Universidade Federal da Paraíba
Centro de Informática
Programa de Pós-Graduação em Modelagem Matemática e Computacional

MEDIDAS DE DIAGNÓSTICO EM MODELOS DE REGRESSÃO BETA PRIME

Maria Eduarda da Cruz Justino

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional da Universidade Federal da Paraíba, UFPB, como parte dos requisitos necessários à obtenção do título de Mestre em Modelagem Matemática e Computacional.

Orientadoras: Tarciana Liberal Pereira de Araújo
Tatiene Correia de Souza

João Pessoa
Agosto de 2023

Ata da Sessão Pública de Defesa de Dissertação de Mestrado de **MARIA EDUARDA DA CRUZ JUSTINO**, candidata ao título de Mestre em Modelagem Matemática e Computacional, realizada no dia 23 de agosto de 2023.

Aos vinte e três dias do mês de agosto do ano de dois mil e vinte e três, às 14h, via videoconferência, reuniram-se os membros da Banca Examinadora constituída para julgar o Trabalho Final da discente MARIA EDUARDA DA CRUZ JUSTINO**,** vinculada a Universidade Federal da Paraíba sob a matrícula nº 20211025883, candidata ao grau de Mestre em "*Modelagem Matemática e Computacional*", na linha de pesquisa *"Modelagem Probabilística"*, do Programa de Pós-Graduação em Modelagem Matemática e Computacional. A comissão examinadora foi composta pelos professores Tarciana Liberal Pereira de Araujo, Orientadora e Presidente da Banca; Tatiene Correia de Souza, Coorientadora; Pedro Rafael Diniz Marinho, Examinador Interno ao Programa; Ana Herminia Andrade e Silva, Examinadora Externa ao Programa; e Francisco Cribari Neto, Examinador Externo à Instituição. Dando início aos trabalhos, a Presidente da Banca cumprimentou os presentes, comunicou a finalidade da reunião e passou a palavra à candidata para que fizesse, oralmente, a exposição do trabalho de dissertação intitulado *"Medidas de diagnóstico em modelos de regressão beta prime"*. Concluída a exposição, a candidata foi arguida pela Banca Examinadora, que emitiu o seguinte parecer: "***aprovada***". Do ocorrido, eu, Gean Paulo P. M. de Barros, secretário do Programa de Pós-Graduação em Modelagem Matemática e Computacional (PPGMMC), lavrei a presente ata, que vai assinada por mim e pelos membros da Banca Examinadora.

João Pessoa, 23 de agosto de 2023.

Gean Paulo Pereira Maurício de Barros
Secretário do PPGMMC
SIAPE 2326476

Profª. Drª. Tarciana Liberal Pereira de Araujo
Orientadora (PPGMMC)

Documento assinado digitalmente
gov.br TARCIANA LIBERAL PEREIRA DE ARAUJO
Data: 24/08/2023 10:23:59-0300
Verifique em https://validar.iti.gov.br

Profª. Drª. Tatiene Correia de Souza
Coorientadora (PPGMMC)

Documento assinado digitalmente
gov.br TATIENE CORREIA DE SOUZA
Data: 30/08/2023 09:55:56-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Pedro Rafael Diniz Marinho
Examinador Interno ao Programa (PPGMMC)

Documento assinado digitalmente
gov.br PEDRO RAFAEL DINIZ MARINHO
Data: 29/08/2023 08:04:11-0300
Verifique em https://validar.iti.gov.br

Profª. Drª. Ana Hermínia Andrade e Silva
Examinadora Externa ao Programa (UFPB)

gov.br ANA HERMINIA ANDRADE E SILVA
Data: 28/08/2023 11:19:08-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Francisco Cribari Neto
Examinador Externo à Instituição (UFPE)

Documento assinado digitalmente
gov.br FRANCISCO CRIBARI NETO
Data: 24/08/2023 13:00:35-0300
Verifique em https://validar.iti.gov.br

MEDIDAS DE DIAGNÓSTICO EM MODELOS DE REGRESSÃO BETA PRIME

Maria Eduarda da Cruz Justino

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL (PPGMMC) DO CENTRO DE INFORMÁTICA DA UNIVERSIDADE FEDERAL DA PARAÍBA COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM MODELAGEM MATEMÁTICA E COMPUTACIONAL.

Examinada por:

Profª. Dra. Tarciana Liberal Pereira de Araújo, UFPB

Profª. Dra. Tatiene Correia de Souza, UFPB

Prof. Dr. Pedro Rafael Diniz Marinho, UFPB

Profª. Dra. Ana Hermínia Andrade e Silva, UFPB

Prof. Dr. Francisco Cribari Neto, UFPE

JOÃO PESSOA, PB – BRASIL
AGOSTO DE 2023

Catalogação na publicação
Seção de Catalogação e Classificação

J96m Justino, Maria Eduarda da Cruz.
Medidas de diagnóstico em modelos de regressão beta prime / Maria Eduarda da Cruz Justino. - João Pessoa, 2023.
85 f. : il.

Orientação: Tarciana Liberal Pereira de Araújo.
Coorientação: Tatiene Correia de Souza.
Dissertação (Mestrado) - UFPB/CI.

1. Matemática computacional. 2. Medidas de predição. 3. Simulação de Monte Carlo. 4. PRESS. 5. Resíduos - Matemática. I. Araújo, Tarciana Liberal Pereira de. II. Souza, Tatiene Correia de. III. Título.

UFPB/BC CDU 519.6(043)

Elaborado por GRACILENE BARBOSA FIGUEIREDO - CRB-15/794

Agradecimentos

A Deus, pelo seu amor incondicional e pelas bênçãos concedidas.

Aos meus pais, pelo amor, apoio e incentivo. Por sempre valorizarem a minha educação e não medirem esforços para que eu chegasse até aqui.

A Denn's, pelo amor, incentivo e compreensão.

À minha família e aos meus amigos, pelo carinho, apoio e torcida.

À minha orientadora Tarciana Liberal e coorientadora Tatiene Souza, por todo suporte dado para o desenvolvimento desta pesquisa, pelas reuniões, correções, disponibilidade e pelos conhecimentos compartilhados. Gratidão por todos os ensinamentos e por serem tão atenciosas.

Aos meus colegas de mestrado, Adriana, Carlos, David, Emannuel e Irineu, pela socialização e companheirismo durante o curso.

Aos professores do PPGMMC, pelos conhecimentos compartilhados e por enriquecerem minha formação acadêmica.

Aos professores do CEEF, por toda a experiência adquirida. Em especial, ao professor Bruno Frascarolli, pelos conselhos e ensinamentos.

Aos professores da banca examinadora Ana Hermínia, Francisco Cribari e Pedro Rafael, pelas contribuições feitas a este trabalho.

À Universidade Federal da Paraíba e ao Programa de Pós-Graduação em Modelagem Matemática Computacional, pela oportunidade concedida.

À Fundação PaqTcPB, pelo apoio financeiro.

Resumo da Dissertação apresentada ao PPGMMC/CI/UFPB como parte dos requisitos necessários para a obtenção do grau de Mestre em Modelagem Matemática e Computacional (M.Sc.)

MEDIDAS DE DIAGNÓSTICO EM MODELOS DE REGRESSÃO BETA PRIME

Maria Eduarda da Cruz Justino

Agosto/2023

Orientadoras: Tarciana Liberal Pereira de Araújo
Tatiene Correia de Souza

Programa: Modelagem Matemática e Computacional

No contexto de modelos para variável resposta contínua positiva, o modelo de regressão beta prime, proposto por Bourguignon et al. (2021), é atrativo para modelar dados positivos assimétricos. Na etapa de validação de um modelo de regressão, uma das técnicas de diagnóstico mais utilizada é a análise de resíduos. Para isso, é importante utilizar resíduos com propriedades conhecidas e que apresentem bom desempenho. Neste trabalho, realizamos um estudo detalhado dos resíduos no modelo de regressão BP. Com esse objetivo, além do resíduo quantílico e do resíduo de Pearson utilizados por Bourguignon et al. (2021), definimos os resíduos ponderado, ponderado padronizado (ESPINHEIRA et al., 2008) e Pearson padronizado (MCCULLAGH; NELDER, 1989) para tal modelo. Em seguida, avaliamos a distribuição empírica desses cinco resíduos em diferentes cenários do modelo de regressão BP e comparamos seus desempenhos para detectar erros de especificação. Simulações de Monte Carlo e aplicações a dados reais são utilizados para esse fim. Adicionalmente, investigamos o desempenho de algumas medidas de predição e de qualidade de ajuste como critérios de seleção de modelos na regressão BP. Nessa perspectiva, propomos o coeficiente de predição P^2, baseado na estatística PRESS, e avaliamos o comportamento dessa medida e dos critérios do tipo pseudo-R^2 através de estudos de simulações de Monte Carlo, considerando diferentes cenários do modelo de regressão BP sob especificação correta e incorreta. Aplicações a dados reais são apresentadas para ilustrar o desempenho das medidas propostas.

Palavras-chaves: Resíduos, Critérios de seleção de modelos, Medidas de predição, PRESS, Simulação de Monte Carlo.

Abstract of Dissertation presented to PPGMMC/CI/UFPB as a partial fulfillment of the requirements for the degree of Master in Modeling Mathematics and Computational (M.Sc.)

DIAGNOSTIC MEASURES IN BETA PRIME REGRESSION MODELS

Maria Eduarda da Cruz Justino

August/2023

Advisor: Tarciana Liberal Pereira de Araújo
Tatiene Correia de Souza

Program: Computational Mathematical Modelling

In the context of models for positive continuous response variables, the beta prime regression model, proposed by Bourguignon et al. (2021), is attractive to model asymmetrical positive data. In the validation stage of a regression model, one of the most commonly used diagnostic techniques is the analysis of residuals. For that, it is important to use residuals with known properties and that present good performance. In the present work, we performed a detailed study of the residuals in the BP regression model. To that end, in addition to the quantile residual and Pearson residual used by Bourguignon et al. (2021), we defined the weighted, standardized weighted (ESPINHEIRA et al., 2008) and standardized Pearson (MCCULLAGH; NELDER, 1989) residuals for this model. Afterward, we evaluate the empirical distribution of those five residuals in different scenarios of the BP regression model and compared their performances to detect misspecification. Studies of Monte Carlo simulations and applications to real data are used for that purpose. Furthermore, we investigated the performance of some prediction and goodness-of-fit measures as model selection criteria in BP regression. In this perspective, we propose the prediction coefficient P^2, based on the PRESS statistic, and we evaluate the behavior of this measure and the pseudo-R^2 type criteria through Monte Carlo simulation studies, considering correct and incorrect specifications of the BP regression model in different scenarios. Applications to real data are presented to illustrate the performance of the proposed measures.

Keywords: Residuals, Model selection criteria, Prediction measures, PRESS, Monte Carlo simulation.

Lista de Ilustrações

Lista de Tabelas

Sumário

1 Introdução

Os modelos de regressão são comumente utilizados em situações cujo objetivo é analisar o comportamento de uma variável de interesse (variável resposta) em relação a uma ou mais variáveis regressoras (variáveis explicativas). No contexto de modelos para variável resposta contínua positiva, Bourguignon et al. (2021) propuseram um modelo de regressão utilizando a distribuição beta prime (BP) reparametrizada em termos da sua média e de um parâmetro de precisão. O modelo proposto baseia-se na suposição de que a variável resposta segue distribuição BP, uma distribuição de probabilidade absolutamente contínua com suporte nos reais positivos, introduzida por Keeping (1962) e McDonald (1984). A utilização da distribuição BP reparametrizada no modelo de regressão BP proposto por Bourguignon et al. (2021) permite a modelagem simultânea da média e da precisão, assumindo que a resposta média e o parâmetro de precisão estão relacionados a um preditor linear por meio de uma função de ligação.

O modelo de regressão BP é atrativo para modelar dados positivos assimétricos, contudo, devido à flexibilidade da distribuição BP, dados simétricos também podem ser modelados. Tal modelo apresenta algumas vantagens em relação aos modelos de regressão baseados em distribuições com mesmo suporte, tais como os Modelos Lineares Generalizados (MLGs) com resposta gama ou gaussiana inversa (NELDER; WEDDERBURN, 1972) e o modelo de regressão Birnbaum-Saunders reparametrizado (SANTOS-NETO et al., 2016). Uma das vantagens refere-se à modelagem de dados mais dispersos, uma vez que a função de variância do modelo de regressão BP é maior que a função de variância do modelo gama. Além disso, a distribuição BP permite obter coeficientes de curtose e assimetria maiores do que as das distribuições gama e gaussiana inversa, o que pode ser desejável na modelagem de dados com assimetria acentuada e valores extremos significativos (BOURGUIGNON et al., 2021).

Na prática, espera-se que o modelo de regressão ajustado seja uma representação adequada da relação entre a variável resposta e as variáveis explicativas. Contudo, os modelos de regressão são aproximações da realidade e em algumas situações o ajuste pode não ser o ideal para os dados em estudo. Dessa forma, torna-se necessária a análise de diagnóstico para validar a qualidade do ajuste do modelo postulado e verificar possíveis afastamentos das suposições admitidas. Uma das técnicas de diagnóstico mais utilizada é a análise de resíduos, que permite identificar observações atípicas e/ou influentes, assim como verificar especificações incorretas do modelo, como a falta de adequação da distribuição considerada para a variável resposta ou da função de ligação, omissão de covariáveis importantes e suposição incorreta de precisão constante quando os dados apresentam precisão variável.

A proximidade da distribuição empírica dos resíduos pela distribuição normal padrão é bastante útil, pois permite que as propriedades dos resíduos sejam conhecidas, facilitando a interpretação dos seus gráficos. Na literatura, encontram-se disponíveis diversos estudos referentes à distribuição empírica de resíduos obtidos em modelos de regressão para a variável

resposta contínua positiva. Cepeda-Cuervo et al. (2016) propuseram dois resíduos baseados nas propriedades da família exponencial para o modelo de regressão gama reparametrizado e concluíram que as distribuições desses resíduos são bem aproximadas pela distribuição normal padrão. Santos-Neto et al. (2016) avaliaram as distribuições empíricas dos resíduos de Pearson, quantílico, escore e desvio para o modelo de regressão Birnbaum-Saunders reparametrizado e concluíram que o resíduo de Pearson apresenta distribuição assimétrica à direita. Scudilio e Pereira (2019) introduziram o resíduo quantílico ajustado e mostraram que sua distribuição é melhor aproximada pela distribuição normal padrão do que a dos outros resíduos investigados nos modelos de regressão gama e gaussiana inversa. Bourguignon et al. (2021) apresentaram o resíduo quantílico e o resíduo de Pearson para o modelo de regressão BP e concluíram que a distribuição do resíduo quantílico é bem aproximada pela distribuição normal padrão, diferentemente do resíduo de Pearson que apresentou distribuição assimétrica à direita.

Ainda no que se refere ao ajuste de modelos de regressão, algumas medidas são úteis para selecionar, dentre o conjunto de modelos candidatos, o modelo mais adequado para explicar o comportamento da variável de interesse. Tais medidas são conhecidas na literatura como critérios de seleção de modelos, tais como o Critério de Informação de Akaike (AIC) (AKAIKE, 1973), o Critério de Informação Bayesiano (BIC) (AKAIKE, 1978) ou Critério Bayesiano de Schwarz (SIC) (SCHWARZ, 1978) e o Critério de Informação de Hannan-Quinn (HQ) (HANNAN; QUINN, 1979). Na comparação entre modelos, o que minimizar essas medidas é o mais recomendado. Adicionalmente, têm-se as medidas baseadas na soma dos quadrados dos resíduos (SQR), como o coeficiente de determinação (R^2) e sua versão ajustada, variações do R^2 conhecidas como pseudo-R^2 (MCFADDEN, 1974; NAGELKERKE, 1991; FERRARI; CRIBARI-NETO, 2004) e versões corrigidas propostas por Bayer e Cribari-Neto (2017).

Entretanto, os critérios de seleção usuais, assim como as medidas do tipo pseudo-R^2, não fornecem informações sobre o desempenho preditivo de um modelo de regressão. Nessa perspectiva, a estatística PRESS (*Prediction Error Sum of Squares*), proposta por Allen (1974), pode ser utilizada como uma medida de indicação do poder preditivo de um modelo. Uma medida semelhante ao R^2 foi proposta por Wold (1982) baseada na PRESS e denominada coeficiente de predição (P^2), que pode ser utilizada como critério para selecionar modelos na perspectiva preditiva, evitando, assim, problemas de *overfitting*, isto é, modelos que se ajustam bem aos dados, mas não conseguem fazer boas previsões para novas observações. Dessa forma, a estatística P^2 mede a qualidade preditiva de um modelo de regressão, enquanto as medidas do tipo pseudo-R^2 estão relacionadas a qualidade do ajuste.

No contexto de seleção de modelos com maior qualidade preditiva, Espinheira et al. (2019) propuseram as estatísticas PRESS e P^2 baseadas em diferentes resíduos para os modelos de regressão beta linear e não linear. Via estudos de simulação de Monte Carlo, os autores avaliaram o desempenho dessas medidas e destacaram que o coeficiente de predição se mostrou útil na detecção de erros de especificação nos modelos de regressão beta.

1.1 Objetivos

O principal objetivo deste trabalho é desenvolver algumas ferramentas de diagnóstico e de qualidade de ajuste para o modelo de regressão BP. Inicialmente, buscamos realizar um estudo mais detalhado de resíduos no modelo de regressão BP. Para isso, além dos resíduos quantílico e de Pearson utilizados por Bourguignon et al. (2021), definimos, para tal modelo, os resíduos ponderado, ponderado padronizado (ESPINHEIRA et al., 2008) e Pearson padronizado (MCCULLAGH; NELDER, 1989). Em particular, objetivamos verificar se a distribuição normal padrão é uma boa aproximação para a distribuição empírica desses cinco resíduos, quando o modelo de regressão BP é corretamente especificado, e analisar seus desempenhos para detectar erros de especificação. Estudos de simulações de Monte Carlo em diferentes cenários sob correta e incorreta especificação do modelo BP são usados para esse fim.

Adicionalmente, objetivamos avaliar os desempenhos de algumas medidas de predição e de qualidade de ajuste como critérios de seleção de modelos na regressão BP. Especificamente, propor versões das estatísticas PRESS e P^2, baseadas em diferentes resíduos para o modelo de regressão BP, e avaliar o comportamento e o desempenho das medidas do tipo P^2 e pseudo-R^2 via estudos de simulações de Monte Carlo em diferentes cenários do modelo de regressão BP sob correta e incorreta especificação.

1.2 Organização da dissertação

Esta dissertação está estruturada em cinco Capítulos. No Capítulo 2, apresentamos a distribuição BP e o modelo de regressão BP associado. As definições dos resíduos utilizados por Bourguignon et al. (2021) e dos resíduos aqui propostos são apresentadas no Capítulo 3. Adicionalmente, apresentamos os resultados referentes a avaliação da distribuição empírica desses resíduos e avaliação dos desempenhos dos mesmos para detectar erros de especificação, finalizando o Capítulo com aplicações a dados reais. No Capítulo 4, definimos as medidas de qualidade de ajuste e as estatísticas de predição PRESS e P^2 para o modelo de regressão BP, realizamos estudos de simulações de Monte Carlo para avaliar o comportamento dessas medidas em diferentes cenários sob correta e incorreta especificação do modelo. Ainda no Capítulo 4, apresentamos aplicações a dados reais para ilustrar o desempenho das medidas propostas. Por fim, no Capítulo 5, apresentamos as conclusões desta dissertação.

1.3 Suporte computacional

As simulações computacionais e a análise gráfica apresentadas neste trabalho foram implementadas no `R` (R Core Team, 2022) em sua versão 4.1.3. O `R` é um ambiente computacional para análise estatística e produção de gráficos, disponível gratuitamente em <http://www.r-project.org>.

2 Modelo de regressão beta prime

2.1 Distribuição beta prime

A distribuição BP, também conhecida como distribuição beta inversa ou distribuição beta do segundo tipo, é uma distribuição de probabilidade absolutamente contínua com suporte nos reais positivos, introduzida por Keeping (1962) e McDonald (1984). Sua densidade pode assumir diferentes formas, dependendo dos valores dos parâmetros que a indexam, o que torna a distribuição BP bastante flexível. Ademais, a depender dos valores dos parâmetros, a função de taxa de falha da distribuição BP pode ter uma forma de "banheira invertida" ou forma crescente (BOURGUIGNON et al., 2021).

Seja Y uma variável aleatória que segue distribuição BP com parâmetros de forma $\alpha > 0$ e $\beta > 0$. Então, a função densidade de probabilidade da variável aleatória Y é escrita na forma

$$f(y;\alpha,\beta) = \frac{y^{\alpha-1}(1+y)^{-(\alpha+\beta)}}{B(\alpha,\beta)}, \qquad y \geq 0, \tag{2.1}$$

em que $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ é a função beta completa e $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ é a função gama. Utiliza-se a notação $Y \sim \text{BP}(\alpha,\beta)$ para denotar que Y é uma variável aleatória com distribuição BP de parâmetros α e β. A função de distribuição acumulada de $Y \sim \text{BP}(\alpha,\beta)$ é da forma

$$F(y;\alpha,\beta) = \mathbb{I}_{\frac{y}{1+y}}(\alpha,\beta), \qquad y \geq 0,$$

em que $\mathbb{I}_y(\alpha,\beta) = B(y;\alpha,\beta)/B(\alpha,\beta)$ é a função beta incompleta regularizada (também conhecida por função beta regularizada) com $B(y;\alpha,\beta) = \int_0^y u^{\alpha-1}(1-u)^{\beta-1}du$. A média e a variância de $Y \sim \text{BP}(\alpha,\beta)$ são dadas, respectivamente, por

$$\mathbb{E}[Y] = \frac{\alpha}{\beta-1}, \quad \beta > 1 \qquad \text{e} \qquad \text{Var}[Y] = \frac{\alpha(\alpha+\beta-1)}{(\beta-2)(\beta-1)^2}, \quad \beta > 2.$$

Usualmente, os modelos de regressão são construídos para modelar a média de uma variável resposta, bem como são definidos para conter uma estrutura de regressão para o parâmetro de precisão ou dispersão. No entanto, a densidade apresentada em (2.1) é indexada por dois parâmetros de forma. Diante disso, Bourguignon et al. (2021) propuseram uma reparametrização da distribuição BP, em termos da sua média e de um parâmetro de precisão, considerando $\mu = \alpha/(\beta-1)$ e $\phi = \beta - 2$. Sob essa reparametrização, a função densidade dada em (2.1) pode ser reescrita substituindo α por $\mu(1+\phi)$ e β por $\phi + 2$. Dessa forma, a função densidade de probabilidade, $f(y;\mu,\phi)$, e a função de distribuição acumulada, $F(y;\mu,\phi)$, de

$Y \sim \text{BP}(\mu, \phi)$ são dadas, respectivamente, por

$$f(y; \mu, \phi) = \frac{y^{\mu(1+\phi)-1}(1+y)^{-[\mu(1+\phi)+\phi+2]}}{B(\mu(1+\phi), \phi+2)} \tag{2.2}$$

e

$$F(y; \mu, \phi) = \mathbb{I}_{\frac{y}{1+y}}(\mu(1+\phi), \phi+2), \tag{2.3}$$

em que $\mu > 0$ e $\phi > 0$. A média e a variância da variável aleatória Y utilizando essa reparametrização são expressas, respectivamente, por

$$\mathbb{E}[Y] = \mu \qquad \text{e} \qquad \text{Var}[Y] = \frac{\mu(1+\mu)}{\phi} = \frac{V(\mu)}{\phi},$$

em que $V(\mu)$ é a função de variância. Nota-se que, para qualquer valor de μ fixo, ϕ é inversamente proporcional à variância da variável aleatória Y, ou seja, quanto maior o valor de ϕ, menor é a variância de Y. Logo, ϕ pode ser interpretado como um parâmetro de precisão e ϕ^{-1} caracteriza-se como um parâmetro de dispersão.

2.1.1 Representação na família exponencial

De acordo com Bourguignon et al. (2021), a distribuição BP reparametrizada, em termos da média e do parâmetro de precisão, pertence à família exponencial de distribuições. A representação da densidade expressa em (2.2) na forma da família exponencial permite obter algumas medidas úteis para as definições de novos resíduos para o modelo de regressão BP.

Seja P_η uma família de distribuições. Dizemos que P_η forma uma família exponencial na forma canônica de dimensão k se cada função densidade em P_η pode ser expressa na forma

$$p_\eta(x) = \exp\left[\sum_{j=1}^{k} \eta_j T_j(x) - \mathcal{A}(\boldsymbol{\eta})\right] h(x),$$

em relação a uma medida comum μ. Em que $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)^\top$ é o vetor de parâmetros naturais e $T_j(x)$ são estatísticas suficientes naturais, $j = i, \ldots, k$, sendo x um ponto no espaço amostral $\mathcal{X} = \{x : p_\eta > 0\}$, o suporte da densidade, $\mathcal{A}(\boldsymbol{\eta})$ é chamada de função geradora de cumulante e $h(x) : \mathcal{X} \to \mathbb{R}^+$. Além disso, o espaço paramétrico natural ($\mathcal{T}$) é definido por

$$\mathcal{T} = \left\{ \boldsymbol{\eta} \in \mathbb{R}^k : \mathcal{A}(\boldsymbol{\eta}) = \log \int \exp\left[\sum_{j=1}^{k} \eta_j T_j(x)\right] h(x) d\mu(x) < \infty \right\},$$

que descreve o conjunto de valores de $\boldsymbol{\eta}$ para os quais para todo $\boldsymbol{\eta} \in \mathcal{T}$ a densidade p_η é definida (LEHMANN; CASELLA, 1998).

Ante o exposto, é possível notar que a densidade da distribuição BP reparametrizada, em termos da média e do parâmetro de precisão, pode ser expressa na forma da família exponencial biparamétrica canônica. Isto é, a densidade apresentada em (2.2) pode ser escrita na forma

$$f(y;\mu,\phi) = \exp[\eta_1 T_1(y) + \eta_2 T_2(y) - \mathcal{A}(\boldsymbol{\eta})]h(y), \tag{2.4}$$

em que $\boldsymbol{\eta} = (\eta_1, \eta_2)^\top = (\mu(1+\phi)-1, -[\mu(1+\phi)+\phi+2])^\top$, $T_1(y) = \log(y)$, $T_2(y) = \log(1+y)$, $h(y) = 1$ e $\mathcal{A}(\boldsymbol{\eta}) = \log[\Gamma(\mu(1+\phi))] + \log[\Gamma(\phi+2)] - \log[\Gamma(\mu(1+\phi)+\phi+2)]$.

Consequentemente, $\mu(1+\phi) = \eta_1 + 1$, $\phi + 2 = -\eta_2 - \eta_1 - 1$, $\mu(1+\phi) + \phi + 2 = -\eta_2$ e $\mathcal{A}(\boldsymbol{\eta}) = \log[\Gamma(\eta_1+1)] + \log[\Gamma(-\eta_2-\eta_1-1)] - \log[\Gamma(-\eta_2)]$. Assim, utilizando as propriedades da família exponencial biparamétrica (LEHMANN; CASELLA, 1998), têm-se que

$$\mathbb{E}(T_1) = \frac{\partial \mathcal{A}(\boldsymbol{\eta})}{\partial \eta_1} = \psi(\eta_1+1) - \psi(-\eta_2-\eta_1-1) = \psi(\mu(1+\phi)) - \psi(\phi+2), \tag{2.5}$$

$$\mathrm{Var}(T_1) = \frac{\partial^2 \mathcal{A}(\boldsymbol{\eta})}{\partial \eta_1^2} = \psi'(\eta_1+1) + \psi'(-\eta_2-\eta_1-1) = \psi'(\mu(1+\phi)) + \psi'(\phi+2), \tag{2.6}$$

$$\mathbb{E}(T_2) = \frac{\partial \mathcal{A}(\boldsymbol{\eta})}{\partial \eta_2} = -\psi(-\eta_2-\eta_1-1) + \psi(-\eta_2) = \psi(\mu(1+\phi)+\phi+2) - \psi(\phi+2), \tag{2.7}$$

$$\mathrm{Var}(T_2) = \frac{\partial^2 \mathcal{A}(\boldsymbol{\eta})}{\partial \eta_2^2} = \psi'(-\eta_2-\eta_1-1) - \psi(-\eta_2) = \psi'(\phi+2) - \psi'(\mu(1+\phi)+\phi+2) \tag{2.8}$$

e

$$\mathrm{Cov}(T_1, T_2) = \frac{\partial^2 \mathcal{A}(\boldsymbol{\eta})}{\partial \eta_1 \partial \eta_2} = \psi'(-\eta_2-\eta_1-1) = \psi'(\phi+2). \tag{2.9}$$

2.2 Modelo de regressão beta prime

Modelos de regressão para variável resposta contínua positiva, considerando a parametrização usual da distribuição BP dada em (2.1), foram apresentados em McDonald e Butler (1990) e Tulupyev et al. (2013). Considerando a reparametrização expressa em (2.2), Bourguignon et al. (2021) propuseram o modelo de regressão BP contendo uma estrutura de regressão para a média e uma estrutura de regressão para o parâmetro de precisão. Tal modelo é mais atrativo em situações práticas, uma vez que permite a modelagem simultânea da média e do parâmetro de precisão, em vez dos parâmetros de forma. O modelo de regressão BP baseia-se na suposição de que a variável resposta segue distribuição BP e que a resposta média e o parâmetro de precisão estão, separadamente, relacionados a um preditor linear que envolve covariáveis e parâmetros desconhecidos por meio de uma função de ligação. Adicionalmente, assume-se linearidade entre os parâmetros.

Bourguignon et al. (2021) definiram o modelo de regressão BP da seguinte forma: Sejam $Y_1, \ldots, Y_n$ variáveis aleatórias independentes, tal que $Y_i \sim \mathrm{BP}(\mu_i, \phi_i)$, para $i = 1, \ldots, n$, ou seja, cada Y_i tem densidade dada em (2.2). O modelo de regressão BP com precisão variável supõe

que a média e o parâmetro de precisão de Y_i satisfazem as seguintes relações funcionais

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} = \eta_{1i} \qquad \text{e} \qquad h(\phi_i) = \mathbf{z}_i^\top \boldsymbol{\nu} = \eta_{2i}, \qquad i = 1, \ldots, n,$$

em que $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ e $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_q)^\top$ são vetores de parâmetros desconhecidos a serem estimados, tal que $\boldsymbol{\beta} \in \mathbb{R}^p$ e $\boldsymbol{\nu} \in \mathbb{R}^q$ com $p + q < n$, η_{1i} e η_{2i} são os preditores lineares, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ e $\mathbf{z}_i = (z_{i1}, \ldots, z_{iq})^\top$ são as observações sobre p e q regressores conhecidos. Dessa forma, as matrizes de covariáveis $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ e $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\top$ tem posto p e q, respectivamente. As funções de ligação $g : \mathbb{R} \to \mathbb{R}^+$ e $h : \mathbb{R} \to \mathbb{R}^+$ devem ser estritamente monótonas, positivas e pelo menos duas vezes diferenciáveis, de forma que $\mu_i = g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})$ e $\phi_i = h^{-1}(\mathbf{z}_i^\top \boldsymbol{\nu})$, com $g^{-1}(\cdot)$ e $h^{-1}(\cdot)$ sendo as funções inversas de $g(\cdot)$ e $h(\cdot)$, nessa ordem. As possíveis escolhas de funções de ligação para μ e ϕ são a logarítmica e raiz quadrada.

Os parâmetros do modelo são estimados pelo método da máxima verossimilhança. Dessa forma, utilizando a função densidade de probabilidade dada em (2.2) e assumindo amostra aleatória de tamanho n, a função log-verossimilhança, $\ell(\boldsymbol{\beta}, \boldsymbol{\nu})$, apresentada por Bourguignon et al. (2021), é expressa como

$$\ell(\boldsymbol{\beta}, \boldsymbol{\nu}) = \sum_{i=1}^{n} \ell(\mu_i, \phi_i), \qquad i = 1, \ldots, n, \tag{2.10}$$

em que

$$\begin{aligned}\ell(\mu_i, \phi_i) = {} & [\mu_i(1+\phi_i) - 1]\log(y_i) - [\mu_i(1+\phi_i) + \phi_i + 2]\log(1+y_i) - \log[\Gamma(\mu_i(1+\phi_i))] \\ & - \log[\Gamma(\phi_i + 2)] + \log[\Gamma(\mu_i(1+\phi_i) + \phi_i + 2)].\end{aligned}$$

A função escore é obtida através da diferenciação da função de log-verossimilhança (2.10) em relação aos parâmetros do modelo. Assim, o vetor escore para $\boldsymbol{\beta}$ e $\boldsymbol{\nu}$, apresentado em Bourguignon et al. (2021), é dado por $\mathbf{U}(\boldsymbol{\beta}, \boldsymbol{\nu}) = (\mathbf{U}_{\boldsymbol{\beta}}, \mathbf{U}_{\boldsymbol{\nu}})^\top$ tal que

$$\begin{aligned}\mathbf{U}_{\boldsymbol{\beta}} &= \mathbf{X}^\top \boldsymbol{\Phi} \mathbf{D}_1 (\boldsymbol{y}^* - \boldsymbol{\mu}^*), \\ \mathbf{U}_{\boldsymbol{\nu}} &= \mathbf{Z}^\top \mathbf{D}_2 (\boldsymbol{y}^\star - \boldsymbol{\mu}^\star),\end{aligned} \tag{2.11}$$

em que $\boldsymbol{\Phi}$, $\mathbf{D}_1$ e $\mathbf{D}_2$ são matrizes diagonais de ordem n, tal que $\boldsymbol{\Phi} = \text{diag}\{(1+\phi_i)\}$, $\mathbf{D}_1 = \text{diag}\left\{\dfrac{1}{g'(\mu_i)}\right\}$, $\mathbf{D}_2 = \text{diag}\left\{\dfrac{1}{h'(\phi_i)}\right\}$, $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)^\top$ em que $y_i^* = \log\left(\dfrac{y_i}{1+y_i}\right)$, $\boldsymbol{y}^\star = (y_1^\star, \ldots, y_n^\star)^\top$ em que $y_i^\star = \log\left(\dfrac{y_i^{\mu_i}}{(1+y_i)^{1+\mu_i}}\right)$, $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_n^*)^\top$ em que $\mu_i^* = \psi(\mu_i(1+\phi_i)) - \psi(\mu_i(1+\phi_i) + \phi_i + 2)$, $\boldsymbol{\mu}^\star = (\mu_1^\star, \ldots, \mu_n^\star)^\top$ em que $\mu_i^\star = \mu_i\left(\mu_i^* - \dfrac{\gamma_i}{\mu_i}\right)$ e $\gamma_i = \psi(\mu_i(1+\phi_i) + \phi_i + 2) - \psi(\phi_i + 2)$, sendo $\psi(\cdot)$ a função digama, definida por $\psi(z) = d\log\Gamma(z)/dz$. Para mais detalhes, ver o Apêndice A.

Com base na representação da distribuição BP na família exponencial (2.4) e utilizando as expressões dadas em (2.5) a (2.9) é possível encontrar o valor esperado e a variância de y_i^* e $y_i^\star$. Tais medidas são expressas, respectivamente, por

$$\mathbb{E}(y_i^*) = \psi(\mu_i(1+\phi_i)) - \psi(\mu_i(1+\phi_i)+\phi_i+2) = \mu_i^*, \tag{2.12}$$

$$\mathrm{Var}(y_i^*) = \psi'(\mu_i(1+\phi_i)) - \psi'(\mu_i(1+\phi_i)+\phi_i+2) = \upsilon_i \tag{2.13}$$

e

$$\begin{aligned}\mathbb{E}(y_i^\star) &= \mu_i[\psi(\mu_i(1+\phi_i)) - \psi(\mu_i(1+\phi_i)+\phi_i+2)] - \psi(\mu_i(1+\phi_i)+\phi_i+2) + \psi(\phi_i+2)\\ &= \mu_i\left(\mu_i^* - \frac{\gamma_i}{\mu_i}\right) = \mu_i^\star, \qquad (2.14)\\ \mathrm{Var}(y_i^\star) &= \mu_i^2[\psi'(\mu_i(1+\phi_i)) + \psi'(\phi_i+2)] + (1+\mu_i)^2[\psi'(\phi_i+2) - \psi'(\mu_i(1+\phi_i)+\phi_i+2)]\\ &\quad - 2\mu_i(1+\mu_i)\psi'(\phi_i+2)\\ &= \mu_i^2\psi'(\mu_i(1+\phi_i)) + \psi'(\phi_i+2) - (1+\mu_i)^2\psi'(\mu_i(1+\phi_i)+\phi_i+2) = \xi_i.\end{aligned}$$

A matriz da informação de Fisher requer as derivadas de segunda ordem da função de log-verossimilhança (2.10) em relação aos parâmetros do modelo. Assim, a matriz da informação de Fisher para o vetor de parâmetros $(\boldsymbol{\beta}, \boldsymbol{\nu})^\top$ do modelo de regressão BP é expressa como

$$\mathbf{K}(\boldsymbol{\beta}, \boldsymbol{\nu}) = \begin{bmatrix} \mathbf{K}_{\boldsymbol{\beta\beta}} & \mathbf{K}_{\boldsymbol{\beta\nu}} \\ \mathbf{K}_{\boldsymbol{\nu\beta}} & \mathbf{K}_{\boldsymbol{\nu\nu}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\top\boldsymbol{\Phi}\mathbf{W}\mathbf{X} & \mathbf{X}^\top\mathbf{D}\mathbf{Z} \\ \mathbf{Z}^\top\mathbf{D}\mathbf{X} & \mathbf{Z}^\top\mathbf{E}\mathbf{Z} \end{bmatrix}, \tag{2.15}$$

em que $\mathbf{K}_{\boldsymbol{\beta\nu}} = (\mathbf{K}_{\boldsymbol{\nu\beta}})^\top$. Além disso, $\mathbf{W}$, $\mathbf{D}$ e $\mathbf{E}$ são matrizes diagonais de ordem n, tal que $\mathbf{W} = \mathrm{diag}\left\{(1+\phi_i)\upsilon_i \dfrac{1}{[g'(\mu_i)]^2}\right\}$ em que $\upsilon_i = \psi'(\mu_i(1+\phi_i)) - \psi'(\mu_i(1+\phi_i)+\phi_i+2)$, $\mathbf{D} = \mathrm{diag}\left\{-(1+\phi_i)\zeta_i \dfrac{1}{g'(\mu_i)}\dfrac{1}{h'(\phi_i)}\right\}$ em que $\zeta_i = \psi'(\mu_i(1+\phi_i)+\phi_i+2) + \mu_i[\psi'(\mu_i(1+\phi_i)+\phi_i+2) - \psi'(\mu_i(1+\phi_i))]$ e $\mathbf{E} = \mathrm{diag}\left\{\xi_i \dfrac{1}{[h'(\phi_i)]^2}\right\}$ em que $\xi_i = \mu_i^2\psi'(\mu_i(1+\phi_i)) + \psi'(\phi_i+2) - (1+\mu_i)^2\psi'(\mu_i(1+\phi_i)+\phi_i+2)$, sendo $\psi'(\cdot)$ a função trigama, definida por $\psi'(z) = d^2\log\Gamma(z)/dz^2$. As respectivas derivadas da função de log-verossimilhança para a obtenção da matriz da informação de Fisher estão apresentadas no Apêndice A.

Em grandes amostras e sob certas condições de regularidades (COX; HINKLEYR, 1979), é possível obter a distribuição assintótica dos estimadores de máxima verossimilhança do modelo. Isto é,

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\nu}} \end{pmatrix} \overset{a}{\sim} \mathcal{N}_{p+q}\left(\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\nu} \end{pmatrix}, \mathbf{K}(\boldsymbol{\beta}, \boldsymbol{\nu})^{-1}\right),$$

em que $\widehat{\boldsymbol{\beta}}$ e $\widehat{\boldsymbol{\nu}}$ são os estimadores de máxima verossimilhança para $\boldsymbol{\beta}$ e $\boldsymbol{\nu}$, respectivamente. Tais

estimadores são obtidos resolvendo simultaneamente o sistema de equações não lineares

$$\begin{cases} \mathbf{U}_{\boldsymbol{\beta}} = 0 \\ \mathbf{U}_{\boldsymbol{\nu}} = 0. \end{cases}$$

No entanto, não apresentam forma fechada, sendo necessário a utilização de um método iterativo de otimização não linear, tais como os métodos de Newton-Raphson, quasi-Newton, escore de Fisher, entre outros. As estimativas dos parâmetros do modelo de regressão BP podem ser obtidas utilizando a função `gamlss` (RIGBY; STASINOPOULOS, 2005) contida no pacote de mesmo nome no ambiente computacional `R` (R Core Team, 2022).

3 Análise de resíduos

Neste Capítulo, apresentamos o resíduo quantílico e o resíduo de Pearson para o modelo de regressão BP e definimos três novos resíduos para esse modelo, denotados aqui por resíduo ponderado, resíduo ponderado padronizado e resíduo de Pearson padronizado. Via estudos de simulação de Monte Carlo, avaliamos as distribuições empíricas desses resíduos e seus desempenhos para detectar erros de especificação do modelo de regressão BP. Por fim, uma aplicação a dados reais é apresentada.

3.1 Resíduos para o modelo de regressão beta prime

3.1.1 Resíduo quantílico

Seja $F(y;\mu,\phi)$ a função de distribuição acumulada da variável resposta contínua Y, em que μ representa a média e ϕ é um parâmetro de precisão ou dispersão. O resíduo quantílico (DUNN; SMYTH, 1996) para modelos de regressão para variável resposta contínua é definido por

$$r_i^Q = \Phi^{-1}\left(F(y_i;\widehat{\mu}_i,\widehat{\phi}_i)\right), \qquad i = 1,\ldots,n, \tag{3.1}$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada normal padrão. Para o modelo de regressão BP, o resíduo quantílico apresentado em Bourguignon et al. (2021) assume que $F(\cdot)$ é dada em (2.3), $\widehat{\mu}_i = g^{-1}(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})$ e $\widehat{\phi}_i = h^{-1}(\mathbf{z}_i^\top\widehat{\boldsymbol{\nu}})$, com $\widehat{\boldsymbol{\beta}}$ e $\widehat{\boldsymbol{\nu}}$ estimadores de máxima verossimilhança de $\boldsymbol{\beta}$ e $\boldsymbol{\nu}$, respectivamente.

De acordo com Dunn e Smyth (1996), o resíduo quantílico possui distribuição assintótica normal padrão se os parâmetros do modelo são consistentemente estimados. Este fato baseia-se no Teorema da transformação integral de probabilidade (ANGUS, 1994).

Teorema 3.1.1 (Teorema da transformação integral de probabilidade)**.** *Se X é uma variável aleatória contínua com função de distribuição acumulada $F_X(\cdot)$, então a transformação $Y = F_X(X)$ é uniformemente distribuída no intervalo unitário.*

Demonstração. Seja $F_Y(y)$ a função de distribuição acumulada da variável aleatória Y. Temos que, $F_Y(y) = P(Y \leq y) = P(F_X(X) \leq y) = P(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y$, $0 < y < 1$. Logo, $Y \sim \mathcal{U}(0,1)$. □

Portanto, dado que $F(y_i;\widehat{\mu}_i,\widehat{\phi}_i)$ é a função de distribuição acumulada da variável resposta (contínua), temos pelo Teorema (3.1.1) que, para μ_i e ϕ_i conhecidos, $F(y_i;\mu_i,\phi_i) \sim \mathcal{U}(0,1)$. Ademais, seja $Z = \Phi^{-1}\left(F(y_i;\mu_i,\phi_i)\right)$, então $P(Z \leq z) = P(\Phi^{-1}\left(F(y_i;\mu_i,\phi_i)\right) \leq z) = P(F(y_i;\mu_i,\phi_i) \leq \Phi(z)) = \Phi(z)$. Logo, $\Phi^{-1}\left(F(y_i;\mu_i,\phi_i)\right) \sim \mathcal{N}(0,1)$.

3.1.2 Resíduo de Pearson

No contexto dos MLGs, introduzidos por Nelder e Wedderburn (1972) com o intuito de estender o modelo normal linear para abranger variáveis respostas com distribuições não normais, o resíduo de Pearson (MCCULLAGH; NELDER, 1989) é definido como a diferença padronizada entre o valor observado da variável resposta e seu respectivo valor ajustado. Nesse sentido, o resíduo de Pearson para o modelo de regressão BP é dado por

$$r_i^P = \frac{\widehat{\phi}_i^{\frac{1}{2}}(y_i - \widehat{\mu}_i)}{\sqrt{\widehat{\mu}_i(1+\widehat{\mu}_i)}}, \qquad i = 1, \ldots, n.$$

Bourguignon et al. (2021) realizaram um estudo de simulação de Monte Carlo para avaliar se as distribuições empíricas dos dois resíduos propostos para o modelo de regressão BP se aproximam da normal padrão. Os autores concluíram que a distribuição do resíduo quantílico é bem aproximada pela distribuição normal padrão, enquanto o resíduo de Pearson apresentou distribuição assimétrica à direita. Adicionalmente, estudos constataram que o resíduo de Pearson para modelos não normais possui assimetria acentuada e mesmo em grandes amostras não é normalmente distribuído (SANTOS-NETO et al., 2016; MEDEIROS, 2019; FENG et al., 2020).

3.1.3 Resíduo ponderado

Uma proposta de resíduo em Espinheira et al. (2008) para o modelo de regressão beta com dispersão constante, posteriormente adaptada por Ferrari et al. (2011) para o modelo beta com dispersão variável, é baseada no método iterativo escore de Fisher para obter a estimativa de máxima verossimilhança para $\boldsymbol{\beta}$. Tal método é definido como

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left[\mathbf{K}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{(m)}\right]^{-1} \mathbf{U}_{\boldsymbol{\beta}}^{(m)},$$

em que m representa as iterações do processo até a convergência. Baseado nessa proposta e utilizando as expressões do vetor escore e da matriz da informação de Fisher dadas, respectivamente, em (2.11) e (2.15), o método iterativo escore de Fisher para estimar $\boldsymbol{\beta}$ no modelo de regressão BP é dado por

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left[\mathbf{X}^\top \boldsymbol{\Phi}^{(m)} \mathbf{W}^{(m)} \mathbf{X}\right]^{-1} \mathbf{X}^\top \boldsymbol{\Phi}^{(m)} \mathbf{D}_1^{(m)} (\boldsymbol{y}^* - \boldsymbol{\mu}^{*(m)}). \tag{3.2}$$

É possível reescrever o processo iterativo (3.2) baseado em mínimos quadrados ponderados, uma extensão do método de mínimos quadrados ordinários para situações em que a hipótese de homocedasticidade não é válida (KUTNER et al., 1996). Assim,

$$\boldsymbol{\beta}^{(m+1)} = \left[\mathbf{X}^\top \boldsymbol{\Phi}^{(m)} \mathbf{W}^{(m)} \mathbf{X}\right]^{-1} \mathbf{X}^\top \boldsymbol{\Phi}^{(m)} \mathbf{W}^{(m)} \mathbf{z}^{(m)},$$

em que $\mathbf{z}^{(m)} = \mathbf{X}\boldsymbol{\beta}^{(m)} + [\mathbf{W}^{(m)}]^{-1}\mathbf{D}_1^{(m)}(\boldsymbol{y}^* - \boldsymbol{\mu}^{*(m)})$ é um vetor $p \times 1$. Após a convergência, obtém-se

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)^{-1} \mathbf{X}^\top \widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\widehat{\mathbf{z}}, \tag{3.3}$$

em que $\widehat{\mathbf{z}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{W}}^{-1}\widehat{\mathbf{D}_1}(\boldsymbol{y}^* - \widehat{\boldsymbol{\mu}}^*)$, sendo $\widehat{\mathbf{W}}$ e $\widehat{\mathbf{D}_1}$ as matrizes $\mathbf{W}$ e $\mathbf{D}_1$, respectivamente, avaliadas sob as estimativas de máxima verossimilhança.

Consequentemente, $\widehat{\boldsymbol{\beta}}$ em (3.3) pode ser visto como um estimador de mínimos quadrados ordinários de $\boldsymbol{\beta}$ através da regressão linear de $\check{\mathbf{y}} = \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\widehat{\mathbf{z}}$ em $\check{\mathbf{X}} = \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}$. O resíduo ordinário dessa regressão é definido por

$$\begin{aligned} \boldsymbol{r}^\beta &= \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\widehat{\mathbf{z}} - \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}\widehat{\boldsymbol{\beta}} \\ &= \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}(\widehat{\mathbf{z}} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{-\frac{1}{2}}\widehat{\mathbf{D}_1}(\boldsymbol{y}^* - \widehat{\boldsymbol{\mu}}^*), \end{aligned} \tag{3.4}$$

utilizando as definições de $\mathbf{W}$ e $\mathbf{D}_1$ apresentadas em (2.15), o resíduo da i-ésima observação é dado por

$$r_i^\beta = (1+\widehat{\phi_i})^{\frac{1}{2}}(1+\widehat{\phi_i})^{-\frac{1}{2}}\widehat{\upsilon_i}^{-\frac{1}{2}}\left[\frac{1}{(g'(\widehat{\mu_i}))^2}\right]^{-1/2}\frac{1}{g'(\widehat{\mu_i})}(y_i^* - \widehat{\mu_i^*}).$$

Segue que

$$r_i^\beta = \frac{y_i^* - \widehat{\mu_i^*}}{\sqrt{\widehat{\upsilon_i}}}, \qquad i = 1, \ldots, n, \tag{3.5}$$

em que $\mu_i^* = \mathbb{E}(y_i^*)$ e $\upsilon_i = \mathrm{Var}(y_i^*)$ são definidas em (2.12) e (2.13), respectivamente. O resíduo em (3.5) é baseado no resíduo ponderado padronizado 1 apresentado em Espinheira et al. (2008), modificado para o modelo de regressão BP e referenciado neste trabalho como resíduo ponderado.

3.1.4 Resíduo ponderado padronizado

Uma possível padronização do resíduo ponderado definido em (3.5) pode ser feita baseada na variância de $\mathbf{z}$, semelhante à proposta de Espinheira et al. (2008) para o resíduo ponderado padronizado 2. É possível reescrever a equação (3.3) como

$$\left(\mathbf{X}^\top \widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top \widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\widehat{\mathbf{z}},$$

e, dado que $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) \approx \left(\mathbf{X}^\top \widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}$, tem-se que

$$\left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)\mathrm{Cov}(\widehat{\boldsymbol{\beta}})\left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)^\top = \left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\right)\mathrm{Cov}(\widehat{\mathbf{z}})\left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\right)^\top$$

$$\left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)\left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right) \approx \left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\right)\mathrm{Cov}(\widehat{\mathbf{z}})\left(\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)$$

$$\mathrm{Cov}(\widehat{\mathbf{z}}) \approx \left(\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\right)^{-1}.$$

Utilizando a expressão dada em (3.3), o resíduo apresentado em (3.4) pode ser escrito na forma

$$\boldsymbol{r}^\beta = \left[\widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}} - \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\right]\widehat{\mathbf{z}}.$$

Segue que

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{r}^\beta) \approx{}& \left[\widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}} - \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\right]\left(\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\right)^{-1} \\
&\times\left[\widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}} - \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\right]^\top \\
={}& \left[\widehat{\boldsymbol{\Phi}}^{-\frac{1}{2}}\widehat{\mathbf{W}}^{-\frac{1}{2}} - \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}^\top\right] \\
&\times\left[\widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}} - \widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\,\mathbf{X}\left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\right] \\
={}& \mathbf{I}_n - \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}} - \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}} \\
&+\left[\widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\right]\left[\widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\right] \\
={}& (\mathbf{I}_n - \widehat{\mathbf{H}}).
\end{aligned}$$

Consequentemente, $\widehat{\mathrm{Cov}(\boldsymbol{r}^\beta)} = (\mathbf{I}_n - \widehat{\mathbf{H}})$, sendo $\mathbf{I}_n$ a matriz identidade de ordem n e

$$\widehat{\mathbf{H}} = \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}^\top\widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}} \tag{3.6}$$

é a matriz de projeção da solução de mínimos quadrados da regressão linear de $\check{\mathbf{y}} = \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\widehat{\mathbf{z}}$ em $\check{\mathbf{X}} = \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}$, tal que $\widehat{\mathbf{H}}$ é simétrica e idempotente. Portanto, o resíduo ponderado padronizado para o modelo de regressão BP, adaptado do resíduo ponderado padronizado 2 apresentado em Espinheira et al. (2008), é definido como

$$r_i^{\beta*} = \frac{r_i^\beta}{\sqrt{\widehat{\mathrm{Cov}(r_i^\beta)}}} = \frac{y_i^* - \widehat{\mu}_i^*}{\sqrt{\widehat{v}_i(1-\widehat{h_{ii}})}}, \qquad i = 1, \ldots, n, \tag{3.7}$$

em que $\widehat{h_{ii}}$ é o i-ésimo elemento da diagonal principal da matriz $\widehat{\mathbf{H}}$ dada em (3.6).

3.1.5 Resíduo de Pearson padronizado

Nos MLGs é comum utilizar a padronização do resíduo de Pearson introduzida por McCullagh e Nelder (1989). Estudos recentes vêm analisando o desempenho do resíduo de Pearson padronizado na análise de diagnósticos em modelos de regressão como, por exemplo, Scudilio e Pereira (2019), Sánchez et al. (2021), Amin et al. (2022a) e Amin et al. (2022b). Baseado na proposta de McCullagh e Nelder (1989), o resíduo de Pearson padronizado para o modelo de regressão BP é definido por

$$r_i^{P*} = \frac{\widehat{\phi}_i^{\frac{1}{2}}(y_i - \widehat{\mu}_i)}{\sqrt{\widehat{\mu}_i(1+\widehat{\mu}_i)(1-\widehat{h_{ii}})}}, \qquad i = 1, \ldots, n.$$

A padronização dos resíduos pela matriz de projeção pode melhorar o desempenho dos resíduos para identificar observações que podem influenciar nas estimativas dos parâmetros do modelo. Contudo, dado que a matriz de projeção é de ordem n, o cálculo dos resíduos padronizados pode ser custoso computacionalmente caso o tamanho da amostra seja grande.

3.2 Avaliação numérica

Nesta seção, apresentamos os resultados dos estudos de simulações de Monte Carlo com enfoque em verificar se as distribuições empíricas dos resíduos r_i^Q, r_i^P, r_i^{P*}, r_i^β e $r_i^{\beta*}$ são bem aproximadas pela distribuição normal padrão e analisar os desempenhos desses resíduos para detectar erros de especificação do modelo de regressão BP. Todas as simulações de Monte Carlo foram realizadas com 10000 réplicas através do ambiente computacional `R`, em que foram utilizados os pacotes `gamlss` (RIGBY; STASINOPOULOS, 2005) para a estimação dos parâmetros e `nortest` (GROSS; LIGGES, 2015) para os testes de normalidade.

3.2.1 Avaliação das distribuições dos resíduos sob especificação correta

O estudo da distribuição empírica dos resíduos obtidos sob especificação correta do modelo de regressão BP foi realizado em diferentes cenários. A estrutura do modelo de regressão BP considerada em todos os cenários é dada por

$$\log(\mu_i) = \beta_1 + \beta_2 x_i \qquad \text{e} \qquad \log(\phi_i) = \nu_1 + \nu_2 z_i, \tag{3.8}$$

com $i = 1, \ldots, n$. Os valores das covariáveis x_i e z_i foram obtidos de forma independente através da distribuição $\mathcal{U}(0,1)$ e mantidos fixos para cada réplica. Consideramos quatro tamanhos de amostras, a saber, $n = 20, 30, 60$ e 120. As simulações foram realizadas considerando quatro cenários do modelo de regressão BP: no cenário I $\mu \in (0{,}1; 24{,}3)$ e $\phi \in (2{,}7; 20{,}0)$; no cenário II $\mu \in (0{,}1; 24{,}3)$ e $\phi \in (20{,}3; 73{,}6)$; no cenário III temos que $\mu \in (20{,}1; 147{,}8)$ e $\phi \in (2{,}7; 20{,}0)$ e por fim, consideramos o cenário IV em que $\mu \in (20{,}1; 147{,}8)$ e $\phi \in (20{,}3; 73{,}6)$. Tais

cenários foram escolhidos com o objetivo de verificar se os valores assumidos para média e para o parâmetro de precisão podem influenciar na distribuição empírica dos resíduos analisados.

Para cada réplica de Monte Carlo calculamos os resíduos r_i^Q, r_i^P, r_i^{P*}, r_i^β e $r_i^{\beta*}$ e para cada observação i das 10000 réplicas calculamos a média, a variância, a assimetria e a curtose para os cinco resíduos avaliados, com $i = 1, \ldots, n$. Por fim, calculamos a média de cada uma dessas medidas para compará-las com os valores teóricos da distribuição normal padrão. Ou seja, buscamos verificar se, em média, a média, a variância, a assimetria e a curtose dos resíduos são próximas de $0, 1, 0$ e 3 (valores teóricos da distribuição normal padrão), respectivamente.

Adicionalmente, aplicamos o teste de Shapiro-Wilk (SW) que testa a hipótese nula $\mathcal{H}_0$: a amostra provém de uma população normalmente distribuída; contra a hipótese alternativa $\mathcal{H}_1$: a amostra não provém de uma população normalmente distribuída; com o intuito de testar a hipótese nula de que os resíduos provêm de uma população com distribuição normal em cada uma das 10000 réplicas de Monte Carlo. Calculamos a proporção de vezes que a hipótese nula não foi rejeitada aos níveis de 1%, 5% e 10% de significância. Vale destacar que o teste de SW possui maior poder para distribuições contínuas assimétricas quando comparado a outros testes de normalidade (RAZALI; WAH, 2011; YAP; SIM, 2011; WIJEKULARATHNA et al., 2020).

As Tabelas 3.1, 3.2, 3.3 e 3.4 apresentam os valores médios da média, variância, assimetria e curtose para os cinco resíduos avaliados e as proporções de vezes que, ao nível de 5% de significância, a hipótese nula dos testes de SW não foi rejeitada, considerando 10000 réplicas de Monte Carlo - SW(5%). As proporções referentes aos níveis de 1% e 10% de significância estão apresentadas no Apêndice B. Ainda nas referidas Tabelas encontram-se os valores dos parâmetros utilizados na geração dos dados para obter os cenários considerados nas simulações.

Observa-se nas Tabelas 3.1, 3.2, 3.3 e 3.4 que os cinco resíduos apresentam média próxima de 0 e variância próxima de 1, valores referenciais da distribuição normal padrão. Em termos de assimetria, o resíduo r_i^Q apresenta os menores valores em relação aos demais resíduos e esses valores se aproximam de 0 à medida que o tamanho da amostra aumenta. Por exemplo, na Tabela 3.1, nota-se que a assimetria do resíduo r_i^Q para $n = 20, 30, 60$ e 120 apresenta valores médios de 0,0799, 0,0616, 0,0366 e 0,0175, respectivamente. Os resíduos r_i^P e r_i^{P*} apresentam assimetria positiva e consideravelmente distantes de 0, principalmente nos cenários I e III em que $\phi \in (2{,}7; 20{,}0)$, como é possível observar nas Tabelas 3.1 e 3.3. Além disso, as assimetrias dos resíduos r_i^P e r_i^{P*} aumentam à medida que consideramos tamanhos de amostras maiores.

Os resíduos r_i^β e $r_i^{\beta*}$ apresentam assimetria negativa em todos os cenários estudados, com valores médios que se distanciam do valor referencial da distribuição normal padrão conforme o tamanho da amostra aumenta. Nota-se também que a assimetria negativa desses resíduos é mais acentuada nos cenários I e III e é razoavelmente próxima de 0 nos cenários II e IV, nos quais consideramos valores maiores para a precisão. Por exemplo, observa-se na Tabela 3.1 que no cenário I, quando $n = 30$, as assimetrias médias dos resíduos r_i^β e $r_i^{\beta*}$ são $-0{,}5174$ e $-0{,}5160$, respectivamente. Mantendo o mesmo intervalo para a média da variável resposta e considerando valores maiores para o parâmetro de precisão (cenário II), a Tabela 3.2 mostra que para $n = 30$

as assimetrias médias dos resíduos r_i^{β} e $r_i^{\beta*}$ são −0,2487 e −0,2480, respectivamente.

Tabela 3.1: Média das medidas descritivas dos resíduos r_i^{Q}, r_i^{P}, r_i^{P*}, r_i^{β} e $r_i^{\beta*}$ obtidos a partir do modelo de regressão BP sob especificação correta e proporções de não rejeição da hipótese nula dos testes de SW - Cenário I: $\mu \in (0{,}1; 24{,}3)$ e $\phi \in (2{,}7; 20{,}0)$, $\boldsymbol{\beta} = (3{,}2; -5{,}3)^\top$ e $\boldsymbol{\nu} = (1{,}0; 2{,}0)^\top$, $i = 1, \ldots, n$.

n	Resíduos	Média	Variância	Assimetria	Curtose	SW (5%)
20	r_i^{Q}	0,0034	0,9926	0,0799	2,5281	96,05
	r_i^{P}	0,0004	0,9554	1,2597	6,2831	55,03
	r_i^{P*}	−0,0008	1,0643	1,2367	6,1429	56,46
	r_i^{β}	0,0051	0,9629	−0,3901	2,6988	89,95
	$r_i^{\beta*}$	0,0047	1,0729	−0,3901	2,6921	90,11
30	r_i^{Q}	−0,0000	0,9979	0,0616	2,6623	95,70
	r_i^{P}	−0,0001	0,9700	1,4841	6,9826	29,98
	r_i^{P*}	−0,0003	1,0405	1,4780	6,9453	30,78
	r_i^{β}	0,0007	0,9732	−0,5174	3,0439	80,91
	$r_i^{\beta*}$	0,0006	1,0439	−0,5160	3,0413	80,99
60	r_i^{Q}	−0,0009	0,9988	0,0366	2,8153	95,13
	r_i^{P}	−0,0006	0,9838	1,6416	8,9373	6,65
	r_i^{P*}	−0,0006	1,0180	1,6400	8,9208	6,74
	r_i^{β}	−0,0004	0,9853	−0,5893	3,3570	60,41
	$r_i^{\beta*}$	−0,0004	1,0195	−0,5890	3,3563	60,76
120	r_i^{Q}	0,0001	1,0001	0,0175	2,9087	95,59
	r_i^{P}	0,0004	0,9952	1,8813	12,0646	0,08
	r_i^{P*}	0,0004	1,0121	1,8809	12,0596	0,09
	r_i^{β}	0,0002	0,9922	−0,6704	3,6452	23,80
	$r_i^{\beta*}$	0,0002	1,0090	−0,6703	3,6451	23,90

Tabela 3.2: Média das medidas descritivas dos resíduos r_i^{Q}, r_i^{P}, r_i^{P*}, r_i^{β} e $r_i^{\beta*}$ obtidos a partir do modelo de regressão BP sob especificação correta e proporções de não rejeição da hipótese nula dos testes de SW - Cenário II: $\mu \in (0{,}1; 24{,}3)$ e $\phi \in (20{,}3; 73{,}6)$, $\boldsymbol{\beta} = (3{,}2; -5{,}3)^\top$ e $\boldsymbol{\nu} = (3{,}0; 1{,}3)^\top$, $i = 1, \ldots, n$.

n	Resíduos	Média	Variância	Assimetria	Curtose	SW (5%)
20	r_i^{Q}	0,0012	0,9996	0,0401	2,5162	95,48
	r_i^{P}	0,0006	0,9941	0,5145	2,9607	85,65
	r_i^{P*}	0,0003	1,1063	0,5078	2,9455	85,99
	r_i^{β}	0,0015	0,9926	−0,1889	2,5585	94,25
	$r_i^{\beta*}$	0,0015	1,1047	−0,1876	2,5551	94,35
30	r_i^{Q}	−0,0001	0,9998	0,0284	2,6545	95,75
	r_i^{P}	−0,0002	0,9950	0,6043	3,3010	76,11
	r_i^{P*}	−0,0002	1,0667	0,6017	3,2943	76,45
	r_i^{β}	−0,0001	0,9945	−0,2487	2,7482	92,92
	$r_i^{\beta*}$	−0,0000	1,0662	−0,2480	2,7473	92,99
60	r_i^{Q}	0,0001	1,0000	0,0126	2,8219	95,39
	r_i^{P}	0,0000	0,9971	0,6438	3,5748	56,07
	r_i^{P*}	0,0000	1,0316	0,6432	3,5730	56,41
	r_i^{β}	0,0001	0,9974	−0,2909	2,9578	86,85
	$r_i^{\beta*}$	0,0001	1,0320	−0,2907	2,9576	86,89
120	r_i^{Q}	−0,0003	0,9998	0,0079	2,9095	95,10
	r_i^{P}	−0,0003	0,9985	0,6961	3,8215	24,35
	r_i^{P*}	−0,0003	1,0154	0,6960	3,8211	24,46
	r_i^{β}	−0,0002	0,9982	−0,3210	3,0760	75,30
	$r_i^{\beta*}$	−0,0002	1,0151	−0,3209	3,0759	75,34

Tabela 3.3: Média das medidas descritivas dos resíduos r_i^Q, r_i^P, r_i^{P*}, r_i^β e $r_i^{\beta*}$ obtidos a partir do modelo de regressão BP sob especificação correta e proporções de não rejeição da hipótese nula dos testes de SW - Cenário III: $\mu \in (20{,}1; 147{,}8)$ e $\phi \in (2{,}7; 20{,}0)$, $\boldsymbol{\beta} = (3{,}0; 2{,}0)^\top$ e $\boldsymbol{\nu} = (1{,}0; 2{,}0)^\top$, $i = 1, \ldots, n$.

n	Resíduos	Média	Variância	Assimetria	Curtose	SW (5%)
20	r_i^Q	0,0046	0,9944	0,0704	2,5117	95,91
	r_i^P	0,0009	0,9531	1,1427	5,1703	60,03
	r_i^{P*}	−0,0002	1,0632	1,1250	5,0946	61,79
	r_i^β	0,0060	0,9686	−0,3704	2,6635	90,25
	$r_i^{\beta*}$	0,0057	1,0804	−0,3700	2,6573	90,50
30	r_i^Q	0,0002	0,9982	0,0583	2,6775	95,63
	r_i^P	−0,0001	0,9750	1,3982	6,9449	36,46
	r_i^{P*}	−0,0003	1,0459	1,3915	6,8923	37,15
	r_i^β	0,0008	0,9795	−0,4618	2,9469	83,86
	$r_i^{\beta*}$	0,0008	1,0508	−0,4604	2,9451	84,10
60	r_i^Q	−0,0002	0,9994	0,0323	2,8169	95,40
	r_i^P	−0,0003	0,9835	1,4770	7,5529	10,49
	r_i^{P*}	−0,0004	1,0176	1,4753	7,5382	10,67
	r_i^β	−0,0000	0,9892	−0,5299	3,2161	65,70
	$r_i^{\beta*}$	−0,0000	1,0234	−0,5296	3,2155	65,99
120	r_i^Q	−0,0002	0,9999	0,0154	2,9052	94,94
	r_i^P	−0,0001	0,9921	1,7031	11,4110	0,48
	r_i^{P*}	−0,0001	1,0090	1,7027	11,4039	0,48
	r_i^β	−0,0001	0,9945	−0,5872	3,4142	32,90
	$r_i^{\beta*}$	−0,0001	1,0114	−0,5871	3,4140	33,24

Tabela 3.4: Média das medidas descritivas dos resíduos r_i^Q, r_i^P, r_i^{P*}, r_i^β e $r_i^{\beta*}$ obtidos a partir do modelo de regressão BP sob especificação correta e proporções de não rejeição da hipótese nula dos testes de SW - Cenário IV: $\mu \in (20{,}1; 147{,}8)$ e $\phi \in (20{,}3; 73{,}6)$, $\boldsymbol{\beta} = (3{,}0; 2{,}0)^\top$ e $\boldsymbol{\nu} = (3{,}0; 1{,}3)^\top$, $i = 1, \ldots, n$.

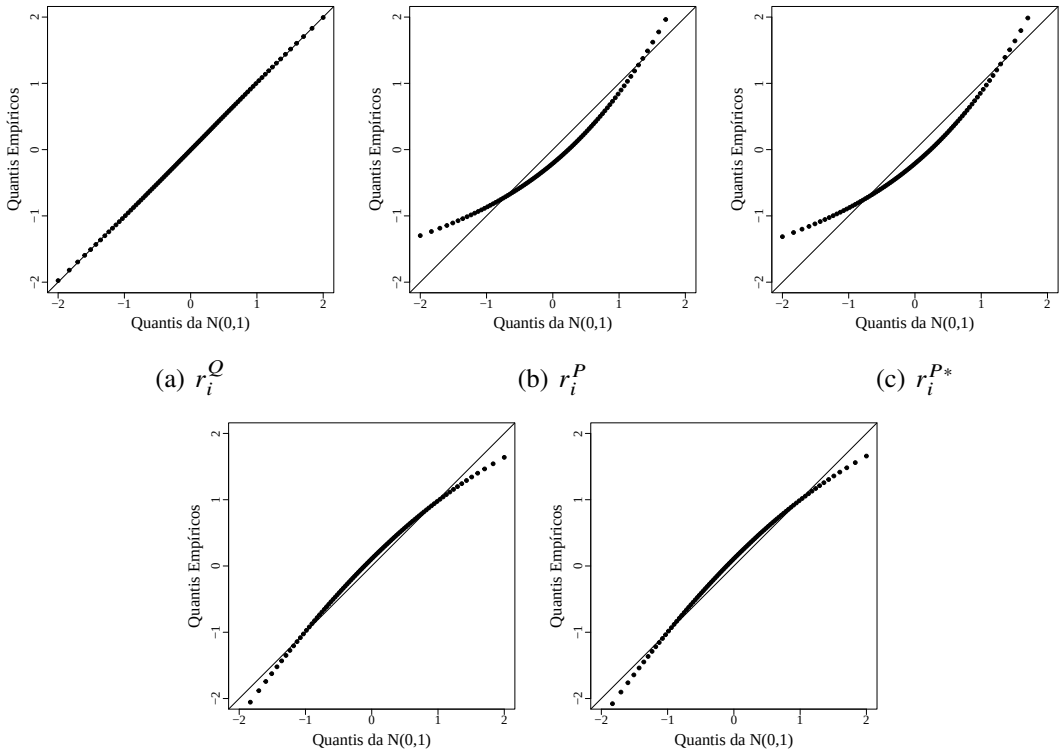
n	Resíduos	Média	Variância	Assimetria	Curtose	SW (5%)
20	r_i^Q	0,0002	0,9992	0,0281	2,5096	95,46
	r_i^P	−0,0006	0,9911	0,4776	2,9118	87,10
	r_i^{P*}	−0,0007	1,1034	0,4714	2,8984	87,27
	r_i^β	0,0006	0,9942	−0,1883	2,5500	94,15
	$r_i^{\beta*}$	0,0007	1,1069	−0,1874	2,5463	94,37
30	r_i^Q	−0,0001	0,9998	0,0302	2,6506	95,87
	r_i^P	−0,0002	0,9959	0,5487	3,1773	79,96
	r_i^{P*}	−0,0003	1,0676	0,5462	3,1710	80,19
	r_i^β	0,0001	0,9950	−0,2184	2,7128	93,42
	$r_i^{\beta*}$	0,0000	1,0668	−0,2176	2,7116	93,35
60	r_i^Q	0,0003	1,0001	0,0140	2,8203	95,72
	r_i^P	0,0003	0,9978	0,5900	3,4633	61,28
	r_i^{P*}	0,0002	1,0323	0,5894	3,4617	61,52
	r_i^β	0,0003	0,9977	−0,2610	2,9161	88,74
	$r_i^{\beta*}$	0,0003	1,0322	−0,2608	2,9159	88,83
120	r_i^Q	0,0003	1,0001	0,0060	2,9047	95,45
	r_i^P	0,0003	0,9987	0,6260	3,6594	33,18
	r_i^{P*}	0,0003	1,0157	0,6259	3,6591	33,32
	r_i^β	0,0003	0,9989	−0,2882	3,0277	79,42
	$r_i^{\beta*}$	0,0003	1,0159	−0,2882	3,0277	79,47

Com relação à curtose, o resíduo r_i^Q apresenta, em todos os cenários considerados, valores médios que se aproximam de 3 à medida que o tamanho da amostra aumenta. Os resíduos r_i^P e r_i^{P*} apresentam coeficientes de curtose consideravelmente maiores que 3 nos cenários I e III, indicando que as distribuições desses resíduos nesses cenários possuem cauda mais pesada que a distribuição normal padrão. Nos cenários II e IV, as curtoses dos resíduos r_i^P e r_i^{P*} são próximas de 3 para $n = 20$ e 30, contudo, se distanciam desse valor referencial para os demais tamanhos de amostra. Nota-se que os resíduos r_i^β e $r_i^{\beta*}$ apresentam valores médios de curtose que se aproximam de 3 ao passo que o tamanho da amostra aumenta nos cenários II e IV, entretanto, nos cenários I e III observa-se um distanciamento do valor referencial quando n aumenta.

No que diz respeito às proporções referentes aos testes de SW, nota-se nos quatro cenários estudados que a hipótese de que o resíduo r_i^Q provém de uma população normalmente distribuída não foi rejeitada, ao nível de 5% de significância, em torno de 95% das vezes. As proporções referentes aos resíduos r_i^P e r_i^{P*} indicam evidências de que a aderência da distribuição desses resíduos à distribuição normal padrão não é satisfatória em todos os cenários considerados. Com relação às proporções referentes aos resíduos r_i^β e $r_i^{\beta*}$, observa-se que nos cenários em que o parâmetro de precisão assume valores maiores (cenários II e IV), as proporções de vezes em que não rejeitamos a hipótese de que os resíduos r_i^β e $r_i^{\beta*}$ provêm de uma população normalmente distribuída são consideravelmente altas, contudo, há um distanciamento do nível nominal com o aumento do tamanho amostral. Uma justificativa é que, como foi observado, a assimetria desses resíduos tende a se distanciar do valor referencial da distribuição normal padrão quando n aumenta. Resultados semelhantes podem ser observados quando consideramos os níveis de 1% e 10% de significância (ver o Apêndice B).

Outra forma de verificar a proximidade da distribuição empírica dos resíduos pela distribuição normal padrão é utilizando análise gráfica. Para isso, construímos gráficos normal de probabilidades para comparar os quantis empíricos dos resíduos com os quantis da normal padrão. As Figuras 3.1, 3.2, 3.3 e 3.4 apresentam os gráficos normal de probabilidades dos resíduos obtidos nos cenários I, II, III e IV, respectivamente, considerando o modelo definido em (3.8) e um tamanho amostral de $n = 90$. Tais gráficos estão em conformidade com a análise das medidas descritivas realizada anteriormente. A distribuição do resíduo r_i^Q apresenta boa aderência à distribuição normal padrão em todos os cenários considerados. Os resíduos r_i^P e r_i^{P*} apresentam comportamentos bastantes similares, ambos se distribuem com assimetria à direita, principalmente nos cenários em que $\phi \in (2{,}7; 20{,}0)$ (Figuras 3.1(b)-(c) e 3.3(b)-(c)). Já a distribuição dos resíduos r_i^β e $r_i^{\beta*}$ é tipicamente bem aproximada pela distribuição normal padrão quando consideramos valores maiores para o parâmetro de precisão (Figuras 3.2(d)-(e) e 3.4(d)-(e)), no entanto, nota-se que esses resíduos se distribuem com leve assimetria negativa nos cenários em que $\phi \in (2{,}7; 20{,}0)$ (Figuras 3.1(d)-(e) e 3.3(d)-(e)).

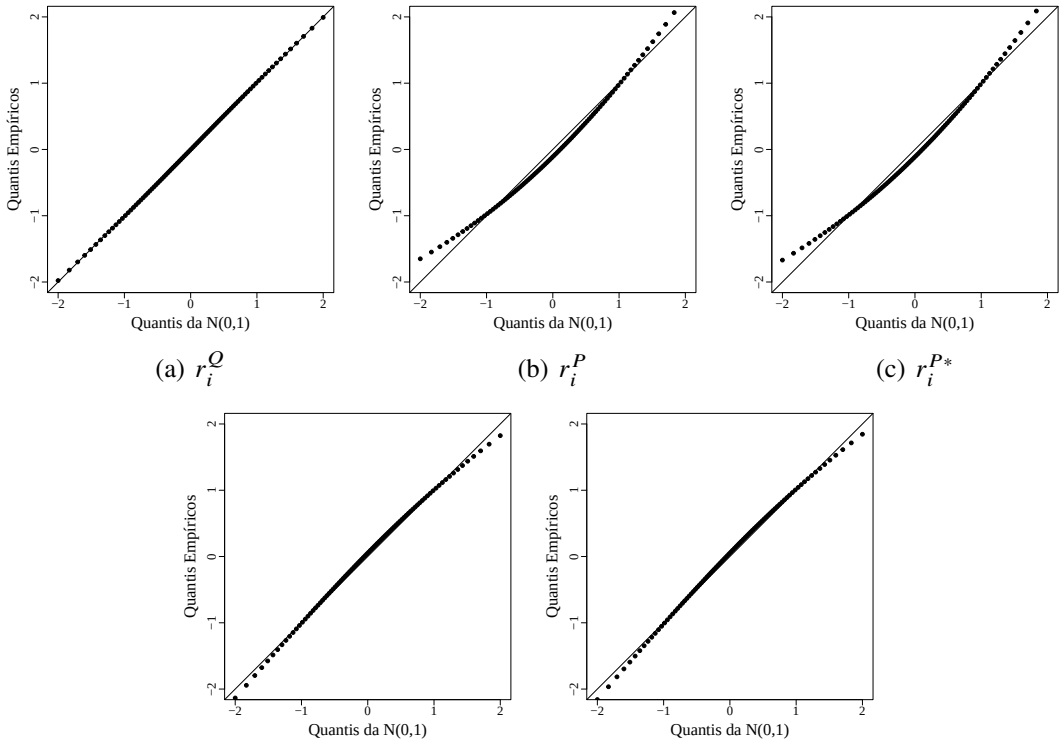
Figura 3.1: Gráficos normal de probabilidades dos resíduos r_i^Q, r_i^P, r_i^{P*}, r_i^β e $r_i^{\beta*}$ obtidos sob especificação correta do modelo de regressão BP - Cenário I com $n = 90$.



(a) r_i^Q (b) r_i^P (c) r_i^{P*}

(d) r_i^β (e) $r_i^{\beta*}$

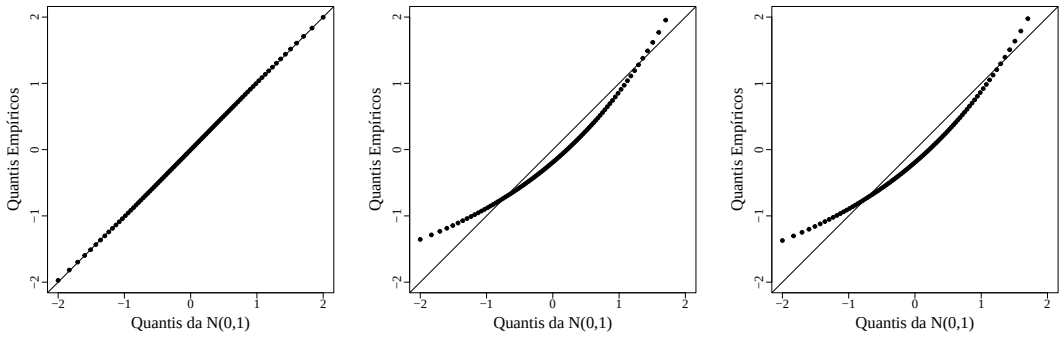
Figura 3.2: Gráficos normal de probabilidades dos resíduos r_i^Q, r_i^P, r_i^{P*}, r_i^β e $r_i^{\beta*}$ obtidos sob especificação correta do modelo de regressão BP - Cenário II com $n = 90$.



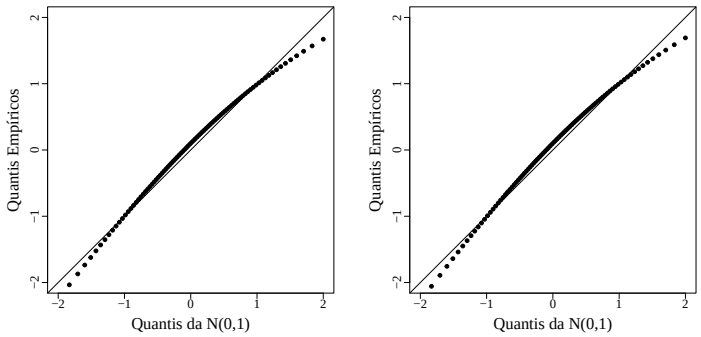
(a) r_i^Q (b) r_i^P (c) r_i^{P*}

(d) r_i^β (e) $r_i^{\beta*}$

Figura 3.3: Gráficos normal de probabilidades dos resíduos r_i^Q, r_i^P, r_i^{P*}, r_i^β e $r_i^{\beta*}$ obtidos sob especificação correta do modelo de regressão BP - Cenário III com $n = 90$.

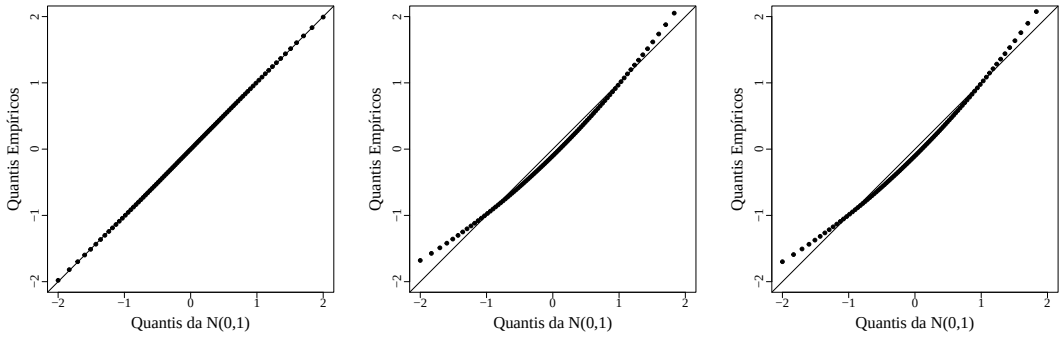


(a) r_i^Q (b) r_i^P (c) r_i^{P*}

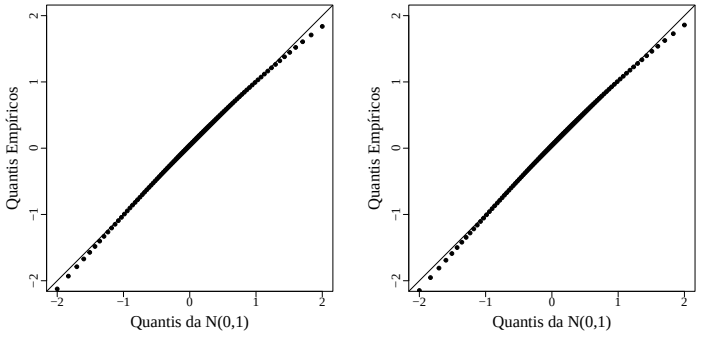


(d) r_i^β (e) $r_i^{\beta*}$

Figura 3.4: Gráficos normal de probabilidades dos resíduos r_i^Q, r_i^P, r_i^{P*}, r_i^β e $r_i^{\beta*}$ obtidos sob especificação correta do modelo de regressão BP - Cenário IV com $n = 90$.



(a) r_i^Q (b) r_i^P (c) r_i^{P*}



(d) r_i^β (e) $r_i^{\beta*}$

Em resumo, os resultados das simulações indicam uma boa concordância da distribuição empírica do resíduo r_i^Q pela distribuição normal padrão quando o modelo de regressão BP é corretamente especificado. Os resíduos r_i^P e r_i^{P*} apresentam comportamento bastante similar, ambos se distribuem com assimetria à direita em conformidade ao estudo apresentado por Bourguignon et al. (2021), além disso, dependendo dos valores do parâmetro de precisão essa assimetria é mais acentuada. Os resíduos r_i^β e $r_i^{\beta*}$ também são bem similares, ambos com distribuição tipicamente bem aproximadas pela distribuição normal padrão quando consideramos valores maiores para a precisão. Contudo, como a aderência da distribuição dos resíduos r_i^β e $r_i^{\beta*}$ pela distribuição normal padrão depende dos valores do parâmetro de precisão, não é conveniente utilizar os limites usuais para detectar pontos atípicos nos gráficos desses resíduos.

Uma proposta para detectar falta de qualidade no ajuste de um modelo de regressão quando os resíduos possuem distribuição desconhecida é utilizar envelopes simulados nos gráficos normal de probabilidades (ATKINSON, 1981). Espera-se, em modelos com bom ajuste, pontos (resíduos) dispersos aleatoriamente entre as bandas de confiança dos gráficos normal de probabilidades com envelope simulado. Quando o modelo é incorretamente ajustado, uma quantidade considerável de pontos deve se apresentar fora dos limites. Os gráficos de envelope simulado para o modelo de regressão BP podem ser obtidos da seguinte forma:

1. Ajustar o modelo de regressão BP aos dados e obter as estimativas $\widehat{\mu_i}$ e $\widehat{\phi_i}$, $i = 1, \ldots, n$;
2. Gerar uma amostra de n observações de $y_i \sim \text{BP}(\widehat{\mu_i}, \widehat{\phi_i})$, $i = 1, \ldots, n$;
3. Ajustar o modelo de regressão BP aos dados simulados e obter ordenadamente os resíduos de interesse;
4. Repetir 100 vezes os passos 2 e 3, obtendo assim n conjuntos de 100 resíduos ordenados;
5. Para cada $i = 1, \ldots, n$, calcular o percentil 2,5%, a média e o percentil 97,5% dos resíduos ordenados simulados para formar as bandas com 95% de confiança do envelope simulado;
6. Construir o gráfico normal de probabilidade das quantidades obtidas no passo 5 com o resíduo de interesse do ajuste original do modelo.

Nos estudos realizados na próxima seção, utilizamos os gráficos de envelopes simulados para verificar a adequação dos modelos de regressão BP. Além disso, como observamos nessas simulações que os resíduos r_i^P e r_i^{P*} apresentam comportamento bastante similar, assim como os resíduos r_i^β e $r_i^{\beta*}$, optamos por apresentar os resultados dos próximos estudos referentes apenas aos resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$.

3.2.2 Avaliação dos resíduos sob especificação incorreta

Nesta seção, apresentamos os resultados referentes aos ajustes de alguns modelos de regressão BP considerando dados simulados, com o intuito de avaliar o desempenho dos resíduos

r_i^Q, r_i^{P*} e $r_i^{\beta*}$ para detectar erros de especificação, assim como analisar o comportamento da distribuição desses resíduos quando o modelo é incorretamente especificado. Os cenários aqui apresentados são baseados em Espinheira et al. (2017) e Feng et al. (2020).

3.2.2.1 Erro de especificação de precisão constante

O primeiro exemplo simulado tem por objetivo analisar os desempenhos dos resíduos para detectar a necessidade da modelagem da precisão quando os dados apresentam precisão variável e são ajustados pelo modelo de regressão BP com precisão constante. Para isso, geramos os dados considerando a seguinte estrutura do modelo de regressão BP

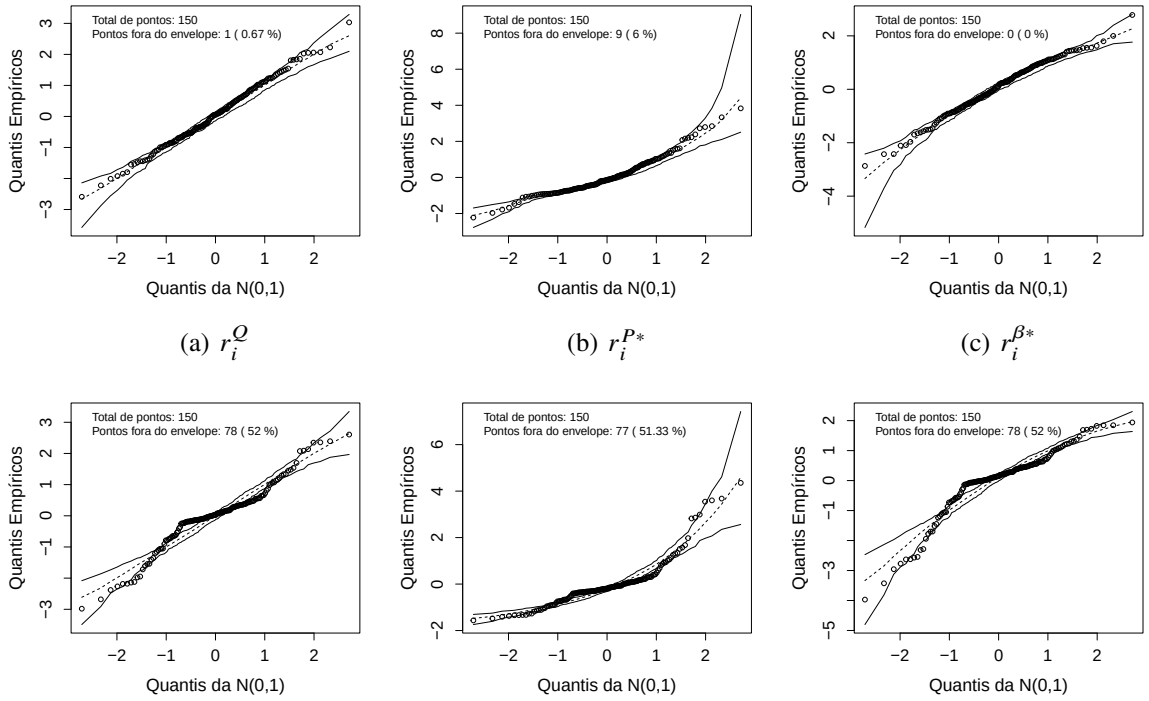
$$\log(\mu_i) = 3 + x_{1i} + 1{,}2x_{2i} \qquad \text{e} \qquad \log(\phi_i) = 1{,}2 + 3{,}2z_i, \tag{3.9}$$

com $i = 1, \ldots, 150$. Os valores das covariáveis x_{1i} e x_{2i} foram obtidos de forma independente a partir da distribuição $\mathcal{U}(0, 1)$ e z_i é uma variável indicadora (*dummy*), em que $z_i = 0$ para $i = 1, \ldots, 75$ e $z_i = 1$ para $i = 76, \ldots, 150$, garantindo, assim, a existência de dois grupos com precisão diferente no conjunto de dados. Ajustamos o modelo sob correta especificação, isto é, considerando o modelo definido em (3.9), e ajustamos também desconsiderando a modelagem da precisão, ou seja, modelando apenas a média da variável resposta.

A Figura 3.5 apresenta os gráficos de envelopes simulados dos resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$ obtidos sob especificação correta e incorreta. Estes indicam que quando o modelo é corretamente especificado (Figura 3.5 (a)-(c)), ambos os resíduos se encontram dispersos entre as bandas de confiança do envelope, como esperado. Contudo, observa-se que, mesmo sob o modelo corretamente especificado, os resíduos r_i^{P*} e $r_i^{\beta*}$ apresentam distribuição assimétrica à direita e distribuição assimétrica à esquerda, respectivamente. Com relação aos gráficos de envelopes sob o modelo incorretamente especificado (Figura 3.5 (d)-(f)), observa-se que os gráficos com base nos três resíduos sugerem uma falta de qualidade de ajuste do modelo quando cometermos o erro de ignorar a precisão variável dos dados. Ambos os gráficos apresentam quantidade considerável de pontos fora dos limites da banda de confiança dos envelopes.

Para verificar se os resíduos conseguem detectar o tipo de erro cometido ao ajustarmos o modelo sob especificação incorreta, construímos gráficos de dispersão dos resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$ contra os índices das observações. Esses são apresentados na Figura 3.6, que apresenta também os gráficos dos resíduos obtidos do modelo corretamente especificado. As Figuras 3.6 (a)-(c) indicam que, sob especificação correta do modelo, os três resíduos apresentam pontos aleatoriamente distribuídos em torno de zero, sem padrão detectável. Quando cometemos o erro de especificação, os gráficos dos resíduos contra os índices das observações (Figuras 3.6 (d)-(f)) evidenciam a existência de dois grupos com precisão diferente no conjunto de dados. Observa-se que, para as 75 primeiras observações, a amplitude dos resíduos é maior que para as observações seguintes, revelando uma característica importante do conjunto de dados. Nota-se que os três resíduos apresentam desempenho satisfatório para detectar o tipo de erro cometido.

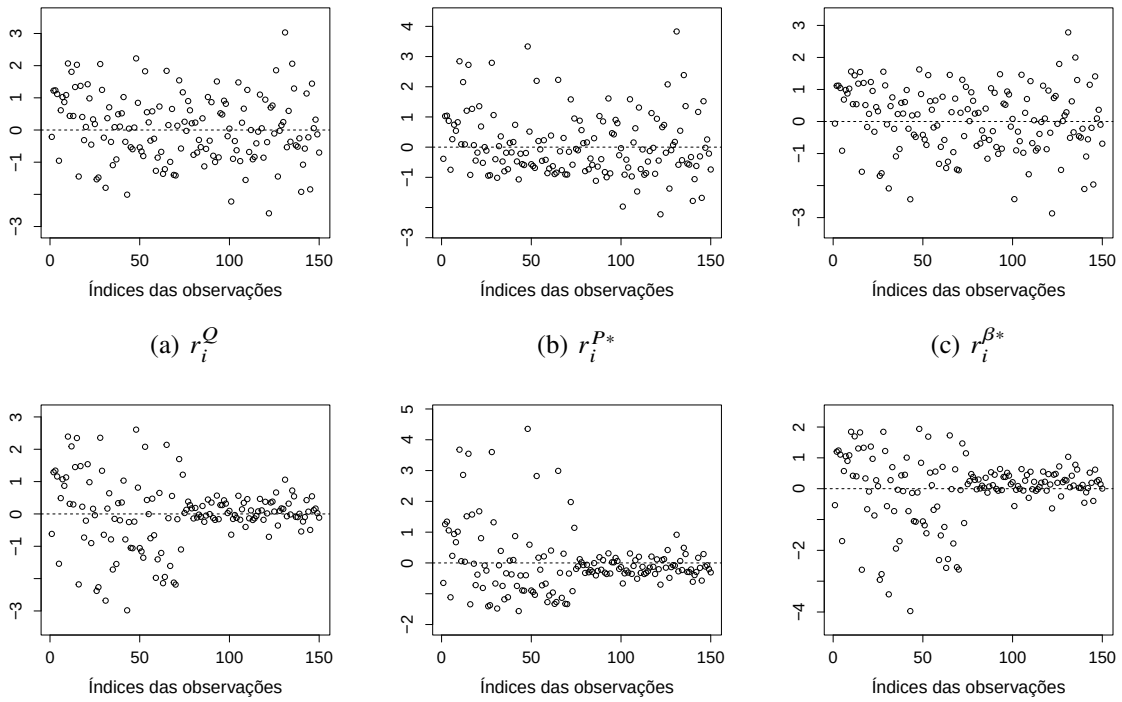
Figura 3.5: Gráficos de envelopes simulados dos resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$ obtidos sob especificação correta (primeira linha) e incorreta (segunda linha) do modelo de regressão BP - Erro de especificação de precisão constante.



(a) r_i^Q (b) r_i^{P*} (c) $r_i^{\beta*}$

(d) r_i^Q (e) r_i^{P*} (f) $r_i^{\beta*}$

Figura 3.6: Gráficos de dispersão dos resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$ obtidos sob especificação correta (primeira linha) e incorreta (segunda linha) do modelo de regressão BP contra índice das observações - Erro de especificação de precisão constante.



(a) r_i^Q (b) r_i^{P*} (c) $r_i^{\beta*}$

(d) r_i^Q (e) r_i^{P*} (f) $r_i^{\beta*}$

Para garantir a confiabilidade dos resultados referentes à distribuição empírica dos resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$, replicamos esse experimento simulando 10000 conjuntos de dados do modelo verdadeiro (3.9). Para cada réplica, ajustamos o modelo sob correta e incorreta especificação e aplicamos o teste de normalidade de SW para cada tipo de resíduo obtido em cada modelo. A Tabela 3.5 apresenta as proporções de vezes que, sob correta e incorreta especificação do modelo e ao nível de 5% de significância, a hipótese nula de que os resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$ provêm de uma população normalmente distribuída não foi rejeitada.

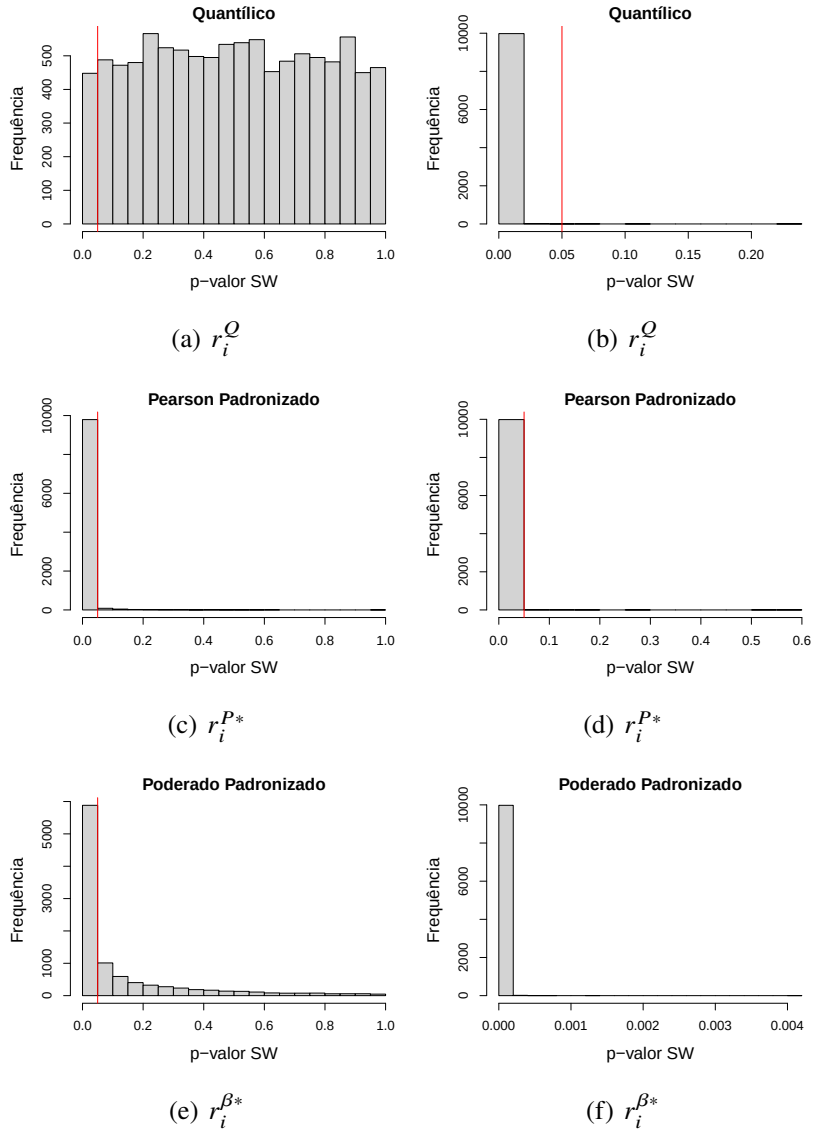
Tabela 3.5: Proporções de não rejeição da hipótese nula dos testes de SW para os resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$ obtidos sob especificação correta e incorreta do modelo de regressão BP - Erro de especificação de precisão constante.

Especificação do modelo	SW (5%)		
	r_i^Q	r_i^{P*}	$r_i^{\beta*}$
Correta	95,52	2,08	41,17
Incorreta	0,08	0,15	0,00

Como esperado, a hipótese de que o resíduo r_i^Q provém de uma população com distribuição normal padrão não é rejeitada em torno de 95% das vezes quando o modelo está corretamente especificado. Quando o modelo é incorretamente especificado, observa-se a não aderência da distribuição do resíduo r_i^Q pela distribuição normal padrão, indicando que a proximidade da distribuição empírica do resíduo r_i^Q pela distribuição normal padrão pode ser usada como um referencial para indicar a adequação ou a falta de qualidade de ajuste do modelo de regressão BP. Por outro lado, as proporções referentes aos testes de SW para os resíduos r_i^{P*} e $r_i^{\beta*}$ revelam que estes, mesmo sob o modelo corretamente especificado, não possuem distribuição empírica bem aproximada pela distribuição normal padrão.

Conclusões semelhantes podem ser feitas observando os histogramas dos p-valores dos testes de SW para os resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$ obtidos sob especificação correta e incorreta do modelo. Nota-se, apenas na Figura 3.7(a), que os p-valores são aproximadamente uniformemente distribuídos, comportamento esperado sob a hipótese nula, cuja distribuição da estatística do teste é contínua (ABRAMOVICH; RITOV, 2013, pág. 69).

Figura 3.7: Histogramas dos p-valores dos testes de SW para os resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$ obtidos sob especificação correta (primeira coluna) e incorreta (segunda coluna) do modelo de regressão BP - Erro de especificação de precisão constante.



(a) r_i^Q (b) r_i^Q

(c) r_i^{P*} (d) r_i^{P*}

(e) $r_i^{\beta*}$ (f) $r_i^{\beta*}$

3.2.2.2 Erro de especificação na covariável do submodelo da média

O segundo exemplo simulado tem por objetivo analisar o desempenho dos resíduos para detectar erro de especificação cometido na covariável do submodelo da média. Simulamos as covariáveis $x_i \sim \mathcal{U}(-1, 1)$ e $z_i \sim \mathcal{U}(0, 1)$, $i = 1, \ldots, 200$, e geramos os dados considerando um termo quadrático na covariável x_i. O modelo de regressão BP é dado por

$$\log(\mu_i) = 2{,}9 - 2{,}5x_i^2 \qquad \text{e} \qquad \log(\phi_i) = 1{,}5 + 1{,}8z_i. \tag{3.10}$$

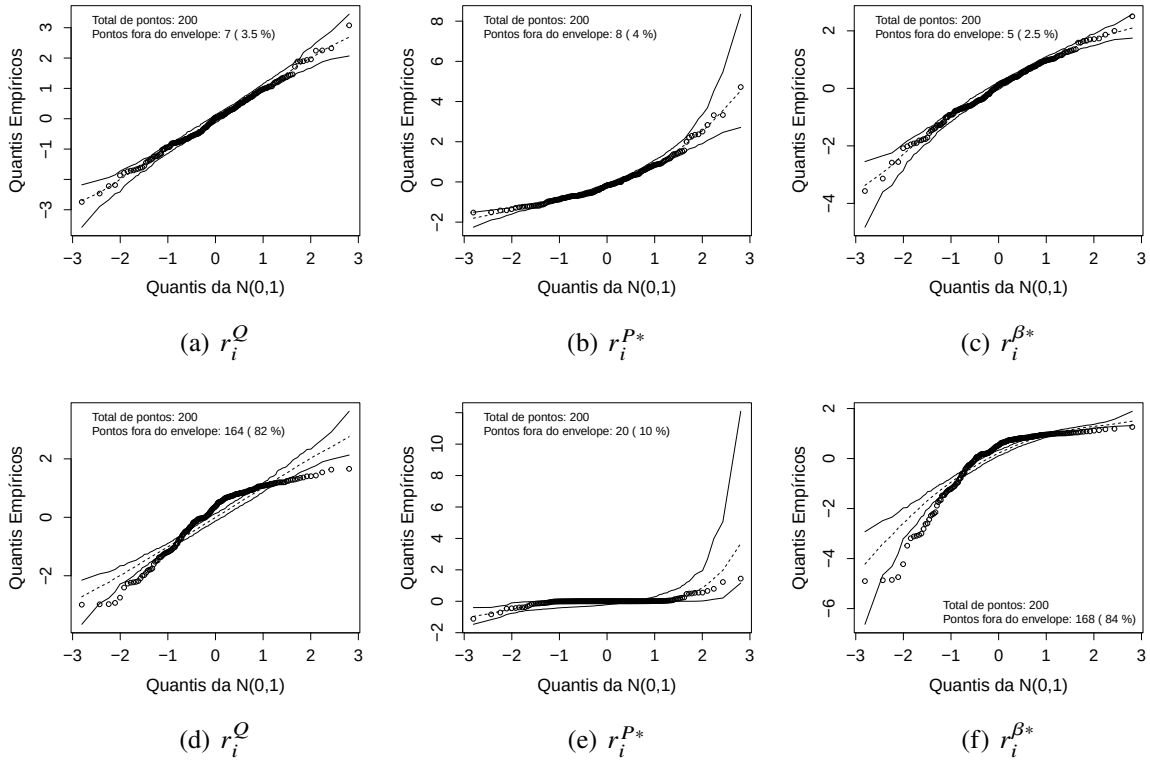
Ajustamos o modelo (3.10) sob correta especificação e também sob incorreta especificação,

desconsiderando o termo quadrático da covariável do submodelo da média, isto é

$$\log(\mu_i) = 2{,}9 - 2{,}5x_i \qquad \text{e} \qquad \log(\phi_i) = 1{,}5 + 1{,}8z_i.$$

A Figura 3.8 apresenta os gráficos de envelopes simulados dos resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$. Quando o modelo é corretamente especificado (Figuras 3.8 (a)-(c)) ambos os resíduos se encontram dispersos entre as bandas de confiança dos seus respectivos envelopes. Novamente observa-se assimetria nas distribuições dos resíduos r_i^{P*} e $r_i^{\beta*}$. Com relação aos gráficos de envelopes sob o modelo incorretamente especificado (Figura 3.8 (d)-(f)), observa-se que os gráficos dos resíduos r_i^Q e $r_i^{\beta*}$ apresentam uma grande quantidade de pontos fora dos limites da banda de confiança dos envelopes, indicando uma falta de ajuste do modelo sob especificação incorreta. Entretanto, o gráfico do resíduo r_i^{P*} apresenta apenas 10% dos pontos fora dos envelopes, quantidade bem inferior quando comparada aos gráficos dos resíduos r_i^Q e $r_i^{\beta*}$, que apresentam cerca de 80% dos pontos fora dos envelopes simulados.

Figura 3.8: Gráficos de envelopes simulados dos resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$ obtidos sob especificação correta (primeira linha) e incorreta (segunda linha) do modelo de regressão BP - Erro de especificação na covariável do submodelo da média.



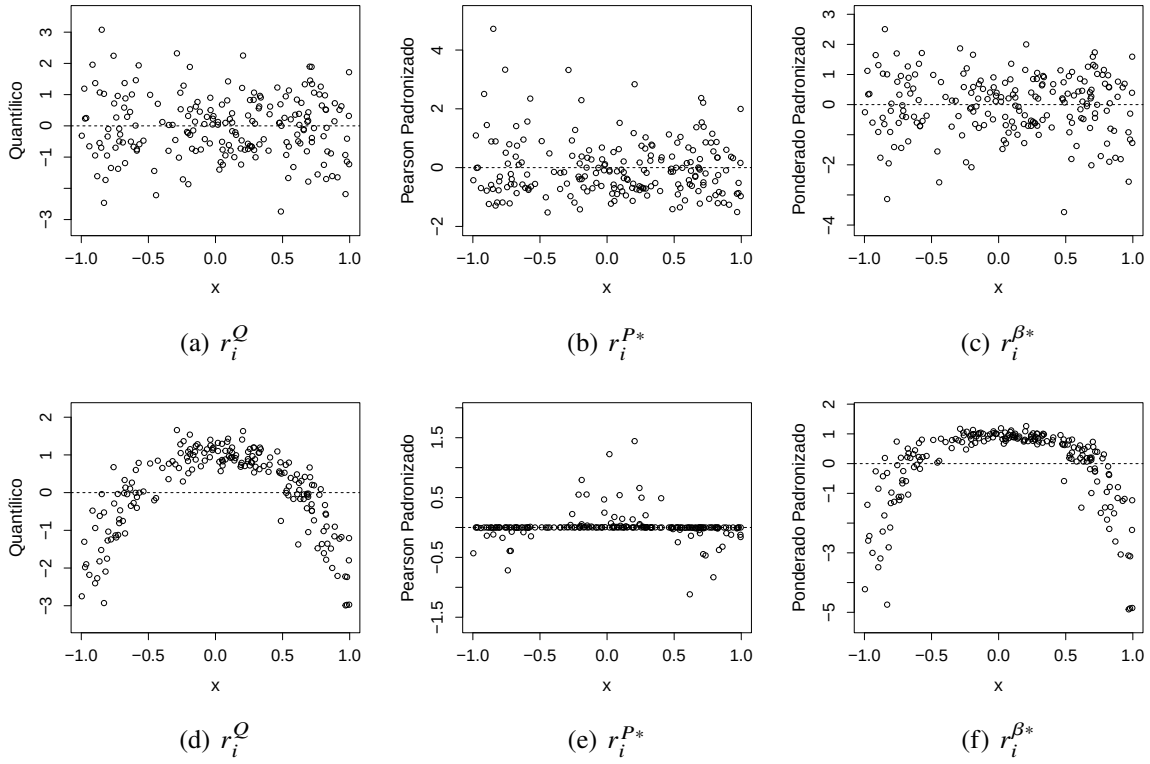
(a) r_i^Q (b) r_i^{P*} (c) $r_i^{\beta*}$

(d) r_i^Q (e) r_i^{P*} (f) $r_i^{\beta*}$

Também construímos gráficos de dispersão dos resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$ contra a covariável x_i para o modelo sob correta e incorreta especificação. Sob especificação correta (Figura 3.9 (a)-(c)), observa-se que os três resíduos apresentam pontos aleatoriamente distribuídos em torno de zero, sem padrão detectável. Contudo, os gráficos sob especificação incorreta (Figura 3.9 (d)-(f)) revelam que o resíduo r_i^{P*} falha ao detectar o tipo de erro cometido no ajuste. Apenas

os resíduos r_i^Q e $r_i^{\beta*}$ indicam a necessidade de considerar um termo quadrático na covariável x_i.

Figura 3.9: Gráficos de dispersão dos resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$ obtidos sob especificação correta (primeira linha) e incorreta (segunda linha) do modelo de regressão BP contra x_i - Erro de especificação na covariável do submodelo da média.



(a) r_i^Q (b) r_i^{P*} (c) $r_i^{\beta*}$

(d) r_i^Q (e) r_i^{P*} (f) $r_i^{\beta*}$

Semelhante ao estudo feito no primeiro tipo de erro de especificação, replicamos 10000 vezes o conjunto de dados do modelo expresso em (3.10) e, para cada réplica, ajustamos o modelo sob especificação correta e incorreta. Calculamos os três tipos de resíduos e aplicamos o teste de normalidade de SW. A Tabela 3.6 apresenta as proporções de vezes que, sob especificação correta e incorreta do modelo e ao nível de 5% de significância, a hipótese nula de que os resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$ provêm de uma população normalmente distribuída não foi rejeitada.

Tabela 3.6: Proporções de não rejeição da hipótese nula dos testes de SW para os resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$ obtidos sob especificação correta e incorreta do modelo de regressão BP - Erro de especificação na covariável do submodelo da média.

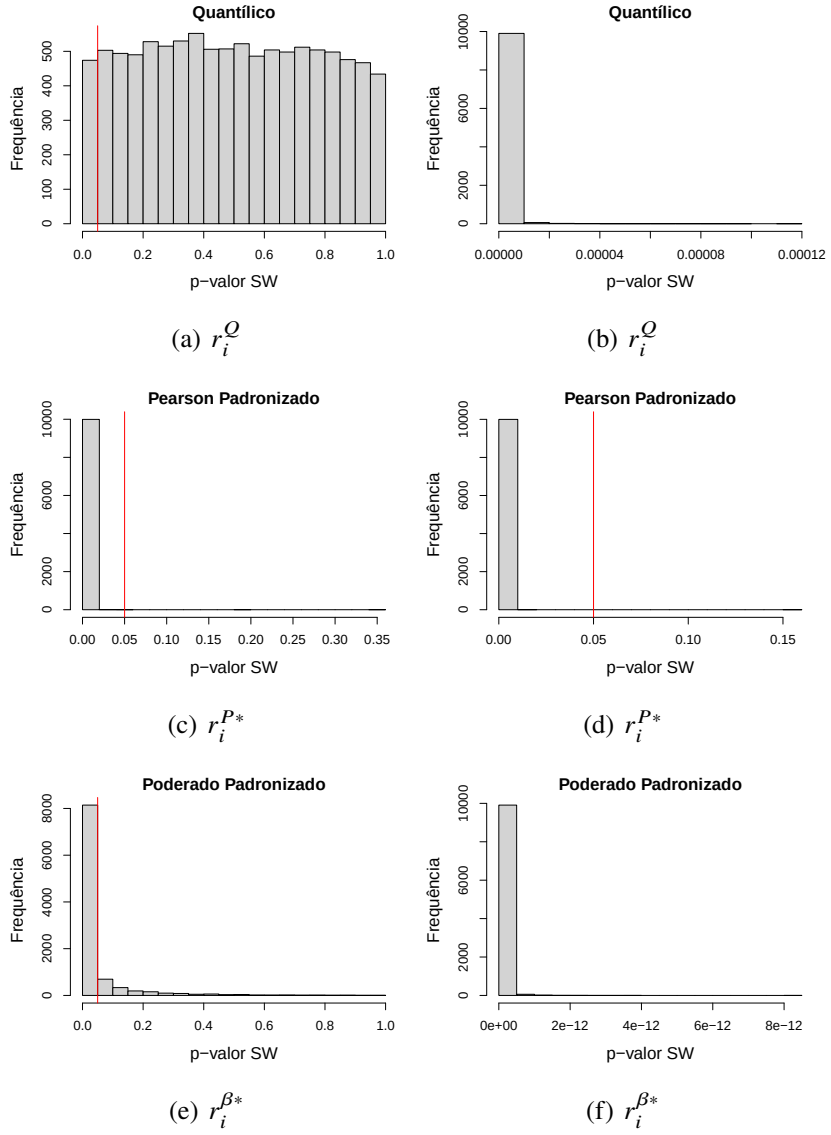
Especificação do modelo	SW (5%)		
	r_i^Q	r_i^{P*}	$r_i^{\beta*}$
Correta	95,34	0,02	18,77
Incorreta	0,00	0,00	0,00

Novamente, observa-se que a aderência da distribuição do resíduo r_i^Q pela distribuição normal padrão pode ser utilizada como um referencial para indicar a qualidade ou a falta de ajuste do modelo de regressão BP. Diferentemente dos resíduos r_i^{P*} e $r_i^{\beta*}$, uma vez que as proporções referentes aos testes de SW para esses resíduos indicam que, mesmo sob o modelo corretamente

especificado, as distribuições desses resíduos podem não ser bem aproximadas pela distribuição normal padrão.

Os histogramas dos p-valores dos testes de SW para os resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$, obtidos sob especificação correta e incorreta do modelo (Figura 3.10) ilustram as conclusões referentes à proximidade da distribuição desses resíduos pela distribuição normal padrão. Mais uma vez, observa-se que apenas o resíduo r_i^Q possui distribuição empírica bem aproximada pela distribuição normal padrão quando o modelo de regressão BP é corretamente especificado.

Figura 3.10: Histogramas dos p-valores dos testes de SW para os resíduos r_i^Q, r_i^{P*} e $r_i^{\beta*}$ obtidos sob especificação correta (primeira coluna) e incorreta (segunda coluna) do modelo de regressão BP - Erro de especificação na covariável do submodelo da média.



(a) r_i^Q (b) r_i^Q

(c) r_i^{P*} (d) r_i^{P*}

(e) $r_i^{\beta*}$ (f) $r_i^{\beta*}$

Em resumo, o resíduo de Pearson padronizado, além possuir distribuição assimétrica à direita, pode não detectar a falta de ajuste do modelo de regressão BP quando o mesmo está incorretamente especificado, como constatamos no exemplo do erro de especificação na covariável do submodelo da média. O resíduo ponderado padronizado apresenta bom desempenho

para detectar os erros de especificação cometidos, porém, este resíduo deverá ser utilizado com cautela, dado que a proximidade da sua distribuição empírica pela distribuição normal padrão depende da precisão dos dados que estão sendo ajustados pelo modelo de regressão BP. Já o resíduo quantílico, além de apresentar desempenho satisfatório para detectar erros de especificação, possui distribuição normal padrão quando o modelo de regressão BP está corretamente especificado, o que pode ser utilizado como um referencial para verificar a adequação do ajuste do modelo. Dessa forma, o resíduo quantílico é um resíduo com boas propriedades para ser utilizado na etapa de análises de diagnósticos de modelos de regressão BP.

Vale ressaltar que as análises de resíduos foram realizadas utilizando os cinco tipos de resíduos apresentados nos estudos de simulações da Seção (3.2.1). Contudo, os resíduos de Pearson e Pearson padronizado, assim como os resíduos ponderado e ponderado padronizado, apresentaram comportamento similar em ambos os gráficos e por isso os resultados dos resíduos de Pearson e ponderado foram aqui omitidos.

3.3 Aplicações

Nesta seção, apresentamos uma aplicação a dados reais para ilustrar o uso do resíduo quantílico para verificar a adequação do modelo de regressão BP. A referida aplicação refere-se aos dados de aluguel de terras agrícolas (WEISBERG, 2005).

3.3.1 Dados de aluguel de terras agrícolas

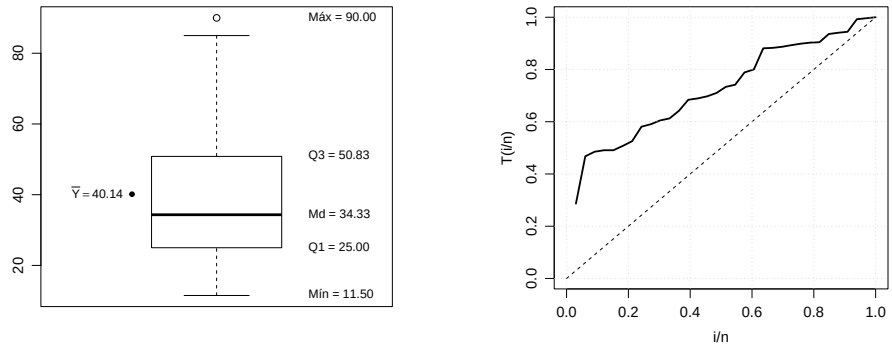
Os dados da aplicação são referentes ao aluguel médio pago (em dólares americanos) por acre plantado com alfafa em 67 condados do estado de Minnesota em 1977, disponíveis em Weisberg (2005). A alfafa é uma leguminosa rica em proteína adequada para a alimentação de vacas leiteiras. Supõe-se que a demanda por terras agrícolas plantadas com alfafa é maior em áreas com alta densidade de vacas leiteiras, o que pode aumentar os preços dos aluguéis. Contudo, os preços dos aluguéis também são afetados se o processo de calagem é necessário para o cultivo da alfafa, devido as despesas adicionais para o preparo do solo. Além disso, se proporção de terras agrícolas usadas para pastagem for baixa, pode haver uma escassez terras disponíveis para alimentação de gados, o que também pode aumentar os preços dos aluguéis das terras plantadas com alfafa.

Santos-Neto et al. (2016) analisaram esses dados visando entender como o aluguel pago por acre plantado com alfafa é afetado pela densidade de vacas leiteiras, quando a calagem é necessária para o cultivo da alfafa, e detectaram precisão variável na regressão Birnbaum-Saunders reparametrizada. Assim como em Santos-Neto et al. (2016), nosso interesse é avaliar o aluguel pago por terras agrícolas plantadas com alfafa que utilizam a calagem para o cultivo dessa cultura. Dessa forma, os dados utilizados referem-se aos 33 condados que fizeram uso de tal procedimento. A variável resposta é o aluguel médio pago por acre plantado com alfafa (y_i) e as variáveis explicativas são: o aluguel médio pago por terras para outros fins agrícolas

(x_{i1}), a densidade de vacas leiteiras em número por milha quadrada (x_{i2}) e a proporção de terras agrícolas usadas para pastagem (x_{i3}), $i = 1, \ldots, 33$.

A forma da função taxa de falha pode ser útil para a escolha da distribuição adequada para a variável em estudo. Uma maneira de verificar a forma da função taxa de falha é utilizando gráfico do tempo total de teste (TTTplot) (AARSET, 1987). A Figura 3.11 apresenta o Boxplot com algumas medidas descritivas e o TTTplot para o aluguel médio pago por acre plantado com alfafa. Utilizamos o pacote `AdequacyModel` (MARINHO et al., 2016) para a construção do gráfico TTTplot. Observa-se uma assimetria positiva na variável resposta (Figura 3.11 (a)). Além disso, o gráfico TTTplot (Figura 3.11 (b)) indica uma função taxa de falha crescente. Essas informações nos levam a supor que a distribuição BP pode ser adequada para esses dados.

Figura 3.11: Gráficos Boxplot e TTTplot para o aluguel médio pago por acre plantado com alfafa.



(a) Boxplot (b) TTTplot

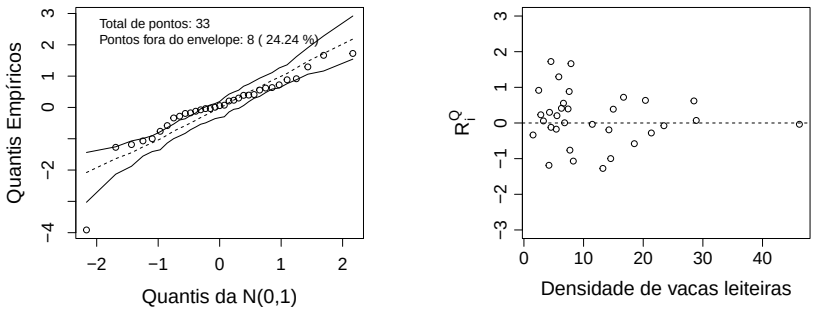
Inicialmente, consideramos o modelo de regressão BP com precisão constante e avaliamos diferentes estruturas para o preditor linear do submodelo da média. Utilizamos a análise residual, fazendo uso do resíduo quantílico, e as medidas AIC e BIC para selecionar o melhor modelo entre os candidatos. O modelo selecionado contém a seguinte estrutura

$$\log(\mu_i) = \beta_0 + \beta_1 \log(x_{i1}) + \beta_2 \log(x_{i2}) \quad \text{e} \quad \log(\phi_i) = \nu_0, \tag{3.11}$$

com $i = 1, \ldots, 33$. As medidas AIC e BIC obtidas para tal modelo foram 212,88 e 218,86, respectivamente.

A Figura 3.12 apresenta o gráfico de envelopes simulados e o gráfico de dispersão do resíduo r_i^Q contra a covariável densidade de vacas leiteiras (x_{i2}). Observa-se uma quantidade considerável de pontos fora dos limites da banda de confiança dos envelopes (Figura 3.12 (a)), sugerindo falta de adequação do modelo com precisão constante. Nota-se, na Figura 3.12 (b), uma dispersão maior dos resíduos para menores valores de x_{i2}. Tal padrão sugere evidências de precisão não constante nos dados e indica que tal covariável pode ser útil para explicar o comportamento do parâmetro de precisão.

Figura 3.12: Gráfico de envelopes simulados (a) e gráfico de dispersão do resíduo r_i^Q versus densidade de vacas leiteiras (b) para os dados de aluguel de terras agrícolas - Modelo de regressão BP com precisão constante.



(a) (b)

Ademais, o teste de SW para r_i^Q apresentou p-valor = 0,0012. Logo, ao nível de 5% de significância, a hipótese nula de que o resíduo quantílico provém de uma população normalmente distribuída foi rejeitada, indicando, assim, a falta de ajuste do modelo com precisão constante.

Diante evidências de precisão não constante, ajustamos o modelo de regressão BP com precisão variável. Com base na Figura 3.12 (b) e nos critérios AIC e BIC, utilizamos a seguinte estrutura

$$\log(\mu_i) = \beta_0 + \beta_1 \log(x_{i1}) + \beta_2 \log(x_{i2}) \quad \text{e} \quad \log(\phi_i) = \nu_0 + \nu_1 x_{i2}, \tag{3.12}$$

com $i = 1, \ldots, 33$. As medidas AIC e BIC obtidas para esse modelo, 204,10 e 211,59, respectivamente, foram inferiores às medidas obtidas para o modelo com precisão constante.

Na Tabela 3.7 são apresentadas as estimativas de máxima verossimilhança dos parâmetros, os erros-padrão e os p-valores associados aos testes de significância dos parâmetros do modelo definido em (3.12). Nota-se que todos os coeficientes são estatisticamente diferentes de zero ao nível de 5% de significância. As estimativas dos parâmetros mostram uma relação positiva entre a média da variável resposta e o aluguel médio pago por toda a terra agrícola, assim como com a densidade de vacas leiteiras, que também está relacionada positivamente com a precisão da variável resposta.

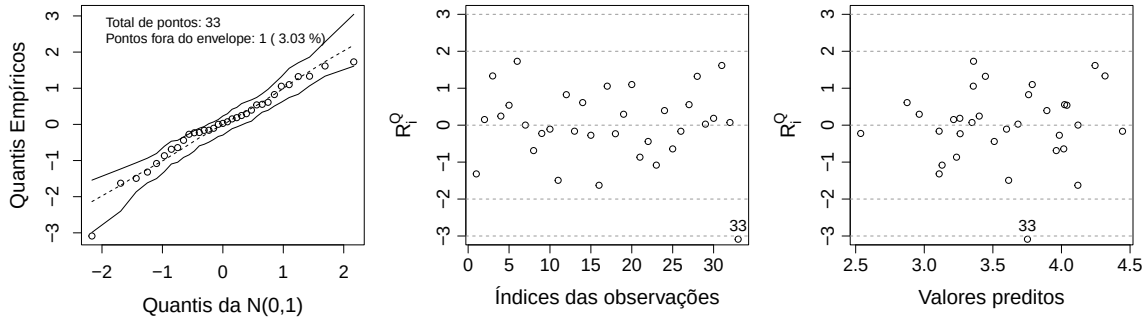
Tabela 3.7: Estimativas dos parâmetros, erros-padrão e p-valores do modelo de regressão BP com precisão variável aplicado aos dados de aluguel de terras agrícolas.

Parâmetro	β_0	β_1	β_2	ν_0	ν_1
Estimativa	−0,6467	1,0108	0,2116	2,5881	0,1302
Erro-padrão	0,1152	0,0319	0,0209	0,4876	0,0346
p-valor	< 0,0000	< 0,0000	< 0,0000	< 0,0000	0,0008

Para a avaliação do ajuste do modelo (3.12) realizamos a análise residual com base no gráfico de envelopes simulados e nos gráficos do resíduo r_i^Q contra o índice das observações e

contra os valores preditos apresentados, respectivamente, na Figura 3.13. O gráfico de envelope simulado sugere que não há indícios de que o modelo de regressão BP com precisão variável não é adequado aos dados. Os resíduos encontram-se aleatoriamente distribuídos em torno de zero nos gráficos de dispersão, apesar de apresentar uma observação fora dos limites. Seria indicada uma análise de influência, para verificar o impacto dessa observação nas estimativas dos parâmetros e nas conclusões inferenciais. Ademais, o teste de SW apresentou p-valor = 0,2645 e assim, ao nível de 5% de significância, a hipótese nula de que o resíduo quantílico provém de uma população normalmente distribuída não foi rejeitada.

Figura 3.13: Gráfico de envelopes simulados (a) e gráficos de dispersão do resíduo r_i^Q versus índice das observações (b) e valores preditos (c) para os dados do aluguel de terras agrícolas - Modelo de regressão BP com precisão variável.



(a) (b) (c)

O teste RESET baseado na estatística escore (SANTOS et al., 2021) apresentou p-valor = 0,8723. Logo, ao nível de 5% de significância, a hipótese nula de que o modelo (3.12) está corretamente especificado não foi rejeitada.

4 Critérios de seleção de modelos

Neste Capítulo, apresentamos algumas medidas de qualidade de ajuste do tipo pseudo-R^2 e propomos as estatísticas PRESS e P^2 como medidas do poder preditivo do modelo de regressão BP. Via estudos de simulação de Monte Carlo, avaliamos o comportamento e o desempenho das medidas do tipo P^2 e pseudo-R^2, em diferentes cenários do modelo de regressão BP sob especificação correta e incorreta. Por fim, duas aplicações a dados reais são apresentadas para ilustrar o desempenho das medidas propostas.

4.1 Medidas de qualidade de ajuste

No que tange à qualidade do ajuste de um modelo de regressão normal linear, o coeficiente de determinação R^2 é utilizado para indicar o quanto da variabilidade da variável resposta é explicada pelo modelo ajustado (NAGELKERKE, 1991). Dentre os modelos candidatos, aquele que maximizar o R^2 é o modelo com melhor qualidade de ajuste. No entanto, uma desvantagem do R^2 é que este, por ser não decrescente em relação à quantidade de parâmetros, favorece a adição de mais covariáveis ao modelo, ainda que essa covariável possua pouco poder explicativo sobre a variável resposta (ALCANTARA et al., 2022). Para garantir a parcimônia de um modelo de regressão, o R^2 ajustado foi proposto com o intuito de incluir um termo penalizador ao R^2 com base na adição de novas covariáveis ao modelo. Versões alternativas ao coeficiente de determinação foram propostas na literatura e são comumente conhecidas como pseudo-R^2.

Ferrari e Cribari-Neto (2004) propuseram o R^2_{FC} como uma medida global da variação explicada para o modelo de regressão beta. Tal medida é definida como o quadrado do coeficiente de correlação amostral entre $\widehat{\boldsymbol{\eta}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ e $g(\boldsymbol{y})$, isto é,

$$R^2_{FC} = \text{Corr}(\widehat{\boldsymbol{\eta}}, g(\boldsymbol{y}))^2,$$

em que $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ e $g(\cdot)$ é a função de ligação da média.

Uma versão corrigida do R^2_{FC} foi proposta por Bayer e Cribari-Neto (2017) como critério de seleção de modelos na regressão beta. Tal correção penaliza o R^2_{FC} com base na adição de novas covariáveis e no tamanho da amostra e é definida por

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

em que $r = p + q$ é a quantidade de parâmetros estimados pelo modelo, sendo p e q o número de parâmetros do submodelo da média e da precisão, respectivamente.

Outra medida de qualidade de ajuste, introduzida por Maddala (1983) e posteriormente generalizada por Nagelkerke (1991) para outros tipos de modelos de regressão, é o R^2_{RV}. Essa

medida é baseada na razão de verossimilhanças e é definida como

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

em que L_{null} e L_{fit} é a função de verossimilhança maximizada do modelo sem regressores e do modelo ajustado, respectivamente.

O critério $R^2_{RV_c}$, uma versão corrigida do R^2_{RV}, foi proposto por Bayer e Cribari-Neto (2017) para o modelo de regressão beta com dispersão variável. A definição do $R^2_{RV_c}$ é dada por

$$R^2_{RV_c} = 1 - \left(1 - R^2_{RV}\right)\left(\frac{n-1}{n-(1+\alpha)p-(1-\alpha)q}\right)^{\delta},$$

em que $\alpha \in [0, 1]$ e $\delta > 0$. Via estudos de simulação de Monte Carlo, os autores sugerem utilizar $\alpha = 0{,}4$ e $\delta = 1$. A vantagem desse critério é que o mesmo leva em consideração, além da estrutura do submodelo da média, o submodelo da dispersão ou precisão, diferentemente dos critérios de seleção usuais que, em sua maioria, utilizam apenas a proximidade entre $\boldsymbol{y}$ e $\widehat{\boldsymbol{\mu}}$ para estabelecer uma medida de qualidade de ajuste (BAYER; CRIBARI-NETO, 2017).

4.2 Estatísticas PRESS e P^2

Os critérios de seleção usuais, assim como os pseudos R^2 e suas versões corrigidas, não fornecem informações sobre a qualidade preditiva de um modelo de regressão (MEDIAVILLA et al., 2008). Nessa perspectiva, a Soma dos Quadrados dos Erros de Previsão (PRESS), proposta por Allen (1974), pode ser utilizada como um critério para selecionar o modelo com maior qualidade preditiva. O cálculo da estatística PRESS é baseado na técnica de validação cruzada *leave-one-out* (STONE, 1974) para avaliar previsões estatísticas, cujo processo consiste em ajustar o modelo candidato n vezes, omitindo uma observação em cada vez e, em cada ajuste, predizer a resposta omitida utilizando o modelo ajustado (MYERS, 1990). Esse processo resulta em n erros de predição, cuja soma dos quadrados desses erros formam a estatística PRESS, uma medida que indica quão bem o modelo de regressão ajustado pode prever novas observações. Modelos com menores valores da PRESS são considerados bons modelos candidatos, dado que estes apresentam pequenos erros de predição (KUTNER et al., 1996).

Assumindo o modelo de regressão linear múltipla, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, em que $\mathbf{y}$ é um vetor $n \times 1$ de variáveis respostas observadas, $\mathbf{X}$ é uma matriz $n \times p$ de covariáveis, $\boldsymbol{\beta}$ é um vetor $p \times 1$ de parâmetros desconhecidos a serem estimados e $\boldsymbol{\varepsilon}$ é um vetor $n \times 1$ dos erros, em que n é o número de observações e p é o número de parâmetros. O erro de predição, e_i, é dado pela diferença entre a resposta observada, y_i, e o seu valor predito baseado no modelo ajustado sem a i-ésima observação, $\widehat{y}_{(-i)}$, ou seja, $e_i = y_i - \widehat{y}_{(-i)}$, em que $\widehat{y}_{(-i)} = \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{(-i)}$, sendo $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ e $\widehat{\boldsymbol{\beta}}_{(-i)}$ é o estimador de mínimos quadrados de $\boldsymbol{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ sem o uso da i-ésima observação, com $i = 1, \ldots, n$. Assim, a estatística PRESS para o modelo de

regressão linear múltipla (ALLEN, 1974) é dada por

$$\text{PRESS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \widehat{y}_{(-i)}\right)^2, \qquad i = 1, \ldots, n.$$

Uma forma alternativa de calcular a estatística PRESS sem precisar reajustar o modelo para cada observação omitida (o que pode ser computacionalmente caro a depender do tamanho da amostra) é utilizar a seguinte expressão equivalente

$$\text{PRESS} = \sum_{i=1}^{n} \left(\frac{r_i}{1 - h_{ii}^*}\right)^2, \qquad i = 1, \ldots, n,$$

em que $r_i = y_i - \widehat{y}_i$ é o resíduo ordinário obtido da regressão linear de $\mathbf{y}$ em $\mathbf{X}$ e h_{ii}^* é o i-ésimo elemento da diagonal principal da matriz de projeção $\mathbf{H}^* = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ (MYERS, 1990).

Uma medida semelhante à abordagem do R^2 foi proposta por Wold (1982) baseada na PRESS e denominado coeficiente de predição (P^2). Este relaciona a estatística PRESS à soma de quadrados totais sem a i-ésima observação. Sua definição para o modelo normal linear é dada por

$$P^2 = 1 - \frac{\text{PRESS}}{\text{SST}_{(-i)}}, \qquad i = 1, \ldots, n,$$

em que $\text{SST}_{(-i)} = \sum_{i=1}^{n} \left(y_i - \overline{y}_{(-i)}\right)^2$ é a soma de quadrados totais sem a i-ésima observação, sendo $\overline{y}_{(-i)}$ a média aritmética dos $n-1$ valores de y_i excluindo a i-ésima observação. Uma expressão equivalente ao $\text{SST}_{(-i)}$, que não requer o ajuste do modelo n vezes, é $\text{SST}_{(-i)} = (n/(n-1))^2\text{SST}$, em que $\text{SST} = \sum_{i=1}^{n}(y_i - \overline{y})^2$ é a soma de quadrados totais calculada com os dados completos. Para mais detalhes, recomendamos Mediavilla et al. (2008). Note que, a estatística PRESS e o $\text{SST}_{(-i)}$ são medidas positivas, logo, o quociente $\text{PRESS}/\text{SST}_{(-i)}$ é positivo. Dessa forma, temos que $P^2 \in (-\infty, 1]$ e quanto mais próxima de 1 for essa medida maior é o poder preditivo do modelo ajustado.

Para o modelo de regressão BP, foi mostrado, na Seção (3.1.3), que o estimador de máxima verossimilhança $\widehat{\boldsymbol{\beta}}$, definido em (3.3), pode ser visto como um estimador de mínimos quadrados ordinários de $\boldsymbol{\beta}$ através da regressão linear de $\check{\mathbf{y}} = \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\widehat{\mathbf{z}}$ em $\check{\mathbf{X}} = \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}$, em que $\boldsymbol{\Phi}$ e $\mathbf{W}$ são as matrizes diagonais cujos elementos estão em (2.11) e (2.15), respectivamente, e $\mathbf{z}$ é o vetor definido em (3.3). Temos que o erro de predição é dado por

$$\check{y}_i - \widehat{\check{y}}_{(-i)} = (1 + \widehat{\phi}_i)^{\frac{1}{2}}\widehat{w}_i^{\frac{1}{2}}\widehat{z}_i - \mathbf{x}_i^\top(1 + \widehat{\phi}_i)^{\frac{1}{2}}\widehat{w}_i^{\frac{1}{2}}\widehat{\boldsymbol{\beta}}_{(-i)}, \tag{4.1}$$

em que $\mathbf{x}_i^\top$ é a i-ésima linha da matriz $\mathbf{X}$. Utilizando o Lema 3.2 apresentado em Miller-Júnior

(1974), tem-se que

$$\widehat{\boldsymbol{\beta}}_{(-i)} = \widehat{\boldsymbol{\beta}} - \frac{\left(\mathbf{X}^\top \widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{x}_i(1+\widehat{\phi}_i)^{\frac{1}{2}}\widehat{w}_i^{\frac{1}{2}}r_i^{\beta}}{1-\widehat{h_{ii}}}, \tag{4.2}$$

em que r_i^{β} é o resíduo ponderado definido em (3.5) e $\widehat{h_{ii}}$ é o i-ésimo elemento da diagonal principal da matriz $\widehat{\mathbf{H}}$ dada em (3.6). Com base na quantidade (4.2) é possível escrever o erro de predição (4.1) na forma

$$\begin{aligned}
\check{y}_i - \widehat{\check{y}}_{(-i)} &= (1+\widehat{\phi}_i)^{\frac{1}{2}}\widehat{w}_i^{\frac{1}{2}}\widehat{z}_i - \mathbf{x}_i^\top(1+\widehat{\phi}_i)^{\frac{1}{2}}\widehat{w}_i^{\frac{1}{2}}\left[\widehat{\boldsymbol{\beta}} - \frac{\left(\mathbf{X}^\top \widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{x}_i(1+\widehat{\phi}_i)^{\frac{1}{2}}\widehat{w}_i^{\frac{1}{2}}r_i^{\beta}}{1-\widehat{h_{ii}}}\right] \\
&= (1+\widehat{\phi}_i)^{\frac{1}{2}}\widehat{w}_i^{\frac{1}{2}}\widehat{z}_i - \mathbf{x}_i^\top(1+\widehat{\phi}_i)^{\frac{1}{2}}\widehat{w}_i^{\frac{1}{2}}\widehat{\boldsymbol{\beta}} + \frac{\mathbf{x}_i^\top(1+\widehat{\phi}_i)^{\frac{1}{2}}\widehat{w}_i^{\frac{1}{2}}\left(\mathbf{X}^\top \widehat{\boldsymbol{\Phi}}\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{x}_i(1+\widehat{\phi}_i)^{\frac{1}{2}}\widehat{w}_i^{\frac{1}{2}}r_i^{\beta}}{1-\widehat{h_{ii}}} \\
&= r_i^{\beta} + \frac{\widehat{h_{ii}}r_i^{\beta}}{1-\widehat{h_{ii}}} = \frac{r_i^{\beta}}{1-\widehat{h_{ii}}}.
\end{aligned}$$

Consequentemente, as estatísticas PRESS e P^2 para o modelo de regressão BP são dadas, respectivamente, por

$$\text{PRESS}_{\beta} = \sum_{i=1}^{n}\left(\check{y}_i - \widehat{\check{y}}_{(-i)}\right)^2 = \sum_{i=1}^{n}\left(\frac{r_i^{\beta}}{1-\widehat{h_{ii}}}\right)^2, \tag{4.3}$$

$$P_{\beta}^2 = 1 - \frac{\text{PRESS}_{\beta}}{\text{SST}_{(-i)}}, \tag{4.4}$$

em que $\text{SST}_{(-i)} = \sum_{i=1}^{n}\left(\check{y}_i - \overline{\check{y}}_{(-i)}\right)^2$, sendo $\overline{\check{y}}_{(-i)}$ a média aritmética dos $n-1$ valores de $\check{\mathbf{y}} = \widehat{\boldsymbol{\Phi}}^{\frac{1}{2}}\widehat{\mathbf{W}}^{\frac{1}{2}}\widehat{\mathbf{z}}$, excluindo a i-ésima observação. Adicionalmente, $\text{SST}_{(-i)} = (n/(n-r))^2\text{SST}$, em que $r = p+q$ é o número de parâmetros estimados pelo modelo e $\text{SST} = \sum_{i=1}^{n}\left(\check{y}_i - \overline{\check{y}}\right)^2$.

Similar à abordagem de Bayer e Cribari-Neto (2017), Espinheira et al. (2019) propuseram uma versão corrigida da estatística P^2. Baseado nessa proposta, temos que a versão corrigida da estatística P_{β}^2 para o modelo de regressão BP é dada por

$$P_{\beta_c}^2 = 1 - \left(1 - P_{\beta}^2\right)\left(\frac{n-1}{n-r}\right), \tag{4.5}$$

em que P_{β}^2 é definido em (4.4). Assim, o coeficiente de predição corrigido pode ser utilizado como critério para selecionar modelos na perspectiva de predição. Dado que a fração $(n-1)/(n-r)$ sempre será positiva e a expressão $(1-P_{\beta}^2)$ pode ser positiva ou negativa, a depender

do valor de P^2_β, temos que $P^2_{\beta_c} \in (-\infty, 1]$ e quanto mais próxima de 1 for essa medida maior é o poder preditivo do modelo ajustado.

Baseado em Cook e Weisberg (1982), Espinheira et al. (2019) consideraram outros tipos de resíduos em versões alternativas das estatísticas de predição para os modelos de regressão beta linear e não linear. No Capítulo 3, vimos que os resíduos de Pearson e Pearson padronizado apresentam desempenhos insatisfatórios em modelos de regressão BP. Diante disso, além das medidas PRESS_β, P^2_β e $P^2_{\beta_c}$ definidas, respectivamente, em (4.3), (4.4) e (4.5), consideramos as versões dessas estatísticas baseadas nos resíduos quantílico e ponderado padronizado. As definições de tais versões são análogas às apresentadas em (4.3), (4.4) e (4.5), porém, utilizando em seus cálculos os resíduos r^Q_i e $r^{\beta*}_i$ definidos em (3.1) e (3.7), respectivamente. Denotamos por P^2_Q, $P^2_{Q_c}$, P^2_β, $P^2_{\beta_c}$, $P^2_{\beta*}$ e $P^2_{\beta*_c}$ as estatísticas de predição e suas versões corrigidas com relação aos resíduos quantílico, ponderado e ponderado padronizado, nesta ordem.

4.3 Avaliação numérica

Nesta seção, apresentamos os resultados dos estudos de simulação de Monte Carlo visando avaliar o comportamento das estatísticas de predição P^2_Q, $P^2_{Q_c}$, P^2_β, $P^2_{\beta_c}$, $P^2_{\beta*}$ e $P^2_{\beta*_c}$ e das medidas de qualidade de ajuste R^2_{FC}, $R^2_{FC_c}$, R^2_{RV} e $R^2_{RV_c}$ no modelo de regressão BP. Verificamos, em diferentes cenários, o comportamento dessas medidas sob correta e incorreta especificação do modelo. Todas as simulações de Monte Carlo foram realizadas com 10000 réplicas através do ambiente computacional `R`.

4.3.1 Avaliação das medidas de predição e de qualidade de ajuste sob especificação correta

Para o estudo do comportamento das estatísticas de predição e de qualidade de ajuste sob especificação correta do modelo de regressão BP com precisão variável, consideramos diferentes cenários. A estrutura do modelo utilizada em todos os cenários é dada por

$$g(\mu_i) = \beta_1 + \beta_2 x_i \qquad \text{e} \qquad h(\phi_i) = \nu_1 + \nu_2 z_i,$$

com $i = 1, \ldots, n$. Os valores das covariáveis x_i e z_i foram obtidos de forma independente através da distribuição uniforme padrão $\mathcal{U}(0, 1)$ e mantidos fixos para cada réplica. Consideramos três tamanhos de amostras, a saber, $n = 30, 60$ e 120.

O objetivo deste estudo é verificar quais valores, em média, as medidas de predição e de qualidade de ajuste podem assumir em diferentes cenários do modelo de regressão BP quando este é corretamente especificado. Além disso, buscamos utilizar tais medidas para investigar a influência do tipo da função de ligação utilizada na modelagem da média e do parâmetro de precisão no que se refere a qualidade do modelo estimado. Para isso, realizamos as simulações utilizando as funções de ligação logarítmica e raiz quadrada nos submodelos da média e da

precisão. Foram considerados dois cenários para a média da variável resposta: $\mu \in (5{,}6; 25{,}0)$ e $\mu \in (25{,}0; 150{,}0)$. Para cada cenário da média consideramos dois cenários diferentes para a precisão: $\phi \in (5{,}4; 26{,}0)$ e $\phi \in (25{,}4; 81{,}0)$. A Tabela 4.1 apresenta os verdadeiros valores dos parâmetros utilizados na geração dos dados para obter os quatro cenários em cada tipo de função de ligação considerada no submodelo da média e da precisão.

Tabela 4.1: Cenários considerados para a avaliação das estatísticas de predição e de qualidade de ajuste sob especificação correta do modelo de regressão BP.

Função de ligação	Cenário	Parâmetros				Intervalos	
		β_1	β_2	ν_1	ν_2	μ	ϕ
	I	5,00	−2,63	2,30	2,80	(5,6; 25,0)	(5,4; 26,0)
$g(\mu_i) = \sqrt{\mu_i}$	II	5,00	−2,63	5,00	4,00	(5,6; 25,0)	(25,4; 81,0)
$h(\phi_i) = \sqrt{\phi_i}$	III	5,00	7,24	2,30	2,80	(25,0; 150,0)	(5,4; 26,0)
	IV	5,00	7,24	5,00	4,00	(25,0; 150,0)	(25,4; 81,0)
	I	5,00	−2,63	1,71	1,55	(5,6; 25,0)	(5,4; 26,0)
$g(\mu_i) = \sqrt{\mu_i}$	II	5,00	−2,63	3,23	1,16	(5,6; 25,0)	(25,4; 81,0)
$h(\phi_i) = \log(\phi_i)$	III	5,00	7,24	1,71	1,55	(25,0; 150,0)	(5,4; 26,0)
	IV	5,00	7,24	3,23	1,16	(25,0; 150,0)	(25,4; 81,0)
	I	3,22	−1,50	2,30	2,80	(5,6; 25,0)	(5,4; 26,0)
$g(\mu_i) = \log(\mu_i)$	II	3,22	−1,50	5,00	4,00	(5,6; 25,0)	(25,4; 81,0)
$h(\phi_i) = \sqrt{\phi_i}$	III	3,22	1,79	2,30	2,80	(25,0; 150,0)	(5,4; 26,0)
	IV	3,22	1,79	5,00	4,00	(25,0; 150,0)	(25,4; 81,0)
	I	3,22	−1,50	1,71	1,55	(5,6; 25,0)	(5,4; 26,0)
$g(\mu_i) = \log(\mu_i)$	II	3,22	−1,50	3,23	1,16	(5,6; 25,0)	(25,4; 81,0)
$h(\phi_i) = \log(\phi_i)$	III	3,22	1,79	1,71	1,55	(25,0; 150,0)	(5,4; 26,0)
	IV	3,22	1,79	3,23	1,16	(25,0; 150,0)	(25,4; 81,0)

As Tabelas 4.2 e 4.3 apresentam os valores médios das medidas obtidas para os quatro cenários, considerando 10000 réplicas de Monte Carlo. A Tabela 4.2 contém os resultados referentes à utilização da função de ligação raiz quadrada no submodelo da média, ou seja, $g(\mu_i) = \sqrt{\mu_i}$, e a Tabela 4.3 apresenta os resultados obtidos utilizando a função de ligação logarítmica, $g(\mu_i) = \log(\mu_i)$. Em ambas Tabelas, altera-se apenas a função de ligação do submodelo da precisão, $h(\phi_i) = \sqrt{\phi_i}$ e $h(\phi_i) = \log(\phi_i)$.

Observa-se nas Tabelas 4.2 e 4.3 que os melhores resultados das medidas avaliadas são obtidos nos cenários em que consideramos maiores valores para a precisão (cenários II e IV), indicando que a qualidade preditiva e a qualidade de ajuste do modelo de regressão BP são influenciadas positivamente por maiores valores de ϕ. Nota-se na Tabela 4.3, por exemplo, que no cenário I para $n = 30$ as estatísticas de predição e de qualidade de ajuste assumem valores médios em torno de 0,87 e 0,70, nessa ordem. Ao considerarmos valores maiores para o parâmetro de precisão (cenário II) essas medidas ficam em torno de 0,95 e 0,90, respectivamente.

Tabela 4.2: Valores médios das estatísticas de predição e de qualidade de ajuste obtidas sob especificação correta do modelo de regressão BP com precisão variável - Modelo gerado: $\sqrt{\mu_i} = \beta_1 + \beta_2 x_i$ e $h(\phi_i) = \nu_1 + \nu_2 z_i$, $i = 1, \ldots, n$.

Função de ligação		$g(\mu_i) = \sqrt{\mu_i}$ $h(\phi_i) = \sqrt{\phi_i}$				$g(\mu_i) = \sqrt{\mu_i}$ $h(\phi_i) = \log(\phi_i)$			
n	Critérios	Cenário				Cenário			
		I	II	III	IV	I	II	III	IV
30	P^2_Q	0,614	0,738	0,621	0,734	0,620	0,742	0,630	0,744
	$P^2_{Q_c}$	0,569	0,707	0,577	0,704	0,576	0,712	0,587	0,714
	P^2_β	0,619	0,739	0,627	0,736	0,626	0,743	0,636	0,745
	$P^2_{\beta_c}$	0,575	0,709	0,583	0,705	0,582	0,713	0,595	0,715
	$P^2_{\beta*}$	0,590	0,719	0,598	0,716	0,597	0,724	0,609	0,726
	$P^2_{\beta*_c}$	0,543	0,687	0,552	0,683	0,551	0,692	0,564	0,694
	R^2_{FC}	0,689	0,886	0,775	0,922	0,671	0,879	0,758	0,918
	$R^2_{FC_c}$	0,653	0,872	0,749	0,913	0,633	0,865	0,730	0,909
	R^2_{RV}	0,747	0,905	0,836	0,943	0,733	0,900	0,824	0,940
	$R^2_{RV_c}$	0,718	0,894	0,818	0,936	0,702	0,888	0,804	0,933
60	P^2_Q	0,615	0,727	0,626	0,730	0,615	0,734	0,631	0,735
	$P^2_{Q_c}$	0,595	0,712	0,606	0,716	0,594	0,720	0,611	0,721
	P^2_β	0,619	0,728	0,629	0,731	0,618	0,735	0,634	0,735
	$P^2_{\beta_c}$	0,598	0,713	0,609	0,716	0,598	0,721	0,615	0,721
	$P^2_{\beta*}$	0,605	0,718	0,616	0,721	0,605	0,726	0,621	0,726
	$P^2_{\beta*_c}$	0,584	0,703	0,596	0,706	0,583	0,711	0,601	0,711
	R^2_{FC}	0,650	0,866	0,742	0,909	0,632	0,861	0,726	0,904
	$R^2_{FC_c}$	0,631	0,859	0,728	0,904	0,612	0,853	0,711	0,899
	R^2_{RV}	0,713	0,887	0,810	0,931	0,697	0,882	0,797	0,927
	$R^2_{RV_c}$	0,698	0,881	0,800	0,927	0,681	0,876	0,787	0,923
120	P^2_Q	0,679	0,796	0,691	0,801	0,677	0,794	0,689	0,799
	$P^2_{Q_c}$	0,670	0,791	0,683	0,796	0,669	0,789	0,681	0,794
	P^2_β	0,680	0,796	0,692	0,801	0,679	0,795	0,690	0,800
	$P^2_{\beta_c}$	0,672	0,791	0,685	0,796	0,670	0,789	0,682	0,794
	$P^2_{\beta*}$	0,675	0,793	0,687	0,798	0,673	0,791	0,685	0,796
	$P^2_{\beta*_c}$	0,666	0,787	0,679	0,792	0,665	0,786	0,677	0,791
	R^2_{FC}	0,649	0,868	0,719	0,901	0,634	0,862	0,705	0,896
	$R^2_{FC_c}$	0,640	0,864	0,712	0,898	0,625	0,858	0,698	0,893
	R^2_{RV}	0,713	0,886	0,800	0,927	0,700	0,881	0,789	0,924
	$R^2_{RV_c}$	0,706	0,883	0,795	0,925	0,693	0,878	0,784	0,922

Com relação às versões das medidas de predição, nota-se que, sob o modelo corretamente especificado, as estatísticas P^2_Q e P^2_β, assim como suas versões corrigidas, apresentam valores bem próximos em todos os cenários considerados e para todos os tamanhos de amostras. As estatísticas de predição do tipo $P^2_{\beta*}$, por utilizarem resíduos padronizados, apresentam valores relativamente menores quando comparadas às outras duas versões das P^2. Os resultados referentes às medidas de qualidade de ajuste indicam que o critério R^2_{RV} e sua versão corrigida apresentam valores maiores que o critério R^2_{FC} e sua versão corrigida, comportamento também

observado no modelo de regressão beta linear (ESPINHEIRA et al., 2019, p. 12).

Tabela 4.3: Valores médios das estatísticas de predição e de qualidade de ajuste obtidas sob especificação correta do modelo de regressão BP com precisão variável - Modelo gerado: $\log(\mu_i) = \beta_1 + \beta_2 x_i$ e $h(\phi_i) = \nu_1 + \nu_2 z_i$, $i = 1, \ldots, n$.

Função de ligação		$g(\mu_i) = \log(\mu_i)$ $h(\phi_i) = \sqrt{\phi_i}$				$g(\mu_i) = \log(\mu_i)$ $h(\phi_i) = \log(\phi_i)$			
n	Critérios	Cenário				Cenário			
		I	II	III	IV	I	II	III	IV
30	P^2_Q	0,880	0,957	0,930	0,975	0,876	0,956	0,929	0,975
	$P^2_{Q_c}$	0,866	0,952	0,922	0,972	0,861	0,950	0,921	0,972
	P^2_β	0,881	0,957	0,931	0,975	0,877	0,956	0,930	0,975
	$P^2_{\beta_c}$	0,867	0,952	0,923	0,972	0,863	0,951	0,922	0,972
	$P^2_{\beta*}$	0,872	0,953	0,926	0,973	0,868	0,952	0,925	0,973
	$P^2_{\beta*_c}$	0,858	0,948	0,918	0,970	0,853	0,947	0,916	0,970
	R^2_{FC}	0,725	0,901	0,802	0,933	0,708	0,896	0,788	0,928
	$R^2_{FC_c}$	0,694	0,890	0,779	0,925	0,674	0,884	0,763	0,920
	R^2_{RV}	0,756	0,909	0,828	0,940	0,740	0,904	0,815	0,936
	$R^2_{RV_c}$	0,727	0,898	0,808	0,933	0,710	0,893	0,794	0,928
60	P^2_Q	0,854	0,943	0,925	0,970	0,852	0,942	0,924	0,969
	$P^2_{Q_c}$	0,846	0,940	0,921	0,969	0,844	0,939	0,920	0,968
	P^2_β	0,855	0,944	0,926	0,970	0,853	0,942	0,924	0,969
	$P^2_{\beta_c}$	0,848	0,941	0,922	0,969	0,845	0,939	0,920	0,968
	$P^2_{\beta*}$	0,850	0,942	0,923	0,969	0,848	0,940	0,921	0,968
	$P^2_{\beta*_c}$	0,842	0,938	0,919	0,968	0,840	0,937	0,917	0,967
	R^2_{FC}	0,689	0,884	0,773	0,921	0,672	0,878	0,760	0,917
	$R^2_{FC_c}$	0,673	0,877	0,761	0,916	0,655	0,872	0,747	0,912
	R^2_{RV}	0,720	0,891	0,803	0,928	0,705	0,886	0,791	0,925
	$R^2_{RV_c}$	0,705	0,886	0,793	0,924	0,689	0,880	0,780	0,920
120	P^2_Q	0,868	0,947	0,934	0,973	0,862	0,945	0,933	0,972
	$P^2_{Q_c}$	0,864	0,945	0,933	0,972	0,858	0,943	0,931	0,971
	P^2_β	0,868	0,947	0,935	0,973	0,862	0,945	0,933	0,972
	$P^2_{\beta_c}$	0,865	0,945	0,933	0,972	0,859	0,944	0,932	0,972
	$P^2_{\beta*}$	0,866	0,946	0,933	0,972	0,860	0,944	0,932	0,972
	$P^2_{\beta*_c}$	0,863	0,944	0,932	0,972	0,856	0,943	0,930	0,971
	R^2_{FC}	0,680	0,879	0,764	0,918	0,665	0,874	0,752	0,914
	$R^2_{FC_c}$	0,672	0,876	0,758	0,916	0,657	0,871	0,746	0,912
	R^2_{RV}	0,717	0,888	0,798	0,927	0,702	0,883	0,788	0,923
	$R^2_{RV_c}$	0,709	0,885	0,793	0,925	0,694	0,880	0,782	0,921

Comparando os resultados apresentados na Tabela 4.2 com os resultados da Tabela 4.3, é possível observar que a utilização da função de ligação logarítmica no submodelo da média aumenta consideravelmente a qualidade preditiva do modelo de regressão BP, dado que sua utilização resulta em maiores valores das estatísticas do tipo P^2. Por exemplo, quando utilizamos a função de ligação raiz quadrada nos submodelos da média e da precisão (ver a Tabela 4.2),

os valores médios das estatísticas de predição P^2_Q, $P^2_{Q_c}$, P^2_β, $P^2_{\beta_c}$, $P^2_{\beta*}$ e $P^2_{\beta*_c}$ no cenário I para $n = 30$ são, respectivamente, 0,614, 0,569, 0,619, 0,575, 0,590 e 0,543. Ao utilizar a função de ligação logarítmica no submodelo da média (ver a Tabela 4.3), tais estatísticas aumentam para 0,880, 0,866, 0,881, 0,867, 0,872 e 0,858, respectivamente.

Com relação às medidas de qualidade de ajuste, os resultados apresentados nas Tabelas 4.2 e 4.3 indicam um pequeno aumento nos valores dos critérios R^2_{FC} e $R^2_{FC_c}$ com o uso da função de ligação logarítmica no submodelo da média. Por exemplo, utilizando a função de ligação raiz quadrada no submodelo da média e da precisão (ver a Tabela 4.2), os valores médios dos critérios R^2_{FC} e $R^2_{FC_c}$ no cenário I para $n = 60$ são, respectivamente, 0,650 e 0,631. Ao utilizar a função de ligação logarítmica no submodelo da média (ver a Tabela 4.3), tais estatísticas aumentam para 0,689 e 0,673, respectivamente. Referente ao critério R^2_{RV} e sua versão corrigida, não observamos grande alteração quando comparamos seus valores médios obtidos utilizando ambas as funções de ligação no submodelo da média.

Dessa forma, observa-se que a utilização da função de ligação logarítmica no submodelo da média resulta em melhor desempenho do modelo de regressão BP, principalmente no que tange à qualidade preditiva do modelo. Ademais, os resultados apresentados indicam que, diferentemente do submodelo da média, a qualidade preditiva e a qualidade de ajuste do modelo de regressão BP não sofrem grandes alterações a depender do tipo de função de ligação utilizada no submodelo da precisão.

Também foi realizado um estudo considerando o modelo de regressão BP com precisão constante para investigar a influência das funções de ligação do submodelo da média no que diz respeito à qualidade do modelo estimado. Foram mantidos os mesmos cenários para a média da variável resposta: $\mu \in (5{,}6; 25{,}0)$ e $\mu \in (25{,}0; 150{,}0)$ e considerado três valores para o parâmetro de precisão constante: $\phi = 20, 50$ e 150. Os valores médios das estatísticas de predição e de qualidade de ajuste obtidas para o modelo de regressão BP com precisão constante, considerando 10000 réplicas de Monte Carlo, estão apresentados na Tabela 4.4.

As mesmas conclusões do estudo considerando o modelo de regressão BP com precisão variável podem ser feitas quando consideramos o modelo com precisão constante. Contudo, no modelo de regressão BP com precisão constante torna-se bem mais evidente a influência da função de ligação logarítmica na qualidade preditiva do modelo. Nota-se na Tabela 4.4 que, quando utilizamos a função de ligação raiz quadrada na modelagem da média, as estatísticas de predição assumem valores distantes de 1 em todos os cenários considerados, inclusive para $\phi = 150$. Observa-se que as medidas de predição aumentam consideravelmente quando utilizamos a função de ligação logarítmica. Além disso, à medida que ϕ aumenta, há melhora na qualidade preditiva e na qualidade de ajuste do modelo de regressão BP. De fato, dado que ϕ é um parâmetro de precisão, quanto maior o valor de ϕ, menor é a variabilidade dos dados.

Tabela 4.4: Valores médios das estatísticas de predição e de qualidade de ajuste obtidas sob especificação correta do modelo de regressão BP com precisão constante - Modelo gerado: $g(\mu_i) = \beta_1 + \beta_2 x_i$, $i = 1, \ldots, n$.

Função de ligação		$g(\mu_i) = \sqrt{\mu_i}$						$g(\mu_i) = \log(\mu_i)$					
n	$\mu \rightarrow$	$\mu \in (5{,}6; 25{,}0)$			$\mu \in (25{,}0; 150{,}0)$			$\mu \in (5{,}6; 25{,}0)$			$\mu \in (25{,}0; 150{,}0)$		
	$\phi \rightarrow$	20	50	150	20	50	150	20	50	150	20	50	150
30	P^2_Q	0,085	0,118	0,201	0,065	0,071	0,082	0,845	0,927	0,974	0,880	0,946	0,981
	$P^2_{Q_c}$	0,017	0,052	0,142	−0,004	0,003	0,014	0,834	0,922	0,972	0,872	0,942	0,979
	P^2_β	0,091	0,120	0,202	0,071	0,074	0,083	0,846	0,928	0,974	0,881	0,946	0,981
	$P^2_{\beta_c}$	0,024	0,055	0,143	0,002	0,005	0,015	0,835	0,922	0,972	0,872	0,942	0,979
	$P^2_{\beta*}$	0,022	0,054	0,142	0,002	0,005	0,015	0,835	0,922	0,972	0,872	0,942	0,979
	$P^2_{\beta*_c}$	−0,050	−0,017	0,078	−0,072	−0,068	−0,058	0,822	0,917	0,970	0,863	0,937	0,978
	R^2_{FC}	0,775	0,893	0,961	0,841	0,928	0,974	0,804	0,907	0,967	0,861	0,937	0,978
	$R^2_{FC_c}$	0,758	0,885	0,958	0,829	0,922	0,972	0,790	0,900	0,964	0,851	0,932	0,976
	R^2_{RV}	0,799	0,905	0,965	0,874	0,943	0,980	0,808	0,908	0,967	0,866	0,939	0,978
	$R^2_{RV_c}$	0,781	0,896	0,962	0,862	0,938	0,978	0,790	0,900	0,964	0,854	0,934	0,977
60	P^2_Q	0,048	0,074	0,145	0,032	0,036	0,044	0,809	0,909	0,967	0,852	0,931	0,975
	$P^2_{Q_c}$	0,015	0,041	0,115	−0,002	0,002	0,010	0,802	0,905	0,966	0,847	0,929	0,975
	P^2_β	0,052	0,075	0,145	0,035	0,037	0,044	0,809	0,909	0,967	0,852	0,931	0,975
	$P^2_{\beta_c}$	0,019	0,043	0,115	0,002	0,004	0,011	0,803	0,905	0,966	0,847	0,929	0,975
	$P^2_{\beta*}$	0,018	0,043	0,115	0,001	0,003	0,010	0,803	0,906	0,966	0,847	0,929	0,975
	$P^2_{\beta*_c}$	−0,016	0,009	0,084	−0,034	−0,032	−0,024	0,796	0,902	0,964	0,841	0,926	0,974
	R^2_{FC}	0,739	0,873	0,953	0,808	0,911	0,968	0,768	0,888	0,959	0,834	0,923	0,973
	$R^2_{FC_c}$	0,730	0,869	0,952	0,801	0,908	0,967	0,760	0,884	0,957	0,829	0,921	0,972
	R^2_{RV}	0,764	0,885	0,958	0,846	0,929	0,975	0,772	0,888	0,959	0,841	0,926	0,973
	$R^2_{RV_c}$	0,754	0,880	0,956	0,840	0,926	0,974	0,762	0,884	0,957	0,834	0,923	0,972
120	P^2_Q	0,034	0,060	0,135	0,016	0,019	0,026	0,805	0,907	0,966	0,848	0,930	0,975
	$P^2_{Q_c}$	0,017	0,044	0,120	−0,000	0,002	0,009	0,802	0,905	0,965	0,845	0,928	0,974
	P^2_β	0,036	0,061	0,135	0,018	0,020	0,026	0,805	0,907	0,966	0,848	0,930	0,975
	$P^2_{\beta_c}$	0,019	0,045	0,120	0,001	0,003	0,009	0,802	0,905	0,965	0,845	0,928	0,974
	$P^2_{\beta*}$	0,019	0,045	0,120	0,001	0,003	0,009	0,802	0,905	0,965	0,845	0,928	0,974
	$P^2_{\beta*_c}$	0,003	0,028	0,105	−0,016	−0,014	−0,008	0,798	0,904	0,965	0,843	0,927	0,974
	R^2_{RV}	0,766	0,886	0,958	0,843	0,927	0,974	0,769	0,887	0,958	0,841	0,926	0,974
	$R^2_{RV_c}$	0,761	0,884	0,957	0,840	0,925	0,973	0,764	0,885	0,957	0,838	0,925	0,973
	R^2_{FC}	0,743	0,876	0,954	0,800	0,907	0,966	0,767	0,887	0,958	0,833	0,923	0,972
	$R^2_{FC_c}$	0,739	0,874	0,954	0,797	0,905	0,966	0,763	0,885	0,958	0,830	0,922	0,972

4.3.2 Avaliação das medidas de predição e de qualidade de ajuste sob especificação incorreta

Nesta seção, apresentamos os resultados referentes aos estudos de simulação de Monte Carlo com o intuito de avaliar o desempenho e o comportamento das medidas de predição e de qualidade de ajuste quando cometemos erros de especificação no modelo de regressão BP. Dessa forma, espera-se que esses critérios identifiquem a ocorrência de erros e selecionem o

modelo utilizado na geração dos dados.

4.3.2.1 Erro de especificação na função de ligação do submodelo da média e omissão da modelagem da precisão

Inicialmente, analisamos o comportamento das estatísticas de predição e de qualidade de ajuste quando cometemos erro de especificação na função de ligação do submodelo da média e de omissão da modelagem da precisão, separadamente. Para isso, os dados foram gerados considerando o modelo de regressão BP com precisão variável, cuja estrutura é dada por

$$\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} \qquad \text{e} \qquad \log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}, \tag{4.6}$$

com $i = 1, \ldots, n$. As covariáveis x_i e z_i foram obtidas de forma independente a partir da distribuição $\mathcal{U}(0, 1)$ e mantidas fixas para cada réplica. Três tamanhos de amostras foram considerados, $n = 30, 60$ e 120. O desempenho das medidas foi avaliado em diferentes cenários do modelo de regressão BP: no cenário I $\mu \in (0{,}7; 16{,}0)$ e $\phi \in (4{,}8; 27{,}0)$; no cenário II $\mu \in (0{,}7; 16{,}0)$ e $\phi \in (27{,}0; 152{,}0)$; no cenário III temos $\mu \in (16{,}0; 170{,}0)$ e $\phi \in (4{,}8; 27{,}0)$ e por fim, consideramos o cenário IV em que $\mu \in (16{,}0; 170{,}0)$ e $\phi \in (27{,}0; 152{,}0)$. Os cenários foram escolhidos com o objetivo de verificar se os valores assumidos para a média da variável resposta e para o parâmetro de precisão podem influenciar no desempenho das medidas para indicar a ocorrência dos erros.

Para verificar o comportamento das medidas de predição e de qualidade de ajuste sob o erro de especificação na função de ligação da média, o modelo expresso em (4.6) foi estimado considerando a função de ligação raiz quadrada no submodelo da média. Ademais, para avaliar o comportamento das medidas sob erro de especificação da omissão da modelagem da precisão, o modelo (4.6) foi estimado considerando precisão constante. A Tabela 4.5 apresenta os valores médios das estatísticas avaliadas sob os erros de especificação nos cenários I e II, considerando 10000 réplicas de Monte Carlo, e na Tabela 4.6 encontram-se os valores referentes aos cenários III e IV. Denotamos os modelos I, II e III para representar, respectivamente, o ajuste do modelo (4.6) sob o erro de especificação na função de ligação do submodelo da média, omissão da modelagem da precisão e modelo corretamente especificado.

Observa-se nas Tabelas 4.5 e 4.6 que ambas as medidas de qualidade de predição e as medidas do tipo R^2_{RV} são maiores quando estimamos o modelo sob especificação correta (Modelo III). A medida de qualidade de ajuste R^2_{FC} e sua versão corrigida é menos sensível para detectar o erro de omissão da modelagem da precisão. Em ambos os cenários, os valores do $R^2_{FC_c}$ são maiores quando desconsideramos a precisão variável dos dados (Modelo II) do que quando estimamos corretamente o modelo (Modelo III).

Nota-se um melhor desempenho nas estatísticas P^2 para detectar os erros cometidos, principalmente quando cometemos o erro na função de ligação da média (Modelo I). Por exemplo, no cenário II para $n = 120$ (ver a Tabela 4.5), as estatísticas de predição apresentam

valores médios em torno de 0,55, enquanto as medidas de qualidade de ajuste estão em torno de 0,92. Além disso, quando consideramos valores maiores para o parâmetro de precisão (Cenários II e IV), ambas as estatísticas são altas quando omitimos a modelagem da precisão (Modelo II), indicando a dificuldade dessas medidas em detectar esse tipo de erro nesses cenários.

Tabela 4.5: Valores médios das estatísticas de predição e de qualidade de ajuste obtidas sob erro na função de ligação da média com $\mu \in (0{,}7; 16{,}0)$ e omissão da modelagem da precisão com ϕ variável - Modelo gerado: $\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$ e $\log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}$, $i = 1, \ldots, n$.

Modelo estimado	Modelo I			Modelo II			Modelo III		
	$\sqrt{\mu_i} = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$ $\log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}$			$\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$			$\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$ $\log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}$		
n	30	60	120	30	60	120	30	60	120
Cenário I: $\mu \in (0{,}7; 16{,}0)$ e $\phi \in (4{,}8; 27{,}0)$; $\boldsymbol{\beta} = (1{,}0; -1{,}6; 2{,}0)^\top$ e $\boldsymbol{\nu} = (2{,}5; -1{,}0; 0{,}9)^\top$									
P^2_Q	0,786	0,684	0,569	0,868	0,849	0,882	0,930	0,903	0,904
$P^2_{Q_c}$	0,741	0,655	0,550	0,853	0,841	0,879	0,916	0,894	0,900
P^2_β	0,790	0,693	0,577	0,868	0,847	0,881	0,932	0,904	0,905
$P^2_{\beta_c}$	0,747	0,664	0,558	0,853	0,839	0,878	0,917	0,895	0,901
$P^2_{\beta*}$	0,763	0,675	0,565	0,854	0,839	0,878	0,923	0,899	0,902
$P^2_{\beta*_c}$	0,714	0,644	0,546	0,837	0,831	0,874	0,907	0,890	0,898
R^2_{FC}	0,785	0,767	0,806	0,821	0,801	0,845	0,819	0,801	0,845
$R^2_{FC_c}$	0,740	0,746	0,797	0,800	0,791	0,841	0,781	0,782	0,838
R^2_{RV}	0,829	0,804	0,841	0,829	0,813	0,859	0,852	0,833	0,870
$R^2_{RV_c}$	0,794	0,786	0,834	0,804	0,800	0,855	0,822	0,817	0,864
Cenário II: $\mu \in (0{,}7; 16{,}0)$ e $\phi \in (27{,}0; 152{,}0)$; $\boldsymbol{\beta} = (1{,}0; -1{,}6; 2{,}0)^\top$ e $\boldsymbol{\nu} = (4{,}2; 0{,}9; -1{,}0)^\top$									
P^2_Q	0,857	0,705	0,629	0,968	0,966	0,968	0,984	0,977	0,975
$P^2_{Q_c}$	0,827	0,677	0,613	0,964	0,964	0,967	0,980	0,975	0,974
P^2_β	0,861	0,721	0,642	0,968	0,966	0,968	0,984	0,977	0,975
$P^2_{\beta_c}$	0,832	0,695	0,627	0,964	0,964	0,967	0,980	0,975	0,974
$P^2_{\beta*}$	0,842	0,699	0,632	0,964	0,964	0,967	0,982	0,976	0,974
$P^2_{\beta*_c}$	0,810	0,671	0,616	0,960	0,962	0,966	0,978	0,974	0,973
R^2_{FC}	0,921	0,916	0,917	0,956	0,955	0,958	0,955	0,955	0,958
$R^2_{FC_c}$	0,905	0,908	0,913	0,951	0,953	0,957	0,946	0,951	0,956
R^2_{RV}	0,935	0,924	0,929	0,958	0,957	0,960	0,964	0,962	0,964
$R^2_{RV_c}$	0,921	0,917	0,926	0,952	0,954	0,959	0,956	0,958	0,962

Com relação às versões das estatísticas de predição, observa-se que o $P^2_{\beta*}$ e sua versão corrigida apresentam os menores valores quando comparados aos valores das outras duas versões obtidas nos modelos em que cometemos erros de especificação, o que indica que tais medidas são mais sensíveis à detecção dos erros cometidos. No entanto, é importante ressaltar que o $P^2_{\beta*}$ e sua versão corrigida utilizam resíduos padronizados pela matriz de projeção de ordem n, tornando custoso o cálculo computacional dessas versões quando o tamanho da amostra é grande. Nessa perspectiva, as estatísticas P^2 baseadas nos resíduos quantílico e ponderado, que são mais facilmente computáveis, uma vez que não envolvem o cálculo da matriz de projeção, mostraram-se versões alternativas competitivas com desempenho satisfatório para serem usadas

como critérios de seleção de modelos na regressão BP.

Tabela 4.6: Valores médios das estatísticas de predição e de qualidade de ajuste obtidas sob erro na função de ligação da média com $\mu \in (16{,}0; 170{,}0)$ e omissão da modelagem da precisão com ϕ variável - Modelo gerado: $\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$ e $\log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}$, $i = 1, \ldots, n$.

Modelo estimado	Modelo I			Modelo II			Modelo III		
	$\sqrt{\mu_i} = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$ $\log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}$			$\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$			$\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$ $\log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}$		
n	30	60	120	30	60	120	30	60	120
Cenário III: $\mu \in (16{,}0; 170{,}0)$ e $\phi \in (4{,}8; 27{,}0)$; $\boldsymbol{\beta} = (3{,}4; 1{,}9; -0{,}7)^\top$ e $\boldsymbol{\nu} = (2{,}5; -1{,}0; 0{,}9)^\top$									
P^2_Q	0,785	0,665	0,568	0,851	0,797	0,823	0,959	0,928	0,923
$P^2_{Q_c}$	0,741	0,634	0,549	0,833	0,786	0,818	0,951	0,922	0,919
P^2_β	0,790	0,672	0,571	0,851	0,795	0,821	0,960	0,929	0,923
$P^2_{\beta_c}$	0,746	0,642	0,552	0,833	0,784	0,816	0,952	0,923	0,920
$P^2_{\beta*}$	0,762	0,654	0,560	0,834	0,784	0,816	0,955	0,925	0,921
$P^2_{\beta*_c}$	0,713	0,621	0,541	0,815	0,772	0,811	0,946	0,918	0,918
R^2_{FC}	0,787	0,729	0,761	0,822	0,771	0,804	0,820	0,770	0,804
$R^2_{FC_c}$	0,743	0,704	0,751	0,801	0,758	0,799	0,783	0,749	0,796
R^2_{RV}	0,839	0,789	0,814	0,831	0,782	0,815	0,854	0,805	0,829
$R^2_{RV_c}$	0,805	0,769	0,806	0,806	0,767	0,809	0,823	0,786	0,822
Cenário IV: $\mu \in (16{,}0; 170{,}0)$ e $\phi \in (27{,}0; 152{,}0)$; $\boldsymbol{\beta} = (3{,}4; 1{,}9; -0{,}7)^\top$ e $\boldsymbol{\nu} = (4{,}2; 0{,}9; -1{,}0)^\top$									
P^2_Q	0,888	0,789	0,699	0,962	0,953	0,950	0,991	0,988	0,986
$P^2_{Q_c}$	0,864	0,769	0,686	0,958	0,950	0,949	0,989	0,987	0,986
P^2_β	0,889	0,794	0,702	0,962	0,953	0,950	0,991	0,988	0,986
$P^2_{\beta_c}$	0,866	0,775	0,689	0,958	0,950	0,948	0,989	0,987	0,986
$P^2_{\beta*}$	0,872	0,779	0,693	0,958	0,950	0,948	0,989	0,988	0,986
$P^2_{\beta*_c}$	0,845	0,759	0,680	0,953	0,948	0,947	0,987	0,986	0,985
R^2_{FC}	0,931	0,922	0,921	0,956	0,947	0,945	0,956	0,947	0,945
$R^2_{FC_c}$	0,917	0,914	0,917	0,951	0,944	0,943	0,947	0,942	0,942
R^2_{RV}	0,948	0,933	0,932	0,958	0,949	0,947	0,964	0,955	0,952
$R^2_{RV_c}$	0,937	0,927	0,929	0,952	0,946	0,946	0,956	0,951	0,950

4.3.2.2 Erro de especificação de omissão de covariáveis no submodelo da média

Também analisamos o comportamento das estatísticas de predição e de qualidade de ajuste sob omissão de covariáveis no submodelo da média. Para isso, os dados foram gerados considerando a seguinte estrutura modelo de regressão BP com precisão variável

$$\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} \qquad \text{e} \qquad \log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}, \quad (4.7)$$

com $i = 1, \ldots, n$. Os valores das covariáveis foram obtidos de forma independente a partir da distribuição $\mathcal{U}(0, 1)$ e foram mantidos fixos para cada réplica. Devido a quantidade maior de parâmetros a serem estimados pelo modelo, consideramos $n = 40$, 80 e 120 para os tamanhos de amostras. Além disso, as simulações foram realizadas considerando quatro cenários do modelo de regressão BP. No cenário I assumimos que $\mu \in (0{,}4; 15{,}0)$ e $\phi \in (7{,}8; 42{,}7)$; para o cenário

II consideramos $\mu \in (0{,}4; 15{,}0)$ e $\phi \in (28{,}9; 160{,}8)$; no cenário III temos que $\mu \in (5{,}9; 156{,}5)$ e $\phi \in (7{,}8; 42{,}7)$ e para o cenário IV consideramos $\mu \in (5{,}9; 156{,}5)$ e $\phi \in (28{,}9; 160{,}8)$.

Para verificar o comportamento das medidas de predição e de qualidade de ajuste sob o erro de especificação de omissão de covariáveis, o modelo expresso em (4.7) foi estimado sob a omissão de 3, 2 e 1 covariáveis no submodelo da média. A Tabela 4.7 apresenta os valores médios das estatísticas avaliadas nos cenários I e II, considerando 10000 réplicas de Monte Carlo, e na Tabela 4.8 encontram-se os resultados para os cenários III e IV. Ambas Tabelas apresentam os resultados referentes aos modelos denotados por I, II e III que representam, respectivamente, a estimação do modelo (4.7) sob a omissão de 3, 2 e 1 covariáveis no submodelo da média. O modelo IV representa o modelo (4.7) sob especificação correta, isto é, foi gerado e estimado com 4 covariáveis no submodelo da média. Ainda, nas Tabelas 4.7 e 4.8 encontram-se os valores dos parâmetros utilizados para obter os cenários considerados nas simulações.

Tabela 4.7: Valores médios das estatísticas de predição e de qualidade de ajuste obtidas sob erro de omissão de covariáveis no submodelo da média com $\mu \in (0{,}4; 15{,}0)$ e ϕ variável - Modelo gerado: $\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}$ e $\log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}$, $i = 1, \ldots, n$.

Modelo estimado	Modelo I			Modelo II			Modelo III			Modelo IV		
	$\log(\mu_i) = \beta_1 + \beta_2 x_{i2}$ $\log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}$			$\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$ $\log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}$			$\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$ $\log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}$			$\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}$ $\log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}$		
n	40	80	120	40	80	120	40	80	120	40	80	120
Cenário I: $\mu \in (0{,}4; 15{,}0)$ e $\phi \in (7{,}8; 42{,}7)$; $\boldsymbol{\beta} = (-1{,}5; 0{,}8; 1{,}4; 1{,}3; 1{,}4)^\top$ e $\boldsymbol{\nu} = (3{,}0; -1{,}0; 0{,}9)^\top$												
P^2_Q	0,476	0,169	0,186	0,573	0,383	0,448	0,713	0,634	0,708	0,944	0,926	0,916
$P^2_{Q_c}$	0,416	0,124	0,157	0,511	0,342	0,424	0,661	0,604	0,692	0,932	0,918	0,911
P^2_β	0,519	0,167	0,114	0,601	0,351	0,418	0,717	0,638	0,691	0,945	0,926	0,917
$P^2_{\beta_c}$	0,464	0,122	0,083	0,542	0,308	0,392	0,666	0,609	0,675	0,933	0,919	0,911
$P^2_{\beta*}$	0,493	0,143	0,099	0,567	0,328	0,402	0,686	0,618	0,680	0,935	0,921	0,913
$P^2_{\beta*_c}$	0,435	0,097	0,068	0,503	0,283	0,376	0,629	0,587	0,663	0,921	0,913	0,907
R^2_{FC}	0,303	0,058	0,064	0,407	0,266	0,338	0,570	0,551	0,623	0,868	0,867	0,874
$R^2_{FC_c}$	0,223	0,008	0,032	0,320	0,216	0,309	0,491	0,514	0,603	0,839	0,854	0,866
R^2_{RV}	0,377	0,107	0,089	0,467	0,271	0,345	0,608	0,561	0,633	0,895	0,887	0,891
$R^2_{RV_c}$	0,313	0,065	0,061	0,388	0,222	0,317	0,531	0,522	0,612	0,868	0,874	0,883
Cenário II: $\mu \in (0{,}4; 15{,}0)$ e $\phi \in (28{,}9; 160{,}8)$; $\boldsymbol{\beta} = (-1{,}5; 0{,}8; 1{,}4; 1{,}3; 1{,}4)^\top$ e $\boldsymbol{\nu} = (4{,}5; -1{,}2; 0{,}7)^\top$												
P^2_Q	0,505	0,154	0,171	0,604	0,392	0,474	0,753	0,672	0,758	0,984	0,977	0,974
$P^2_{Q_c}$	0,449	0,109	0,143	0,546	0,351	0,451	0,709	0,645	0,745	0,980	0,975	0,973
P^2_β	0,557	0,178	0,120	0,638	0,373	0,455	0,760	0,685	0,745	0,984	0,978	0,974
$P^2_{\beta_c}$	0,506	0,134	0,089	0,585	0,331	0,431	0,716	0,659	0,731	0,980	0,975	0,973
$P^2_{\beta*}$	0,533	0,154	0,105	0,608	0,351	0,441	0,734	0,668	0,735	0,981	0,976	0,973
$P^2_{\beta*_c}$	0,480	0,109	0,074	0,550	0,307	0,416	0,686	0,640	0,721	0,977	0,974	0,971
R^2_{FC}	0,331	0,063	0,070	0,443	0,291	0,369	0,621	0,606	0,683	0,957	0,957	0,959
$R^2_{FC_c}$	0,254	0,013	0,038	0,361	0,243	0,341	0,552	0,573	0,666	0,947	0,953	0,957
R^2_{RV}	0,413	0,111	0,089	0,506	0,289	0,370	0,659	0,608	0,684	0,966	0,963	0,965
$R^2_{RV_c}$	0,353	0,068	0,060	0,434	0,241	0,342	0,592	0,574	0,666	0,957	0,959	0,963

Tabela 4.8: Valores médios das estatísticas de predição e de qualidade de ajuste obtidas sob erro de omissão de covariáveis no submodelo da média com $\mu \in (5{,}9; 156{,}5)$ e ϕ variável - Modelo gerado: $\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}$ e $\log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}$, $i = 1, \ldots, n$.

Modelo estimado	Modelo I: $\log(\mu_i) = \beta_1 + \beta_2 x_{i2}$; $\log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}$			Modelo II: $\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$; $\log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}$			Modelo III: $\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$; $\log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}$			Modelo IV: $\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}$; $\log(\phi_i) = \nu_1 + \nu_2 z_{i2} + \nu_3 z_{i3}$		
n	40	80	120	40	80	120	40	80	120	40	80	120
Cenário III: $\mu \in (5{,}9; 156{,}5)$ e $\phi \in (7{,}8; 42{,}7)$; $\boldsymbol{\beta} = (1{,}1; 0{,}8; 1{,}3; 1{,}2; 1{,}3)^\top$ e $\boldsymbol{\nu} = (3{,}0; -1{,}0; 0{,}9)^\top$												
P^2_Q	0,632	0,391	0,358	0,722	0,455	0,490	0,828	0,687	0,717	0,973	0,960	0,952
$P^2_{Q_c}$	0,590	0,358	0,336	0,681	0,418	0,468	0,797	0,661	0,702	0,967	0,956	0,949
P^2_β	0,656	0,363	0,244	0,737	0,406	0,442	0,830	0,688	0,697	0,973	0,960	0,952
$P^2_{\beta_c}$	0,617	0,329	0,218	0,698	0,366	0,418	0,799	0,663	0,681	0,967	0,956	0,949
$P^2_{\beta*}$	0,637	0,345	0,232	0,714	0,384	0,428	0,812	0,671	0,687	0,969	0,957	0,950
$P^2_{\beta*_c}$	0,596	0,310	0,205	0,672	0,343	0,403	0,778	0,644	0,670	0,962	0,953	0,947
R^2_{FC}	0,325	0,070	0,077	0,427	0,280	0,354	0,586	0,567	0,641	0,887	0,889	0,895
$R^2_{FC_c}$	0,248	0,021	0,045	0,343	0,231	0,326	0,510	0,532	0,622	0,862	0,878	0,889
R^2_{RV}	0,425	0,141	0,103	0,510	0,305	0,373	0,643	0,598	0,654	0,911	0,906	0,909
$R^2_{RV_c}$	0,366	0,100	0,075	0,438	0,258	0,346	0,573	0,563	0,634	0,889	0,895	0,903
Cenário IV: $\mu \in (5{,}9; 156{,}5)$ e $\phi \in (28{,}9; 160{,}8)$; $\boldsymbol{\beta} = (1{,}1; 0{,}8; 1{,}3; 1{,}2; 1{,}3)^\top$ e $\boldsymbol{\nu} = (4{,}5; -1{,}2; 0{,}7)^\top$												
P^2_Q	0,697	0,403	0,349	0,769	0,454	0,511	0,877	0,714	0,752	0,992	0,989	0,987
$P^2_{Q_c}$	0,662	0,371	0,326	0,735	0,417	0,489	0,855	0,690	0,739	0,991	0,988	0,986
P^2_β	0,722	0,392	0,247	0,787	0,414	0,472	0,880	0,722	0,736	0,993	0,989	0,987
$P^2_{\beta_c}$	0,690	0,359	0,221	0,755	0,375	0,449	0,858	0,699	0,722	0,991	0,988	0,986
$P^2_{\beta*}$	0,707	0,374	0,234	0,768	0,393	0,459	0,867	0,707	0,726	0,991	0,988	0,986
$P^2_{\beta*_c}$	0,673	0,340	0,208	0,734	0,352	0,435	0,842	0,683	0,712	0,989	0,987	0,985
R^2_{FC}	0,352	0,076	0,082	0,460	0,304	0,381	0,631	0,614	0,691	0,963	0,965	0,967
$R^2_{FC_c}$	0,278	0,026	0,050	0,381	0,256	0,353	0,564	0,582	0,674	0,955	0,961	0,965
R^2_{RV}	0,460	0,147	0,104	0,547	0,323	0,394	0,689	0,640	0,697	0,971	0,970	0,971
$R^2_{RV_c}$	0,405	0,106	0,076	0,480	0,277	0,368	0,628	0,608	0,680	0,964	0,967	0,969

Observa-se nas Tabelas 4.7 e 4.8 que ambas versões das estatísticas de predição e de qualidade de ajuste aumentam à medida que as covariáveis são inseridas no submodelo da média, como esperado. Novamente, nota-se que, quando o modelo de regressão BP é corretamente especificado (ver a Tabela 4.7 - Modelo IV), o R^2_{RV} e sua versão corrigida apresentam valores maiores que o R^2_{FC} e sua versão corrigida. Isso justifica o fato de que, sob a omissão de covariáveis, os critérios R^2_{FC} e $R^2_{FC_c}$ apresentam valores menores quando comparados aos R^2_{RV} e $R^2_{RV_c}$. Além disso, dado que o critério R^2_{FC} e sua versão corrigida levam em consideração a correlação entre $\widehat{\boldsymbol{\eta}}$ e $g(\boldsymbol{y})$, pequenos valores desse critério indicam que essa correlação é baixa, evidenciando um erro de especificação no submodelo da média, neste caso, a omissão de covariáveis.

Também realizamos o estudo do comportamento das medidas sob o erro de omissão de covariáveis considerando o modelo de regressão BP com precisão constante. Para isso,

consideramos a mesma estrutura para o submodelo da média, apresentada no modelo definido em (4.7), e para o parâmetro da precisão consideramos $\phi = 20, 50$ e 150. A Tabela 4.9 apresenta os resultados referentes aos cenários em que $\mu \in (0{,}4; 15{,}0)$ e a Tabela 4.10 contém os resultados para os cenários em que $\mu \in (5{,}9; 156{,}5)$.

Tabela 4.9: Valores médios das estatísticas de predição e de qualidade de ajuste obtidas sob erro de omissão de covariáveis no submodelo da média com $\mu \in (0{,}4; 15{,}0)$ e ϕ constante - Modelo gerado: $\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}$, $i = 1, \ldots, n$.

Modelo estimado	Modelo I			Modelo II			Modelo III			Modelo IV		
	$\log(\mu_i) = \beta_1 + \beta_2 x_{i2}$			$\log(\mu_i) = \beta_1 + \beta_2 x_{i2}$ $+\beta_3 x_{i3}$			$\log(\mu_i) = \beta_1 + \beta_2 x_{i2}$ $+\beta_3 x_{i3} + \beta_4 x_{i4}$			$\log(\mu_i) = \beta_1 + \beta_2 x_{i2}$ $+\beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}$		
n	40	80	120	40	80	120	40	80	120	40	80	120
$\mu \in (0{,}4; 15{,}0)$; $\boldsymbol{\beta} = (-1{,}5; 0{,}8; 1{,}4; 1{,}3; 1{,}4)^\top$; $\phi = 20$												
P^2_Q	0,393	0,121	0,176	0,498	0,351	0,422	0,659	0,608	0,697	0,918	0,904	0,913
$P^2_{Q_c}$	0,360	0,099	0,162	0,456	0,325	0,407	0,621	0,587	0,687	0,906	0,897	0,909
P^2_β	0,417	0,110	0,079	0,495	0,312	0,385	0,634	0,609	0,676	0,919	0,904	0,913
$P^2_{\beta_c}$	0,386	0,087	0,063	0,453	0,285	0,369	0,592	0,589	0,665	0,907	0,897	0,910
$P^2_{\beta*}$	0,387	0,084	0,063	0,456	0,287	0,368	0,600	0,588	0,664	0,907	0,897	0,910
$P^2_{\beta*_c}$	0,354	0,060	0,047	0,411	0,259	0,352	0,555	0,566	0,652	0,893	0,890	0,906
R^2_{FC}	0,305	0,059	0,065	0,410	0,268	0,341	0,581	0,555	0,631	0,890	0,875	0,888
$R^2_{FC_c}$	0,268	0,035	0,049	0,361	0,239	0,323	0,533	0,532	0,618	0,874	0,867	0,883
R^2_{RV}	0,339	0,078	0,055	0,424	0,264	0,335	0,573	0,557	0,632	0,896	0,881	0,893
$R^2_{RV_c}$	0,295	0,049	0,036	0,362	0,227	0,313	0,508	0,525	0,616	0,875	0,870	0,887
$\mu \in (0{,}4; 15{,}0)$; $\boldsymbol{\beta} = (-1{,}5; 0{,}8; 1{,}4; 1{,}3; 1{,}4)^\top$; $\phi = 50$												
P^2_Q	0,409	0,105	0,163	0,521	0,359	0,436	0,691	0,642	0,732	0,964	0,957	0,961
$P^2_{Q_c}$	0,377	0,081	0,149	0,481	0,334	0,422	0,656	0,623	0,723	0,958	0,954	0,960
P^2_β	0,448	0,117	0,084	0,526	0,335	0,412	0,669	0,652	0,717	0,964	0,957	0,961
$P^2_{\beta_c}$	0,418	0,094	0,068	0,486	0,308	0,397	0,631	0,633	0,707	0,958	0,954	0,960
$P^2_{\beta*}$	0,419	0,091	0,068	0,490	0,311	0,397	0,639	0,632	0,706	0,958	0,954	0,960
$P^2_{\beta*_c}$	0,388	0,068	0,052	0,448	0,283	0,381	0,598	0,613	0,696	0,952	0,951	0,958
R^2_{FC}	0,328	0,063	0,069	0,439	0,287	0,364	0,621	0,597	0,675	0,951	0,943	0,949
$R^2_{FC_c}$	0,291	0,039	0,053	0,392	0,259	0,348	0,578	0,576	0,664	0,943	0,939	0,947
R^2_{RV}	0,363	0,083	0,058	0,452	0,281	0,357	0,610	0,595	0,674	0,953	0,946	0,952
$R^2_{RV_c}$	0,321	0,054	0,039	0,393	0,245	0,336	0,550	0,567	0,659	0,944	0,941	0,949
$\mu \in (0{,}4; 15{,}0)$; $\boldsymbol{\beta} = (-1{,}5; 0{,}8; 1{,}4; 1{,}3; 1{,}4)^\top$; $\phi = 150$												
P^2_Q	0,417	0,096	0,155	0,532	0,365	0,443	0,707	0,661	0,751	0,987	0,985	0,986
$P^2_{Q_c}$	0,385	0,073	0,141	0,493	0,340	0,429	0,673	0,643	0,742	0,985	0,984	0,986
P^2_β	0,464	0,121	0,086	0,542	0,348	0,427	0,687	0,676	0,738	0,987	0,985	0,986
$P^2_{\beta_c}$	0,435	0,098	0,070	0,504	0,322	0,412	0,651	0,658	0,729	0,985	0,984	0,986
$P^2_{\beta*}$	0,436	0,095	0,071	0,508	0,324	0,412	0,659	0,657	0,728	0,985	0,984	0,986
$P^2_{\beta*_c}$	0,406	0,072	0,055	0,467	0,298	0,397	0,620	0,639	0,719	0,983	0,983	0,985
R^2_{FC}	0,339	0,065	0,071	0,453	0,298	0,377	0,641	0,620	0,699	0,983	0,980	0,982
$R^2_{FC_c}$	0,303	0,041	0,056	0,407	0,270	0,360	0,600	0,600	0,688	0,980	0,979	0,981
R^2_{RV}	0,374	0,085	0,060	0,466	0,292	0,368	0,629	0,618	0,695	0,983	0,981	0,983
$R^2_{RV_c}$	0,333	0,056	0,041	0,408	0,256	0,347	0,571	0,591	0,682	0,980	0,979	0,982

Tabela 4.10: Valores médios das estatísticas de predição e de qualidade de ajuste obtidas sob erro de omissão de covariáveis no submodelo da média com $\mu \in (5{,}9; 156{,}5)$ e ϕ constante - Modelo gerado: $\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}$, $i = 1, \ldots, n$.

Modelo estimado	Modelo I			Modelo II			Modelo III			Modelo IV		
	$\log(\mu_i) = \beta_1 + \beta_2 x_{i2}$			$\log(\mu_i) = \beta_1 + \beta_2 x_{i2}$ $+\beta_3 x_{i3}$			$\log(\mu_i) = \beta_1 + \beta_2 x_{i2}$ $+\beta_3 x_{i3} + \beta_4 x_{i4}$			$\log(\mu_i) = \beta_1 + \beta_2 x_{i2}$ $+\beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}$		
n	40	80	120	40	80	120	40	80	120	40	80	120
$\mu \in (5{,}9; 156{,}5)$; $\boldsymbol{\beta} = (1{,}1; 0{,}8; 1{,}3; 1{,}2; 1{,}3)^\top$; $\phi = 20$												
P^2_Q	0,413	0,173	0,244	0,511	0,368	0,431	0,660	0,610	0,689	0,921	0,909	0,918
$P^2_{Q_c}$	0,381	0,152	0,231	0,470	0,343	0,416	0,621	0,589	0,678	0,909	0,903	0,915
P^2_β	0,424	0,123	0,077	0,498	0,307	0,374	0,629	0,608	0,664	0,921	0,910	0,918
$P^2_{\beta_c}$	0,393	0,101	0,062	0,456	0,280	0,358	0,587	0,587	0,652	0,910	0,904	0,915
$P^2_{\beta*}$	0,393	0,097	0,062	0,460	0,281	0,358	0,596	0,587	0,651	0,909	0,903	0,915
$P^2_{\beta*_c}$	0,360	0,074	0,046	0,415	0,253	0,341	0,550	0,564	0,639	0,896	0,897	0,911
R^2_{FC}	0,330	0,071	0,077	0,432	0,282	0,356	0,599	0,570	0,647	0,906	0,895	0,906
$R^2_{FC_c}$	0,293	0,047	0,062	0,385	0,253	0,339	0,553	0,547	0,634	0,892	0,888	0,902
R^2_{RV}	0,387	0,109	0,066	0,469	0,298	0,363	0,607	0,593	0,653	0,912	0,901	0,911
$R^2_{RV_c}$	0,347	0,081	0,047	0,412	0,262	0,342	0,547	0,564	0,637	0,894	0,892	0,906
$\mu \in (5{,}9; 156{,}5)$; $\boldsymbol{\beta} = (1{,}1; 0{,}8; 1{,}3; 1{,}2; 1{,}3)^\top$; $\phi = 50$												
P^2_Q	0,427	0,165	0,237	0,531	0,377	0,442	0,688	0,642	0,720	0,965	0,959	0,964
$P^2_{Q_c}$	0,396	0,143	0,224	0,492	0,352	0,427	0,653	0,622	0,710	0,959	0,957	0,962
P^2_β	0,450	0,130	0,081	0,525	0,326	0,397	0,661	0,648	0,700	0,965	0,959	0,964
$P^2_{\beta_c}$	0,420	0,108	0,065	0,486	0,300	0,381	0,622	0,629	0,690	0,959	0,957	0,962
$P^2_{\beta*}$	0,421	0,104	0,066	0,490	0,301	0,381	0,631	0,628	0,689	0,959	0,957	0,962
$P^2_{\beta*_c}$	0,389	0,081	0,050	0,447	0,274	0,365	0,589	0,608	0,678	0,953	0,954	0,961
R^2_{FC}	0,349	0,075	0,082	0,457	0,300	0,376	0,634	0,607	0,684	0,958	0,953	0,959
$R^2_{FC_c}$	0,314	0,051	0,066	0,412	0,272	0,360	0,592	0,586	0,673	0,952	0,950	0,957
R^2_{RV}	0,409	0,115	0,069	0,495	0,316	0,382	0,640	0,629	0,689	0,961	0,956	0,961
$R^2_{RV_c}$	0,370	0,087	0,050	0,440	0,281	0,362	0,585	0,603	0,674	0,953	0,952	0,958
$\mu \in (5{,}9; 156{,}5)$; $\boldsymbol{\beta} = (1{,}1; 0{,}8; 1{,}3; 1{,}2; 1{,}3)^\top$; $\phi = 150$												
P^2_Q	0,434	0,160	0,232	0,541	0,381	0,448	0,703	0,658	0,736	0,988	0,986	0,987
$P^2_{Q_c}$	0,404	0,138	0,219	0,503	0,357	0,433	0,669	0,640	0,726	0,986	0,985	0,987
P^2_β	0,464	0,134	0,083	0,540	0,337	0,409	0,677	0,669	0,719	0,988	0,986	0,987
$P^2_{\beta_c}$	0,435	0,112	0,067	0,502	0,310	0,394	0,640	0,651	0,709	0,986	0,985	0,987
$P^2_{\beta*}$	0,436	0,108	0,068	0,506	0,312	0,394	0,649	0,650	0,709	0,986	0,985	0,987
$P^2_{\beta*_c}$	0,406	0,085	0,052	0,465	0,285	0,378	0,609	0,631	0,699	0,984	0,984	0,986
R^2_{FC}	0,360	0,077	0,084	0,471	0,309	0,387	0,651	0,626	0,703	0,985	0,984	0,986
$R^2_{FC_c}$	0,325	0,053	0,068	0,427	0,282	0,371	0,612	0,606	0,693	0,983	0,982	0,985
R^2_{RV}	0,421	0,118	0,071	0,509	0,325	0,393	0,657	0,647	0,707	0,986	0,984	0,986
$R^2_{RV_c}$	0,383	0,091	0,052	0,456	0,290	0,373	0,605	0,623	0,693	0,983	0,983	0,985

Nota-se que as conclusões são similares as do estudo feito considerando o modelo com precisão variável. As estatísticas de predição e de qualidade de ajuste conseguem identificar o erro de omissão de covariáveis no submodelo da média, dado que à medida que as covariáveis omitidas são inseridas no modelo, ambas as estatísticas aumentam.

Em resumo, observamos que a qualidade preditiva do modelo de regressão BP é afetada positivamente quando utilizamos a função de ligação logarítmica no submodelo da média e, sob especificação correta do modelo, quanto maior a precisão dos dados mais próximos de 1 são os valores de tais medidas. Nota-se que as versões das medidas de predição com base nos resíduos quantílico, ponderado e ponderado padronizado apresentam desempenho satisfatório na identificação de especificação incorreta do modelo de regressão BP. Com relação aos critérios do tipo pseudo-R^2, o R^2_{RV} é mais sensível à identificação da necessidade da modelagem da precisão quando os dados apresentarem precisão variável e são ajustados com precisão constante do que o R^2_{FC}. Ressalta-se, ainda, a influência do parâmetro de precisão. Vimos que, quanto maior o valor de ϕ, mais próximos de 1 são os valores das medidas, mesmo quando o modelo é incorretamente especificado.

4.4 Aplicações

Nesta seção, apresentamos duas aplicações a dados reais para ilustrar o uso das medidas de predição e de qualidade de ajuste na seleção de modelos na regressão BP. A primeira aplicação refere-se aos dados do milho (GRIFFITHS et al., 1993) e a segunda aplicação refere-se aos dados de aluguel de terras agrícolas (WEISBERG, 2005).

4.4.1 Dados do milho

Os dados da primeira aplicação são referentes a um experimento descrito em Griffiths et al. (1993), em que a produtividade do milho é estudada considerando combinações de nitrogênio e fosfato, que são dois dos principais nutrientes essenciais para o desenvolvimento das plantas. A variável resposta (y_i) é a produtividade do milho (libras/acre) e as variáveis explicativas são as combinações de nitrogênio (x_{i1}) e fosfato (x_{i2}) correspondente à i-ésima experimentação, $i = 1, \ldots, 30$. Bourguignon et al. (2021) concluíram que o modelo de regressão BP apresentou melhor ajuste para esse conjunto de dados quando comparados a alguns modelos de regressão para dados positivos assimétricos.

Ajustamos diversos candidatos a preditores lineares para esses dados e utilizamos as medidas de predição e de qualidade de ajuste para selecionar o melhor modelo na perspectiva de poder preditivo e de maior variabilidade explicada. A Tabela 4.11 apresenta os valores das medidas obtidas nos ajustes dos modelos M1, M2, M3 e M4, cujas estruturas são dadas, respectivamente, por

$$
\begin{aligned}
&\text{M1:} \quad \log(\mu_i) = \beta_0 + \beta_1 \log(x_{i1}),\\
&\text{M2:} \quad \sqrt{\mu_i} = \beta_0 + \beta_1 \log(x_{i1}) + \beta_2 \log(x_{i2}),\\
&\text{M3:} \quad \log(\mu_i) = \beta_0 + \beta_1 \log(x_{i1}) + \beta_2 \log(x_{i2}),\\
&\text{M4:} \quad \log(\mu_i) = \beta_0 + \beta_1 \log(x_{i1}) + \beta_2 \log(x_{i2}) \quad \text{e} \quad \log(\phi_i) = \nu_1 + \nu_2 x_{i2},
\end{aligned}
$$

com $i = 1, \ldots, 30$.

Observa-se que ambas medidas sugerem que a função de ligação logarítmica é a mais indicada para a modelagem da média, com destaque para as medidas de predição que enfatizam a adequação dessa função de ligação. Além disso, o critério $R^2_{FC_c}$ não seleciona o modelo de regressão com precisão variável (M4) como o melhor modelo, achados em concordância com os resultados obtidos nas simulações, em que concluímos que esse critério tende a desconsiderar a modelagem da precisão, ainda que os dados apresentem precisão variável. As demais medidas sugerem o modelo M4 como melhor modelo para o conjunto de dados do milho, dado que os valores dessas medidas são mais próximos de 1 quando comparados aos valores dos demais ajustes.

Tabela 4.11: Valores das estatísticas de predição e de qualidade de ajuste - Modelo de regressão BP aplicado aos dados do milho.

Modelo	P^2_Q	$P^2_{Q_c}$	P^2_β	$P^2_{\beta_c}$	$P^2_{\beta*}$	$P^2_{\beta*_c}$	R^2_{FC}	$R^2_{FC_c}$	R^2_{RV}	$R^2_{RV_c}$
M1	0,480	0,441	0,394	0,349	0,355	0,307	0,390	0,345	0,371	0,314
M2	−0,082	−0,207	−0,058	−0,180	−0,323	−0,475	0,849	0,832	0,852	0,829
M3	0,892	0,879	0,891	0,879	0,878	0,864	0,875	0,861	0,881	0,864
M4	0,996	0,995	0,996	0,995	0,995	0,994	0,875	0,855	0,897	0,878

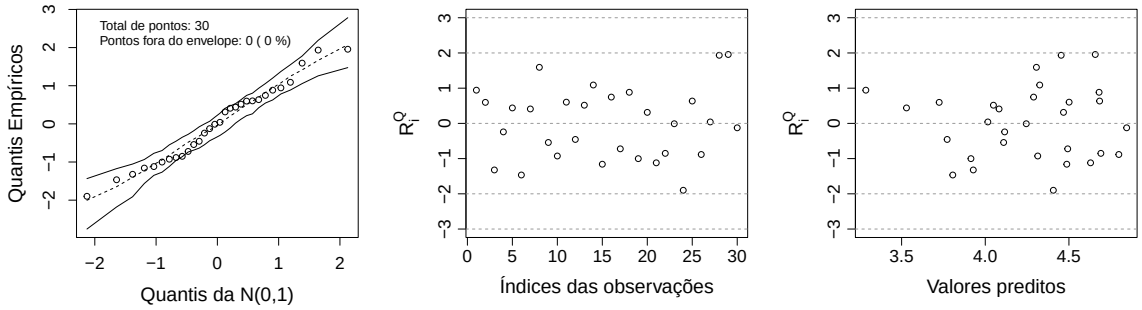
Na Tabela 4.12 são apresentadas as estimativas de máxima verossimilhança dos parâmetros, os erros-padrão e os p-valores do modelo M4. Nota-se que todas as covariadas envolvidas são estatisticamente significativas ao nível de 5%. Ademais, o teste RESET (SANTOS et al., 2021) apresentou p-valor = 0,9262, logo, ao nível de 5% de significância, a hipótese nula de que o modelo está corretamente especificado não foi rejeitada.

Tabela 4.12: Estimativas dos parâmetros, erros-padrão e p-valores do modelo de regressão BP com precisão variável aplicado aos dados do milho.

Parâmetro	β_0	β_1	β_2	ν_0	ν_1
Estimativa	0,5207	0,3506	0,3990	2,7027	0,0072
Erro-padrão	0,2788	0,0330	0,0423	0,6650	0,0034
p-valor	0,0736	$< 0{,}0000$	$< 0{,}0000$	0,0004	0,0432

A análise residual (ver a Figura 4.1) valida a qualidade do ajuste modelo de regressão BP com precisão variável selecionado pelas medidas de predição e de qualidade de ajuste, indicando que o modelo M4 é adequado para explicar o comportamento da produtividade do milho. Nota-se que o resíduo r_i^Q encontra-se disperso entre as bandas de confiança do envelope e apresenta pontos aleatoriamente distribuídos em torno do zero nos gráficos de dispersão, sem observações atípicas. Ademais, o teste de SW apresentou p-valor = 0,5994 e assim, ao nível de 5% de significância, não rejeitamos a hipótese nula de que o resíduo quantílico provém de uma população normalmente distribuída.

Figura 4.1: Gráfico de envelopes simulados (a) e gráficos de dispersão do resíduo r_i^Q versus índice das observações (b) e valores preditos (c) para os dados do milho - Modelo de regressão BP com precisão variável.



(a) (b) (c)

Vale ressaltar que o modelo M4 selecionado pelas medidas de predição e de qualidade de ajuste é o mesmo utilizado por Bourguignon et al. (2021) e considerado por (SANTOS et al., 2021), que confirmaram pelo teste RESET a especificação correta do modelo de regressão BP com precisão variável. Essa aplicação evidencia os resultados obtidos nas simulações e ressaltam a utilidade dos critérios aqui propostos na seleção de modelos na regressão BP.

4.4.2 Dados de aluguel de terras agrícolas

Nesta aplicação utilizamos o conjunto de dados descrito na Seção 3.3.1, referente ao aluguel de terras agrícolas plantadas com alfafa em condados do estado de Minnesota em 1977 (WEISBERG, 2005). A variável resposta é o aluguel médio pago por acre plantado com alfafa (y_i) e as variáveis explicativas são: o aluguel médio pago por terras para outros fins agrícolas (x_{i1}) e a densidade de vacas leiteiras (x_{i2}) para o grupo em que a calagem é necessária para o cultivo de alfafa, $i = 1, \ldots, 33$.

Ajustamos diversos modelos de regressão BP ao conjunto de dados e o modelo selecionado pelas medidas de predição e de qualidade de ajuste contém a seguinte estrutura

$$\log(\mu_i) = \beta_0 + \beta_1 \log(x_{i1}) + \beta_2 \log(x_{i2}) \quad \text{e} \quad \log(\phi_i) = \nu_1 + \nu_2 x_{i2},$$

com $i = 1, \ldots, 33$. A Tabela 4.13 apresenta os valores dos critérios obtidos pelo ajuste do modelo. Ambas as medidas são próximas de 1, indicando que o modelo ajustado apresenta alta qualidade preditiva e de ajuste. Vale ressaltar que para esse ajuste, as estimativas para o parâmetro de precisão assumem valores de $\widehat{\phi_i} \in (16{,}24; 5441{,}93)$ e, como visto nos estudos de simulação, ambas as medidas de predição e qualidade de ajuste apresentam valores próximos de 1 quando ϕ assume valores altos. Além disso, tais valores também indicam indícios de especificação correta do modelo.

Tabela 4.13: Valores das estatísticas de predição e de qualidade de ajuste - Modelo de regressão BP aplicado aos dados de aluguel de terras agrícolas.

Estatísticas	P^2_Q	$P^2_{Q_c}$	P^2_β	$P^2_{\beta_c}$	$P^2_{\beta*}$	$P^2_{\beta*_c}$	R^2_{FC}	$R^2_{FC_c}$	R^2_{RV}	$R^2_{RV_c}$
Valores	0,999	0,999	0,999	0,999	0,998	0,998	0,910	0,897	0,932	0,921

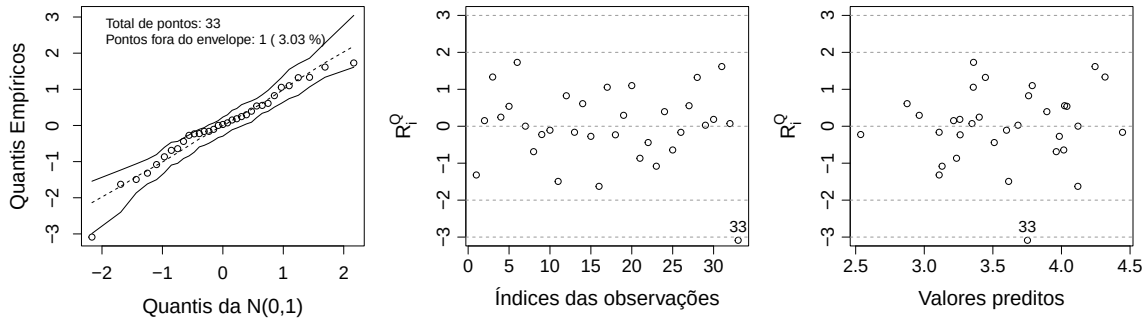
Na Tabela 4.14 são apresentadas as estimativas de máxima verossimilhança dos parâmetros, os erros-padrão e os p-valores associados aos testes de significância dos parâmetros do modelo. Observa-se que todos os parâmetros são estatisticamente significativos ao nível de 5% de significância.

Tabela 4.14: Estimativas dos parâmetros, erros-padrão e p-valores do modelo de regressão BP com precisão variável aplicado aos dados de aluguel de terras agrícolas.

Parâmetro	β_0	β_1	β_2	ν_0	ν_1
Estimativa	−0,6467	1,0108	0,2116	2,5881	0,1302
Erro-padrão	0,1152	0,0319	0,0209	0,4876	0,0346
p-valor	$< 0{,}0000$	$< 0{,}0000$	$< 0{,}0000$	$< 0{,}0000$	0,0008

Ademais, a análise residual (ver a Figura 4.2) valida a qualidade do ajuste do modelo selecionado pelas medidas de predição e de qualidade de ajuste. Nota-se que o resíduo r_i^Q encontra-se disperso entre as bandas de confiança do envelope e apresenta pontos aleatoriamente distribuídos em torno do zero nos gráficos de dispersão, apesar de apresentar uma observação fora dos limites de detecção de pontos atípicos. Aplicamos o teste de SW e este apresentou p-valor = 0,2645, logo, ao nível de 5% de significância, não rejeitamos a hipótese nula de que o resíduo r_i^Q provém de uma população normalmente distribuída. Referente ao teste RESET, este apresentou p-valor = 0,8723 e assim, ao nível de 5% de significância, a hipótese nula de que o modelo está corretamente especificado não foi rejeitada.

Figura 4.2: Gráfico de envelopes simulados (a) e gráficos de dispersão do resíduo r_i^Q versus índice das observações (b) e valores preditos (c) para os dados do aluguel de terras agrícolas - Modelo de regressão BP com precisão variável.



(a) (b) (c)

5 Conclusões

Neste trabalho, propomos e avaliamos algumas medidas para análise de dignóstico em modelos de regressão BP. Propomos os resíduos ponderado, ponderado padronizado e Pearson padronizado para o modelo de regressão BP. Via estudos de simulações de Monte Carlo, avaliamos a distribuição empírica e o desempenho desses resíduos e dos resíduos quantílico e de Pearson, já propostos para tal modelo. Verificamos em diferentes cenários e tamanhos de amostra que, dentre os resíduos analisados, o resíduo quantílico é o resíduo que possui distribuição empírica que melhor se aproxima da distribuição normal padrão quando o modelo de regressão BP é corretamente especificado.

Observamos que a boa aderência da distribuição empiríca do resíduo quantílico pela distribuição normal padrão pode ser utilizada como um referencial para indicar a qualidade ou a falta de ajuste do modelo de regressão BP, dado que essa característica permite que esse resíduo possua propriedades conhecidas, cujos gráficos são mais facilmente interpretáveis. Além disso, vimos que o resíduo quantílico apresenta desempenho satisfatório para detectar erros de especificação no modelo BP. Aplicações a dados reais ilustram o desempenho desse resíduo. Dessa forma, concluímos que o resíduo quantílico apresenta boas propriedades para ser utilizado em análises de diagnósticos do modelo de regressão BP.

Adicionalmente, com base nos resíduos quantílico, ponderado e ponderado padronizado, propomos e avaliamos versões do coeficiente de predição P^2 como medida do poder preditivo do modelo de regressão BP. Os resultados dos estudos de simulações de Monte Carlo, em diferentes cenários com e sem erros de especificação, mostraram que ambas as versões das estatísticas P^2 apresentam desempenho satisfatório para serem utilizadas como critérios de seleção de modelos na regressão BP. Devido ao uso da matriz de projeção, as medidas baseadas nos resíduos ponderado padronizado podem ser mais custosas computacionalmente quando o tamanho da amostra é grande. Nessa perspectiva, as medidas P^2 baseadas nos resíduos quantílico e ponderado se mostraram versões competitivas e com desempenho satisfatório para serem utilizadas na classe de modelos de regressão BP.

Avaliamos também as medidas R^2_{FC} e R^2_{RV} e suas versões corrigidas como medidas de qualidade de ajuste do modelo de regressão BP. Vimos que a qualidade preditiva e a qualidade de ajuste do modelo são afetadas positivamente quando os dados apresentam precisão alta. Observamos que, quando a precisão dos dados é alta e em modelos corretamente especificados, os valores das estatísticas devem estar próximos de um. Além disso, observamos que a utilização da função de ligação logarítmica no submodelo da média propicia uma melhor qualidade preditiva do modelo de regressão BP. Ademais, com relação aos critérios do tipo pseudo-R^2, vimos que o critério R^2_{RV} e sua versão corrigida são mais sensíveis à especificação correta da modelagem da precisão. Aplicações a dados reais evidenciaram a utilidade das medidas de predição aqui propostas e das medidas de qualidade de ajuste como critérios de seleção de modelos na regressão

BP.

Os critérios do tipo pseudo-R^2 e P^2 fornecem conclusões distintas, uma vez que os pseudos-R^2 podem ser utilizados para selecionar modelos com maior variabilidade explicada e os critérios do tipo P^2 selecionam modelos com maior qualidade de predição. Dessa forma, essas medidas devem ser utilizadas conjuntamente na seleção de modelos de regressão BP com maior variância explicada e alta qualidade preditiva.

Em suma, este trabalho contribuiu para avanços significativos na classe de modelos de regressão BP. Avaliamos resíduos e propomos medidas de predição para um modelo de regressão que pode ser aplicado em diversas situações práticas. Com base nos resultados aqui apresentados, recomendamos a utilização do resíduo quantílico e das medidas de predição aqui propostas na etapa de análise de diagnóstico do modelo de regressão BP, a fim de validar a qualidade do ajuste e verificar possíveis especificações incorretas do modelo.

5.1 Sugestões para trabalhos futuros

Em decorrência deste trabalho, apresentam-se como sugestões para futuras pesquisas:

- Avaliação das medidas de predição e dos critérios de seleção AIC, BIC, HQ na seleção de modelos na regressão BP no contexto de modelos fácil e dificilmente identificável, tal como em Giampaoli et al. (2009) e Bayer e Cribari-Neto (2017);
- Análise de Influência em modelos de regressão BP;
- Intervalos de Predição Bootstrap em modelos de regressão BP.

Referências

AARSET, M. V. How to identify a bathtub hazard rate. **IEEE Transactions on Reliability**, IEEE, v. 36, n. 1, p. 106–108, 1987. DOI:10.1109/TR.1987.5222310.

ABRAMOVICH, F.; RITOV, Y. **Statistical Theory: A Concise Introduction**. New York: Hoboken Chapman & Hall/CRC, 2013.

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: **Second International Symposium on Information Theory**. Hungary: Akadémiai Kiadó, 1973. p. 267–281.

AKAIKE, H. A Bayesian analysis of the minimum AIC procedure. **Annals of the Institute of Statistical Mathematics**, Springer, v. 30, n. 1, p. 9–14, 1978. DOI:10.1007/BF02480194.

ALCANTARA, I. M. et al. Model selection using PRESS statistic. **Computational Statistics**, Springer, v. 38, n. 1, p. 285–298, 2022. DOI:10.1007/s00180-022-01228-1.

ALLEN, D. M. The relationship between variable selection and data agumentation and a method for prediction. **Technometrics**, Taylor & Francis, v. 16, n. 1, p. 125–127, 1974. DOI:10.2307/1267500.

AMIN, M. et al. Empirical examination of the Poisson regression residuals for the evaluation of influential points. **Mathematical Problems in Engineering**, Hindawi, v. 2022, n. 1, p. 1–9, 2022. DOI:10.1155/2022/6995911.

AMIN, M. et al. Outlier detection in gamma regression using pearson residuals: Simulation and an application. **AIMS Mathematics**, v. 7, n. 8, p. 15331–15347, 2022. DOI:10.3934/math.2022840.

ANGUS, J. E. The probability integral transform and related results. **SIAM review**, SIAM, v. 36, n. 4, p. 652–654, 1994. DOI:10.1137/1036146.

ATKINSON, A. C. Two graphical displays for outlying and influential observations in regression. **Biometrika**, Oxford University Press, v. 68, n. 1, p. 13–20, 1981. DOI:10.1093/biomet/68.1.13.

BAYER, F. M.; CRIBARI-NETO, F. Model selection criteria in beta regression with varying dispersion. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 46, n. 1, p. 729–746, 2017. DOI:10.1080/03610918.2014.977918.

BOURGUIGNON, M. et al. A new regression model for positive random variables with skewed and long tail. **Metron**, Springer, v. 79, p. 33–55, 2021. DOI:10.1007/s40300-021-00203-y.

CEPEDA-CUERVO, E. et al. On gamma regression residuals. **Journal of the Iranian Statistical Society**, v. 15, n. 1, p. 29–44, 2016.

COOK, R. D.; WEISBERG, S. **Residuals and Influence in Regression**. New York: Chapman and Hall, 1982.

COX, D. R.; HINKLEYR, D. V. **Theoretical Statistics**. London: Chapman and Hall, 1979.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and graphical statistics**, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996. DOI:10.2307/1390802.

ESPINHEIRA, P. L. et al. On beta regression residuals. **Journal of Applied Statistics**, Taylor & Francis, v. 35, n. 4, p. 407–419, 2008. DOI:10.1080/02664760701834931.

ESPINHEIRA, P. L. et al. On nonlinear beta regression residuals. **Biometrical Journal**, Wiley Online Library, v. 59, n. 3, p. 445–461, 2017. DOI:10.1002/bimj.201600136.

ESPINHEIRA, P. L. et al. Model selection criteria on beta regression for machine learning. **Machine Learning and Knowledge Extraction**, MDPI, v. 1, n. 1, p. 427–449, 2019. DOI:10.3390/make1010026.

FENG, C. et al. A comparison of residual diagnosis tools for diagnosing regression models for count data. **BMC Medical Research Methodology**, BioMed Central, v. 20, n. 1, p. 1–21, 2020. DOI:10.1186/s12874-020-01055-2.

FERRARI, S. L. P.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of applied statistics**, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004. DOI:10.1080/0266476042000214501.

FERRARI, S. L. P. et al. Diagnostic tools in beta regression with varying dispersion. **Statistica Neerlandica**, Wiley Online Library, v. 65, n. 3, p. 337–351, 2011. DOI:10.1111/j.1467-9574.2011.00488.x.

GIAMPAOLI, V. et al. Critérios de seleção de modelos para o modelo de regressão beta. **Revista Brasileira de Estatística**, v. 70, n. 232, p. 7–27, 2009.

GRIFFITHS, W. E. et al. **Learning and Practicing Econometrics**. New York: John Wiley & Sons, 1993.

GROSS, J.; LIGGES, U. **nortest: Tests for Normality**. [S.l.], 2015. R package version 1.0-4. Disponível em: <https://CRAN.R-project.org/package=nortest>.

HANNAN, E. J.; QUINN, B. G. The determination of the order of an autoregression. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 41, n. 2, p. 190–195, 1979. DOI:10.1111/j.2517-6161.1979.tb01072.x.

KEEPING, E. S. **Introduction to Statistical Inference**. Publisher. Princeton: D. Van Nostrand Company, 1962.

KUTNER, M. H. et al. **Applied Linear Statistical Models**. 4. ed. New York: WCB McGraw-Hill, 1996.

LEHMANN, E. L.; CASELLA, G. **Theory of Point Estimation**. 2. ed. New York: Springer Science & Business Media, 1998.

MADDALA, G. S. **Limited-Dependent and Qualitative Variables in Econometrics**. 3. ed. Cambridge: Cambridge University Press, 1983.

MARINHO, P. R. D. et al. **AdequacyModel: Adequacy of Probabilistic Models and General Purpose Optimization**. [S.l.], 2016. R package version 2.0.0. Disponível em: <https://CRAN.R-project.org/package=AdequacyModel>.

MCCULLAGH, P.; NELDER, J. A. **Generalized Linear Models**. 2. ed. London: Chapman and Hill, 1989.

MCDONALD, J. B. Some generalized functions for the size distribution of income. **Econometrica**, The Econometric Society, v. 52, n. 3, p. 647–665, 1984. DOI:10.2307/1913469.

MCDONALD, J. B.; BUTLER, R. J. Regression models for positive random variables. **Journal of Econometrics**, Elsevier, v. 43, n. 1-2, p. 227–251, 1990. DOI:10.1016/0304-4076(90)90118-D.

MCFADDEN, D. Conditional logit analysis of qualitative choice behavior. In: ZAREMBKA, P. (Ed.). **Fontiers in Econometrics**. New York: Academic press, 1974. p. 105–142.

MEDEIROS, R. M. R. de. **Um novo modelo de regressão para dados em Z**. 132 f. Dissertação (Mestrado em Matemática Aplicada e Estatística) — Universidade Federal do Rio Grande do Norte, Natal, 2019.

MEDIAVILLA, F. A. M. et al. A comparison of the coefficient of predictive power, the coefficient of determination and AIC for linear regression. **The Journal of Applied Business and Economics**, North American Business Press, v. 8, n. 4, p. 1–12, 2008.

MILLER-JÚNIOR, R. G. An Unbalanced Jackknife. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 2, n. 5, p. 880–891, 1974. DOI:10.1214/aos/1176342811.

MYERS, R. H. **Classical and Modern Regression with Applications**. 2. ed. California: Duxbury Press Belmont, 1990.

NAGELKERKE, N. J. D. A note on a general definition of the coefficient of determination. **Biometrika**, Oxford University Press, v. 78, n. 3, p. 691–692, 1991. DOI:10.1093/biomet/78.3.691.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society: Series A (General)**, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972. DOI:10.2307/2344614.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2022. Disponível em: <https://www.R-project.org/>.

RAZALI, N. M.; WAH, Y. B. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. **Journal of Statistical Modeling and Analytics**, v. 2, n. 1, p. 21–33, 2011.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 54, n. 3, p. 507–554, 2005. DOI:10.1111/j.1467-9876.2005.00510.x.

SÁNCHEZ, L. et al. Birnbaum-Saunders quantile regression and its diagnostics with application to economic data. **Applied Stochastic Models in Business and Industry**, Wiley Online Library, v. 37, n. 1, p. 53–73, 2021. DOI:10.1002/asmb.2556.

SANTOS, K. H. dos et al. A misspecification test for beta prime regression models. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 52, n. 8, p. 1–14, 2021. DOI:10.1080/03610918.2021.1971244.

SANTOS-NETO, M. et al. Reparameterized Birnbaum-Saunders regression models with varying precision. **Electronic Journal of Statistics**, Institute of Mathematical Statistics and Bernoulli Society, v. 10, n. 2, p. 2825–2855, 2016. DOI:10.1214/16-EJS1187.

SCHWARZ, G. Estimating the dimension of a model. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. DOI:10.1214/aos/1176344136.

SCUDILIO, J.; PEREIRA, G. H. A. Adjusted quantile residual for generalized linear models. **Computational Statistics**, Springer, v. 35, n. 1, p. 399–421, 2019. DOI:0.1007/s00180-019-00896-w.

STONE, M. Cross-validatory choice and assessment of statistical predictions. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 36, n. 2, p. 111–147, 1974. DOI:10.1111/j.2517-6161.1974.tb00994.x.

TULUPYEV, A. et al. Beta prime regression with application to risky behavior frequency screening. **Statistics in Medicine**, Wiley Online Library, v. 32, n. 23, p. 4044–4056, 2013. DOI:10.1002/sim.5820.

WEISBERG, S. **Applied Linear Regression**. Third. Nova Jersey: John Wiley & Sons, 2005. v. 528.

WIJEKULARATHNA, D. K. et al. Power analysis of several normality tests: A Monte Carlo simulation study. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 51, n. 3, p. 757–773, 2020. DOI:10.1080/03610918.2019.1658780.

WOLD, H. O. A. Soft modelling: The basic design and some extensions. **Systems Under Indirect Observation, Part II**, North Holland, v. 2, n. 1, p. 1–54, 1982.

YAP, B. W.; SIM, C. H. Comparisons of various types of normality tests. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 81, n. 12, p. 2141–2155, 2011. DOI:10.1080/00949655.2010.520163.

Apêndice A

Vetor Escore e Matriz da Informação de Fisher

A função escore é obtida através da diferenciação da função de log-verossimilhança (2.10) com relação a β_j e ν_k. Assim, os componentes da função escore para $\boldsymbol{\beta}$ e $\boldsymbol{\nu}$ são dados, respectivamente, por

$$U_{\beta_j} = \frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\nu})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i(\mu_i, \phi_i)}{\partial \mu_i} \frac{d\mu_i}{d\eta_{1i}} \frac{\partial \eta_{1i}}{\partial \beta_j}, \qquad j = 1, ..., p,$$
$$U_{\nu_j} = \frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\nu})}{\partial \nu_k} = \sum_{i=1}^{n} \frac{\partial \ell_i(\mu_i, \phi_i)}{\partial \phi_i} \frac{d\phi_i}{d\eta_{2i}} \frac{\partial \eta_{2i}}{\partial \nu_k}, \qquad k = 1, ..., q,$$

em que $d\mu_i/d\eta_{1i} = 1/g'(\mu_i)$, $\partial\eta_{1i}/\partial\beta_j = x_{ij}$, $d\phi_i/d\eta_{2i} = 1/h'(\phi_i)$ e $\partial\eta_{2i}/\partial\nu_k = z_{ik}$. Além disso, de (2.10) têm-se que

$$\frac{\partial \ell_i(\mu_i, \phi_i)}{\partial \mu_i} = (1+\phi_i)\left\{\log\left(\frac{y_i}{1+y_i}\right) - [\psi(\mu_i(1+\phi_i)) - \psi(\mu_i(1+\phi_i)+\phi_i+2)]\right\}, \tag{5.1}$$
$$\frac{\partial \ell_i(\mu_i, \phi_i)}{\partial \phi_i} = \log\left(\frac{y_i^{\mu_i}}{(1+y_i)^{1+\mu_i}}\right) - [\mu_i\psi(\mu_i(1+\phi_i)) + \psi(\phi_i+2) - (1+\mu_i)\psi(\mu_i(1+\phi_i)+\phi_i+2)].$$

Logo,

$$U_{\beta_j} = \sum_{i=1}^{n}(1+\phi_i)(y_i^* - \mu_i^*)\frac{1}{g'(\mu_i)}x_{ij},$$
$$U_{\nu_j} = \sum_{i=1}^{n}(y_i^\star - \mu_i^\star)\frac{1}{h'(\phi_i)}z_{ik},$$

com $y_i^* = \log\left(\frac{y_i}{1+y_i}\right)$, $\mu_i^* = [\psi(\mu_i(1+\phi_i)) - \psi(\mu_i(1+\phi_i)+\phi_i+2)]$, $y_i^\star = \log\left(\frac{y_i^{\mu_i}}{(1+y_i)^{1+\mu_i}}\right)$ e $\mu_i^\star = \mu_i\psi(\mu_i(1+\phi_i)) + \psi(\phi_i+2) - (1+\mu_i)\psi(\mu_i(1+\phi_i)+\phi_i+2)$.

Assim, a função escore para $\boldsymbol{\beta}$ e $\boldsymbol{\nu}$ na forma matricial é dada, respectivamente, por

$$\mathbf{U}_{\boldsymbol{\beta}} = \mathbf{X}^\top\boldsymbol{\Phi}\mathbf{D}_1(\boldsymbol{y}^* - \boldsymbol{\mu}^*),$$
$$\mathbf{U}_{\boldsymbol{\nu}} = \mathbf{Z}^\top\mathbf{D}_2(\boldsymbol{y}^\star - \boldsymbol{\mu}^\star),$$

em que $\mathbf{D}_1 = \text{diag}\{1/g'(\mu_1), \ldots, 1/g'(\mu_n)\}$, $\boldsymbol{\Phi} = \text{diag}\{(1+\phi_1), \ldots, (1+\phi_n)\}$, $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)^\top$, $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_n^*)^\top$, $\mathbf{D}_2 = \text{diag}\{1/h'(\phi_1), \ldots, 1/g'(\phi_n)\}$, $\boldsymbol{y}^\star = (y_1^\star, \ldots, y_n^\star)^\top$ e $\boldsymbol{\mu}^\star = (\mu_1^\star, \ldots, \mu_n^\star)^\top$.

A matriz da informação de Fisher conjunta para $\boldsymbol{\beta}$ e $\boldsymbol{\nu}$ é obtida calculando as derivadas de segunda ordem da função de log-verossimilhança (2.10). As derivadas em relação a β_j e β_l,

para $j, l = 1, \ldots, p$, são

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\beta}, \boldsymbol{\nu})}{\partial \beta_j \partial \beta_l} &= \sum_{i=1}^{n} \frac{\partial}{\partial \mu_i}\left[\frac{\partial \ell_i(\mu_i, \phi_i)}{\partial \mu_i}\frac{d\mu_i}{d\eta_{1i}}\frac{\partial \eta_{1i}}{\partial \beta_j}\right]\frac{d\mu_i}{d\eta_{1i}}\frac{\partial \eta_{1i}}{\partial \beta_l} \\ &= \sum_{i=1}^{n}\left[\frac{\partial^2 \ell_i(\mu_i, \phi_i)}{\partial \mu_i^2}\frac{d\mu_i}{d\eta_{1i}} + \frac{\partial \ell_i(\mu_i, \phi_i)}{\partial \mu_i}\frac{\partial}{\partial \mu_i}\left(\frac{d\mu_i}{d\eta_{1i}}\right)\right]\frac{d\mu_i}{d\eta_{1i}} x_{ij} x_{il}.\end{aligned}$$

Como $\mathbb{E}(y_i^*) = \mu_i^*$ (2.12), segue que $\mathbb{E}\left(\frac{\partial \ell_i(\mu_i, \phi_i)}{\partial \mu_i}\right) = 0$. Portanto, tomando o valor esperado

$$\mathbb{E}\left(\frac{\partial^2 \ell(\boldsymbol{\beta}, \boldsymbol{\nu})}{\partial \beta_j \partial \beta_l}\right) = \sum_{i=1}^{n} \mathbb{E}\left(\frac{\partial^2 \ell_i(\mu_i, \phi_i)}{\partial \mu_i^2}\right)\left(\frac{d\mu_i}{d\eta_{1i}}\right)^2 x_{ij} x_{il}.$$

De (5.1), tem-se que

$$\frac{\partial^2 \ell_i(\mu_i, \phi_i)}{\partial \mu_i^2} = -(1+\phi_i)^2[\psi'(\mu_i(1+\phi_i)) - \psi'(\mu_i(1+\phi_i) + \phi_i + 2)].$$

Então,

$$\mathbb{E}\left(\frac{\partial^2 \ell(\boldsymbol{\beta}, \boldsymbol{\nu})}{\partial \beta_j \partial \beta_l}\right) = -\sum_{i=1}^{n}(1+\phi_i) w_i x_{ij} x_{il},$$

com $w_i = (1+\phi_i)\upsilon_i[1/(g'(\mu_i))^2]$ sendo $\upsilon_i = \psi'(\mu_i(1+\phi_i)) - \psi'(\mu_i(1+\phi_i) + \phi_i + 2)$. Em forma matricial, tem-se

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

em que $\mathbf{W} = \text{diag}\{w_1, \ldots, w_n\}$.

As derivadas de segunda ordem da função de log-verossimilhança em relação a β_j e ν_k, para $j = 1, \ldots, p$ e $k = 1, \ldots, q$, são

$$\frac{\partial^2 \ell(\boldsymbol{\beta}, \boldsymbol{\nu})}{\partial \beta_j \partial \nu_k} = \sum_{i=1}^{n} \frac{\partial}{\partial \phi_i}\left[\frac{\partial \ell_i(\mu_i, \phi_i)}{\partial \mu_i}\frac{d\mu_i}{d\eta_{1i}}\frac{\partial \eta_{1i}}{\partial \beta_j}\right]\frac{d\phi_i}{d\eta_{2i}}\frac{\partial \eta_{2i}}{\partial \nu_k} = \sum_{i=1}^{n}\frac{\partial^2 \ell_i(\mu_i, \phi_i)}{\partial \mu_i \partial \phi_i}\frac{d\mu_i}{d\eta_{1i}}\frac{d\phi_i}{d\eta_{2i}} x_{ij} z_{ik}.$$

Tomando o valor esperado,

$$\mathbb{E}\left(\frac{\partial^2 \ell(\boldsymbol{\beta}, \boldsymbol{\nu})}{\partial \beta_j \partial \nu_k}\right) = \sum_{i=1}^{n}\mathbb{E}\left(\frac{\partial^2 \ell_i(\mu_i, \phi_i)}{\partial \mu_i \partial \phi_i}\right)\frac{d\mu_i}{d\eta_{1i}}\frac{d\phi_i}{d\eta_{2i}} x_{ij} z_{ik}.$$

De (5.1), obtém-se que

$$\begin{aligned}\frac{\partial^2 \ell_i(\mu_i, \phi_i)}{\partial \mu_i \partial \nu_k} &= (y_i^* - \mu_i^*) + (1+\phi_i)\psi'(\mu_i(1+\phi_i) + \phi_i + 2) \\ &\quad + \mu_i(1+\phi_i)[\psi'(\mu_i(1+\phi_i) + \phi_i + 2) - \psi'(\mu_i(1+\phi_i))].\end{aligned}$$

Como $\mathbb{E}(y_i^* - \mu_i^*) = 0$, então,

$$\mathbb{E}\left(\frac{\partial^2 \ell(\boldsymbol{\beta}, \boldsymbol{\nu})}{\partial \beta_j \partial \nu_k}\right) = -\sum_{i=1}^{n} d_i x_{ij} z_{ik}$$

com $d_i = -(1+\phi_i)\zeta_i[1/(g'(\mu_i))][1/(h'(\phi_i))]$ sendo $\zeta_i = \psi'(\mu_i(1+\phi_i)+\phi_i+2)+\mu_i[\psi'(\mu_i(1+\phi_i)+\phi_i+2)-\psi'(\mu_i(1+\phi_i))]$. Em forma matricial,

$$\mathbb{E}\left(\frac{\partial^2 \ell(\boldsymbol{\beta}, \boldsymbol{\nu})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\nu}^\top}\right) = -\mathbf{X}^\top \mathbf{D} \mathbf{Z}$$

em que $\mathbf{D} = \text{diag}\{d_1, \ldots, d_n\}$. Por fim, as derivadas em relação a ν_k e ν_m, para $k, m = 1, \ldots, q$, são

$$\frac{\partial^2 \ell(\boldsymbol{\beta}, \boldsymbol{\nu})}{\partial \nu_k \partial \nu_m} = \sum_{i=1}^{n} \left[\frac{\partial^2 \ell_i(\mu_i, \phi_i)}{\partial \phi_i^2}\frac{d\phi_i}{d\eta_{2i}} + \frac{\partial \ell_i(\mu_i, \phi_i)}{\partial \phi_i}\frac{\partial}{\partial \phi_i}\left(\frac{d\phi_i}{d\eta_{2i}}\right)\right]\frac{d\phi_i}{d\eta_{2i}} z_{ik} z_{im}.$$

Como $\mathbb{E}(y_i^\star) = \mu_i^\star$ (2.14), segue que $\mathbb{E}\left(\frac{\partial \ell_i(\mu_i, \phi_i)}{\partial \phi_i}\right) = 0$. Portanto,

$$\mathbb{E}\left(\frac{\partial^2 \ell(\boldsymbol{\beta}, \boldsymbol{\nu})}{\partial \nu_k \partial \nu_m}\right) = \sum_{i=1}^{n} \mathbb{E}\left(\frac{\partial^2 \ell_i(\mu_i, \phi_i)}{\partial \phi_i^2}\right)\left(\frac{d\phi_i}{d\eta_{2i}}\right)^2 z_{ik} z_{im}.$$

De (5.1), obtém-se que

$$\frac{\partial^2 \ell_i(\mu_i, \phi_i)}{\partial \phi_i^2} = -[\mu_i^2 \psi'(\mu_i(1+\phi_i)) + \psi'(\phi_i+2) - (1+\mu_i)^2 \psi'(\mu_i(1+\phi_i)+\phi_i+2)].$$

Então,

$$\mathbb{E}\left(\frac{\partial^2 \ell(\boldsymbol{\beta}, \boldsymbol{\nu})}{\partial \nu_k \partial \nu_m}\right) = -\sum_{i=1}^{n} e_i z_{ik} z_{im},$$

com $e_i = \xi_i[1/(h'(\phi_i))^2]$ sendo $\xi_i = \mu_i^2 \psi'(\mu_i(1+\phi_i)) + \psi'(\phi_i+2) - (1+\mu_i)^2 \psi'(\mu_i(1+\phi_i)+\phi_i+2)$. Em forma matricial,

$$\mathbb{E}\left(\frac{\partial^2 \ell(\boldsymbol{\beta}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu} \partial \boldsymbol{\nu}^\top}\right) = -\mathbf{Z}^\top \mathbf{E} \mathbf{Z},$$

em que $\mathbf{E} = \text{diag}\{e_1, \ldots, e_n\}$.

Finalmente, a matriz da informação de Fisher conjunta para $\boldsymbol{\beta}$ e $\boldsymbol{\nu}$, é dada por

$$\mathbf{K}(\boldsymbol{\beta}, \boldsymbol{\nu}) = \begin{bmatrix} \mathbf{K}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \mathbf{K}_{\boldsymbol{\beta}\boldsymbol{\nu}} \\ \mathbf{K}_{\boldsymbol{\nu}\boldsymbol{\beta}} & \mathbf{K}_{\boldsymbol{\nu}\boldsymbol{\nu}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\top \boldsymbol{\Phi} \mathbf{W} \mathbf{X} & \mathbf{X}^\top \mathbf{D} \mathbf{Z} \\ \mathbf{Z}^\top \mathbf{D} \mathbf{X} & \mathbf{Z}^\top \mathbf{E} \mathbf{Z} \end{bmatrix} \tag{5.2}$$

Apêndice B

Tabelas complementares da avaliação numérica dos resíduos

Tabela 5.1: Proporções de não rejeição da hipótese nula dos testes de SW para os resíduos r_i^Q, r_i^P, r_i^{P*}, r_i^β e $r_i^{\beta*}$ aos níveis de 1% e 10% de significância.

n	Resíduos	Cenário I		Cenário II		Cenário III		Cenário IV	
		1%	10%	1%	10%	1%	10%	1%	10%
20	r_i^Q	99,34	90,96	99,28	90,55	99,36	90,99	99,30	90,28
	r_i^P	74,53	43,76	95,48	76,87	78,49	48,88	96,05	79,28
	r_i^{P*}	75,79	45,13	95,72	77,35	79,54	50,25	96,09	79,56
	r_i^β	97,45	82,24	98,97	88,10	97,69	83,07	98,94	88,46
	$r_i^{\beta*}$	97,65	82,35	99,01	88,36	97,87	83,00	99,00	88,54
30	r_i^Q	99,32	90,69	99,39	91,33	99,32	90,63	99,27	91,04
	r_i^P	49,48	21,31	90,53	65,65	56,65	27,19	92,24	69,15
	r_i^{P*}	50,14	22,02	90,67	66,02	57,02	28,01	92,38	69,34
	r_i^β	93,01	70,76	98,10	86,39	94,63	74,43	98,57	87,32
	$r_i^{\beta*}$	93,14	71,12	98,12	86,38	94,80	74,81	98,67	87,33
60	r_i^Q	99,20	90,53	99,27	90,64	99,07	90,33	99,22	90,93
	r_i^P	15,80	3,91	76,10	45,10	22,50	6,55	80,74	50,40
	r_i^{P*}	16,07	4,06	76,40	45,35	22,69	6,60	80,88	50,50
	r_i^β	79,91	48,38	95,76	79,64	84,29	53,75	96,84	81,37
	$r_i^{\beta*}$	80,22	48,89	95,77	79,72	84,43	54,26	96,81	81,32
120	r_i^Q	99,19	91,25	98,92	90,20	99,00	90,19	99,21	90,57
	r_i^P	0,44	0,02	43,74	16,47	1,78	0,26	53,67	23,81
	r_i^{P*}	0,45	0,02	43,95	16,69	1,82	0,26	53,86	23,93
	r_i^β	44,43	15,74	89,30	64,96	54,83	22,81	91,95	69,36
	$r_i^{\beta*}$	44,75	15,95	89,34	65,13	55,24	23,05	91,96	69,48

Apêndice C

Implementação em R

```
####################################
###    Matriz H - Alavancagem    ####
####################################

influence.BP <- function(model)
{
  n <- model$N
  y <- model$y
  X <- model$mu.x
  beta <- model$mu.coefficients
  mu <- model$mu.fv
  phi <- model$sigma.fv
  linkmu <- model$mu.link
  yast <- log(y/(1+y))
  muast <- digamma(mu*(1+phi)) - digamma(mu*(1+phi)+phi+2)

  #Derivando a função de ligação
  if(linkmu=="log"){
    explink<-expression(log(mu))
    DlinkDmu<-D(explink,"mu")
  } else if(linkmu=="sqrt"){
    explink<-expression(sqrt(mu))
    DlinkDmu<-D(explink,"mu")
  }else{warning("Função link diferente")
  }

  Dlink <- (1/eval(DlinkDmu))
  DLINK <- diag(Dlink,ncol=n,nrow = n) # matrix d mu/d eta
  phii <- (1+phi)
  PHI <- diag(phii,ncol=n,nrow = n) # matrix PHI
  Vi <- trigamma(mu*(1+phi)) - trigamma(mu*(1+phi)+phi+2)
  wi <- (1+phi)*Vi*Dlink^2
  W <- diag(wi,ncol=n,nrow = n) # matrix W
  Inv <- solve(t(X)%*%PHI%*%W%*%X)

  Hat <- (PHI^(1/2))%*%(W^(1/2))%*%X%*%Inv%*%t(X)%*%(PHI^(1/2))%*%(W^(1/2))
  hat <- diag(Hat)

  z <- X%*%beta + solve(W)%*%DLINK%*%(yast-muast)
  y1 <- (PHI^(1/2))%*%(W^(1/2))%*%z
```

```
  Result <- list(hat=hat,H=Hat, ychap = y1, zi = z)
  Result
}

####################################
###   RESIDUALS BP REGRESSION    ####
####################################

residuals.BP <- function(model, type = c("quantile", "pearson", "pearson P",
                                         "sweighted1", "sweighted2"), ...)
{
  y <- model$y
  mu <- model$mu.fv
  phi <- model$sigma.fv
  yast <- log(y/(1+y))
  muast <- digamma(mu*(1+phi)) - digamma(mu*(1+phi)+phi+2)
  v <- trigamma(mu*(1+phi)) - trigamma(mu*(1+phi)+phi+2)
  hat <-influence.BP(model)$hat

  type <- match.arg(type)

  ## ajuste de erro
  wts <- weights(model)
  if(is.null(wts)) wts <- 1

  res <- switch(type,
                "quantile" = {
                  sqrt(wts) * qnorm(pBP(y,mu,phi))
                },
                "pearson" = {
                  sqrt(wts) * sqrt(phi)*(y-mu) / sqrt(mu*(1+mu))
                },
                "pearson P" = {
                  sqrt(wts) * sqrt(phi)*(y-mu) / sqrt(mu*(1+mu)*(1-hat))
                },
                "sweighted1" = {
                  sqrt(wts) * (yast-muast) / sqrt(v)
                },
                "sweighted2" = {
                  sqrt(wts) * (yast-muast) / sqrt(v*(1-hat))
                })
  return(res)
}

####################################
######R2 AND P2 - BP REGRESSION######
####################################
```

```
criterios <- function(model, model0)
{
  y <- model$y
  n <- model$N
  p <- length(model$mu.coefficients)
  q <- length(model$sigma.coefficients)
  r <- p + q
  mu <- model$mu.fv
  sigma <- model$sigma.fv
  eta <- model$mu.lp
  hat <- influence.BP(model)$hat
  ychap <- influence.BP(model)$ychap
  z <- influence.BP(model)$zi

  ############### PRESS ###############
  press <- function(res) sum((res/(1-hat))^2)

  press_quantile <- press(residuals.BP(model, type="quantile"))
  press_pearson <- press(residuals.BP(model, type="pearson"))
  press_pearson_p <- press(residuals.BP(model, type="pearson P"))
  press_sweighted1 <- press(residuals.BP(model, type="sweighted1"))
  press_sweighted2 <- press(residuals.BP(model, type="sweighted2"))

  ############### P2 ###############
  sst <- sum((ychap-mean(ychap))^2)

  p2_quantile <- 1 - press_quantile/((n/(n-r))^2*sst)
  p2_pearson <- 1 - press_pearson/((n/(n-r))^2*sst)
  p2_pearson_p <- 1 - press_pearson_p/((n/(n-r))^2*sst)
  p2_sweighted1 <- 1 - press_sweighted1/((n/(n-r))^2*sst)
  p2_sweighted2 <- 1 - press_sweighted2/((n/(n-r))^2*sst)

  ###### P2 corrigido - Espinheira et. al (2019) ######
  p2_quantile_c <- 1-(1-p2_quantile)*((n-1)/(n-r))
  p2_pearson_c <- 1-(1-p2_pearson)*((n-1)/(n-r))
  p2_pearson_p_c <- 1-(1-p2_pearson_p)*((n-1)/(n-r))
  p2_sweighted1_c <- 1-(1-p2_sweighted1)*((n-1)/(n-r))
  p2_sweighted2_c <- 1-(1-p2_sweighted2)*((n-1)/(n-r))

  ########### Função de log verosimilhança ##############
  log.like = function(y,mu,sigma) {
    a <- mu*(1+sigma)
    b <- 2 + sigma
    lmd.i = (a-1)*log(y) - (a+b)*log(1+y) - lbeta(a,b)
    log.theta = sum(lmd.i)
    return(log.theta)
```

```
  }

  ######### R2 Razão de verossimilhança #################
  mu0 <- model0$mu.fv
  sigma0 <- model0$sigma.fv
  L0 <- exp(log.like(y,mu0,sigma0)) #função verossimilhanca do modelo nulo
  L1 <- exp(log.like(y,mu,sigma)) #função verossimilhanca do modelo ajustado

  r2_RV <- 1-(L0/L1)^(2/n)

  ## R2 Razão de verossimilhança corrigido - Bayer e Cribari-Neto (2017) ##
  alpha <- 0.4
  delta <- 1

  r2_RV_c <- 1-(1-r2_RV)*((n-1)/(n-(1+alpha)*p-(1-alpha)*q))^delta

  ############## R2 Ferrari e Cribari-Neto (2004)  #################
  if(model$mu.link=="log"){
    g <- log(y)
  } else if(model$mu.link=="sqrt"){
    g <- sqrt(y)
  }else{warning("Função link diferente")
  }

  r2_FC <- cor(eta,g)^2

  ## R2 Ferrari e Cribari-Neto corrigido - Bayer e Cribari-Neto (2017) ##

  r2_FC_c <- 1-(1-r2_FC)*((n-1)/(n-r))

  ##########################################################################
  medidas = c(p2_quantile, p2_quantile_c, p2_pearson, p2_pearson_c,
              p2_pearson_p, p2_pearson_p_c, p2_sweighted1, p2_sweighted1_c
              p2_sweighted2, p2_sweighted2_c, r2_RV, r2_RV_c, r2_FC, r2_FC_c)
  ##########################################################################
}
```

Simulações de Monte Carlo - Resíduos

```
##########################################################################
###### SIMULAÇÂO MEDIDAS DESCRITIVAS RESÍDUOS - ESPECIFICAÇÃO CORRETA ######

rm(list=ls())
setwd("C:/Users/euedu/Dropbox/Critérios Seleção de Modelos Regressão BP/
      Simulação/simulação")
source("gamlss_BP.R")
source("glmBP.R")
```

```
source("Estimation.R")
source("Residual_H_Log_Like_BP.R")

# Pacotes
library(extraDistr)
library(gamlss)
library(nortest)
library(moments)

NREP <- 10000     # Replicas de Monte Carlo
n <- 20           # Tamanho da amostra (20, 30, 60, 120)

# Covariaveis geradas a partir da distribuiçao uniforme (0,1)
set.seed(2022)
x0 <- rep(1,n)  #Intercepto
x1 <- runif(n)
z0 <- rep(1,n)  #Intercepto
z1 <- runif(n)

# Matrizes de regressores
X <- cbind(x0,x1)
Z <- cbind(z0,z1)

# Matrizes para armazenar os resíduos e os p-valores da simulação
pvalue <- array(NA,dim=c(NREP,5,4))
contp1 <- array(0,dim=c(NREP,5,4))
contp5 <- array(0,dim=c(NREP,5,4))
contp10 <- array(0,dim=c(NREP,5,4))

quantile <- array(NA,dim=c(NREP,n,4))
pearson <- array(NA,dim=c(NREP,n,4))
pearson.p <- array(NA,dim=c(NREP,n,4))
sweighted1 <- array(NA,dim=c(NREP,n,4))
sweighted2 <- array(NA,dim=c(NREP,n,4))

# Matrizes com os valores reais dos parametros par = (beta0, beta1, nu0, nu1)
par <- matrix(0, 4, 4)

# 4 cenarios
par[1,] <- c(3.2, -5.3, 1, 2)
par[2,] <- c(3.2, -5.3, 3, 1.3)
par[3,] <- c(3, 2, 1, 2)
par[4,] <- c(3, 2, 3, 1.3)

time_i <- Sys.time()
for(k in 1:4){
  cont <- 0   # contagem de não convergência
```

```
# Preditor linear
eta1 <- X%*%par[k,1:2]
eta2 <-  Z%*%par[k,3:4]

# exp() -> usando link log
mu <- as.vector(exp(eta1))
phi <- as.vector(exp(eta2))

# Variancia da variavel resposta
Vmu <-  mu*(1+mu)
vary <-  Vmu/phi

cat("Cenário:", k, "\n")
cat("Summary mu\n"); print(summary(mu))
cat("Summary phi\n"); print(summary(phi))
cat("Summary variancia de y\n"); print(summary(vary))

# Simulacao de Monte Carlo

i <- 1
while(i <= NREP) {
  # Gerando a variavel resposta
  y=rBP(n,mu,phi)

  # estimando o modelo
  fit <- gamlss(y~x1, sigma.formula=~z1,
                family = BP(mu.link="log",sigma.link="log"), trace=F)

  # se convergir
  if(fit$converged == TRUE)
  {
    # Resíduos usando as funçõees em: Residual_H_Log_Like_BP.R
    quantile[i,,k] <- residuals.BP(fit, type = "quantile")
    pearson[i,,k] <- residuals.BP(fit, type = "pearson")
    pearson.p[i,,k] <- residuals.BP(fit, type = "pearson P")
    sweighted1[i,,k] <- residuals.BP(fit, type = "sweighted1")
    sweighted2[i,,k] <- residuals.BP(fit, type = "sweighted2")

    #Testes de normalidade de Shapiro-Wilk
    pvalue[i,1,k] <- shapiro.test(quantile[i,,k])$p.value
    pvalue[i,2,k] <- shapiro.test(pearson[i,,k])$p.value
    pvalue[i,3,k] <- shapiro.test(pearson.p[i,,k])$p.value
    pvalue[i,4,k] <- shapiro.test(sweighted1[i,,k])$p.value
    pvalue[i,5,k] <- shapiro.test(sweighted2[i,,k])$p.value

    if(pvalue[i,1,k]>0.01) contp1[i,1,k] <- 1
```

```
      if(pvalue[i,2,k]>0.01) contp1[i,2,k] <- 1
      if(pvalue[i,3,k]>0.01) contp1[i,3,k] <- 1
      if(pvalue[i,4,k]>0.01) contp1[i,4,k] <- 1
      if(pvalue[i,5,k]>0.01) contp1[i,5,k] <- 1

      if(pvalue[i,1,k]>0.05) contp5[i,1,k] <- 1
      if(pvalue[i,2,k]>0.05) contp5[i,2,k] <- 1
      if(pvalue[i,3,k]>0.05) contp5[i,3,k] <- 1
      if(pvalue[i,4,k]>0.05) contp5[i,4,k] <- 1
      if(pvalue[i,5,k]>0.05) contp5[i,5,k] <- 1

      if(pvalue[i,1,k]>0.10) contp10[i,1,k] <- 1
      if(pvalue[i,2,k]>0.10) contp10[i,2,k] <- 1
      if(pvalue[i,3,k]>0.10) contp10[i,3,k] <- 1
      if(pvalue[i,4,k]>0.10) contp10[i,4,k] <- 1
      if(pvalue[i,5,k]>0.10) contp10[i,5,k] <- 1

      i <- i + 1

    }else{ # se não convergir - refaça
      cont <- cont + 1
      print(c("Nao convergiu",i,cont))
      i <- i - 1
    }
  } # fim do laço de monte carlo
}

time_f <- Sys.time()
time_exe <- time_f - time_i
time_exe  # tempo total para todos os n: 01h38min

################ Medidas descritivas dos resíduos ##################

desc <- function(residuos){
  medidas = apply(residuos, 2,
                  function(x) c(mean(x), var(x), skewness(x), kurtosis(x)))
  result = c(mean(medidas[1,]), mean(medidas[2,]), mean(medidas[3,]),
             mean(medidas[4,]))
  result
}

medidas <- array(NA, dim=c(5,7,4))

for (l in 1:4) {
  medidas[,,l] <- cbind(rbind(desc(quantile[,,l]), desc(pearson[,,l]),
                              desc(pearson.p[,,l]), desc(sweighted1[,,l]),
                              desc(sweighted2[,,l])),
```

```
                        colSums(contp10[,,l])/NREP*100,
                        colSums(contp5[,,l])/NREP*100,
                        colSums(contp1[,,l])/NREP*100)
}

colnames(medidas)<- c("Média", "Variância", "Assimetria", "Curtose",
                      "SW 10%", "SW 5%", "SW 1%")
rownames(medidas)<- c("Quantilico", "Pearson", "Pearson P", "Ponderado",
                      "Ponderado P")

cat("Médias das estatítiticas descritivas dos resíduos para o cenário I \n")
round(medidas[,,1],4)
cat("Médias das estatítiticas descritivas dos resíduos para o cenário II \n")
round(medidas[,,2],4)
cat("Médias das estatítiticas descritivas dos resíduos para o cenário III \n")
round(medidas[,,3],4)
cat("Médias das estatítiticas descritivas dos resíduos para o cenário IV \n")
round(medidas[,,4],4)
```

Simulações de Monte Carlo - P^2 e R^2

```
###############################################################################
###### SIMULAÇÃO CRITERIOS DE SELEÇÃO DE MODELOS - Especificação correta ######

rm(list=ls())
setwd("C:/Users/euedu/Dropbox/Critérios Seleção de Modelos Regressão BP/
      Simulação/simulação")
source("gamlss_BP.R")
source("glmBP.R")
source("Estimation.R")
source("Residual_H_Log_Like_BP.R")
source("criterios.R")

# Pacotes
library(extraDistr)
library(gamlss)

NREP <- 10000     # Replicas de Monte Carlo
n <- 30           # Tamanho da amostra (30, 60, 120)

# Covariaveis geradas a partir da distribuiçao uniforme (0,1)
set.seed(2022)
x0 <- rep(1,n)  #Intercepto
x1 <- runif(n)
z0 <- rep(1,n)  #Intercepto
z1 <- runif(n)
```

```
# Matrizes de regressores
X <- cbind(x0,x1)
Z <- cbind(z0,z1)

# Matrizes para armazenar os critérios
crit <- array(0, dim=c(NREP,14,16))

# Matrizes com os valores reais dos parametros par = (beta0, beta1, nu0, nu1)
par <- matrix(0, 16, 4)

# 4 cenarios para g(mu)= sqrt e h(phi)= sqrt
par[1,] <- c(5, -2.63, 2.3, 2.8)
par[2,] <- c(5, -2.63, 5, 4)
par[3,] <- c(5, 7.24, 2.3, 2.8)
par[4,] <- c(5, 7.24, 5, 4)
# 4 cenarios para g(mu)= sqrt e h(phi)= log
par[5,] <- c(5, -2.63, 1.71, 1.55)
par[6,] <- c(5, -2.63, 3.23, 1.16)
par[7,] <- c(5, 7.24, 1.71, 1.55)
par[8,] <- c(5, 7.24, 3.23, 1.16)
# 4 cenarios para g(mu)= log e h(phi)= sqrt
par[9,] <- c(3.22, -1.5, 2.3, 2.8)
par[10,] <- c(3.22, -1.5, 5, 4)
par[11,] <- c(3.22, 1.79, 2.3, 2.8)
par[12,] <- c(3.22, 1.79, 5, 4)
# 4 cenarios para g(mu)= log e h(phi)= log
par[13,] <- c(3.22, -1.5, 1.71, 1.55)
par[14,] <- c(3.22, -1.5, 3.23, 1.16)
par[15,] <- c(3.22, 1.79, 1.71, 1.55)
par[16,] <- c(3.22, 1.79, 3.23, 1.16)

time_i <- Sys.time()
for (k in 1:16) {
  cont <- 0         # contagem de não convergência

  #Preditor linear
  eta1 <- X%*%par[k,1:2]
  eta2 <- Z%*%par[k,3:4]

  if(k >= 1 && k <= 4){ # usando g(mu)= sqrt e h(phi)= sqrt
    mu <- as.vector((eta1)^2)
    phi <- as.vector((eta2)^2)
  }

  if(k >= 5 && k <= 8){ # usando g(mu)= sqrt e h(phi)= log
    mu <- as.vector((eta1)^2)
    phi <- as.vector(exp(eta2))
```

```
}

if(k >= 9 && k <= 12){ # usando g(mu)= log e h(phi)= sqrt
  mu <- as.vector(exp(eta1))
  phi <- as.vector((eta2)^2)
}

if(k >= 13 && k <= 16){ # usando g(mu)= log e h(phi)= log
  mu <- as.vector(exp(eta1))
  phi <- as.vector(exp(eta2))
}

# Variancia da variavel resposta
Vmu <- mu*(1+mu)
vary <- Vmu/phi

cat("Cenário:", k, "\n")
cat("Summary mu\n");  print(summary(mu))
cat("Summary phi\n"); print(summary(phi))
cat("Summary variancia de y\n"); print(summary(vary))

# Simulacao de Monte Carlo

i <- 1
while(i <= NREP)
{
  # Gerando a variavel resposta
  y=rBP(n,mu,phi)

  if(k >= 1 && k <= 4){ # usando funcao de ligacao = sqrt e sqrt
    # modelo nulo
    fit0 <- gamlss(y~1, sigma.formula=~1,
                   family = BP(mu.link="sqrt",sigma.link="sqrt"), trace=F)
    fit1 <- gamlss(y~x1, sigma.formula=~z1,
                   family = BP(mu.link="sqrt",sigma.link="sqrt"), trace=F)
  }

  if(k >= 5 && k <= 8){ # usando funcao de ligacao = sqrt e log
    # modelo nulo
    fit0 <- gamlss(y~1, sigma.formula=~1,
                   family = BP(mu.link="sqrt",sigma.link="log"), trace=F)
    fit1 <- gamlss(y~x1, sigma.formula=~z1,
                   family = BP(mu.link="sqrt",sigma.link="log"), trace=F)
  }

  if(k >= 9 && k <= 12){ # usando funcao de ligacao = log e sqrt
    # modelo nulo
```

```
      fit0 <- gamlss(y~1, sigma.formula=~1,
                     family = BP(mu.link="log",sigma.link="sqrt"), trace=F)
      fit1 <- gamlss(y~x1, sigma.formula=~z1,
                     family = BP(mu.link="log",sigma.link="sqrt"), trace=F)
    }

    if(k >= 13 && k <= 16){ # usando funcao de ligacao = log e log
      # modelo nulo
      fit0 <- gamlss(y~1, sigma.formula=~1,
                     family = BP(mu.link="log",sigma.link="log"), trace=F)
      fit1 <- gamlss(y~x1, sigma.formula=~z1,
                     family = BP(mu.link="log",sigma.link="log"), trace=F)
    }

    # se convergir
    if(fit1$converged == TRUE && fit0$converged == TRUE){
      crit[i,,k] <- criterios(fit1,fit0)
      i <- i + 1
    }else{
      # se não convergir - refaça
      cont <- cont + 1
      print(c("Nao convergiu",i,cont))
      i <- i - 1
    }
  }
} # Fim da simulação de Monte Carlo

time_f <- Sys.time()
time_exe <- time_f - time_i
time_exe   #  tempo total para todos os n: 05h46min

mean_criterios <- colMeans(crit[,,])
row.names(mean_criterios) <- c("P2Q", "P2Qc", "P2P", "P2Pc", "P2Pp", "P2Ppc",
                               "P2B", "P2Bc", "P2Bp", "P2Bc", "R2RV", "R2RVc",
                               "R2FC", "R2FCc")
colnames(mean_criterios) <- rep(c("C1", "C2", "C3", "C4"), 4)

cat("Médias dos critérios para os 4 cenarios usando link sqrt e sqrt \n")
mean_criterios[,1:4]
cat("Médias dos critérios para os 4 cenarios usando link sqrt e log \n")
mean_criterios[,5:8]
cat("Médias dos critérios para os 4 cenarios usando link log e sqrt \n")
mean_criterios[,9:12]
cat("Médias dos critérios para os 4 cenarios usando link log e log \n")
mean_criterios[,13:16]
```