eta_{2t}} \frac{\partial \eta_{2t}}{\partial \gamma_j}, \tag{2.16}$$

com

$$\begin{aligned} \frac{\partial \ell_t(\mu_t,\phi_t)}{\partial \phi_t} =& \mu_t \left[\log\left(\frac{y_t}{1-y_t}\right) - \psi(\mu_t\phi_t) + \psi((1-\mu_t)\phi_t) \right] \\ &+ \log(1-y_t) - \psi((1-\mu_t)\phi_t) + \psi(\phi_t), \end{aligned} \tag{2.17}$$

Além disto, $\mathrm{d}\phi_t/\mathrm{d}\eta_{2t} = 1/h'(\phi_t)$ e $\partial\eta_{2t}/\partial\gamma_j = z_{tj}$.

Utilizando (2.15), a função escore para cada um dos parâmetros γ_j é escrita como

$$\frac{\partial \ell(\beta,\gamma)}{\partial \gamma_j} = \sum_{t=1}^{n} \left[\mu_t(y_t^* - \mu_t^*) + \log(1-y_t) - \psi((1-\mu_t)\phi_t) + \psi(\phi_t)\right] \frac{1}{h'(\phi_t)} z_{tj}, \tag{2.18}$$

podendo ser expressa por

$$\frac{\partial \ell(\beta,\gamma)}{\partial \gamma_j} = \sum_{t=1}^{n} a_t \frac{1}{h'(\phi_t)} z_{tj}, \tag{2.19}$$

em que $a_t = \mu_t(y_t^* - \mu_t^*) + \log(1-y_t) - \psi((1-\mu_t)\phi_t) + \psi(\phi_t)$. Na forma matricial, (2.19) é escrita como

$$U_\gamma(\beta,\gamma) = Z^\top H a,$$

em que Z é uma matriz $n \times q$ com $z_t^\top$ sendo a t-ésima linha, H =diag$\{1/h'(\phi_1), \ldots, h'(\phi_n)\}$ e $a = (a_1, \ldots, a_n)$.

A composição da matriz de informação de Fisher para β e ϕ necessita das derivadas de segunda ordem do logaritmo da função de verossimilhança em relação aos parâmetros. Assim, por meio da Equação (2.12), tem-se que, para i,$p = 1, \ldots, k$,

$$\begin{aligned}\frac{\partial^2 \ell(\beta,\gamma)}{\partial \beta_i \partial \beta_p} &= \sum_{t=1}^{n} \frac{\partial}{\partial \mu_t}\left(\frac{\partial \ell_t(\mu_t,\phi_t)}{\partial \mu_t}\frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_{1t}}\frac{\partial \eta_{1t}}{\partial \beta_i}\right)\frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_{1t}}\frac{\partial \eta_{1t}}{\partial \beta_p} \\ &= \sum_{t=1}^{n}\left\{\frac{\partial^2 \ell_t(\mu_t,\phi_t)}{\partial \mu_t^2}\frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_{1t}} + \frac{\partial \ell_t(\mu_t,\phi_t)}{\partial \mu_t}\frac{\partial}{\partial \mu_t}\left(\frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_{1t}}\right)\right\}\frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_{1t}} x_{ti}x_{tp}.\end{aligned}$$

Admitindo μ_t^*, apresentado em (2.15), segue que $\mathbb{E}(\partial \ell(\mu_t,\phi_t)/\partial \mu_t) = 0$. Logo,

$$\mathbb{E}\left(\frac{\partial^2 \ell(\beta,\phi)}{\partial \beta_i \partial \beta_p}\right) = \sum_{t=1}^{n} \mathbb{E}\left(\frac{\partial^2 \ell_t(\mu_t,\phi_t)}{\partial \mu_t^2}\right)\left(\frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_t}\right)^2 x_{ti}x_{tp}.$$

De (2.13), obtém-se que

$$\frac{\partial^2 \ell_t(\mu_t,\phi_t)}{\partial \mu_t^2} = -\phi_t^2\left\{\psi'(\mu_t\phi_t) + \psi'((1-\mu_t)\phi_t)\right\},$$

em que $\psi'(.)$ é a função trigama. Então,

$$\mathbb{E}\left(\frac{\partial^2 \ell(\beta,\phi_t)}{\partial \beta_i \partial \beta_p}\right) = -\sum_{t=1}^{n} \phi_t w_t x_{ti}x_{tp},$$

com

$$w_t = \phi_t\left[\psi'(\mu_t\phi_t) + \psi'((1-\mu_t)\phi_t)\right]\frac{1}{\{g'(\mu_t)\}^2}.$$

A forma matricial pode ser escrita como

$$\mathbb{E}\left(\frac{\partial^2 \ell(\beta,\phi)}{\partial \beta \partial \beta^\top}\right) = -X^\top \Phi W X,$$

de forma que W=diag$\{w_1, \ldots, w_n\}$. De (2.14), as segundas derivadas de $\ell(\beta,\gamma)$ em relação a β_i e γ_j, para $i = 1, \ldots, k$ e $j = 1, \ldots, q$, são expressas por

$$\frac{\partial^2 \ell(\beta,\gamma)}{\partial \beta_i \partial \gamma_j} = \sum_{t=1}^{n} \frac{\partial}{\partial \phi_t}\left(\phi_t(y_t^* - \mu_t^*)\frac{1}{g'(\mu_t)}x_{ti}\right)\frac{\mathrm{d}\phi_t}{\mathrm{d}\eta_{2t}}\frac{\partial \eta_{2t}}{\partial \gamma_j}$$

de modo que

$$\frac{\partial}{\partial \phi_t}\left(\phi_t(y_t^* - \mu_t^*)\frac{1}{g'(\mu_t)}x_{ti}\right) = \left\{(y_t^* - \mu_t^*) - \phi_t\left[\psi'(\mu_t\phi_t)\mu_t - \psi'((1-\mu_t)\phi_t)(1-\mu_t)\right]\right\}\frac{1}{g'(\mu_t)}x_{ti}.$$

Além disso, $\mathrm{d}\phi_t/\mathrm{d}\eta_{2t} = 1/h'(\phi_t)$ e $\partial \eta_{2t}/\partial \gamma_j = z_{tj}$. Tendo em vista que $\mathbb{E} = (y_t^* - \mu_t^*) = 0$, então

$$\mathbb{E}\left(\frac{\partial^2 \ell(\beta,\gamma)}{\partial \beta_i \partial \gamma_j}\right) = -\sum_{t=1}^{n} c_t \frac{1}{g'(\mu_t)}\frac{1}{h'(\phi_t)}x_{ti}z_{tj},$$

com $c_t = \phi_t \{\mu_t \psi'(\mu_t \phi_t) - (1-\mu_t)\psi'((1-\mu_t)\phi_t)\}$. A forma matricial é expressa por

$$\mathbb{E}\left(\frac{\partial^2 \ell(\beta,\gamma)}{\partial \beta \partial \gamma^\top}\right) = X^\top CTHZ,$$

em que C =diag$\{c_1,\ldots,c_n\}$. Tomando (2.16), as derivadas segundas de $\ell(\beta,\gamma)$ com relação a γ_j e γ_l, para $j,l = 1,\ldots,q$, são dadas por

$$\begin{aligned}\frac{\partial^2 \ell(\beta,\gamma)}{\partial \gamma j \partial \gamma_l} &= \sum_{t=1}^{n} \frac{\partial}{\partial \phi_t}\left(\frac{\partial \ell_t(\mu_t,\phi_t)}{\partial \phi_t}\frac{\mathrm{d}\phi_t}{\mathrm{d}\eta_{2t}}\frac{\partial \eta_{2t}}{\partial \gamma_j}\right)\frac{\mathrm{d}\phi_t}{\mathrm{d}\eta_{2t}}\frac{\partial \eta_{2t}}{\partial \gamma_l} \\ &= \sum_{t=1}^{n}\left(\frac{\partial^2 \ell_t(\mu_t,\phi_t)}{\partial \phi_t^2}\frac{\mathrm{d}\phi_t}{\mathrm{d}\eta_{2t}} + \frac{\partial \ell_t(\mu_t,\phi_t)}{\partial \phi_t}\frac{\partial}{\partial \phi_t}\left(\frac{\mathrm{d}\phi_t}{\mathrm{d}\eta_{2t}}\right)\right)\frac{\mathrm{d}\phi_t}{\mathrm{d}\eta_{2t}} z_{tl} z_{tj}.\end{aligned}$$

Sendo $\mathbb{E}(\partial \ell(\mu_t,\phi_t)/\partial \phi_t) = 0$, então

$$\mathbb{E}\left(\frac{\partial^2 \ell(\beta,\gamma)}{\partial \gamma_j \partial \gamma_l}\right) = \sum_{t=1}^{n} \mathbb{E}\left(\frac{\partial^2 \ell_t(\mu_t\phi_t)}{\partial \phi_t^2}\right)\left(\frac{\mathrm{d}\phi_t}{\mathrm{d}\eta_{2t}}\right)^2 z_{tl} z_{tj}.$$

Utilizando (2.17), tem-se que

$$\mathbb{E}\left(\frac{\partial^2 \ell(\beta,\gamma)}{\partial \gamma_j \partial \gamma_l}\right) = -\sum_{t=1}^{n}\left[\psi'(\mu_t\phi_t)\mu_t^2 + \psi'((1-\mu_t)\phi_t)(1-\mu_t^2) - \psi'(\phi_t)\right]\frac{1}{\{h'(\phi_t)^2\}} z_{tl} z_{tj},$$

podendo ser escrito na seguinte forma matricial:

$$\mathbb{E}\left(\frac{\partial^2 \ell(\beta,\gamma)}{\partial \gamma \partial \gamma^\top}\right) = -Z^\top D^* Z,$$

de forma que D^* =diag$\{d_1^*,\ldots,d_n^*\}$ e com

$$d_t^* = \left[\psi'(\mu_t\phi_t)\mu_t^2 + \psi'((1-\mu_t)\phi_t)(1-\mu_t^2) - \psi'(\phi_t)\right]\frac{1}{\{h'(\phi_t)^2\}}.$$

Por fim, a matriz de informação de Fisher para β e γ é dada por

$$K^* = K^*(\beta,\gamma) = \begin{pmatrix} K^*_{\beta\beta} & K^*_{\beta\gamma} \\ K^*_{\gamma\beta} & K^*_{\gamma\gamma} \end{pmatrix} = \begin{pmatrix} \phi X^\top \Phi W X & X^\top CTHZ \\ X^\top CTHZ & -Z^\top D^* Z \end{pmatrix}.$$

Em grandes amostras e sob certas condições de regularidade (SEN; SINGER,1993),

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} \sim N_{k+q}\left(\begin{pmatrix} \beta \\ \gamma \end{pmatrix}, K^{-1}\right),$$

aproximadamente, em que $\hat{\beta}$ e $\hat{\gamma}$ são os estimadores de máxima verossimilhança de β e γ, respectivamente. A obtenção destes estimadores se dá através do seguinte sistema

$$\begin{cases} U_\beta(\beta,\gamma) = 0 \\ U_\gamma(\beta,\gamma) = 0 \end{cases}.$$

Analogamente ao modelo de regressão beta com dispersão fixa, aqui os estimadores de máxima verossimilhança não podem ser escritos por meio de uma forma fechada, sendo necessário o uso de algoritmos de otimização não linear. Dessa forma, devem ser especificados os valores iniciais a serem utilizados no esquema iterativo. Para tal, Ferrari *et al.* (2011) sugerem que o valor inicial para β seja o mesmo utilizado no modelo com dispersão constante, ou seja, $\beta^{(0)} = (X^\top X)^{-1} X^\top g(y_t)$. Enquanto para ϕ_t, $t = 1, ..., n$, o chute inicial é dado por

$$\phi^{(0)} = \frac{\breve{\mu}_t(1-\breve{\mu}_t)}{\breve{\sigma}_t^2} - 1,$$

de modo que $\breve{\mu}_t = g^{-1}(x_t^\top (X^\top X)^{-1} X^\top z)$ e $\breve{\sigma}_t^2 = \frac{\breve{e}^\top \breve{e}}{(n-k)g'(\breve{\mu}_t)^2}$, com $\breve{e} = z - X(X^\top X)^{-1} X^\top z$ e $z = g(y_t)$. Considerou-se que $h(\phi_t) = \sum_{g=1}^{1} z_{tj}\gamma_j$ e a estimativa inicial de γ expressa por $(Z^\top Z)^{-1} Z^\top k$, de forma que $k = (h(\phi_1^{(0)}), ..., h(\phi_1^{(n)}))^\top$.

2.5 RESÍDUOS

Tomando os resíduos propostos por Espinheira *et al.* (2008), para o modelo de regressão beta com dispersão constante, os mesmos foram foi estendido por Ferrari *et al.* (2011) para os casos em que a dispersão varia ao longo das observações, além de apresentarem um novo resíduo para os modelos de regressão beta com dispersão variável. Nas seções seguintes, serão apresentados o resíduo ponderado, resíduo ponderado padronizado e resíduo combinado.

2.5.1 Resíduo Ponderado

Para o modelo de regressão beta com dispersão variável, o processo iterativo Scoring de Fisher para a estimativa de β é dado por

$$\beta^{(m+1)} = \beta^{(m)} + \left(X^\top \Phi^{(m)} W^{(m)} X\right)^{-1} X^\top \Phi^{(m)} T^{(m)} (y^* - \mu^{*(m)}),$$

Assumindo a forma de um processo iterativo de mínimos quadrados reponderados o processo é escrito como

$$\beta^{(m+1)} = \left(X^\top \Phi^{(m)} W^{(m)} X\right)^{-1} X^\top \Phi^{(m)} W^{(m)} u_1^{(m)}, \tag{2.20}$$

em que $u_1^{(m)} = \eta_1^{(m)} + W^{-1(m)} T^{(m)} \left(y^* - \mu^{*(m)}\right)$. Após a convergência, obtém-se

$$\hat{\beta} = \left(X^\top \hat{\Phi} \hat{W} X\right)^{-1} X^\top \hat{\Phi} \hat{W} u_1, \tag{2.21}$$

com

$$u_1 = \hat{\eta}_1 + \hat{W}^{-1} \hat{T} (y^* - \hat{\mu}^*),$$

que pode ser observado como um estimador de mínimos quadrados considerado a regressão linear de $\Phi^{1/2}W^{1/2}u_1$ em $\Phi^{1/2}W^{1/2}X$. Assim, o resíduo de mínimos quadrados dessa regressão é definido por

$$r^{\beta} = \hat{\Phi}^{1/2}\hat{W}^{1/2}u_1 - \hat{\Phi}^{1/2}\hat{W}^{1/2}X\hat{\beta} = \hat{\Phi}^{1/2}\hat{W}^{1/2}(u_1 - X\hat{\beta}) = \hat{\Phi}^{1/2}\hat{W}^{-1/2}\hat{T}(y^* - \hat{\mu}^*). \quad (2.22)$$

Então, o resíduo da t-ésima observação é dado por

$$r_t^{\beta} = \frac{(y_t^* - \hat{\mu}_t^*)}{\sqrt{\hat{\upsilon}_t}}, \quad (2.23)$$

sendo

$$\upsilon_t = \psi'(\mu_t\phi) + \psi'((1-\mu_t)\phi). \quad (2.24)$$

2.5.2 Resíduo Ponderado Padronizado

Semelhante ao resíduo proposto para o modelo de regressão beta com dispersão constante, a extensão proposta por Ferrari *et al.* (2011) para resíduo ponderado padronizado baseia-se na variância de u_1 para apresentar a padronização do resíduo (2.23). Sendo $\mathrm{Cov}(\hat{\beta}) \approx (X^\top \Phi W X)^{-1}$, $\hat{W} \approx W$, $\hat{\Phi} \approx \Phi$ e reescrevendo a equação (2.21) como

$$\left(X^\top \hat{\Phi}\hat{W}X\right)\hat{\beta} = X^\top \hat{\Phi}\hat{W}u_1, \quad (2.25)$$

obtém-se

$$\begin{aligned}
\left(X^\top \Phi W X\right)\mathrm{Cov}(\hat{\beta})\left(X^\top \Phi W X\right)^\top &\approx \left(X^\top \Phi W\right)\mathrm{Cov}(u_1)\left(X^\top \Phi W\right)^\top \\
\left(X^\top \Phi W X\right)\left(X^\top \Phi W X\right)^{-1}\left(X^\top \Phi W X\right)^\top &\approx \left(X^\top \Phi W\right)\mathrm{Cov}(u_1)\left(W\Phi X\right)^\top \\
\mathrm{Cov}(u_1) &\approx W^{-1}\Phi^{-1}.
\end{aligned}$$

Então, $\widehat{\mathrm{Cov}}(u_1) \approx \hat{W}^{-1}\hat{\Phi}^{-1}$. Tomando (2.22), o mesmo pode ser expresso como

$$\begin{aligned}
r^{\beta} &= \hat{\Phi}^{1/2}\hat{W}^{1/2}\left(u - 1 - X\left(X^\top \hat{\Phi}\hat{W}X\right)^{-1}X^\top \hat{\Phi}\hat{W}u_1\right) \\
&= \left(\hat{\Phi}^{1/2}\hat{W}^{1/2} - \hat{\Phi}^{1/2}\hat{W}^{1/2}X\left(X^\top \hat{\Phi}\hat{W}X\right)^{-1}X^\top \hat{\Phi}\hat{W}\right)u_1.
\end{aligned}$$

Admitindo $\widehat{\text{Cov}}(u_1) \approx \hat{W}^{-1}\hat{\Phi}^{-1}$, $\hat{W} \approx W$ e $\hat{\Phi} \approx \Phi$, segue que

$$\begin{aligned}
\text{Cov}(r^{\beta}) &\approx \left(\Phi^{1/2}W^{1/2} - \Phi^{1/2}W^{1/2}X\left(X^{\top}\Phi WX\right)^{-1}X^{\top}\Phi W\right)\text{Cov}(u_1) \\
&\times \left(\Phi^{1/2}W^{1/2} - \Phi^{1/2}W^{1/2}X\left(X^{\top}\Phi WX\right)^{-1}X^{\top}\Phi W\right)^{\top} \\
&\approx \left(I - \Phi^{1/2}W^{1/2}X\left(X^{\top}\Phi^{1/2}WX\right)^{-1}X^{\top}\Phi^{1/2}W^{1/2}\right) \\
&\approx (I-G).
\end{aligned}$$

Assim, $\widehat{\text{Cov}}(r^{\beta}) \approx I - \hat{G}$, com $\hat{G} = \hat{\Phi}^{1/2}\hat{W}^{1/2}X\left(X^{\top}\hat{\Phi}^{1/2}\hat{W}X\right)^{-1}X^{\top}\hat{\Phi}^{1/2}\hat{W}^{1/2}$, com G sendo a matriz de projeção de $\Phi^{1/2}W^{1/2}u_1$ contra $\Phi^{1/2}W^{1/2}X$, sendo uma matriz simétrica e idempotente. Além disso, ressalta-se que $\Phi = \{\phi_1, \ldots, \phi_n\}$. Por fim, a padronização do resíduo é definida por

$$r_{pt}^{\beta} = \frac{r_t^{\beta}}{\sqrt{\widehat{\text{Cov}}(r_t^{\beta})}} = \frac{y_t^* - \hat{\mu}_t^*}{\sqrt{(\hat{\upsilon}_t(1-\hat{g}_{tt}^*))}},$$

de modo que g_{tt}^* é o t-ésimo elemento da diagonal principal de G e υ_t encontra-se definido em (2.24).

2.5.3 Resíduo *Variance*

Nossa proposta de resíduo está baseada no processo iterativo Scoring de Fisher para γ. Assim, o resíduo *variance* considera o seguinte processo iterativo para γ:

$$\gamma^{(m+1)} = \gamma^{(m)} + \left(K_{\gamma\gamma}^*\right)^{-1} U_{\gamma}^{(m)}(\beta, \gamma),$$

com m indicando os passos necessários até a convergência do processo. Assim, processo iterativo de interesse é dado por

$$\gamma^{(m+1)} = \gamma^{(m)} + \left(Z^{\top}D^{*(m)}Z\right)^{-1} Z^{\top}H^{(m)}a^{(m)} \tag{2.26}$$

de modo que $a = (a, \ldots, a_n)^{\top}$ com

$$a_t = \mu_t(y^* - \mu_t^*) + \log(1-y_t) - \psi((1-\mu_t)\phi_t) + \psi(\phi_t). \tag{2.27}$$

Considerando um processo de mínimos quadrados reponderados, o processo em (2.26) tem a seguinte forma:

$$\gamma^{(m+1)} = \left(Z^{\top}D^{*(m)}Z\right)^{-1} Z^{\top}D^{*(m)}u_2^{(m)}$$

com $u_2^{(m)} = \eta_2^{(m)}{}_2 + D^{*(m)}H^{(m)}a^{(m)}$ e $\eta_2^{(m)} = (\eta_{21}, \ldots, \eta_{2n}) = Z\gamma$. Na convergência do processo

$$\hat{\gamma} = \left(Z^\top \hat{D}^* Z\right)^{-1} Z^\top \hat{D}^* \hat{u_2}$$

em que $\hat{u_2} = \hat{\eta}_2 + \hat{D}^{*-1}\hat{H}\hat{a}$, podendo ser observado como um estimador de mínimos quadrados considerado a regressão linear de $D^{*1/2}Z$ em $D^{*1/2}u_2$. Assim, o resíduo *variance* ordinário dessa regressão é definido por

$$r^\gamma = D^{*1/2}u_2 - D^{*1/2}Z\gamma = D^{*1/2}(\hat{\eta}_2 + \hat{D}^*\hat{H}\hat{a} - \hat{\eta}_2) = (\hat{D}^*)^{-1/2}\hat{H}\hat{a}.$$

Desta forma, o resíduo da t-ésima observação é

$$r_t^\gamma = \frac{\hat{a}_t}{\sqrt{\hat{\xi}_t}}, \tag{2.28}$$

em que $\xi_t = \mathrm{Var}(a_t) = \mu_t^2\psi'(\mu_t\phi_t) + (1-\mu_t)^2\psi'((1-\mu_t)\phi_t)$ e a_t está definido em (2.27).

2.5.4 Resíduo Combinado

Carregando a informação sobre todos os parâmetros do modelo, o resíduo combinado é construído tomando por base o processo iterativo Scoring de Fisher de β e γ, apresentados nas Equações (2.26) e (2.20), respectivamente. Logo, a t-ésima observação do resíduo combinado é expressa por

$$r_t^{\beta\gamma} = (y_t^* - \hat{\mu}^*) + \hat{a}_t, \tag{2.29}$$

com $\hat{a}_t$ definido em (2.27). Além disso, a padronização do resíduo da t-ésima observação é expressa como

$$r_t^{\beta\gamma} = \frac{(y_t^* - \hat{\mu}^*) + \hat{a}_t}{\sqrt{\hat{\zeta}_t}}, \tag{2.30}$$

em que $\zeta_t = (1+\mu_t)^2\psi'(\mu_t\phi_t) + \mu_t^2\psi'((1-\mu_t)\phi_t) - \psi'(\phi_t)$, com $\hat{\zeta}_t$ dado por ζ_t, sendo este avaliado em $\hat{\mu}_t$ e $\hat{\phi}_t$. Quando admitido um modelo de dispersão constante, considera-se que $\phi_1 = \cdots = \phi_n = \phi$.

3 NOVOS RESÍDUOS PARA O MODELO DE REGRESSÃO BETA

3.1 FAMÍLIAS EXPONENCIAIS

Uma família $P = \left\{ p_\theta, \theta \in \Theta \subset \mathbb{R}^k \right\}$ de distribuições compõem uma classe exponencial s-dimensional caso as densidades em P são escritas como

$$p_\theta(x) = \exp\left\{ \sum_{j=1}^{s} \eta_j(\theta) T_j(x) - B(\theta) \right\} h(x),$$

com respeito a uma medida μ. Nesse sentido, $\eta_j(.)$, $j = 1, \ldots, s$, e $B(.)$ são funções reais de θ, além de $T_j(x)$ serem estatísticas com valor real com x sendo um ponto no espaço amostral χ, o suporte da função de densidade, ou seja, valores x de X onde p_θ é não nula, e h(X) é determinada sobre χ a valores reais não negativos.

Frequentemente, é mais conveniente usar $\eta_j(.)$, $j = 1, \ldots, s$, como parâmetros e escrever a densidade de X em sua forma canônica

$$p_\eta(x) = \exp\left\{ \sum_{j=1}^{s} \eta_j T_j(x) - \mathrm{A}(\eta) \right\} h(x).$$

Essa representação também é denominada como família exponencial canônica, de modo que os parâmetros η_j são denominados de parâmetros naturais. Nesse sentido, o espaço paramétrico natural, sendo denotado por Δ, é definido como

$$\Delta = \left\{ \eta = (\eta_1, \ldots, \eta_s) : e^{\mathrm{A}(\eta)} = \int e^{\sum_{j=1}^{s} \eta_j T_j(x)} h(x) \mathrm{d}\mu(x) < \infty \right\},$$

de modo que $\forall \eta \in \Delta$, p_η é uma densidade (LEHMANN; CASELLA,1998). O seguinte teorema apresenta uma propriedade útil das famílias exponenciais, sendo provado em Nielsen (1978) e Lehmann (1986).

Teorema 3.1 (Teorema da Função Integrável). *Para qualquer função integrável f e qualquer* η *ponto interior de* Δ *a integral*

$$\int f(x) \exp\left\{ \sum_{j=1}^{s} \eta_j T_j(x) - \mathrm{A}(\eta) \right\} h(x) \mathrm{d}\mu(x)$$

é contínua e tem derivadas de todas as ordens com respeito aos $\eta's$*, podendo estas ser obtidas diferenciando-se sob o sinal da integral.*

Diante do exposto, a função de densidade beta pode ser expressa nos termos da família exponencial biparamétrica canônica. Então,

$$f(y_t; \mu_t, \phi_t) = \exp\{\tau_1 T_1 + \tau_2 T_2 - \mathrm{A}(\tau)\} \left(\frac{1}{y_t(1-y_t)} \right),$$

em que $\tau = (\tau_1, \tau_2) = (\mu_t\phi_t, \phi_t)$,

$$(T_1, T_2) = (Y_t^*, Y_t^{**}) = \left(\log\left(\frac{y_t}{1-y_t}\right), \log(1-y_t)\right) \qquad \text{e}$$
$$\mathrm{A}(\tau) = \{-\log\Gamma(\phi_t) + \log\Gamma(\mu_t\phi_t) + \log\Gamma((1-\mu_t)\phi_t)\}.$$

Dessa forma, tem-se que

$$\mathbb{E}(T_1) = \mathbb{E}(Y_t^*) = \partial\mathrm{A}(\tau)/\partial\tau_1 = \psi(\mu_t\phi_t) - \psi((1-\mu_t)\phi_t) = \varepsilon_t^*, \tag{3.1}$$

$$\mathbb{E}(T_2) = \mathbb{E}(\log(1-Y_t)) = \partial\mathrm{A}(\tau)/\partial\tau_2 = \psi((1-\mu_t)\phi_t) - \psi(\phi_t) = \varepsilon_t^{**}, \tag{3.2}$$

$$\mathrm{Var}(T_1) = \mathrm{Var}(Y_t^*) = \partial^2\mathrm{A}(\tau)/\partial\tau_1^2 = \psi'(\mu_t\phi_t) + \psi'((1-\mu_t)\phi_t) = v_t, \tag{3.3}$$

$$\mathrm{Var}(T_2) = \mathrm{Var}(\log(1-Y_t)) = \partial^2\mathrm{A}(\tau)/\partial\tau_1^2 = \psi'((1-\mu_t)\phi_t) - \psi'(\phi_t) = \xi_t \tag{3.4}$$

e

$$\mathrm{Cov}(T_1, T_2) = \partial^2\mathrm{A}(\tau)/\partial\tau_1\partial\tau_2 = -\psi'((1-\mu_t)\phi_t). \tag{3.5}$$

Como dito anteriormente, tais quantidades serão utilizadas para a definição dos novos resíduos. Assim, é notável que essas quantidades estão relacionadas a estimação por máxima verossimilhança.

3.2 NOVOS RESÍDUOS

Apresentamos, para a estrutura de regressão beta, estimadores de mínimos quadrados ordinários da regressão linear da variável resposta y_t transformada pelas estatísticas da família beta (T_1, T_2) para os submodelos da média e dispersão. Para β e γ os estimadores são

$$\begin{aligned} \hat{\beta} &= (X^\top X)^{-1} X^\top Y^* \\ \hat{\gamma} &= (Z^\top Z)^{-1} Z^\top Y^{**}, \end{aligned} \tag{3.6}$$

de tal forma que $Y^* = (Y_1^*, ..., Y_n^*)^\top$, com $Y_t^* = (\log\{Y_t/(1-Y_t)\})$ e $Y^{**} = (Y_1^{**}, ..., Y_n^{**})^\top$, com $Y_t^{**} = (\log(1-Y_t))$.

Uma vez definido o estimador de mínimos quadrados de β, o resíduo ordinário da regressão de Y^* contra X é dado por

$$\begin{aligned} Y^* - X\hat{\beta} &= Y^* - X\left[(X^\top X)^{-1}X^\top Y^*\right] \\ &= \left[\mathbb{I}_n - X(X^\top X)^{-1}X^\top\right]Y^* \\ &= Y^* - H_1 Y^*, \end{aligned} \tag{3.7}$$

com $H_1 = X(X^\top X)^{-1}X^\top$ sendo a matriz de projeção e $\mathbb{I}_n$ a matriz identidade de ordem n. O resíduo ordinário (3.7) tem a forma dada por $r^\beta = Y^* - \hat{\varepsilon}^*$ e assumindo as estatísticas apresentadas em (3.1) e (3.3), para a t-ésima observação o resíduo padronizado do submodelo da média é dado por

$$r_t^\beta = \frac{\log\{Y_t/(1-Y_t)\} - \psi(\hat{\mu}_t\hat{\phi}_t) - \psi((1-\hat{\mu}_t)\hat{\phi}_t)}{\sqrt{\psi'(\hat{\mu}_t\hat{\phi}_t) + \psi'((1-\hat{\mu}_t)\hat{\phi}_t)}} \quad = \frac{Y_t^* - \hat{\varepsilon}_t^*}{\sqrt{\hat{v}_t}}. \tag{3.8}$$

Para o submodelo da precisão (γ), o resíduo ordinário da regressão de Y^{**} contra Z é dado por

$$\begin{aligned} Y^{**} - Z\hat{\gamma} &= Y^{**} - Z\left[(Z^\top Z)^{-1}Z^\top Y^{**}\right] \\ &= \left[\mathbb{I}_n - Z(Z^\top Z)^{-1}Z^\top\right]Y^{**} \\ &= Y^{**} - H_2 Y^{**}, \end{aligned} \tag{3.9}$$

com $H_2 = Z(Z^\top Z)^{-1}Z^\top$ sendo a matriz de projeção e $\mathbb{I}_n$ a matriz identidade de ordem n. Tal resíduo (3.9) tem a forma dada por $r^\gamma = Y^{**} - \hat{\varepsilon}^{**}$ e assumindo as estatísticas apresentadas em (3.2) e (3.4), para a t-ésima observação o resíduo padronizado do submodelo da precisão é dado por

$$r_t^\gamma = \frac{\log(1-Y_t) - \psi((1-\hat{\mu}_t)\hat{\phi}_t) - \psi(\hat{\phi}_t)}{\sqrt{\psi'((1-\hat{\mu}_t)\hat{\phi}_t) - \psi'(\hat{\phi}_t)}} \quad = \frac{(Y_t^{**} - \hat{\varepsilon}_t^{**})}{\sqrt{\hat{\xi}_t}}. \tag{3.10}$$

A esse resíduo damos o nome de *Resíduo Variance 2*.

Finalmente, propomos um terceiro resíduo está relacionado com os submodelos da média e dispersão (β, γ). Vamos admitir que $Y^\dagger = (Y^* + Y^{**})$ e, assim, a regressão de $Y^\dagger$ contra M produz o resíduo ordinário de mínimos quadrados $Y^\dagger - H_3 Y^\dagger$, com a matriz de projeção denotada por $H_3 = M(M^\top M)^{-1}M^\top$. Tomando as equações (3.1) e (3.2), verifica-se que

$$H_3 Y^\dagger = \mathbb{E}(Y^* + Y^{**}) = \mathbb{E}(\log\{Y_t/(1-Y_t)\} + \log(1-Y_t)) = (\hat{\varepsilon}^* + \hat{\varepsilon}^{**}). \tag{3.11}$$

Além disso,

$$\begin{aligned}\mathrm{Var}(Y^* + Y^{**}) &= \mathrm{Var}(\log\{Y_t/(1-Y_t)\} + \log(1-Y_t)) \\ &= \mathrm{Var}(\log\{Y_t/(1-Y_t)\}) + \mathrm{Var}(\log(1-Y_t)) + 2 \times \mathrm{Cov}(\log\{Y_t/(1-Y_t)\}, \log(1-Y_t)) \\ &= (\psi'(\mu_t\phi_t) - \psi'(\phi_t)) \\ &= \varphi_t.\end{aligned} \tag{3.12}$$

A padronização do resíduo é dada por

$$r_t^{\beta\gamma} = \frac{(Y_t^* - \hat{\varepsilon}_t^*) + (Y_t^{**} - \hat{\varepsilon}_t^{**})}{\sqrt{\hat{\varphi}_t}}. \tag{3.13}$$

Desta forma, apresentamos o *Resíduo Combinado 2*.

4 CRITÉRIOS DE SELEÇÃO DE MODELOS

No processo de seleção de modelos no contexto da regressão beta busca-se selecionar modelos de regressão com alto poder preditivo e variância bem explicada. É notado que a omissão de covariáveis importantes em um modelo de regressão comumente resulta em viés sobre as estimativas dos coeficientes do modelo, bem como quando uma covariável é inserida erroneamente nos preditores. Nesse contexto, ressalta-se que a existência das seguintes questões associadas à qualidade dos modelos, pode-se citar, a exemplo, dispersão erroneamente modelada, abandono da suposição da não linearidade nos preditores, entre outras. Finalmente, um aspecto fundamental sobre osmodelos de regressão é a especificação correta da distribuição de probabilidade da variável resposta. Por sua vez, todos os problemas acima citados afetam a variabilidade na modelagem da regressão. Além disso, o último erro de especificação supracitado é a causa mais provável para a sensibilidade das estimativas de máxima verossimilhança a casos influentes.Caso os resultados inferenciais dependam de apenas uma observação ou de um pequeno conjunto de observação, de forma que o esquema de estimação de modelo possui falta de robustez, então, verifica-se que as conclusões inferenciais mudam significativamente em função da ocorrência de tais casos influentes no conjunto de dados. Tal informação pode ser avaliada, de forma empírica, através da análise de influência local e, mais profundamente, fazendo o estudo da função de influência da distribuição.

Comumente, diversas versões da estatística R^2 não conseguem apontar se as predições são tendenciosas. Esta informação é tipicamente estudada por meio da análise visual dos gráficos de resíduos. Ademais, medidas baseadas na estatítica PRESS (*Predictive Residual Sum of Squares*), apresentada por Allen (1974), levam em conta os resíduos e são úteis no fornecimento de informações sobre o viés de $\hat{\mu}_t$. Baseada na PRESS, a estatística P^2, tipicamente denominada como a estatística R^2 de predição, é mais fácil de ser interpretada. Essa medida, P^2, foi definida por Quan (1988) como um critério Q^2 capaz de validar o poder preditivo do modelo, sem que ocorra a seleção de amostras adicionais, ou considerar subconjuntos dos dados em conjuntos de formação e validação, tal como é requerido em um esquema de validação cruzada. Com isso, a avaliação da medida P^2 pode evitar a ocorrência de modelos que aparentam apresentar um bom ajuste para um conjunto de dados específico, mas que fornece predições imprecisas para novas observações. Em tais casos, diz-se que o modelo apresentado possui *Overfitting*. Tipicamente, um cenário de *overfitting* pode ser verificado quando uma versão de R^2 apresenta um valor bastante alto, próximo de um, enquanto o valor da medida P^2 chega

a apresentar valores próximos de zero. Ademais, conforme apresentado por Espinheira *et al.* (2019), a estatística P^2 também fornece informação sobre o esquema de estimação por máxima verossimilhança, ocasionada pela ocorrência de observações influentes no conjunto de dados.

Quanto à questão da variabilidade, a mesma é bem identificada pelos coeficientes de determinação e suas versões corrigidas. Dentre as várias versões de tais medidas, serão abordadas as propostas de apresentadas por Nagelkerke (1991) e Ferrari e Cribari-Neto (2004), R^2_{LR} e R^2_{FC}, respectivamente. O R^2_{LR} é expresso por um menos a razão de L_{null} (função de máxima verossimilhança sem regressores) e L_{fit} (função de máxima verossimilhança do modelo investigado). Nesse sentido, a razão de verossimilhanças indica o nível de melhoria em relação ao modelo nulo, que é oferecido pelo modelo proposto. Assim, caso o modelo proposto possua uma verossimilhança elevada, tem-se um indicio de que a estrutura investigada é consideravelmente melhor do que o modelo nulo. Nagelkerke (1991) aponta que a medida R^2 pode ser interpretada como a proporção da variabilidade explicada pelo modelo, enquanto que, alternativamente, 1-R^2 vem a ser a variação não explicada pelo mesmo. Ainda, Nagelkerke (1991) também evidencia que, frequentemente, a variação deve ser explicada como qualquer medida extendida para qual uma distribuição não é degenerada. No entanto, a interpretação apresentada para a medida R^2 se mostra bastante superficial em termos de variabilidade do modelo.

Geralmente, o R^2_{LR} tente a ser superior a outras versões R^2, a medida que a comparação do modelo proposto é realizada com o modelo nulo. Abstraindo-se da questão da variabilidade do modelo, o R^2_{LR} é uma métrica importante, uma vez que leva em conta a probabilidade do modelo estimado ser o verdadeiro processo de geração dos dados. Adicionalmente, a mesma tem uma relação com critérios de relação, os quais, pode-se citar o Critério de Informação de Akaike (AIC), apresentado por Akaike (1973), o Critério Baysiano de Schwarz (BIC), proposto por Schwarz (1978). Bayer e Cribari-Neto (2017) fornecem um extenso estudo, em que avaliam vários critérios de seleção e informação para modelos de regressão beta. Em adição, os autores definiram um esquema de seleção rápida em duas fases, assumindo tanto os submodelos da média, quanto os da dispersão.

Ainda no que se refere à variabilidade, Ferrari e Cribari-Neto (2004) propuseram uma medida intuitiva da variabilidade explicada pelo modelo estimado, denominada R^2_{FC}. Essa medida é baseada no coeficiente de correlação entre $g(y)$ e $\hat{\eta}$, de modo que, quanto maior é esta correlação, maior vem ser a covariância de $g(y)$ e $\hat{\eta}$, indicando que mudanças expressivas ocorrem em $g(y)$ e $\hat{\eta}$ na mesma direção. Adicionalmente, vale notar que quando R^2_{FC}= 1 representa o alinhamento perfeito $g(y)$ e $\hat{\eta}$, e portanto entre $\hat{\mu}$ e y. Além do mais, os valores da

correlação dependem da variância de $g(y)$ e da variância de $\hat{\eta}$, evidenciando que quanto menor a variância do modelo estimado (Var($\hat{\eta}$)), maiores são as chances do critério R^2_{FC} apresentar valores elevados, embora o R^2_{FC} pode ser penalizado caso a variância de $g(y)$ seja alta, fator importante a ser considerado durante a etapa de ajuste do modelo. Devido a isso, é observado que o critério R^2_{FC} assume valores tipicamente menores que outras versões de R^2. Por outro lado, isto pode indicar que o R^2_{FC} é uma estatística mais rigorosa e eficaz para o processo de escolha de um melhor modelo, em termos de variabilidade (ou variância da resposta bem estimada).

Uma relação a ser destacada entre as medidas apresentadas é que quando R^2_{FC} é consideravelmente menor que R^2_{LR}, tem-se forte indício de erro na especificação do submodelo da média, tendo em vista que a correlação entre $g(y)$ e $\hat{\eta}$ é baixa e η pode estar mal especificado. Outra relação a ser abordada é entre a variabilidade e o viés. Esse aspecto pode ser notado quando P^2 for alto e não considerando o baixo valor de R^2_{FC}, tem-se que modelagem apresenta um bom desempenho quanto ao poder preditivoIndividualmente, baixos valores de R^2_{FC} evidenciam uma má especificação no submodelo da média que pode ser resultante, por exemplo, da falta de covariadas importantes, ou devido a utilização de uma função de ligação inadequada, ou uma má especificação sobre a expressão matemática usada para definir o modelo da média.

Uma vez que o R^2_{FC} não é pequeno, $\hat{\mu}$ e y estão próximos e o problema não está localizado no submodelo da média. Tem-se uma evidência de que a má especificação está localizada no modelo postulado para a dispersão. Como o gráfico normal de probabilidades com envelope simulado possui uma relação com a variância do modelo postulado, o mesmo pode ser utilizado para confirmar a presença de má especificação no modelo. Sem dúvida, a análise residual contém informações que dizem respeito tanto ao viés como à variabilidade do modelo ajustado.

4.1 CRITÉRIOS R^2 PARA AVALIAÇÃO DA QUALIDADE DO AJUSTE

Tendo em vista a avaliação da qualidade do ajuste do modelo de regressão beta, Ferrari e Cribari-Neto (2004) apresentaram uma proposta de R^2, denominada R^2_{FC}, que consiste no quadrado do coeficiente de correlação amostral entre $g(y)$ e $\hat{\eta} = X\hat{\beta}$, sendo $\hat{\beta}$ dado pelo estimador de máxima verossimilhança de β. A relação entre a correlação de $g(y)$ e $\hat{\eta}$ depende de suas variâncias, dessa forma, caso o modelo estimado possua pouca variabilidade, espera-se que a medida R^2_{FC} apresente valores altos. É importante apontar para a possibilidade da variância de $g(y)$ ser alta, de modo a penalizar o critério e fazendo com que este venha a assumir valores

consideravelmente mais baixos do que outras versões apresentadas. Um segundo aspecto a ser notado é que altos valores da correlação implicam alta covariância, assim, os valores das medidas de $g(y)$ e $\hat{\eta}$ alteram-se conjuntamente no mesmo sentido. Com a finalidade de comparar a qualidade do ajuste de modelos com diferentes números de regressões, Bayer e Cribari-Neto (2017) propuseram versão corrigida do R^2_{FC}, inserindo um termo de penalização para a inclusão de novas covariáveis no modelo, similar ao R^2 ajustado para o modelo de regressão linear. A versão corrigida, $R^2_{FC_c}$, tem a seguinte forma:

$$R^2_{FC_c} = 1 - \left(1 - R^2_{FC}\right)\left(\frac{n-1}{n-p}\right),$$

de modo que $p = k + r$ é o quantitativo de parâmetros do modelo, com k e r o número de parâmetros do modelo da média e da precisão, respectivamente.

Além do critério R^2_{FC} e sua versão corrigida, uma outra generalização do R^2 é apresentada em Nagelkerke (1971). A medida R^2_{LR} tem por base a utilização da razão de verossimilhanças e é definida por

$$R^2_{LR} = 1 - \left(\frac{L_{null}}{L_{fit}}\right),$$

em que L_{null} é a função de máxima verossimilhança sem regressores e L_{fit} é a função de máxima verossimilhança do modelo investigado. Nesse sentido, quanto maior for L_{fit} em relação a L_{null}, observa-se o indicativo da qualidade do modelo ajustado em relação ao modelo nulo. Por sua vez, menor será o valor da razão de verossimilhanças, fazendo com que a medida R^2_{LR} tenha valores próximos de um. Com ênfase no modelo de regressão beta, tomando a proposta de Bayer e Cribari-Neto (2017), pode-se considerar a seguinte expressão para a versão corrigida de R^2_{LR}

$$R^2_{LR_c} = 1 - \left(1 - R^2_{LR}\right)\left(\frac{n-1}{n-p}\right),$$

para os casos em que a dispersão é constante e

$$R^2_{LR_c} = 1 - \left(1 - R^2_{LR}\right)\left[\frac{n-1}{n-(1+\alpha)k-(1-\alpha)r}\right],$$

quando a modelagem trata da dispersão variável, com $\alpha \in (0,1)$ e $\delta > 0$. Ainda, levando em consideração estudos de simulação, Bayer e Cribari-Neto (2017) evidenciam que $\alpha = 0,4$ e $\delta = 1$.

4.1.1 Novos critérios R^2 para avaliação da qualidade do ajuste

Tomando por base a proposta apresentada por Ferrari e Cribari-Neto (2004) e Bayer e Cribari-Neto (2017), nosso critério consiste na utilização do quadrado do coeficiente de

correlação de Pearson para os submodelos da média e da dispersão. Desta forma, buscando avaliar qual a proporção da variabilidade total da resposta é correspondente a regressão, propomos para o submodelo da média o seguinte critério R^2:

$$\begin{aligned}
\omega^2_{y^*,\hat{\varepsilon}^*} &= \left(\frac{\sum_{t=1}^{n}(y_t^* - \bar{y}^*)(\hat{\varepsilon}_t^* - \bar{y}^*)}{\sqrt{\sum_{t=1}^{n}(y_t^* - \bar{y}^*)^2 \sum_{t=1}^{n}(\hat{\varepsilon}_t^* - \bar{y}^*)^2}} \right)^2 \\
&= \left(\frac{\left(\sum_{t=1}^{n}(\hat{\varepsilon}_t^* y_t^* - \hat{\varepsilon}_t^* \bar{y}^* - \bar{y}^* y_t^* + \bar{y}^*)\right)^2}{\sum_{t=1}^{n}(y_t^* - \bar{y}^*)^2 \sum_{t=1}^{n}(\hat{\varepsilon}_t^* - \bar{y}^*)^2} \right) \\
&= \left(\frac{\left(\sum_{t=1}^{n} \hat{\varepsilon}_t^* y_t^* - \sum_{t=1}^{n} \hat{\varepsilon}_t^* \bar{y}^* - \sum_{t=1}^{n} \bar{y}^* y_t^* + \sum_{t=1}^{n} \bar{y}^{*^2}\right)^2}{\sum_{t=1}^{n}(y_t^* - \bar{y}^*)^2 \sum_{t=1}^{n}(\hat{\varepsilon}_t^* - \bar{y}^*)^2} \right) \\
&= \left(\frac{\left(\sum_{t=1}^{n} \hat{\varepsilon}_t^* y_t^* - \bar{y}^* \sum_{t=1}^{n} \hat{\varepsilon}_t^* - n\bar{y}^2 + n\bar{y}^2\right)^2}{\sum_{t=1}^{n}(y_t^* - \bar{y}^*)^2 \sum_{t=1}^{n}(\hat{\varepsilon}_t^* - \bar{y}^*)^2} \right) \\
&= \left(\frac{\left(\sum_{t=1}^{n} \hat{\varepsilon}_t^* y_t^* - \bar{y}^* \sum_{t=1}^{n} \hat{\varepsilon}_t^*\right)^2}{\sum_{t=1}^{n}(y_t^* - \bar{y}^*)^2 \sum_{t=1}^{n}(\hat{\varepsilon}_t^* - \bar{y}^*)^2} \right) \\
&= \left(\frac{\left(\sum_{t=1}^{n} \hat{\varepsilon}_t^* y_t^* - \bar{y}^* \sum_{t=1}^{n} \hat{\varepsilon}_t^*\right)^2}{\sum_{t=1}^{n}(y_t^* - \bar{y}^*)^2 \sum_{t=1}^{n}(\hat{\varepsilon}_t^* - \bar{y}^*)^2} \right) = \mathrm{R}^2_{EC_m}
\end{aligned} \tag{4.1}$$

Analogamente, para o submodelo da dispersão temos que

$$\omega^2_{y^{**},\hat{\varepsilon}^{**}} = \left(\frac{\left(\sum_{t=1}^{n} \hat{\varepsilon}_t^{**} y_t^{**} - \bar{y}^{**} \sum_{t=1}^{n} \hat{\varepsilon}_t^{**}\right)^2}{\sum_{t=1}^{n}(y_t^* - \bar{y}^{**})^2 \sum_{t=1}^{n}(\hat{\varepsilon}_t^{**} - \bar{y}^{**})^2} \right) = \mathrm{R}^2_{EC_v}. \tag{4.2}$$

Diante disso, observa-se que, para ambos os critérios, quanto mais próximo a um é o seu valor melhor é a qualidade do ajuste do modelo aos dados utilizados. Os critérios definidos em (4.1) e (4.2) não podem ser utilizados para expressar a qualidade do ajuste no que se refere à variância pelo fato de que tais medidas sempre apresentam acréscimo ou pelo menos não diminuem quando uma nova covariada é incluída no modelo. Nesses casos, se a covariada não é importante para a explicação da resposta é possível que esse aumento seja muito pequeno, mas não ocorre a diminuição do valor do R2. Similar ao R^2 ajustado para os modelos de regressão linear, propomos as versões corrigidas dos critérios, inserindo um termo de penalização para a inclusão de novas covariadas no modelo. A versão corrigida, $R^2_{*_c}$, e tem a seguinte forma

$$R^2_{EC_{*c}} = 1 - \left(1 - R^2_*\right)\left(\frac{n-1}{n-p}\right),$$

de modo que $p = k + r$ é o quantitativo de parâmetros do modelo, com k e r o número de parâmetros do modelo da média e da precisão, respectivamente, com * podendo ser os critérios apresentados em (4.1) e (4.2).

4.2 ESTATÍSTICAS PRESS, P^2 E R^2 BASEADAS NOS NOVOS RESÍDUOS

Para a verificação da qualidade do ajuste modelo frente à previsões futuras utilizaremos a estatística PRESS. O cálculo consiste no ajuste do modelo, repetidas vezes, quando excluída a t-ésima observação. No modelo de regressão beta, assumindo que $\hat{\delta}$ é uma solução de mínimos para a regressão de $y^{\dagger}$ sobre X. Dessa forma, o erro de predição é expresso por

$$y_t - \hat{y}_{(t)} = y_t^{\dagger} - x_t^{\top}\hat{\delta}_{(t)}. \tag{4.3}$$

Em linha com Pregibon (1981), tem-se que

$$\hat{\delta}_{(t)} = \hat{\delta} - \left\{ \frac{\left(X^{\top}X\right)^{-1} r_t^{\delta}}{1-h_{tt}} \right\}.$$

Logo, a Equação (4.3) pode ser escrita como

$$\begin{aligned}
\breve{y}_t - \hat{\breve{y}}_{(t)} =& y_t^{\dagger} - x_t^{\top}\hat{\delta} - \left\{ \frac{\left(X^{\top}X\right)^{-1} r_t^{\delta}}{1-h_{tt}} \right\} \\
=& y_t^{\dagger} - x_t^{\top}\delta + \left\{ \frac{x_t^{\top}\left(X^{\top}X\right)^{-1} x_t r_t^{\delta}}{1-h_{tt}} \right\} \\
=& y_t^{\dagger} - \hat{\eta}_t + \frac{h_{tt} r_t^{\delta}}{1-h_{tt}} \\
=& \frac{r_t^{\delta}(1-h_{tt}) + h_{tt} r_t^{\delta}}{1-h_{tt}} \\
=& \frac{r_t^{\delta} - h_{tt} r_t^{\delta} + h_{tt} r_t^{\delta}}{1-h_{tt}} \\
=& \frac{r_t^{\delta}}{1-h_{tt}}.
\end{aligned}$$

Portanto, sendo a PRESS definida como o somatório dos erros de predição quando removida uma observação, tem-se

$$PRESS = \sum_{t=1}^{n} \left(\breve{y}_t - \hat{\breve{y}}_{(t)}\right)^2 = \sum_{t=1}^{n} \left(\frac{r_t^{\delta}}{1-h_{tt}^{\dagger}} \right)^2,$$

com $h_{tt}^{\dagger}$ o t-ésimo elemento da diagonal da matriz de projeção $H^{\dagger} = X(X^{\top}X)^{-1}X^{\top}$. Baseada na PRESS, a estatítica P^2, apresentada por Quan (1988), tem a seguinte forma

$$P^2 = 1 - \frac{PRESS}{(n/(n-k))^2 SST}.$$

Pode-se mostrar que $SST_{(t)} = [n/(n-k)]^2 SST$, com k sendo a quantidade de parâmetros estimados.

Nos modelos de regressão linear clássicos, a análise de variância busca mensurar o quanto da variabilidade da variável resposta pode ser explicada através do modelo de regressão proposto e estimado. A decomposição da variabilidade total da resposta em torno de sua média é expressa pela quantidade SST (Soma de Quadrados Totais), de modo que esta soma de quadrados pode ser decomposta em duas somas de quadrados, a saber: SSR (Soma de Quadrados da Regressão) e SSE (Soma de Quadrados dos Erros), conforme a Equação (4.4):

$$\sum_{t=1}^{n}(y_t-\bar{y})^2=\sum_{t=1}^{n}(\hat{\mu}_t-\bar{y})^2+\sum_{t=1}^{n}(y_t-\hat{\mu}_t)^2. \tag{4.4}$$

Nossa proposta consiste em apresentar a decomposição da variabilidade total da resposta em torno da média, fazendo o uso da SST para cada submodelo, ou seja, uma decomposição da soma de quadrados totais para o submodelos da média e da dispersão, conforme apresentada na Equação (4.4) para a obtenção de novos critérios de aderência do tipo R^2. Assim, a decomposição proposta tem a forma geral dada por

$$\sum_{t=1}^{n}\left(y_t^{\dagger}-\bar{y}^{\dagger}\right)^2=\sum_{t=1}^{n}\left(\hat{\mu}_t{}^{\dagger}-\bar{y}^{\dagger}\right)^2+\sum_{t=1}^{n}\left(y_t^{\dagger}-\hat{\mu}_t{}^{\dagger}\right)^2. \tag{4.5}$$

Deve-se observar que $y^{\dagger}$ será definido por y^{*} quando estivermos tratando do submodelo da média e y^{**} quando estivermos tratando do submodelo da dispersão, sem perda de generalidade. Para os novos resíduos, apresentados em (3.8) e (3.10), os critérios de aderência R^2 propostos possuem forma geral é expressa por

$$R^2_{AD}=1-\frac{SSE}{SST}=1-\frac{\sum_{t=1}^{n}(y_t^{\dagger}-\hat{\mu}_t{}^{\dagger})^2}{\sum_{t=1}^{n}(y_t^{\dagger}-\bar{y}^{\dagger})^2}, \tag{4.6}$$

com

$$SSE=\sum_{t=1}^{n}(y_t^{\dagger}-\hat{\mu}_t{}^{\dagger})^2=(y^{\dagger}-\hat{\mu}^{\dagger})^{\top}(y^{\dagger}-\hat{\mu}^{\dagger})= \quad y^{\top\dagger}y^{\dagger}-\hat{\mu}^{\top\dagger}y^{\dagger}-y^{\dagger}\hat{\mu}^{\top\dagger}+\hat{\mu}^{\top\dagger}\hat{\mu}^{\dagger}=y^{\top\dagger}(\mathbb{I}-\mathrm{H}\dagger)\mathrm{y}^{\dagger}$$

e

$$SST=\sum_{t=1}^{n}(y_t^{\dagger}-\bar{y}^{\dagger})^2=y^{\top\dagger}y^{\dagger}-n\frac{y^{\top\dagger}\ell}{n}\frac{y^{\ddagger}\ell^{\top}}{n}=\frac{y^{\top\dagger}y^{\dagger}-y^{\top\dagger}\ell\ell^{\top}y^{\dagger}}{n}= \quad y^{\top\dagger}\left(\mathbb{I}-\frac{\ell\ell^{\top}}{n}\right)y^{\dagger},$$

para $\hat{\delta}=(X^{\top}X)^{-1}X^{\top}y$, sendo a solução de mínimos quadrados de uma regressão de y sobre X, $\mathbb{I}$ a matriz identidade e ℓ o vetor de uns.

Dessa forma, a Equação (4.6) pode ser reescrita como:

$$R^2_{AD} = 1 - \frac{SSE}{SST} = 1 - \left[\frac{y^{\top\dagger}(\mathbb{I} - \mathrm{H})y^{\dagger}}{y^{\top\dagger}\left(\mathbb{I} - \frac{\ell\ell^{\top}}{n}\right)y^{\dagger}}\right].$$

Tomando o resíduo de mínimos quadrados padronizado da para o submodelo da média, ou seja, o resíduo padronizado da regressão de Y^* sobre X apresentado em (3.8), propomos as estatísticas PRESS e P^2, sendo a primeira definida como

$$PRESS = \sum_{t=1}^{n}\left(\frac{r_t^{\beta}}{1 - h_{tt}^*}\right)^2,$$

com r_t^{β} o *resíduo da média* (3.8) e h_{tt}^* o t-ésimo elemento diagonal da matriz de projeção definido em (3.7). Propomos o seguinte coeficiente de predição baseado na PRESS:

$$P^2 = 1 - \frac{PRESS}{(n/(n-p)^2)SST},$$

em que $SST = \sum_{t=1}^{n}(Y_t^* - \bar{Y}^*)^2$, com $\bar{Y}^*$ sendo a média aritmética de $Y_t^* = \log(Y_t/(1-Y_t))$.

Assumindo $Y^* = (Y_1^*, ..., Y_n^*)^{\top}$, com $Y_t^* = T_1 = \log(Y_t/(1-Y_t))$ e r_t^{β} apresentado em (3.8). A proposta para o critério de aderência R^2 envolvendo o submodelo da média é dada por

$$R^2_{AD_m} = 1 - \frac{SSE_{\mu}}{SST_{\mu}},$$

em que $SST_{\mu} = \sum_{t=1}^{n}(Y_t^* - \bar{Y}^*)^2$ e $SSE_{\mu} = \sum_{t=1}^{n}(y_t^* - \hat{\varepsilon}_t^*)^2$. O novo critério de seleção tem sua versão corrigida definida como

$$R^2_{AD_{mc}} = 1 - \left(1 - R^2_{PM}\right)\left(\frac{n-1}{n-p}\right).$$

Analogamente, para o resíduo de mínimos quadrados padronizado para o submodelo da dispersão, ou seja, o resíduo padronizado da regressão de Y^{**} sobre Z, apresentado em (3.10), propomos as estatísticas PRESS e P^2, sendo a primeira definida como

$$\mathrm{PRESS} = \sum_{t=1}^{n}\left(\frac{r_t^{\gamma}}{1 - h_{tt}^{**}}\right)^2,$$

com r_t^{γ} sendo o *resíduo variance 2* dado em (3.10) e h_{tt}^{**} o t-ésimo elemento diagonal da matriz de projeção definidos em (3.9). A proposta para o coeficiente de predição baseado na PRESS é

$$P^2 = 1 - \frac{\mathrm{PRESS}}{(n/(n-p)^2)SST},$$

em que $SST = \sum_{t=1}^{n}(Y_t^{**} - \bar{Y}^{**})^2$, com $\bar{Y}^{**}$ sendo a média aritmética de $Y_t^{**} = \log(1-Y_t)$.

Admitindo $Y^{**} = (Y_1^{**}, ..., Y_n^{**})^\top$, com $Y_t^{**} = T_2 = \log(1 - Y_t)$ e r_t^γ dado em (3.10), a proposta para o critério de predição R^2 envolvendo o submodelo da precisão é

$$R^2_{AD_v} = 1 - \frac{SSE_\phi}{SST_\phi},$$

em que $SST_\phi = \sum_{t=1}^n (Y_t^{**} - \bar{Y}^{**})^2$ e $SSE_\phi = \sum_{t=1}^n (y_t^{**} - \hat{\varepsilon}_t^{**})^2$. O novo critério de seleção tem sua versão corrigida definida como

$$R^2_{AD_{vc}} = 1 - \left(1 - R^2_{PV}\right) \left(\frac{n-1}{n-p}\right).$$

Finalmente, utilizando o *resíduo variance* baseado no processo iterativo Scoring de Fisher apresentado na seção (2.5.3), propomos medidas PRESS e P^2. Em linha com Espinheira *et al.* (2019), para o modelo de regressão beta, a medida PRESS é dada por

$$\text{PRESS} = \sum_{t=1}^{n} \left(\frac{r_t^\gamma}{1 - h_{4tt}}\right)^2,$$

em que r_t^γ é o *resíduo variance* definido em (2.28). Além disso, h_{4tt} é o t-ésimo elemento da diagonal principal de

$$H_4 = (\hat{D}^*)^{1/2} Z (Z^\top \hat{D}^* Z)^{-1} Z^\top (\hat{D}^*)^{1/2}.$$

De modo que H_4 é uma matriz simétrica de ordem n $\times$ n, sendo Z uma matriz de posto completo de ordem n $\times$ k, com a i-ésima linha denotada por $\mathbf{z}_t^\top$ e $D^* = \text{diag}\{d_1^* \dots d_n^*\}$ com

$$d_t^* = \left[\psi'(\mu_t\phi_t)\mu_t^2 + \psi'((1-\mu_t)\phi_t)(1-\mu_t^2) - \psi'(\phi_t)\right] \frac{1}{\{h'(\phi_t)^2\}}.$$

Fazendo o uso da PRESS, apresentamos a estatística P^2, definida como

$$P^2 = 1 - \frac{\text{PRESS}}{SST_{(t)}},$$

em que $SST = (y_t^\dagger - \bar{y}_t^\dagger)$, sendo $\bar{y}_t^\dagger$ a média aritmética sem a t-ésima observação dos valores de

$$y_t^\dagger = \hat{d}^{*1/2} u_{2,t},$$

com $u_{2,t}$ sendo o t-ésimo elemento do vetor

$$u_2 = \hat{\eta}_2 + \hat{D}^{*-1} \hat{H} \hat{a},$$

em que H=diag$\{\frac{1}{h'(\phi_1)}, \dots, \frac{1}{h'(\phi_n)}\}$ e $\hat{a}$ definido conforme a expressão 2.27 .

5 APLICAÇÕES

5.1 APRESENTAÇÃO DOS DADOS - RISCO A DOENÇAS CARDÍACAS

Doenças coronarianas são apontadas como as causas mais comuns de morte entre pessoas pretas. Este fenômeno pode estar relacionado, em parte, às taxas mais altas de fatores de risco para doenças cardíacas entre pretos, destacando-se as diferenças no acesso aos cuidados de saúde e as diferenças nos perfis dos fatores de risco. Além disso, é observado que diabetes e hipertensão são mais prevalentes entre populações negras do que brancas e que o índice de massa corporal é significativamente associado a prevalências mais altas de hipertensão, diabetes e colesterol HDL baixo. No presente capítulo, apresentaremos três aplicações sobre um conjunto de dados reais (Risco à doenças cardíacas). Os dados são apresentados por Willems et al. (1997), que trata de 381 observações, referentes aos fatores de risco de doenças coronárias entre uma amostra populacional de negros da região rurais no estado da Virgínia, nos Estados Unidos. Neste conjunto de dados, propomos modelos de regressão com distribuição beta a fim de estudaremos o comportamento dos resíduos e dos critérios de seleção de modelos propostos nesta dissertação. Utilizamos o chute inicial apresentado no Capítulo 2 para obter os parâmetros da regressão por meio do método da máxima verossimilhança utilizando BFGS.

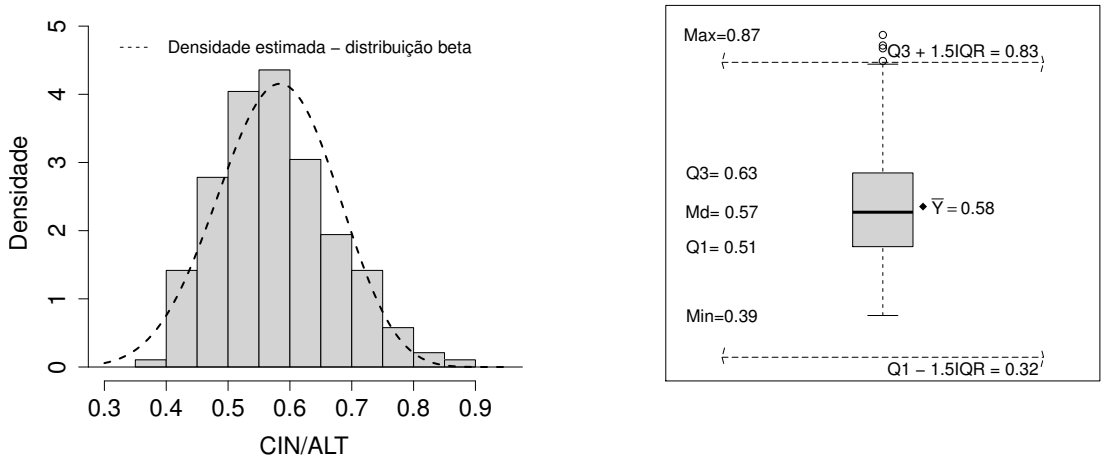
Nas Seções (5.2), (5.3) e (5.4) apresentaremos os resultados da aplicações . Para a Aplicação I, quando a média da variável resposta (y_1) - razão entre a circunferência da cintura (cm) e a altura (cm), CIN/QUA - se encontra dispersa no intervalo da unidade padrão, utilizamos as covariadas hemoglobina glicosilada - HBGLI (x_2), índice de massa corporal - IMC (x_3), circunferência do quadril (cm) - QUA (x_4), uma iteração entre a hemoglobina glicosilada e o gênero - HBGLI*GEN (x_5), idade (x_6) e a razão ente o colesterol total e o colesterol HDL - RCH (x_7), em que utilizamos a função log - log para ajustar o modelo proposto. Destacamos que esta variável resposta é considerada um indicador de obesidade abdominal. Na Figura 1, observamos no painel (a) que o histograma dos dados se assemelha à distribuição beta e que, através do painel (b), a variável apresenta valores próximos ao valor médio unitário , 0.50, sendo que 75% estão acima de 0.51, o valor mínimo é de 0.39, o valor máximo é 0.87, a média e mediana são 0.58 e 0.57, respectivamente. Vale salientar que existem observações *outliers* no conjunto de dados. Uma vez que esse indicador de obesidade abdominal aponta que para valores maiores que 0.5 o risco de obesidade é alto, verificamos que a amostra apresenta uma propensão à obesidade.

Para a Aplicação II, quando a média da variável resposta (y_2) - proporção entre o

colesterol de alta densidade e o colesterol total, HDL/CT - está perto de zero, utilizamos as covariadas índice de massa corpórea - IMC (x_2) e o percentual entre a medida da circunferência da cintura e a altura - (CIN/ALT) (x_3) e hemoglobina glicosilada - HBGLI (x_4). Com isto, utilizamos a função de ligação log - log para o ajuste do modelo proposto. Na Figura 2, observamos no painel (a) que o histograma dos dados apresenta semelhança à distribuição beta e que, através do painel (b), a variável resposta apresenta 75% dos valores acima de 0.18, o valor mínimo é de 0.05, o valor máximo é 0.67, a média e mediana são 0.25 e 0.24, respectivamente, além de existem observações *outliers* no conjunto de dados. Esta variável resposta evidencia que a amostra utilizada apresenta baixos valores de colesterol do tipo HDL, responsável por remover o colesterol das artérias, evitando o seu acúmulo, indicando que os entrevistados possuem alta propensão a doenças.

Figura 1 – Histograma e Box Plot da variável resposta. Aplicação I: dados CIN/ALT.

(a) **(b)**

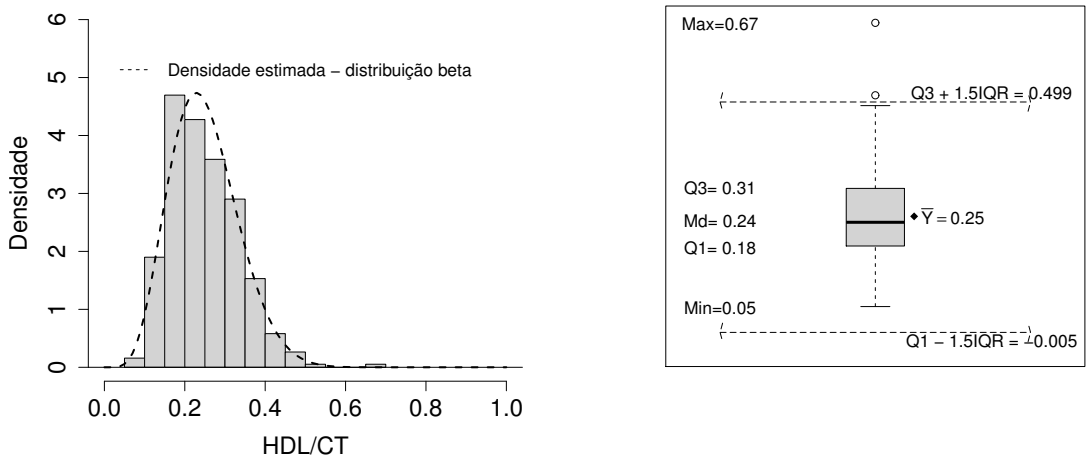


Fonte: do autor.

Finalmente, na Aplicação III, quando a média da variável resposta (y_3) - proporção entre o colesterol não HDL e o colesterol total, NHDL/CT - se encontra próxima de um, para o ajuste do modelo adotamos a funções de ligação do tipo complemento log - log, em que a covariadas utilizadas foram hemoglobina glicada - HBLGI (x_2), circunferência da cintura - CIN (x_3), índice de massa corpórea - IMC (x_4), circunferência do quadril - QUA (x_5), peso - PSO (x_6) e idade - ID (x_7). Na Figura 3, observamos no painel (a) que o histograma dos dados apresenta similaridade com a forma da distribuição beta e que, através do painel (b), a variável apresenta 75% dos valores acima de 0.70, o valor mínimo é de 0.33, o valor máximo é 0.95, a

Figura 2 – Histograma e Box Plot da variável resposta. Aplicação II: dados HDL/CT.

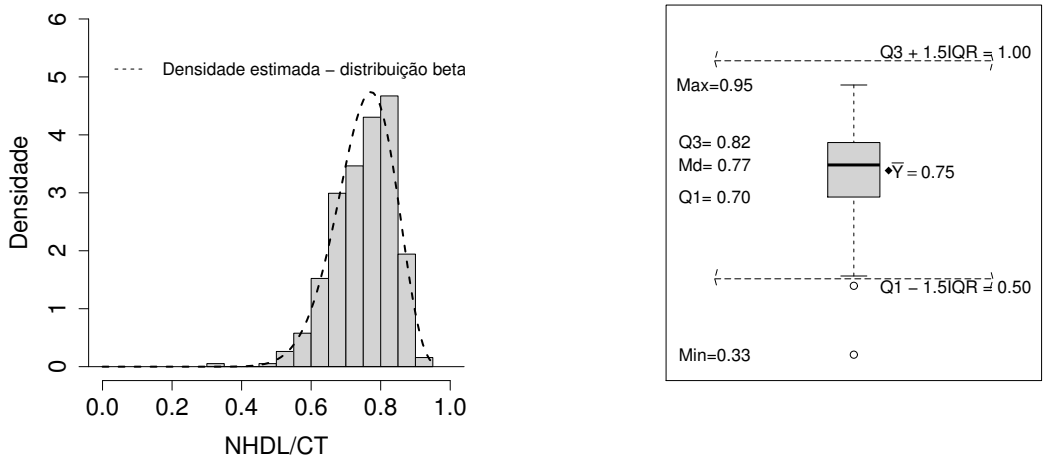
(a) **(b)**



Fonte: do autor.

Figura 3 – Histograma e Box Plot da variável resposta. Aplicação III: dados NHDL/CT.

(a) **(b)**



Fonte: do autor.

média e mediana são 0.75 e 0.77, respectivamente, além de existirem observações *outliers* no conjunto de dados. A variável resposta da Aplicação III indica que a amostra utilizada apresenta acentuada concentração de colesterol de baixa densidade (LDL) e colesterol de densidade muito baixa (VLDL). Estas lipoproteínas são responsáveis pelo acumulo de gorduras nas artérias, ocasionando o aumento do risco de doenças cardíacas.

Na Figura 4 apresentamos as medidas resumo das variáveis utilizadas nas aplicações supracitadas que se referem aos aspectos clínicos dos entrevistados. Verificamos que a amostra populacional que compõem o conjunto de dados apresenta uma alta concentração de diabéticos,

evidenciados pelos *outliers* no painel (a). Além disso, 75% dos valores registrados para a variável HBGLI (hemoglobina glicosilada) estão acima de 4.39, com média 5.58. O valor de referência para que um indivíduo não seja diabético pode variar entre 4.5 e 5.6. Quanto à presença de colesterol HDL (resposável por impedir o acumulo de colesterol nas artérias e reduzindo o risco de doenças cardíacas) - painel (b)-, podemos notar que os entrevistados apresentam, em média valores próximos aos níveis recomendados, ou seja, uma concentração de HDL maior que 50. Conforme o painel (c), apresentam em média valores superiores ao recomendado para a concentração de colesterol total (o desejável são valores menores que 200), indicando uma propensão ao acumulo de colesterol na corrente sanguínea. A capacidade de remoção do colesterol total por parte do colesterol HDL é apresentada no painel (d) em que observamos que 75% dos valores registrados para a variável RCH estão acima dos valores de referência ideais (menores que 3). Em todas as variáveis observadas notamos a presença de *outliers*.

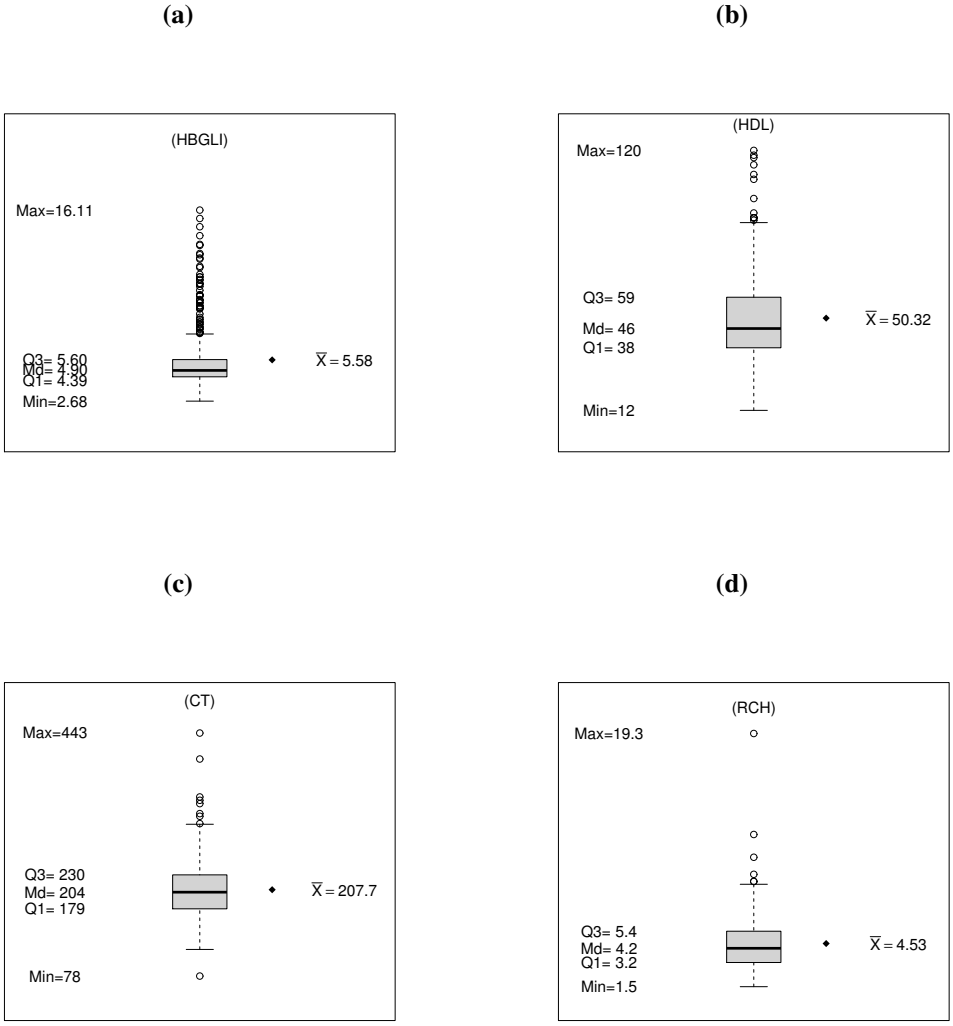
Na Figura 5 apresentamos medidas resumo das variáveis utilizadas nas aplicações supracitadas que se referem aos aspectos físicos dos entrevistados. A amostra populacional que compõe o conjunto de dados apresenta concentração de participantes com sobrepeso. No painel (a), o valor médio da variável IMC é de 28.84 e 75% dos valores registrados estão acima de 24.13. Para essa variável os valores de referência para a classificação de sobrepeso estão entre 25 e 29.9. Além disso, os valores superiores a 30 indicam a presença de obesos na amostra coletada. Quanto ao peso, apresentado no painel (b), a variável resposta apresenta 75% dos valores acima de 68.49, o valor mínimo é de 44.91, o valor máximo é 147.42, a média e mediana são 80.7 e 78.93, respectivamente. A circunferência do quadril - painel (c) - apresentada na amostra destaca um valor médio de 109.4, caracterizando uma propensão ao acumulo de gordura na região abdominal. Sobre a idade dos entrevistados, notamos que 75% dos respondentes possuem idade superior a 34 anos, configurando uma amostra composta por adultos.

5.2 APLICAÇÃO I

Para esta aplicação a variável de interesse (resposta) é o percentual entre a medida da circunferência da cintura e a altura (CIN/ALT) de 381 pacientes. Esta é uma medida que busca avaliar o risco de doenças cardíacas.

Para esta aplicação a variável de interesse (resposta) é o percentual entre a medida da circunferência da cintura e a altura (CIN/ALT) de 381 pacientes. Esta é uma medida que busca avaliar o risco de doenças cardíacas. As covariáveis são hemoglobina glicosilada - HBGLI

Figura 4 – Box plot das covariadas HBGLI, HDL, CT, RCH.



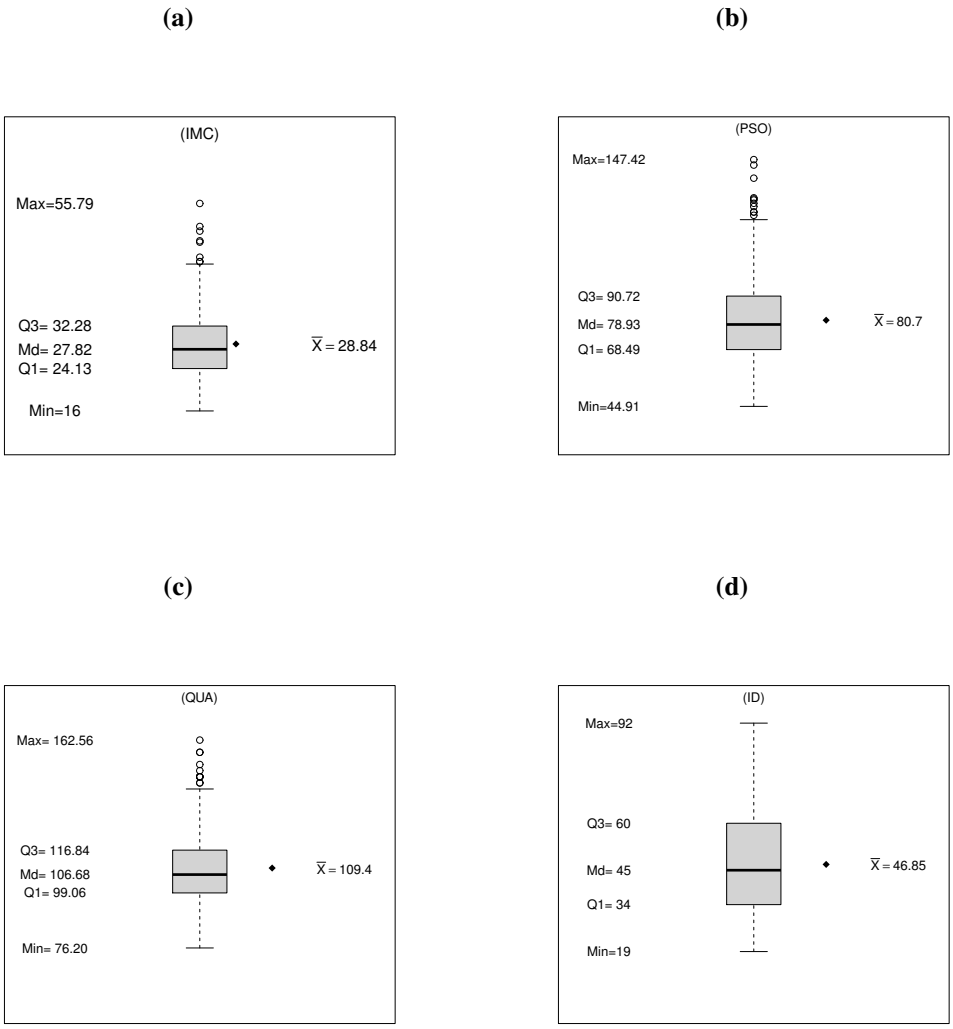
Fonte: do autor.

(x_1), índice de massa corporal - IMC (x_2), circunferência do quadril - QUA (x_3) e uma iteração entre a hemoglobina glicosilada e o gênero - HBGLI*GEN (x_4). Inicialmente, consideramos o seguinte modelo de regressão beta com dispersão fixa e função de ligação do tipo loglog:

$$g(\mu_t) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 \left(x_{t5} * x_{t2}\right)^2, \qquad t = 1, \ldots, 381. \tag{5.1}$$

Na Tabela 1 são apresentadas as estimativas dos parâmetros, erros-padrão e p-valores, em que todos os parâmetros são estatisticamente diferentes de zero a um nível de significância de 5%. Buscando verificar a adequação do modelo da dispersão fixa (5.1), utilizamos o gráfico de probabilidade normal com envelope simulado e conforme a Figura 6 que apontaram para a presença de uma dispersão não constante. Por meio dos envelopes, tanto os resíduos baseados no processo iterativo quanto os derivados das estatísticas suficientes e completas apresentam pontos fora das bandas do envelope (atípicos) indicando que os resultados inferenciais tirados do

Figura 5 – Box plot das covariadas IMC, PSO, QUA E ID.



Fonte: do autor.

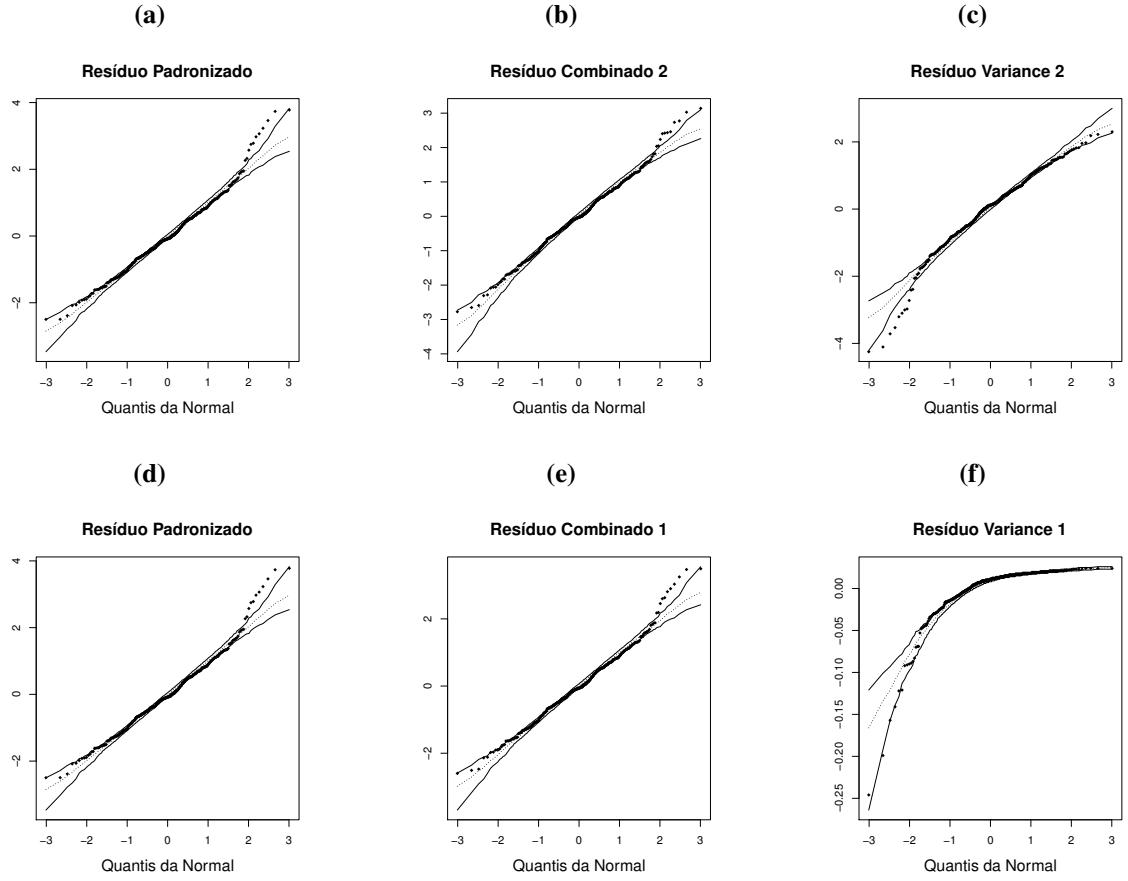
modelo ajustado podem ser imprecisos. O *resíduo combinado 2* destaca resíduos com valores altos, enquanto que os dois resíduos *variance* apontam para erros na especificação da dispersão. Sendo tais resíduos derivados do submodelo da dispersão, os mesmos apresentam uma boa capacidade (sensibilidade) para identificar erros de especificação da dispersão.

Tabela 1 – Estimativas, erros-padrão e p-valores dos parâmetros. Aplicação I: dados cintura/altura.

Modelo		Dispersão fixa - loglog					
Descrição dos Parâmetros		β_1 (Const)	β_2 (HBGLI)	β_3 (IMC)	β_4 (QUA)	β_5 (GEN*HBGLI)^2	ϕ
Dataset completo	estimat.	-0.928	0.019	0.029	0.006	-0.001	4.883
	e.p.	0.068	0.004	0.002	0.001	0.000	0.072
	p-valor	0.000	0.000	0.000	0.000	0.002	0.000

Fonte: do autor.

Figura 6 – Gráficos de envelope simulado. Aplicação I: dados CIN/ALT - dispersão fixa (loglog).



Fonte: do autor.

Uma vez que há evidências de dispersão variável e com base na análise dos resíduos contra as covariadas, propomos o seguinte modelo de regressão beta com dispersão variável e função de ligação do tipo loglog:

$$\begin{aligned} g(\mu_t) &= \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 \left(x_{t5} * x_{t2}\right)^2 \\ \log(\phi_t) &= \gamma_1 + \gamma_2 (\exp(x_{t2}))^{-1} + \gamma_3 x_{t3} + \gamma_4 x_{t6} + \gamma_5 \left(x_{t5} * x_{t7}\right), \end{aligned} \qquad t = 1, \ldots, 381. \tag{5.2}$$

Os parâmetros foram estimados por máxima verossimilhança e o grau de heterocedasticidade estimado é de $\hat{\lambda} = 8.57$ com mínimo=0.0004 e máximo=0.0035. Para a avaliação do ajuste do modelo, plotamos os resíduos contra os índices das observações. Os resíduos parecem estar espalhados aleatoriamente em torno de zero, apresentando observações atípicas. Na Figura 7 - painéis (a), (b) e (c) - destacam-se os seguintes pontos atípicos: (24,33,170,232) em ambos os resíduos, no entanto o *resíduo variance 2* indica mais enfaticamente tais pontos. Analogamente aos resíduos baseados nas estatísticas suficientes e completas da distribuição beta, na Figura 8 - painéis (a), (b), (c) - destacam-se os seguintes pontos atípicos: (24,33,170,232) em ambos os resíduos obtidos do processo iterativo Scoring de Fisher. Notamos também que o caso 192 apenas

é apontado como potencialmente atípico no gráfico do *resíduo variance 1* (painel c). Mais uma vez, os dois novos *resíduos variance* fornecem informações complementares. Os gráficos dos resíduos versus o preditor linear - painéis (d), (e) e (f) das Figuras (7 e 8) - não indicam violação severa nas suposições do modelo. O ganho com a modelagem da dispersão pode ser verificado através dos envelopes simulados que apontam para veracidade de que a distribuição beta é adequada para a modelagem da variável resposta. Conforme os painéis (g), (h) e (i) é possível detectar alguns pontos próximos as bandas, tal como o caso 192, destacado exclusivamente pelo *resíduo variance 1*. Além disso, o envelope do *resíduo variance 2* destaca dois novos pontos potencialmente atípicos. Para verificar a suposição de que o impacto inferencial dos pontos potencialmente atípicos (24,33,34,43,170,192,203,232) é atenuado pela modelagem da dispersão, na Tabela 2 apresentamos as estimativas dos parâmetros, erros padrão (e.p), mudanças relativas nas estimativas dos parâmetros (%) e nos erros padrão (%) devido à exclusão das observações e os respectivos p-valores.

Através da Tabela 2 observamos que no submodelo da média as estimativas pontuais e p-valores não apresentaram alterações consideráveis. Quanto ao submodelo da dispersão, podemos notar que a remoção individual dos pontos 24, 33, 43, 192 - enfaticamente destacados pelos resíduos *variance* - exerce uma influência substancial nas estimativas do parâmetro da dispersão. Verificamos que a suposição sobre a redução no impacto inferencial das estimativas pela modelagem da dispersão não pode ser descartada, uma vez que mesmo com a influência dos pontos sobre o modelo da dispersão os parâmetros são estatisticamente diferentes de zero a um nível de significância de 10%. A exclusão conjunta dos casos $\{24, 33, 170\}$ e $\{24, 33, 34, 43, 170, 192, 203, 232\}$ provoca mudanças nos erros-padrão dos coeficientes estimados para o submodelo da média e promovem significativa alteração na significância dos parâmetros do submodelo da dispersão, evidenciando que os mesmo são conjuntamente prejudiciais.

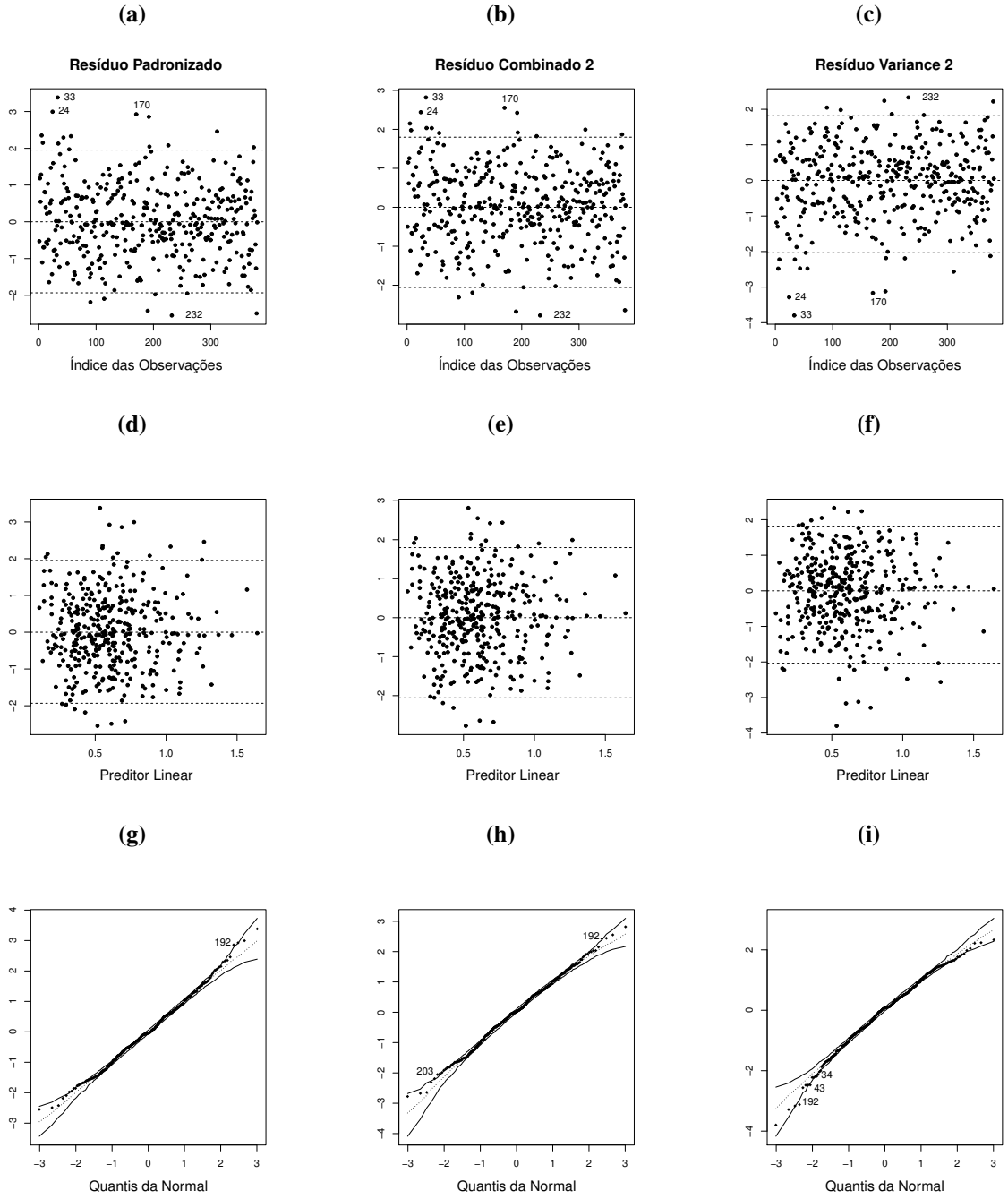
Avaliamos a qualidade do modelo em termos de variabilidade e predição através dos critérios de seleção R^2_{FC},R^2_{LR}, PRESS, P^2. Adicionalmente, apresentamos os os novos critérios R^2_{EC} e R^2_{AD} para variabilidade e aderência. Por meio da Tabela 3 podemos verificar que o desempenho conjunto das medidas apontam para uma bom ajuste do modelo em termos de variabilidade. Também podemos notar o efeito que a remoção das observações exerce sobre a qualidade do modelo, principalmente quanto ocorre a remoção das doze observações destacadas na análise dos resíduos. As retiradas conjuntas dos casos $\{24, 33, 170\}$ e $\{24, 33, 34, 43, 170, 192, 203, 232\}$, além da observação 33, também podem ser destacadas, evidenciando o efeito que tais pontos atípicos possuem sobre a variabilidade modelo postulado. Sobre as novas medidas R^2_{EC}, a

Tabela 2 – Estimativas, erros-padrão e p-valores dos parâmetros. Aplicação I: dados cintura/altura.

Modelo		Dispersão variável - loglog									
Descrição dos Parâmetros		β_1 (Const)	β_2 (HBGLI)	β_3 (IMC)	β_4 (QUA)	β_5 (GEN*HBGLI)^2	γ_1 (Const)	γ_2 exp(HBGLI)$^{-1}$	γ_3 (IMC)	γ_4 (ID)	γ_5 (GEN*RCH)
Dataset completo	Estimat.	-0.936	0.019	0.028	0.006	-0.001	5.813	20.194	-0.026	-0.009	0.076
	e.p	0.063	0.005	0.002	0.001	0.000	0.458	9.240	0.011	0.005	0.028
	p-valor	0.000	0.000	0.000	0.000	0.003	0.000	0.029	0.021	0.048	0.007
obs. 24 deletada	Estimat.	0.003	-0.010	-0.011	0.017	-0.111	-0.012	-0.001	-0.065	-0.109	-0.075
	e.p	-0.008	-0.022	-0.045	0.000	0.000	0.001	0.000	0.000	0.000	0.000
	p-valor	0.000	0.000	0.000	0.000	0.004	0.000	0.029	0.031	0.082	0.013
obs. 33 deletada	Estimat.	-0.013	-0.026	0.014	-0.034	-0.111	0.027	-0.143	0.169	-0.098	-0.130
	e.p	-0.005	-0.043	-0.045	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	p-valor	0.000	0.000	0.000	0.000	0.004	0.000	0.061	0.007	0.075	0.019
obs. 34 deletada	Estimat.	0.018	0.021	-0.004	0.017	0.000	0.010	0.072	0.042	0.087	0.026
	e.p	-0.003	0.000	-0.045	0.000	0.000	0.001	0.002	0.000	0.000	0.000
	p-valor	0.000	0.000	0.000	0.000	0.003	0.000	0.019	0.016	0.033	0.006
obs. 43 deletada	Estimat.	0.002	-0.021	-0.004	0.017	-0.111	0.003	-0.031	0.042	-0.076	-0.050
	e.p	-0.002	-0.022	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	p-valor	0.000	0.000	0.000	0.000	0.004	0.000	0.034	0.016	0.071	0.011
obs. 170 deletada	Estimat.	-0.013	0.041	0.011	-0.034	0.000	-0.010	0.027	0.027	-0.174	0.092
	e.p	-0.008	-0.022	-0.045	0.000	0.000	0.001	0.001	0.000	0.000	0.000
	p-valor	0.000	0.000	0.000	0.000	0.002	0.000	0.025	0.018	0.106	0.003
obs. 192 deletada	Estimat.	-0.004	0.000	-0.004	0.000	-0.111	-0.006	0.032	-0.015	-0.098	-0.072
	e.p	-0.008	-0.022	-0.045	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	p-valor	0.000	0.000	0.000	0.000	0.003	0.000	0.024	0.023	0.077	0.013
obs. 203 deletada	Estimat.	-0.016	-0.010	0.007	-0.034	-0.111	0.011	-0.004	0.050	0.043	-0.004
	e.p	0.002	-0.022	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000
	p-valor	0.000	0.000	0.000	0.000	0.003	0.000	0.030	0.015	0.041	0.007
obs. 232 deletada	Estimat.	0.008	0.005	-0.014	0.034	0.000	0.019	-0.037	0.069	0.065	-0.062
	e.p	-0.005	-0.022	0.000	0.000	0.000	0.003	0.001	0.000	0.000	0.000
	p-valor	0.000	0.000	0.000	0.000	0.002	0.000	0.035	0.014	0.035	0.012
obs. $\{34, 43, 192\}$ deletadas	Estimat.	0.016	0.000	-0.007	0.034	-0.111	0.007	0.071	0.065	-0.109	-0.101
	e.p	-0.014	-0.022	-0.045	0.000	0.000	0.002	0.003	0.000	0.000	0.004
	p-valor	0.000	0.000	0.000	0.000	0.004	0.000	0.020	0.014	0.083	0.016
obs. $\{24, 33, 170\}$ deletadas	Estimat.	-0.024	0.000	0.018	-0.068	-0.111	0.006	-0.120	0.134	-0.402	-0.116
	e.p	-0.022	-0.065	-0.045	0.000	0.000	0.002	0.001	0.000	0.000	0.004
	p-valor	0.000	0.000	0.000	0.000	0.002	0.000	0.055	0.009	0.242	0.018
todas obs. deletadas	Estimat.	-0.017	-0.016	0.007	-0.034	-0.111	0.042	-0.102	0.322	-0.467	-0.319
	e.p	-0.042	-0.109	-0.045	-0.100	0.000	0.009	0.005	0.009	0.000	0.007
	p-valor	0.000	0.000	0.000	0.000	0.002	0.000	0.051	0.002	0.305	0.069

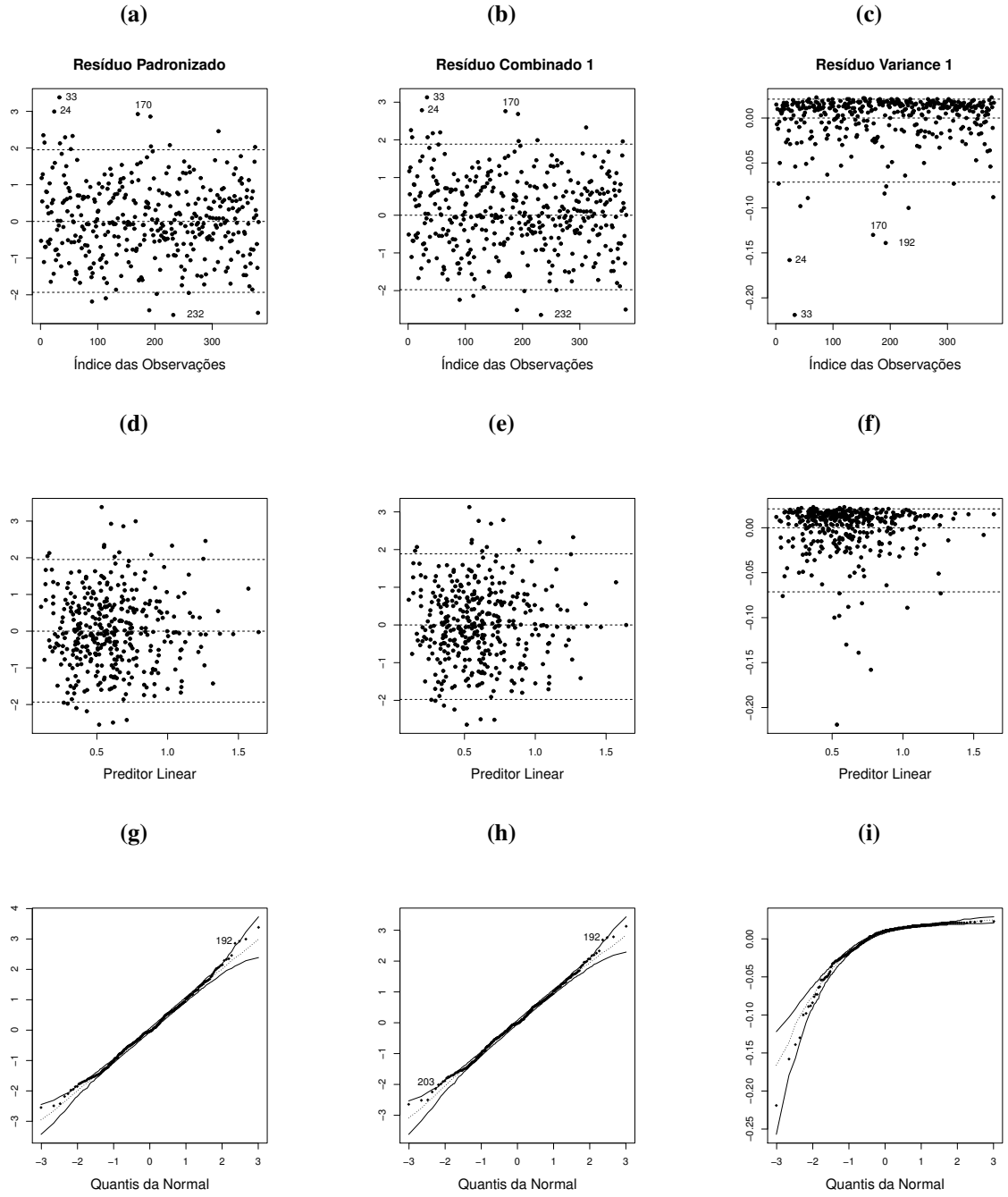
Fonte: do autor.

Figura 7 – Gráficos dos resíduos. Estatísticas suficientes e completas. Aplicação I: dados CIN/ALT - dispersão variável (loglog).



Fonte: do autor.

Figura 8 – Gráficos dos resíduos. Processo iterativo Scoring de Fisher. Aplicação I: dados CIN/ALT - dispersão variável (loglog).



Fonte: do autor.

Tabela 3 – Critérios para a variabilidade. Aplicação I: dados cintura/altura.

	R^2_{FC}	$R^2_{FC_c}$	R^2_{LR}	$R^2_{LR_c}$	$R^2_{EC_m}$	$R^2_{EC_{mc}}$	$R^2_{EC_v}$	$R^2_{EC_{vc}}$
Dataset completo	0.793	0.788	0.813	0.807	0.794	0.788	0.787	0.782
obs. 24 deletada	0.798	0.793	0.815	0.809	0.798	0.793	0.793	0.788
obs. 33 deletada	0.800	0.795	0.817	0.812	0.800	0.795	0.793	0.788
obs. 34 deletada	0.793	0.788	0.814	0.809	0.794	0.789	0.787	0.782
obs. 43 deletada	0.797	0.792	0.815	0.809	0.797	0.792	0.790	0.785
obs. 170 deletada	0.797	0.792	0.816	0.811	0.797	0.792	0.790	0.785
obs. 192 deletada	0.797	0.792	0.815	0.810	0.797	0.792	0.791	0.786
obs. 203 deletada	0.793	0.788	0.813	0.807	0.793	0.788	0.787	0.782
obs. 232 deletada	0.795	0.790	0.815	0.809	0.795	0.790	0.788	0.783
obs. $\{34, 43, 192\}$ deletadas	0.801	0.796	0.819	0.814	0.801	0.796	0.795	0.790
obs. $\{24, 33, 170\}$ deletadas	0.808	0.804	0.823	0.818	0.808	0.803	0.803	0.798
todas obs. deletadas	0.818	0.814	0.833	0.828	0.817	0.813	0.812	0.807

Fonte: do autor.

Tabela 4 – Critérios para a predição e aderência. Aplicação I: dados cintura/altura.

	Est. Suficientes e Completas									Processo Iterativo Scoring de Fisher				
	$PRESS_M$	P^2_M	$P^2_{M_c}$	P^2_V	$P^2_{V_c}$	$R^2_{AD_m}$	$R^2_{AD_{mc}}$	$R^2_{AD_v}$	$R^2_{AD_{vc}}$	$PRESS_M$	P^2_M	$P^2_{M_c}$	P^2_V	$P^2_{V_c}$
Dataset completo	12.762	0.876	0.875	0.876	0.875	0.874	0.871	0.873	0.870	382.370	0.546	0.535	0.981	0.981
obs. 24 deletada	12.300	0.878	0.877	0.878	0.877	0.876	0.873	0.875	0.872	381.160	0.557	0.547	0.981	0.981
obs. 33 deletada	12.317	0.879	0.878	0.879	0.877	0.876	0.873	0.875	0.872	381.180	0.522	0.511	0.982	0.981
obs. 34 deletada	12.709	0.876	0.875	0.876	0.874	0.873	0.870	0.872	0.869	381.310	0.550	0.539	0.981	0.980
obs. 43 deletada	12.524	0.878	0.876	0.877	0.876	0.875	0.872	0.874	0.871	381.430	0.542	0.530	0.981	0.980
obs. 170 deletada	12.518	0.878	0.876	0.877	0.876	0.875	0.872	0.874	0.871	381.320	0.551	0.540	0.982	0.981
obs. 192 deletada	12.428	0.878	0.877	0.878	0.876	0.875	0.872	0.875	0.871	381.270	0.550	0.539	0.982	0.981
obs. 203 deletada	12.689	0.876	0.875	0.876	0.875	0.874	0.871	0.873	0.870	381.340	0.539	0.527	0.981	0.980
obs. 232 deletada	12.588	0.877	0.876	0.876	0.875	0.874	0.871	0.873	0.870	381.390	0.538	0.527	0.981	0.981
obs. $\{34, 43, 192\}$deletadas	12.133	0.879	0.878	0.879	0.877	0.877	0.874	0.875	0.872	379.250	0.549	0.538	0.982	0.982
obs. $\{24, 33, 170\}$ deletadas	11.611	0.882	0.881	0.882	0.881	0.880	0.877	0.879	0.876	378.850	0.540	0.529	0.986	0.985
todas obs. deletadas	10.738	0.886	0.885	0.886	0.885	0.884	0.881	0.883	0.880	373.570	0.525	0.513	0.986	0.986

Fonte: do autor.

qualidade do ajuste para o submodelo da média - $R^2_{EC_m}$ e $R^2_{EC_{mc}}$ - apresenta similaridade com o critério R^2_{FC} e sua forma corrigida. Para o submodelo da dispersão, as medidas $R^2_{EC_v}$ e $R^2_{EC_{vc}}$ apresentam valores mais baixos que os apontados pelo critérios R^2_{FC}.

Seguindo para a avaliação da qualidade preditiva e aderência, na Tabela 4 a utilização das estatísticas suficientes e completas da distribuição beta revela baixos valores para a estatítica PRESS, indicando uma qualidade preditiva quando trabalhamos com a distribuição da variável resposta. Outro aspecto a ser destacado são os valores da medida P^2 - maiores que 0.8 em ambos os submodelos - corroborando com a qualidade já evidenciada pela estatística PRESS. Adicionalmente, o novo critério R^2_{AC} permite avaliar a qualidade da aderência da distribuição aos dados. Finalmente, podemos concluir favoravelmente ao modelo postulado. As estimativas dos parâmetros que compões o submodelo da média não sofreram severos efeitos das observações atípicas, o impacto inferencial foi atenuado pela modelagem da dispersão, os novos resíduos

variance destacaram pontos sensíveis ao submodelo da dispersão, o modelo apresenta boa qualidade em termos e variabilidade, predição e aderência.

5.3 APLICAÇÃO II

Nesta aplicação buscamos modelar a proporção do HDL sobre o CT através das covariáveis referentes ao índice de massa corpórea - IMC (x_2) e o percentual entre a medida da circunferência da cintura e a altura - (CIN/ALT) (x_3). Assumimos o seguinte modelo de regressão beta com dispersão fixa e função de ligação do tipo loglog:

$$g(\mu_t) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3}, \qquad t = 1, \ldots, 379. \tag{5.3}$$

A Tabela 5 apresenta as informações relativa aos parâmetros estimados. Podemos observar a significância de todas as estimativas, eventualmente indicando que o modelo da dispersão fixa está bem ajustado. Porém, a análise dos gráficos de envelope simulados - apresentados na Figura 9 - indica a necessidade de modelagem da dispersão, além de destacar em todos os resíduos a presença de pontos fora do envelope. Da mesma forma que na aplicação anterior (5.2), ambos os resíduos *variance* apresentaram melhor desempenho ao enfatizar pontos fora do envelope, favorecendo à hipótese de modelagem da dispersão.

Tabela 5 – Estimativas, erros-padrão e p-valores dos parâmetros. Aplicação II: dados HDL/CT.

Modelo		Dispersão fixa - loglog			
Descrição dos Parâmetros		β_1 (Const)	β_2 (IMC)	β_3 (CIN/QUA)	ϕ
Dataset completo	estimat.	0.544	-0.010	-0.677	3.363
	e.p.	0.149	0.002	0.164	0.072
	p-valor	0.000	0.000	0.000	0.000

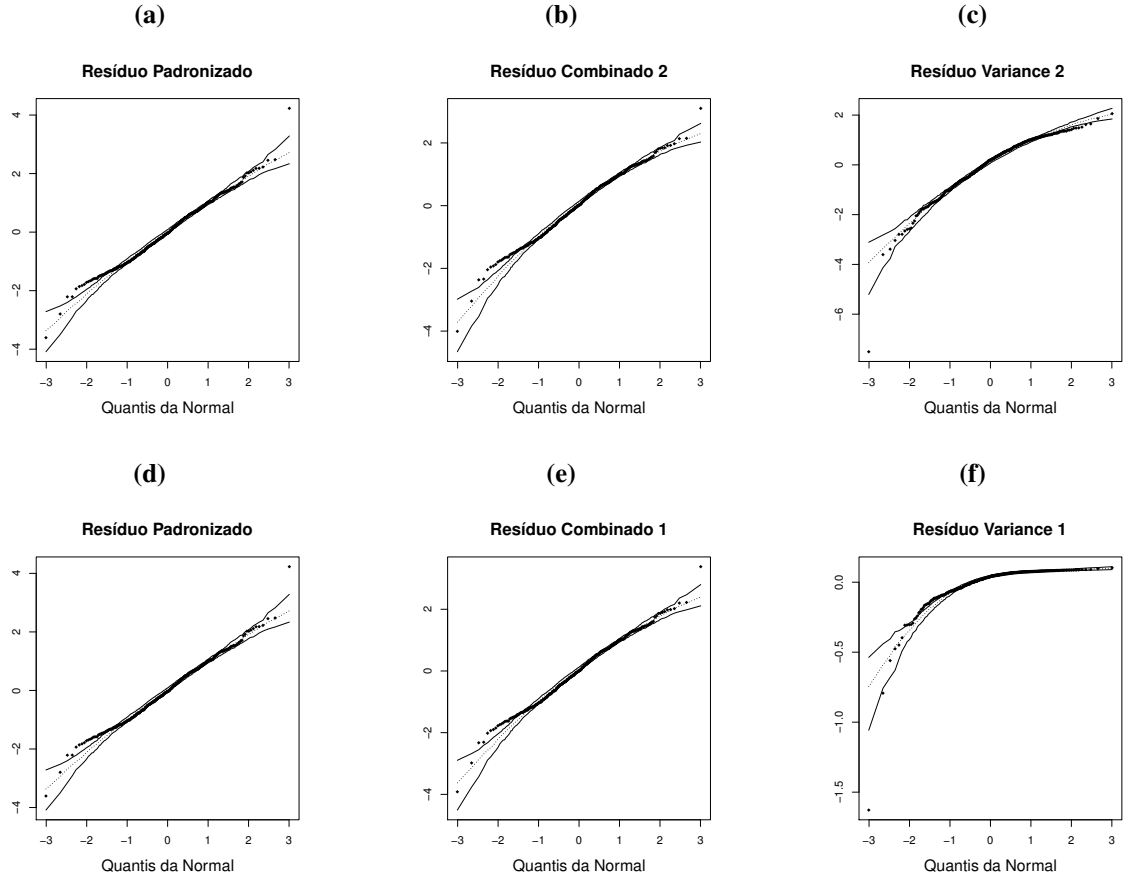
Fonte: do autor.

Com isso, assumimos o seguinte modelo de regressão beta com dispersão variável e função de ligação do tipo loglog apresentado em (5.4).

$$\begin{aligned} g(\mu_t) &= \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} \\ \log(\phi_t) &= \gamma_1 + \gamma_2 x_{t2} + \gamma_3 x_{t3} + \gamma_4 x_{t4}, \end{aligned} \qquad t = 1, \ldots, 379. \tag{5.4}$$

Para o submodelo da dispersão adicionamos a covariada hemoglobina glicosilada - HBGLI (x_4) e verificamos um grau de heterocedasticidade de $\hat{\lambda} = 7.71$ com mínimo=0.002 e máximo=0.016. O submodelo da média não apresentou alterações na significância, além de apresentar parâmetros significativos no submodelo da dispersão. Tais informações são apresentadas na Tabela 6. Os

Figura 9 – Gráficos de envelope simulado. Aplicação II: dados HDL/CT - dispersão fixa (loglog).

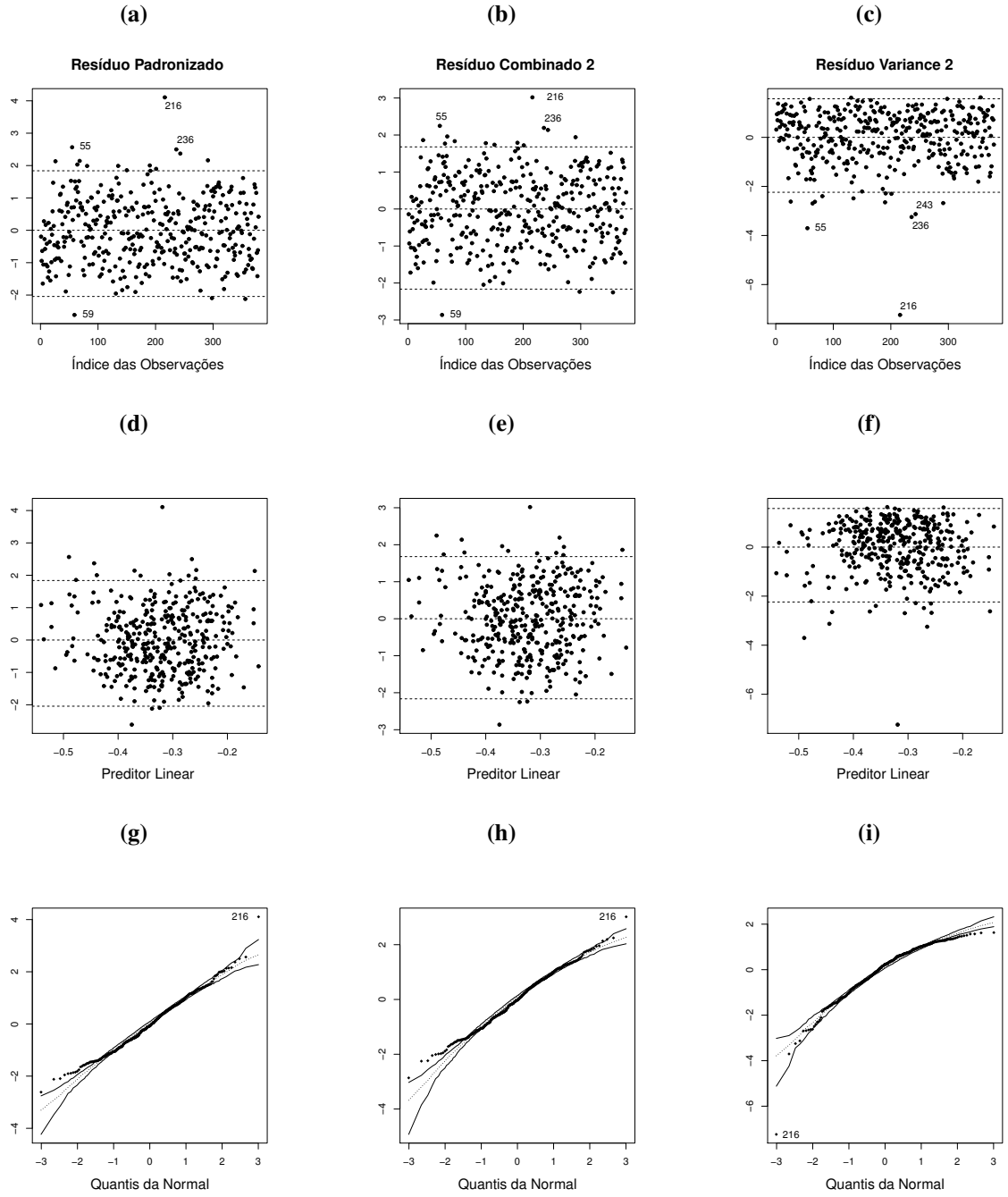


Fonte: do autor.

gráficos de resíduos do modelo são apresentados nas Figuras 10 e 11. Os gráficos dos painéis (a), (b) e (c) indicam que os resíduos *versus* observações possuem um comportamento aleatório e uniforme em torno de zero, além de destacar pontos atípicos (55, 59, 216, 236, 243). Os gráficos de resíduos *versus* o preditor linear - painéis (d), (e) e (f) - não apontaram para violações nas suposições de modelo. Por meio dos envelopes simulados - painéis (d), (e) e (f) - verificamos que o caso 216 é altamente atípico, além da ocorrência de pontos fora do envelope, evidenciando a inadequação da distribuição beta para a modelagem da variável resposta. Uma vez que a má especificação da distribuição de probabilidade pode gerar sensibilidade nas estimativas e casos influentes, também apresentamos na Tabela 6 as estimativas dos parâmetros, erros-padrão (e.p), mudanças relativas nas estimativas dos parâmetros (%) e nos erros-padrão (%) devido à exclusão das observações e os respectivos p-valores. Notamos que os pontos exercem um efeito considerável nas estimativas dos parâmetros γ_3 e γ_4 do submodelo da dispersão e que o caso 216 não apresenta indícios de ser influente.

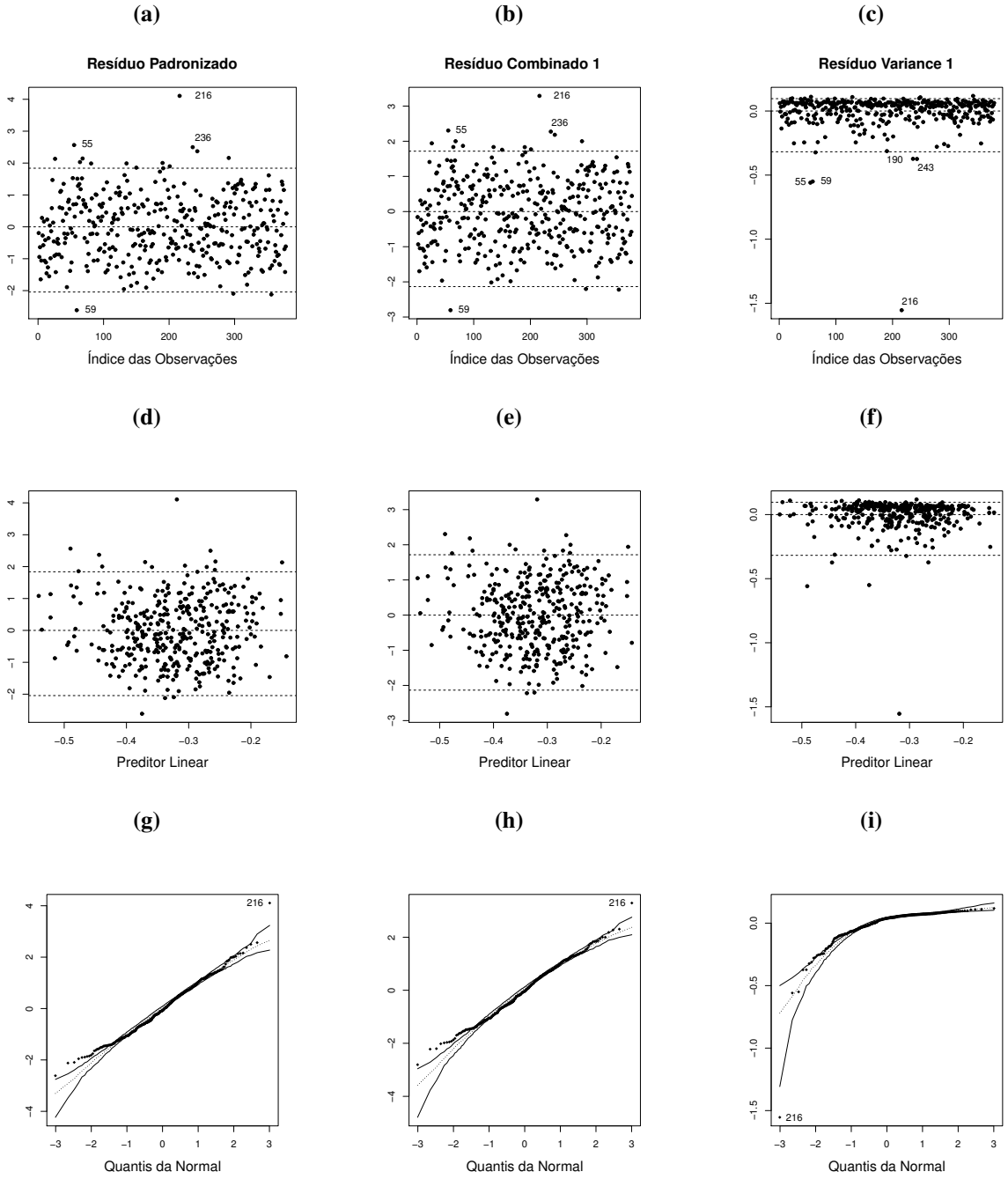
Na Tabela 7 são apresentados os valores das medidas para a avaliação da qualidade

Figura 10 – Gráficos dos resíduos. Estatísticas suficientes e completas. Aplicação II: dados HDL/CT - dispersão variável (loglog).



Fonte: do autor.

Figura 11 – Gráficos dos resíduos. Estatísticas suficientes e completas. Aplicação II: dados HDL/CT - dispersão variável (loglog).



Fonte: do autor.

Tabela 6 – Estimativas, erros-padrão e p-valores dos parâmetros. Aplicação II: dados HDL/CT.

Modelo		Dispersão fixa -loglog						
Descrição dos Parâmetros		β_1 (Const)	β_2 (IMC)	β_3 (CIN/QUA)	γ_1 (Const)	γ_2 (IMC)	γ_3 (CIN/QUA)	γ_4 (HBGLI)
Dataset completo	Estimat.	0.461	-0.009	-0.607	4.663	0.026	-1.778	-0.082
	e.p	0.147	0.002	0.162	0.920	0.011	1.032	0.033
	p-valor	0.002	0.000	0.000	0.000	0.017	0.085	0.012
obs. 55 deletada	Estimat.	0.098	0.023	0.081	-0.102	0.155	-0.261	0.062
	e.p	-0.004	0.000	-0.010	0.006	0.009	0.004	0.000
	p-valor	0.001	0.000	0.000	0.000	0.006	0.205	0.007
obs. 59 deletada	Estimat.	-0.008	-0.012	0.007	-0.012	0.053	0.091	-0.381
	e.p	-0.004	0.000	-0.006	0.001	0.000	0.001	0.028
	p-valor	0.002	0.000	0.000	0.000	0.012	0.061	0.128
obs. 216 deletada	Estimat.	0.033	0.000	0.040	-0.021	-0.091	-0.139	-0.012
	e.p	-0.025	0.000	-0.025	0.001	0.000	0.001	0.003
	p-valor	0.001	0.000	0.000	0.000	0.030	0.139	0.013
obs. 236 deletada	Estimat.	-0.055	0.000	-0.042	0.047	0.004	0.118	0.046
	e.p	-0.007	0.000	-0.007	0.002	0.000	0.001	0.000
	p-valor	0.003	0.000	0.000	0.000	0.016	0.054	0.008
obs. 243 deletada	Estimat.	0.028	0.035	0.015	-0.024	0.136	-0.027	0.034
	e.p	-0.007	0.000	-0.010	0.001	0.009	0.000	0.000
	p-valor	0.001	0.000	0.000	0.000	0.007	0.094	0.009
obs. $\{59,216,236\}$ deletadas	Estimat.	-0.026	-0.035	0.008	0.012	-0.038	0.062	-0.343
	e.p	-0.036	-0.063	-0.036	0.003	0.000	0.003	0.031
	p-valor	0.002	0.000	0.000	0.000	0.021	0.068	0.107
todas obs. deletadas	Estimat.	0.096	0.012	0.102	-0.123	0.273	-0.246	-0.233
	e.p	-0.050	-0.063	-0.059	0.010	0.009	0.008	0.031
	p-valor	0.000	0.000	0.000	0.000	0.003	0.198	0.060

Fonte: do autor.

do modelo em termos de variabilidade, e predição. Podemos verificar que o desempenho conjunto das medidas apontam para uma baixa qualidade do ajuste do modelo em termos de variabilidade. Também podemos notar o efeito pouco significativo que a remoção das observações exerce sobre a qualidade do modelo. Sobre as novas medidas R^2_{EC}, notamos a baixa qualidade do ajuste para o submodelo da média - $R^2_{EC_m}$ e $R^2_{EC_{mc}}$ - além da similaridade com o critério R^2_{FC} e sua forma corrigida. Para o submodelo da dispersão, as medidas $R^2_{EC_v}$ e $R^2_{EC_{vc}}$ apresentam valores mais baixos que os apresentados pelos critérios R^2_{FC}. Neste ponto, é digno de nota que a distribuição beta apresenta ajuste ruim quando os valores da variável resposta estão próximos dos limites do intervalo (0,1) devido a variância da distribuição.

Na Tabela 8 evidenciamos os valores das medidas referentes à qualidade preditiva e aderência. O uso das estatísticas suficientes e completas da distribuição beta revela valores intermediários para a estatítica PRESS, indicando relativa qualidade do modelo em termos de predição ao trabalhamos com a distribuição da variável resposta. Destacamos os valores da medida P^2 - maiores que 0.7 em ambos os submodelos. Adicionalmente, o novo critério R^2_{AC}

Tabela 7 – Critérios para a variabilidade. Aplicação II: dados HDL/CT.

	R^2_{FC}	$R^2_{FC_c}$	R^2_{LR}	$R^2_{LR_c}$	$R^2_{EC_m}$	$R^2_{EC_{mc}}$	$R^2_{EC_v}$	$R^2_{EC_{vc}}$
Dataset completo	0.116	0.101	0.150	0.132	0.124	0.110	0.108	0.093
obs. 55 deletada	0.126	0.112	0.161	0.144	0.134	0.120	0.118	0.104
obs. 59 deletada	0.116	0.101	0.141	0.124	0.118	0.104	0.109	0.094
obs. 216 deletada	0.121	0.107	0.153	0.135	0.128	0.114	0.122	0.108
obs. 236 deletada	0.114	0.100	0.152	0.134	0.123	0.109	0.107	0.092
obs. 243 deletada	0.122	0.108	0.159	0.141	0.130	0.116	0.114	0.100
obs. $\{59, 216, 236\}$ deletadas	0.119	0.105	0.145	0.128	0.122	0.107	0.122	0.108
todas obs. deletadas	0.137	0.123	0.168	0.151	0.139	0.125	0.142	0.128

Fonte: do autor.

Tabela 8 – Critérios para a predição e aderência. Aplicação II: dados HDL/CT.

	Est. Suficientes e Completas									Processo Iterativo Scoring de Fisher				
	$PRESS_M$	P^2_M	$P^2_{M_c}$	P^2_V	$P^2_{V_c}$	$R^2_{AD_m}$	$R^2_{AD_{mc}}$	$R^2_{AD_v}$	$R^2_{AD_{vc}}$	$PRESS_M$	P^2_M	$P^2_{M_c}$	P^2_V	$P^2_{V_c}$
Dataset completo	72.168	0.762	0.761	0.764	0.762	0.760	0.756	0.759	0.755	365.470	0.149	0.136	0.183	0.170
obs. 55 deletada	70.758	0.763	0.762	0.765	0.763	0.761	0.757	0.760	0.756	365.180	0.177	0.164	0.193	0.180
obs. 59 deletada	69.852	0.762	0.760	0.764	0.762	0.759	0.755	0.759	0.755	364.130	0.161	0.147	0.184	0.171
obs. 216 deletada	68.722	0.762	0.761	0.765	0.763	0.760	0.756	0.760	0.756	367.740	0.147	0.133	0.264	0.252
obs. 236 deletada	71.261	0.762	0.761	0.764	0.762	0.760	0.756	0.759	0.755	364.830	0.149	0.135	0.179	0.165
obs. 243 deletada	71.277	0.763	0.762	0.764	0.762	0.760	0.756	0.759	0.755	364.840	0.171	0.157	0.188	0.175
obs. $\{59, 216, 236\}$deletadas	65.490	0.762	0.761	0.765	0.763	0.759	0.755	0.760	0.756	365.820	0.158	0.144	0.273	0.261
todas obs. deletadas	63.204	0.764	0.762	0.767	0.765	0.761	0.757	0.762	0.758	365.250	0.211	0.198	0.313	0.302

Fonte: do autor.

permite avaliar a qualidade da aderência da distribuição aos dados e apresenta similaridades com os valores da medida P^2. Quanto às medidas derivadas do processo iterativo, notamos que a alta variabilidade do modelo estimado também impacta na qualidade preditiva, sendo notado pelos baixos valores das medidas de P^2 para o submodelo da média e altos valores para o submodelo da dispersão.

Podemos concluir que o modelo postulado apresenta baixa qualidade. Tomando os gráficos de envelope simulado, concluímos que a distribuição beta é inadequada para a modelagem da variável resposta. Casos atípicos foram detectados pelos resíduos, em especial pelos novos resíduos *variance*. Quanto à variabilidade, as medidas analisadas destacam baixa capacidade explicativa do modelo. Ao analisarmos a qualidade preditiva e aderência derivada da distribuição beta, observamos uma proximidade entre os valores, sendo estes maiores que 0.7. Neste caso, se considerarmos apenas os valores da medida P^2, temos indícios que a utilização da distribuição beta apresenta qualidade preditiva apenas para esse conjunto de dados. Analogamente, a aderência da distribuição para a modelagem também apresenta certa adequação ao conjunto de dados. Já, ao observamos a qualidade preditiva do modelo derivado do processo iterativo Scoring de Fisher, notamos que há baixa qualidade preditiva do modelo postulado.

5.4 APLICAÇÃO III

Buscando modelar a proporção do NHDL sobre o CT através das covariáveis hemoglobina glicada - HBLGI (x_2), circunferência da cintura - CIN (x_3), índice de massa corpórea - IMC (x_4) e circunferência do quadril - QUA (x_5). Assumimos o seguinte modelo de regressão beta com dispersão fixa e função de ligação do tipo loglog:

$$g(\mu_t) = \beta_1 + \beta_2 \log(x_{t2}) + \beta_3 x_{t3} + \beta_4 \log(x_{t4}) + \beta_5 x_{t5}^2, \qquad t = 1, \ldots, 381. \tag{5.5}$$

Na Tabela 9 apresentamos informações relativa aos parâmetros estimados pelo modelo. A análise dos gráficos de envelope simulados - apresentados na Figura 12 - indica a necessidade de modelagem da dispersão, além de destacar em todos os resíduos a presença de pontos fora do envelope, com destaque para o envelope do *resíduo combinado 2* - obtido a partir das estatísticas suficientes e completas da distribuição beta - que destacou a presença de um ponto potencialmente influente. Dada a necessidade da modelagem da dispersão, assumimos o seguinte modelo de

Tabela 9 – Estimativas, erros-padrão e p-valores dos parâmetros. Aplicação III: dados NHDL/CT.

Modelo		Dispersão fixa - C-loglog					
Descrição dos Parâmetros		β_1 (Const)	β_2 (log(HBGLI))	β_3 (CIN)	β_4 (log(IMC))	β_5 $(QUA)^2$	ϕ
Dataset completo	estimat.	-1.126	0.168	0.005	0.288	-0.000	3.448
	e.p.	0.269	0.037	0.002	0.106	0.000	0.072
	p-valor	0.000	0.000	0.002	0.007	0.011	0.000

Fonte: do autor.

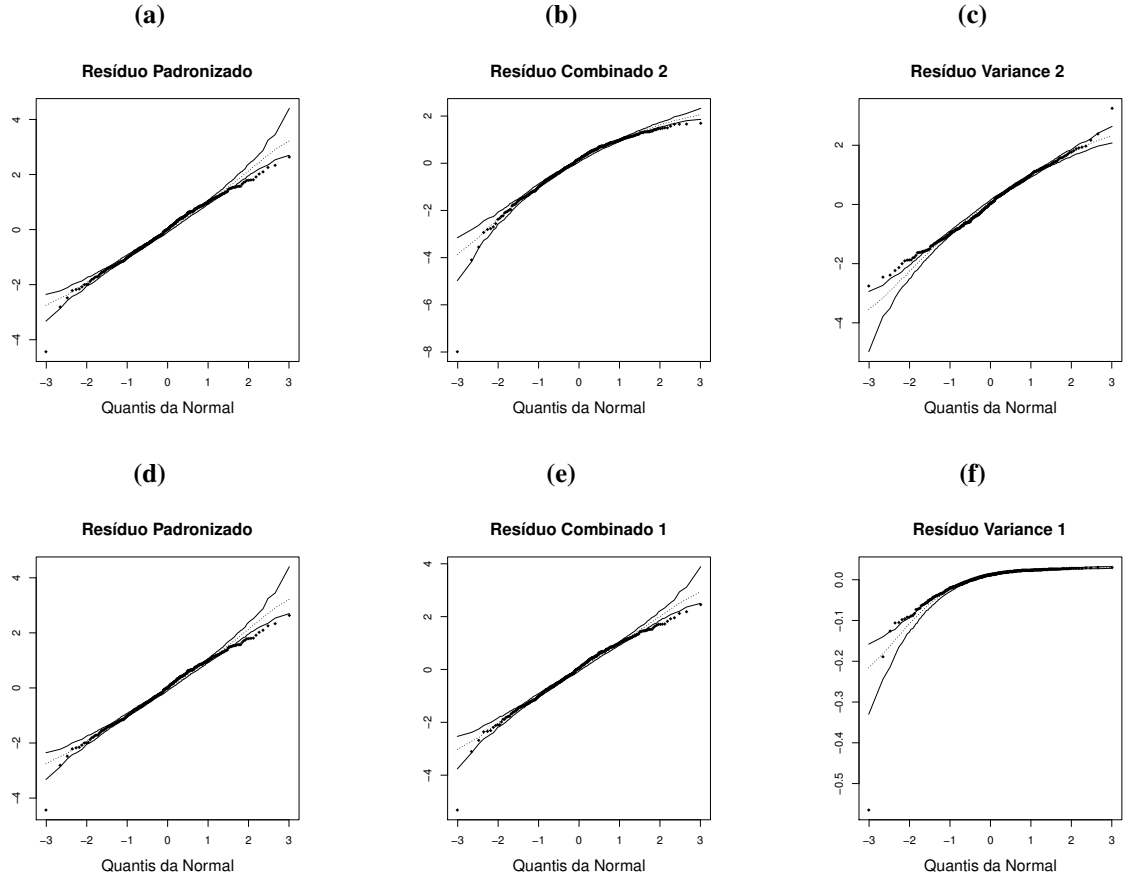
regressão beta com dispersão variável e função de ligação do tipo loglog apresentado em (5.6):

$$\begin{aligned} g(\mu_t) &= \beta_1 + \beta_2 \log(x_{t2}) + \beta_3 x_{t3} + \beta_4 \log(x_{t4}) + \beta_5 x_{t5}^2 \\ \log(\phi_t) &= \gamma_1 + \gamma_2 \log(x_{t2}) + \gamma_3 (\exp(x_{t3}))^{-1} + \gamma_4 x_{t6}^2 + \gamma_5 x_{t7}, \end{aligned} \qquad t = 1, \ldots, 381. \tag{5.6}$$

Adicionalmente, no submodelo da dispersão consideramos duas novas covariadas (peso - PSO (x_6) e idade - ID (x_7)) e observamos um grau de heterocedasticidade de $\hat{\lambda} = 2.53$ com mínimo=0.003 e máximo=0.008. O submodelo da média não apresentou alterações na significância, além de apresentar parâmetros significativos no submodelo da dispersão. Tais informações são apresentadas na Tabela 10.

Os gráficos de resíduos do modelo são apresentados nas Figuras 13 e 14. Os gráficos dos painéis (a), (b) e (c) indicam que os resíduos *versus* observações possuem comportamento aleatório em torno de zero, destacando pontos atípicos (56, 60, 65, 192), com destaque para o

Figura 12 – Gráficos de envelope simulado. Aplicação III: dados NHDL/CT - dispersão fixa (c-loglog).



Fonte: do autor.

caso 218. Os gráficos de resíduos *versus* o preditor linear - painéis (d), (e) e (f) - não apontaram para violações no modelo. Por meio dos envelopes simulados - painéis (d), (e) e (f) - verificamos que o caso 218 é altamente atípico - especialmente no envelope do *resíduo combinado 2* - , além da ocorrência de pontos fora do envelope, evidenciando a inadequação da distribuição beta para a modelagem da variável resposta. Para verificar o efeito dos destes pontos atípicos, apresentamos na Tabela 10 as estimativas dos parâmetros, erros-padrão (e.p), mudanças relativas nas estimativas dos parâmetros (%) e nos erros-padrão (%) devido à exclusão das observações e os respectivos p-valores. Notamos que os pontos exercem um efeito considerável na estimativa do parâmetro γ_5 do submodelo da dispersão, o caso 216 não apresenta indícios de ser influente e tomando os gráficos dos resíduos, a modelagem da dispersão apresentou redução no impacto inferencial do modelo.

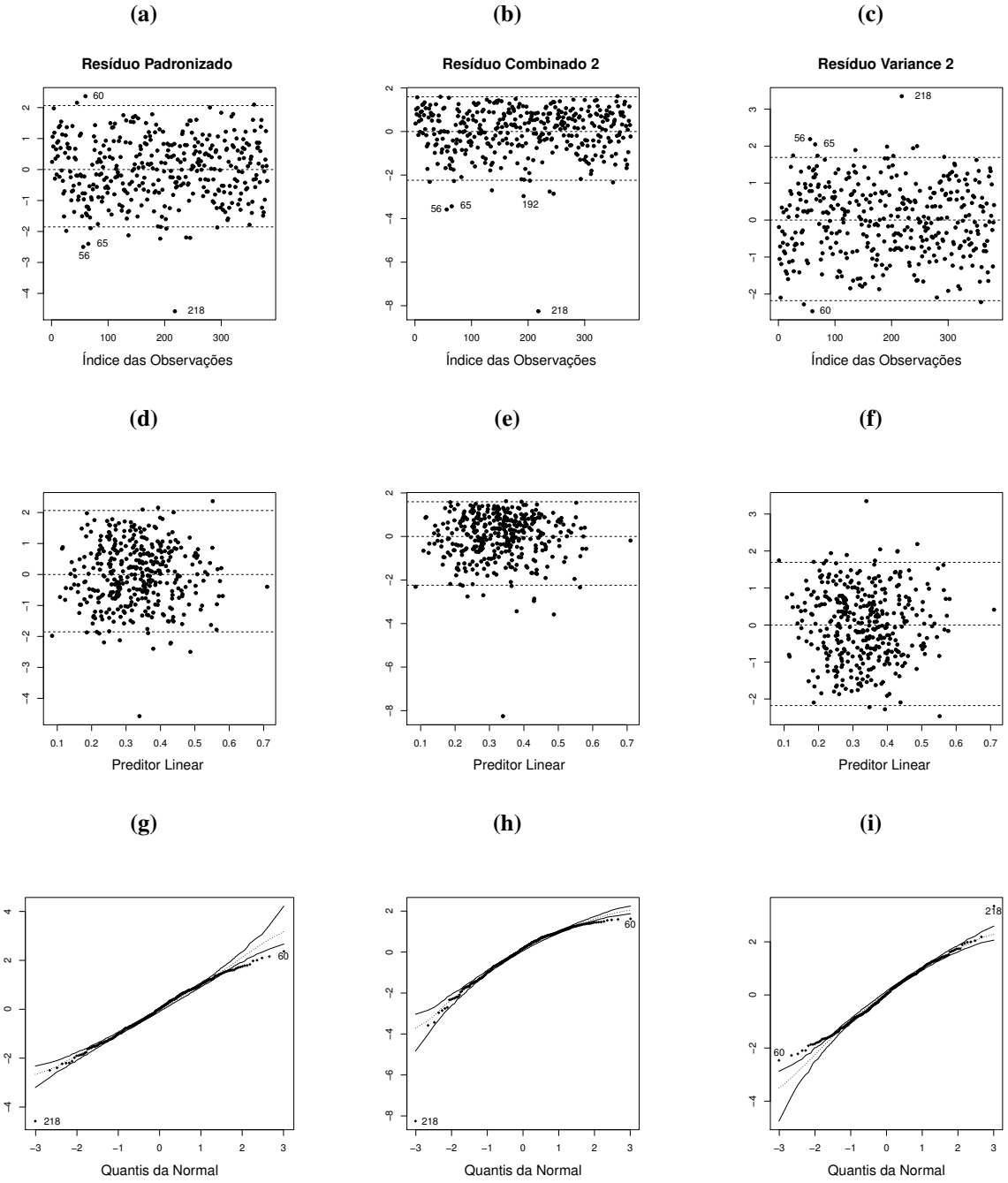
Na Tabela 11 são apresentados os valores das medidas para a avaliação da qualidade do modelo em termos de variabilidade e predição. Analogamente ao que observamos na aplicação 5.3, a avaliação das medidas aponta para baixa qualidade do ajuste do modelo em termos de

Tabela 10 – Estimativas, erros-padrão e p-valores dos parâmetros. Aplicação III: dados NHDL/CT.

Modelo		Dispersão fixa -loglog									
Descrição dos Parâmetros		β_1 (Const)	β_2 (log(HBGLI))	β_3 (CIN)	β_4 (log(IMC))	β_5 $(QUA)^2$	γ_1 (Const)	γ_2 (log(HBGLI))	γ_3 $(\exp(CIN))^{-1}$	γ_4 $(PSO)^2$	γ_5 $(ID)^2$
Dataset completo	Estimat.	-1.122	0.161	0.005	0.292	0.000	3.734	-0.318	0.000	0.004	-0.002
	e.p	0.268	0.039	0.002	0.105	0.000	0.000	0.000	0.000	0.000	0.000
	p-valor	0.000	0.000	0.002	0.006	0.007	0.000	0.000	0.000	0.000	0.000
obs. 56 deletada	Estimat.	0.008	-0.035	0.087	-0.008	0.000	-0.034	0.271	0.000	0.429	-1.632
	e.p	-0.008	0.000	0.000	-0.009	0.000	0.000	0.000	0.000	0.000	0.000
	p-valor	0.000	0.000	0.001	0.006	0.004	0.000	0.000	0.000	0.000	0.000
obs. 60 deletada	Estimat.	-0.027	-0.080	0.022	-0.015	0.000	-0.066	-0.410	0.000	0.167	0.211
	e.p	-0.001	-0.015	0.000	-0.002	0.000	0.000	0.000	0.000	0.000	0.000
	p-valor	0.000	0.000	0.001	0.006	0.005	0.000	0.000	0.000	0.000	0.000
obs. 65 deletada	Estimat.	0.004	0.066	0.000	-0.004	0.000	-0.027	-0.243	0.000	-0.238	-0.842
	e.p	-0.007	-0.020	0.000	-0.007	0.000	0.000	0.000	0.000	0.000	0.000
	p-valor	0.000	0.000	0.002	0.006	0.007	0.000	0.000	0.000	0.000	0.000
obs. 192 deletada	Estimat.	-0.005	-0.030	0.109	-0.027	0.000	0.005	0.094	0.000	0.048	-0.368
	e.p	-0.006	0.000	0.000	-0.006	0.000	0.000	0.000	0.000	0.000	0.000
	p-valor	0.000	0.000	0.001	0.007	0.004	0.000	0.000	0.000	0.000	0.000
obs. 218 deletada	Estimat.	0.025	0.039	0.043	0.022	0.000	0.038	-0.529	0.000	-0.095	3.526
	e.p	-0.039	-0.053	-0.067	-0.038	0.000	0.000	0.000	0.000	0.000	0.000
	p-valor	0.000	0.000	0.001	0.003	0.002	0.000	0.000	0.000	0.000	0.000
todas obs. deletadas	Estimat.	-0.008	-0.058	0.217	-0.038	0.000	-0.086	-0.825	0.000	0.286	0.842
	e.p	-0.058	-0.089	-0.067	-0.058	0.000	0.000	0.000	0.000	0.000	0.000
	p-valor	0.000	0.000	0.000	0.005	0.001	0.000	0.000	0.000	0.000	0.000

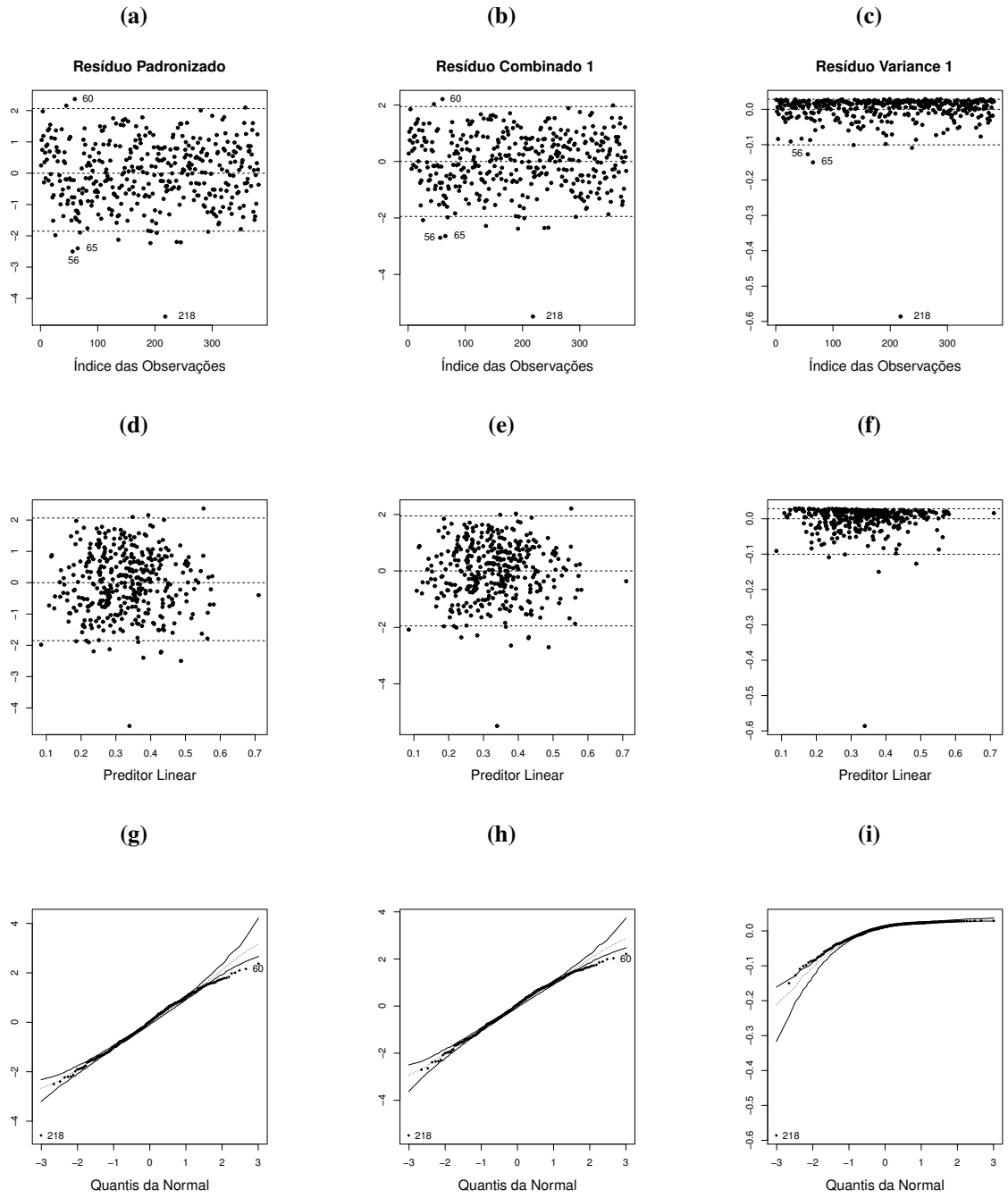
Fonte: do autor.

Figura 13 – Gráficos dos resíduos. Estatísticas suficientes e completas. Aplicação III: dados NHDL/CT - dispersão variável (c-loglog).



Fonte: do autor.

Figura 14 – Gráficos dos resíduos. Processo iterativo Scoring de Fisher. Aplicação III: dados NHDL/CT - dispersão variável (c-loglog).



Fonte: do autor.

Tabela 11 – Critérios para a variabilidade. Aplicação III: dados NHDL/CT.

	R^2_{FC}	$R^2_{FC_c}$	R^2_{LR}	$R^2_{LR_c}$	$R^2_{EC_m}$	$R^2_{EC_{mc}}$	$R^2_{EC_v}$	$R^2_{EC_{vc}}$
Dataset completo	0.187	0.168	0.190	0.165	0.184	0.164	0.187	0.167
obs. 56 deletada	0.196	0.177	0.200	0.176	0.193	0.173	0.195	0.175
obs. 60 deletada	0.178	0.158	0.178	0.154	0.174	0.154	0.176	0.156
obs. 65 deletada	0.193	0.174	0.192	0.167	0.190	0.170	0.192	0.173
obs. 192 deletada	0.193	0.173	0.195	0.171	0.190	0.170	0.192	0.173
obs. 218 deletada	0.196	0.177	0.209	0.185	0.192	0.172	0.191	0.171
todas obs. deletadas	0.211	0.191	0.212	0.189	0.206	0.187	0.202	0.183

Fonte: do autor.

Tabela 12 – Critérios para a predição e aderência. Aplicação III: dados NHDL/CT.

	Est. Suficientes e Completas									Processo Iterativo Scoring de Fisher				
	$PRESS_M$	P^2_M	$P^2_{M_c}$	P^2_V	$P^2_{V_c}$	$R^2_{AD_m}$	$R^2_{AD_{mc}}$	$R^2_{AD_v}$	$R^2_{AD_{vc}}$	$PRESS_M$	P^2_M	$P^2_{M_c}$	P^2_V	$P^2_{V_c}$
Dataset completo	67.415	0.768	0.765	0.769	0.766	0.763	0.757	0.762	0.757	67.415	0.567	0.556	0.890	0.887
obs. 56 deletada	66.138	0.769	0.766	0.770	0.767	0.764	0.758	0.763	0.758	66.138	0.605	0.595	0.895	0.892
obs. 60 deletada	65.702	0.767	0.765	0.768	0.765	0.762	0.756	0.762	0.756	65.702	0.587	0.577	0.884	0.881
obs. 65 deletada	65.892	0.768	0.766	0.769	0.767	0.763	0.758	0.763	0.757	65.892	0.581	0.571	0.897	0.894
obs. 192 deletada	66.450	0.768	0.766	0.769	0.767	0.763	0.758	0.763	0.757	66.450	0.579	0.569	0.893	0.890
obs. 218 deletada	63.944	0.768	0.766	0.768	0.766	0.763	0.757	0.762	0.756	63.944	0.610	0.601	0.944	0.942
todas obs. deletadas	58.290	0.770	0.768	0.770	0.768	0.765	0.759	0.764	0.758	58.290	0.677	0.670	0.957	0.955

Fonte: do autor.

variabilidade. Também podemos notar o efeito pouco significativo que a remoção das observações exerce sobre a qualidade do modelo. As novas medidas R^2_{EC} indicam a baixa qualidade do ajuste para o submodelo da média - $R^2_{EC_m}$ e $R^2_{EC_{mc}}$ - além da similaridade, já observada nas aplicações anteriores, com o critério R^2_{FC} e sua forma corrigida. No submodelo da dispersão, as novas medidas R^2_{EC} apresentam valores mais baixos que os apresentados pelos critérios R^2_{FC}.

Na Tabela 12 evidenciamos os valores das medidas referentes à qualidade preditiva e aderência. O uso das estatísticas suficientes e completas da distribuição beta apresenta valores intermediários para a estatítica PRESS, indicando relativa qualidade do modelo em termos de predição ao trabalhamos com a distribuição da variável resposta. Destacamos os valores da medida P^2 - maiores que 0.7 em ambos os submodelos. O novo critério R^2_{AC} avalia a qualidade da aderência da distribuição aos dados e indica similaridades com os valores da medida P^2. Quanto às medidas derivadas do processo iterativo, é digno de nota que os valores apresentados pela estatística PRESS foram os mesmos que os obtidos por meio das estatísticas suficientes e completas da distribuição beta. Além disso, as medidas de predição obtidas dos resíduos do processo iterativo Scoring de Fisher indicam razoável qualidade, podendo ser verificado através dos valores das medidas de P^2 para o submodelo da média e da dispersão.

Concluímos que o modelo postulado apresenta baixa qualidade. Tomando os gráficos de envelope simulado, concluímos que a distribuição beta apresenta algumas inadequações para

a modelagem da variável resposta. Casos atípicos foram detectados pelos resíduos, em especial pelo novo resíduo *combinado*. Quanto à variabilidade, as medidas analisadas destacam baixa capacidade explicativa do modelo. Ao analisarmos a qualidade preditiva e aderência derivada da distribuição beta, observamos proximidade entre os valores, sendo estes maiores que 0.7. Neste caso, se considerarmos apenas os valores da medida P^2, temos indícios que a utilização da distribuição beta atua bem no poder preditivo apenas para esse conjunto de dados. Analogamente, a aderência da distribuição para a modelagem também apresenta certa adequação ao conjunto de dados. Já, ao observamos a qualidade preditiva do modelo derivado do processo iterarivo Scoring de Fisher, notamos que há qualidade preditiva razoável do modelo postulado.

6 CONSIDERAÇÕES FINAIS

Neste estudo apresentamos o modelo de regressão beta proposto por Ferrari e Cribari-Neto (2004) e sua extensão para o caso em que o modelo considera a dispersão variável proposta por Smithson e Verkulien (2006). Tomando por base Espinheira *et al.* (2008), apresentamos um novo resíduo baseado no processo iterativo Scoring de Fisher, o qual denominamos de *resíduo variance 1*. Considerndo as covariadas, buscamos retirar o peso do modelo escolhido em prol da distribuição de probabilidade. Utilizamos o fato de que a distribuição beta forma uma família exponencial bidimensional e por meio do *Teorema da Função Integrável* - demonstrado por Barndorff-Nielsen (1978) e Lehmann (1986) - propomos uma nova classe de resíduos, dentre os quais destacamos o resíduo *variance 2*, e critérios do tipo pseudo-R^2 baseados nas estatísticas suficientes e completas com a finalidade de avaliar a variabilidade e aderência, além de realizar diagnósticos em modelos de aprendizado de máquina (*machine learning*) com distribuição beta. Além disso, utilizando os resíduos, propomos critérios de seleção dos modelos em termos de variabilidade e aderência. Quanto à predição, utilizamos a estatística PRESS e o coeficiente de predição P^2, introduzido por Espinheira *et al.* (2019). Para a avaliação das propostas apresentas no presente estudo fizemos o uso de três aplicações relacionadas ao risco de doenças cardíacas. Nelas avaliamos o desempenho da média da variável resposta (y_1) - razão entre a circunferência da cintura (cm) e a altura (cm), CIN/QUA - se encontra dispersa no intervalo da unidade padrão; média da variável resposta (y_2) - proporção entre o colesterol de alta densidade e o colesterol total, HDL/CT - está próxima de zero e média da variável resposta (y_3) - proporção entre o colesterol não HDL e o colesterol total, NHDL/CT - se encontra próxima de um.

E todas as aplicações o impacto inferencial foi atenuado pela modelagem da dispersão e que os novos resíduos *variance* destacaram pontos sensíveis ao submodelo da dispersão. Sendo tais resíduos derivados do submodelo da dispersão, os mesmos apresentam boa capacidade (sensibilidade) para identificar erros de especificação da dispersão. Esta sensibilidade pode ser observada em termos da capacidade da variabilidade explicada, por meio das novas medidas de R^2_{EC} para o submodelo da dispersão, uma vez que as mesmas apresentaram valores mais baixos que o R^2_{FC}. A qualidade da aderência dos dados à distribuição pode ser numericamente avaliada por meio das medidas R^2_{AD}. Além disso, quando baseados nas estatísticas suficientes e completas, as medidas R^2_{AD} e P^2 apresentam proximidade. Observarmos que isso é indício de que há uma relação entre as duas medidas, uma vez que a modelagem apresenta boas predições. Além disso, a medida R^2_{AD} também está relacionada com a qualidade dos envelopes simulados

quanto à suposição da adequabilidade da distribuição beta para modelagem da variável resposta. As aplicações próximas aos extremos do intervalo unitário apontaram para uma baixa qualidade dos modelos postulados, em especial quando observados os critérios de variabilidade. É digna de nota a comparação das medidas de predição derivadas do processo iterativo com as provenientes das estatísticas suficientes e completas. Nestas últimas, as informações derivadas das matrizes de pesos, em especial a matriz W que carrega informações sobre o preditor e sobre a variância, não são consideradas. Por tanto, observamos que a função do modelo utilizada para a obtenção de predições não é boa, ao passo que a distribuição de probabilidade apresenta-se adequada, indicando que a utilização da distribuição beta atua bem no poder preditivo apenas para esse conjunto de dados.

Finalmente, como sugestão de trabalhos futuros, buscaremos avaliar via simulações o comportamento de tais medidas, considerar os seguintes métodos de estimação: máxima verossimilhança aproximada e máxima verossimilhança perfilada, além de estudar o desempenho das medidas propostas quando a hipótese de não linearidade dos estimadores é assumida.

REFERÊNCIAS

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. Proc. of the 2nd Int. Symp. on Information Theory, p. 267–281, 1991.

ALLEN, D. M. The relationship between variable selection and data augmentation and a method for prediction. Technometrics, v. 16, p. 125–127, 1974.

BARNDORFF-NIELSEN, O. **Information and Exponential Families in Statistical Theory**. [S.l.]: New York: Wiley, 1978.

BAYER, F. M.; CRIBARI-NETO, F. Model selection criteria in beta regression with varying dispersion. Communications in Statistics – Simulation and Computation, v. 46, 2017.

DOORNIK, J. A. **An Object-Oriented Matrix Programming Language Ox**. [S.l.]: London: Timberlake Consultants Ltd, 2009. v. 6.

ESPINHEIRA, P. L.; FERRARI, S. L. P.; CRIBARI-NETO, F. On beta regression residuals. Journal of Applied Statistics, v. 35, n. 4, p. 407–419, 2008.

ESPINHEIRA, P. L.; SILVA, L. C. M.; SILVA, A. O.; OSPINA, R. Model selection criteria on beta regression for machine learning. Machine Learning and Knowledge Extraction, p. 427–449, 2019.

FERRARI, S.; NETO, F. C. Beta regression for modelling rates and proportions. Journal of Applied Statistics, v. 31, n. 7, p. 799–815, 2004.

FERRARI, S. L. P.; ESPINHEIRA, P. L.; CRIBARI-NETO, F. Diagnostic tools in beta regression with varying dispersion. Statistica Neerlandica, v. 65, n. 3, p. 337–351, 2011.

LAMPORT, L. **A Document Preparation System**. [S.l.]: Massachusetts, Addison- Wesley, 1994. v. 2.

LEHMANN, E. **Testing Statistical Hypotheses**. [S.l.]: New York: Springer-Verlag, 1986.

LEHMANN, E. L.; CASELLA, G. **Theory of Point Estimation**. [S.l.]: New York: Springerl, 1998. v. 2.

MEDIAVILLA, F.; LANDRUM, F.; SHAH, V. A. A comparison of the coeffcient of predictive power, the coeffcient of determination and aic for linear regression. Decision Sciences Institute, p. 1261–1266, 2008.

NAGELKERKE, N. J. D. A note on a general definition of the coefficient of determination. Biometrika, v. 78, n. 3, p. 691–692, 1991.

PAULA, G. A. **Modelos de Regressão: com apoio computacional**. [S.l.]: Instituto de Matemática e Estatística. Universidade de São Paulo, 2013.

PREGIBON, D. Logistic regression diagnostics. Annals of Statistics, v. 9, p. 705–724, 1981.

QUAN, T. N. The prediction sum of squares as a general measure for regression diagnostics. Journal of Business and Economic Statistics, v. 9, p. 501–504, 1988.

SCHWARZ, G. Estimating the dimension of a model. Annals of Statistics, v. 6, n. 2, p. 461–464, 1978.

SEN, P.; SINGER, J. M. **Large Sample Methods in Statistics: An Introduction With Applications**. [S.l.]: New York: Chapman and Hall, 1994. v. 6.

SMITHSON, M.; VERKULIEN, J. A better lemonsqueezer? Maximum likelihood regression with beta distribuited dependent variables. Psychological Methods, v. 11, n. 1, p. 54–71, 2006.

WILLEMS, J.; SAUNDERS, T.; HUNT, D.; J.B., S. Prevalence of coronary heart disease risk factors among rural blacks: a community-based study. Southern Medical Journal, v. 90, n. 8, p. 814–820, 1997.