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

Fernanda Clotilde da Silva

Um Critério de Seleção Para Modelos Beta Baseado no Trade-off Predição e Variabilidade

Recife
2022

Fernanda Clotilde da Silva

Um Critério de Seleção Para Modelos Beta Baseado no Trade-off Predição e Variabilidade

Tese apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco como requisito parcial à obtenção do título de Doutor em estatística.
Área de Concentração: Estatística Aplicada.

Orientadora: Profª. Drª. Patrícia Leone Espinheira Ospina

Recife
2022

Catalogação na fonte
Bibliotecária Nataly Soares Leite Moro, CRB4-1722

S586c Silva, Fernanda Clotilde da
Um critério de seleção para modelos beta baseado no trade-off predição e variabilidade / Fernanda Clotilde da Silva. – 2022.
115 f.: il., fig., tab.

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

1. Estatística Aplicada. 2. Overfitting. 3. Predição. 4. Variabilidade. 5. Regressão beta I. Ospina, Patrícia Leone Espinheira (orientadora). II. Título.

310 CDD (23. ed.) UFPE- CCEN 2022 - 42

FERNANDA CLOTILDE DA SILVA

UM CRITÉRIO DE SELEÇÃO PARA MODELOS BETA BASEADO NO TRADE-OFF PREDIÇÃO E VARIABILIDADE

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

Aprovada em: 16 de fevereiro de 2022.

BANCA EXAMINADORA

Prof. Dr.Francisco Cribari Neto
UFPE

Prof.Dr. Michel Helcias Montoril
UFSCAR

Profª Drª. Patrícia Leone Espinheira Ospina
UFPE

Prof. Dr. Rafael Izbicki
UFSCAR

Prof. Dr. Raydonal Ospina Martínez
UFPE

Dedico esse trabalho aos meus pais, Berenice Clotilde e Edenilson Ferreira, por serem a minha maior motivação, e ao meu esposo, Fábio, por estar sempre ao meu lado.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por ter me sustentado até aqui, e ter me capacitado para realizar esse projeto que tem sido o doutorado, sei que foi um plano DEle e não meu, apenas fui um instrumento em suas mãos.

Agradeço a minha orientadora Patrícia, pela maravilhosa orientação ao longo desse trabalho, e também por ter me abraçado e acolhido desde o início da minha pós-graduação, compreendendo as minhas limitações, e com muita paciência e carinho me ajudado a prosseguir. A senhora é uma inspiração para mim, não apenas como profissional, mais também como pessoa. Foi um privilégio ser sua aluna.

Sou muito grata aos meus pais, por sempre estarem comigo, me apoiando, incentivando e aconselhando, vocês são o maior presente e motivação de minha vida.

Agradeço também ao meu esposo, por compartilhar comigo momentos tão importantes em minha vida, e por todo o seu amor e companheirismo, sempre me apoiando, ajudando e incentivando. Posso garantir que ao seu lado a bagagem tornou-se mais leve.

Também sinto-me grata a toda minha família, em especial: às minhas irmãs, Dayana e Clarinha, que sempre alegram a minha vida, e ao meu avô, o Sr. Nei, por seu carinho e amabilidade.

Aos meus tios, Michelli e Horácio, responsáveis pelo início de minha trajetória na área acadêmica, me oferecendo condições para cursar uma graduação, e sempre me ajudando em tudo, jamais esquecerei o que fizeram por mim.

Aos meus tios de Recife, Preta e Nildo, os quais me acolheram todas as vezes que precisei, durante o início do mestrado e do doutorado.

Agradeço às minhas amigas, Terezinha, Verinha e Wanessa, estão sempre em meu coração, apesar da distância. Aos amigos que fiz no doutorado, Ioneris, Ana Cristina, César Diogo, Jonas, Pedro e Jodavid, acredito no futuro brilhante de cada um.

A todos os professores do Programa de Pós-graduação em Estatística da UFPE, por contribuírem para minha formação acadêmica, e por serem profissionais excepcionais.

A todos os funcionários do Departamento de Estatística, pela dedicação, carinho e paciência para com todos os alunos da Pós-graduação em Estatística.

Aos professores da Banca de Avaliações, pelas sugestões e contribuições, desde o exame de qualificação de nossa pesquisa.

À CAPES, pelo apoio financeiro.

"Para que todos vejam, e saibam, e considerem, e juntamente entendam que a mão do Senhor fez isso, e o Santo de Israel o criou." (Is 41, 20)

RESUMO

Muitas vezes surge a necessidade de estudar dados cujos valores pertencem ao intervalo $(0, 1)$, nessas situações podemos optar pelo uso do modelo de regressão beta, proposto por Ferrari e Cribari-Neto (2004), que se baseia em supor que a variável resposta segue uma distribuição beta, sob uma nova parametrização. Alguns métodos de análise de diagnóstico foram desenvolvidos para essa classe de modelos, buscando verificar a adequabilidade do ajuste, identificando possíveis afastamentos das suposições feitas. Entretanto, o uso desses métodos geralmente segue após a escolha de um conjunto de variáveis explicativas relevantes para o modelo, esse procedimento é conhecido como seleção de modelos, e algumas medidas usadas como critérios de seleção têm sido desenvolvidas. Dentre elas destacam-se os pseudos R^2, que visam avaliar a proporção de variação da resposta explicada pelo modelo ajustado, essas medidas foram estudadas e implementadas por Bayer e Cribari-Neto (2017) para a classe de modelos beta. Além dessas quantidades, também dispomos do critério de seleção P^2, que busca avaliar a habilidade do modelo em predizer valores consistentes da variável resposta, com base na estatística PRESS (*Predictive Residual Sum of Squares*), proposta por Allen (1971) para o modelo normal linear, e introduzida aos modelos beta por Espinheira et al. (2019). Uma vez que a definição desses critérios baseia-se no poder de explicação da variabilidade ou no poder de predição, esse trabalho tem como objetivo propor um processo de seleção para a classe de modelos beta considerando ambos os interesses, isto é, apresentar uma medida capaz de indicar modelos que consigam explicar bem a variabilidade da resposta, logo apresentam bons ajustes, e também consigam predizer bons valores. Esse processo consiste em determinar uma constante chamada "$\hat{\alpha}$", obtido com base no método *bootstrap* paramétrico, usada para construir uma nova estatística chamada "*BV*" (*Bias and Variability*), que diz respeito ao balanceamento viés e variância. Além disso, $\hat{\alpha}$ muitas vezes consiste em um bom indicador do viés da estimativa para a precisão. Dessa forma, avaliamos o desempenho das nossas medidas por meio de estudos de simulações de Monte Carlo, e aplicamos alguns bancos de dados reais, comprovando na prática a eficácia dessas estatísticas. Notamos que a avaliação conjunta das estatística "*BV*" e "$\hat{\alpha}$" é o que determina o processo de seleção que estamos propondo.

Palavras-chave: *overfitting*; predição; variabilidade; regressão beta.

ABSTRACT

Many times the need arises to study data whose values belong to the interval $(0,1)$, in these situations we can opt to use the beta regression model, proposed by Ferrari and Cribari-Neto (2004), which is based on assuming that the response variable follows a beta distribution, under a new parameterization. Some diagnostic analysis methods have been developed for this class of models, seeking to verify the adequacy of the fit, identifying possible deviations from the assumptions made. However, the use of these methods usually follows after the choice of a set of explanatory variables relevant to the model, this procedure is known as model selection, and some measures used as selection criteria have been developed. Among them we highlight the pseudo R^2, which aims to evaluate the proportion of variation of the response explained by the fitted model, these measures have been studied and implemented by Bayer and Cribari-Neto (2017) for the class of beta models. In addition these quantities, we also have the selection criterion P^2, which seeks to evaluate the ability of the model to predict consistent values of the response variable, based on the PRESS statistic (Predictive Residual Sum of Squares), proposed by Allen (1971) for the linear normal model, and introduced to beta models by Espinheira et al. (2019). Since the definition of these criteria is based on the power to explain the variability or in the power of prediction, this work aims to propose a selection process for the class of beta models considering both interests, that is, to present a measure capable of indicating models that can explain well the variability of the response, therefore present good fits, and also manage to predict good values. This process consists in determining a constant called "$\hat{\alpha}$", obtained from the parametric bootstrap method, used to construct a new statistic called "*BV*" (Bias and Variability), which is concerned with balancing bias and variance. In addition, "$\hat{\alpha}$" is often a good indicator of the bias of the estimate for precision. Thus, we evaluate the performance of our measures through Monte Carlo simulation studies, and apply some real databases, proving in practice the effectiveness of these statistics. We note that the joint evaluation of the "*BV*" and "$\hat{\alpha}$" statistics is what determines the selection process we are proposing.

Keywords: overfitting; prediction; variability; beta regression.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Geralmente, nos estudos de análise de dados surge o interesse de avaliar possíveis relações entre variáveis, de modo que muitas vezes essa relação é representada por meio de modelos matemáticos, que associam um conjunto de variáveis (chamadas de variáveis explicativas ou covariáveis) a uma única variável (chamada de variável resposta ou dependente). Esses modelos são conhecidos como "modelos de regressão", podendo ser utilizados para representar relações lineares ou não lineares. Por muito tempo, o modelo de regressão linear foi utilizado como único método de associar a média da variável resposta com uma função linear das variáveis explicativas (essa função é dita preditor linear). De maneira que, cada valor da resposta está associado a uma função linear dos respectivos valores das variáveis explicativas, juntamente a um erro não observável, e além disso, é suposto que esse erro possui uma distribuição normal, de média zero e variância constante. Dessa forma, sob tais suposições, tem-se que a variável resposta segue uma distribuição normal, sendo então, definido o modelo de regressão normal linear.

Entretanto, existem casos em que os valores da variável resposta são restritos ao intervalo $(0, 1)$, por exemplo, quando desejamos modelar taxas ou proporções. Nesses casos, o uso do modelo de regressão normal linear pode fornecer valores ajustados que excedem os limites do intervalo $(0, 1)$. Uma alternativa é transformar a variável resposta, de modo que ela assuma valores reais e a sua média possa ser modelada. No entanto, esse procedimento apresenta algumas desvantagens, como por exemplo, pode não ser tão fácil interpretar os parâmetros do modelo em termos da resposta original, além disso, medidas de proporções geralmente exibem assimetria, logo, uma análise baseada no pressuposto de normalidade pode fornecer resultados não confiáveis. Sob esse contexto, Ferrari e Cribari-Neto (2004) desenvolveram o modelo de regressão beta, que se baseia na suposição de que a variável reposta segue distribuição beta, com uma nova parametrização, que indexa a média da resposta e um parâmetro de precisão. Dessa forma, supondo que a média está relacionada a uma estrutura de regressão linear, tem-se que os parâmetros do modelo podem ser interpretados em termos da média da variável resposta, por meio de uma função de ligação, semelhante aos modelos lineares generalizados (MLGs), apresentados por Nelder e Wedderburm (1972).

O uso do modelo de regressão beta tem crescido consideravelmente nos últimos anos, e muitos trabalhos propondo algumas extensões de tal modelo têm sido desenvolvidos. Paolino (2001) e Smithson e Verkuilen (2006) apresentam esse modelo de regressão, cuja média e precisão são modeladas simultaneamente, utilizando máxima verossimilhança para estimar os parâmetros de ambos os modelos. Simas et al. (2010) abordam o modelo de regressão beta considerando uma estrutura de regressão não linear tanto para o modelo da média como para o da precisão. Ospina e Ferrari (2012) propuseram uma classe de modelos beta inflacionados em zero ou um, com base nas distribuições de mistura entre uma distribuição beta e uma distribuição de Bernoulli degenerada em zero e/ou um. Pereira et al. (2012) abordam a classe de modelos de regressão beta inflacionados truncados. Carrasco et al. (2014) propõem uma classe de modelos beta considerando erros nas variáveis.

Um procedimento bastante importante realizado após ajustar um modelo de regressão a um determinado conjunto de dados é a análise de diagnóstico, que busca avaliar a adequabilidade do modelo ajustado, verificando possíveis afastamentos das suposições feitas para o modelo. Encontramos na literatura, diversas técnicas de diagnósticos, propostas inicialmente para o modelo normal linear clássico e expandidas para demais classes de modelos. Muitas dessas técnicas são desenvolvidas com base na identificação de observações com comportamentos atípicos das demais. Cox e Snell (1968) propuseram a princípio identificar esse tipo de observação realizando análises de resíduos, uma vez que observações com altos valores de resíduos (chamadas de pontos aberrates ou *outlier*) podem afetar a média dos valores preditos da resposta e também as estimativas dos parâmetros do modelo. Algumas padronizações dos resíduos foram realizadas por Belsley et al. (1980) e por Cook e Weisberg (1982) para o modelo de regressão normal linear, enquanto uma nova padronização foi sugerida por Pregibon (1981) para os modelos lineares generalizados.

Além disso, outra forma de identificar observações com comportamentos atípicos é por meio da matriz de projeção, apresentada por Hoaglin e Welsch (1978). Nesse caso, as observações indicadas são chamadas de alavancas, e com base em suas localizações podem exercerem pesos desproporcionais nos valores ajustados e também nas estimativas dos parâmetros, quando isso ocorre, dizemos tratar-se também de um ponto influente, uma vez que a sua inclusão ou retirada influencia fortemente nos valores estimados. Existem outras medidas capazes de identificar pontos influentes, a mais conhecida e utilizada é a distância de Cook, proposta por Cook (1977) para o modelo normal linear e estendida para diversas classes de modelos. Um dos métodos mais modernos de análise de diagnóstico é o método de influência local, que busca avaliar o comportamento de uma determinada medida de influência, quando são inseridas leves perturbações aos dados ou ao modelo, dessa forma, é possível identificar a presença de observações que sob pequenas alterações, modificam consideravelmente os resultados do ajuste, para mais detalhes ver Paula (2013).

No que se refere às técnicas de diagnósticos tradicionais, dispomos ainda de um gráfico de análise bastante importante e muito utilizado, desenvolvido por Atkinson (1981) para o modelo normal linear, esse gráfico é comumente conhecido como "gráfico de envelope", sendo utilizado para detectar afastamentos da distribuição postulada para a variável resposta. A sua metodologia está baseada na definição de bandas de confiança para os resíduos, obtidas via simulação de Monte Carlo, de modo que, se a distribuição assumida não for adequada, é esperado que os pontos ou parte deles estejam fora dessas bandas de confiança, a metodologia desse gráfico foi estendida para diversos outros modelos. Todos esses métodos citados também foram estendidos para avaliar a adequabilidade do modelo de regressão beta, e ademais, diversos outros estudos têm sido realizados nos últimos anos especificamente para esse modelo. Smithson e Verkuilen (2006) apresentam testes de hipóteses desenvolvidos para os modelos da média e da precisão, para o modelo de regressão beta. Ospina et al. (2006) avaliam melhores métodos de estimação, tanto pontual como intervalar, para os parâmetros do modelo beta. Espinheira (2007) desenvolve diversos aspectos de inferência e análise para essa classe de modelos, sob cenários em que o parâmetro de precisão é fixo e também variável, propondo assim, diversas medidas de

grande valia na validação do modelo.

Geralmente, a análise de diagnóstico procede a escolha de um conjunto de covariáveis consideradas relevantes, dentre um conjunto mais amplo de regressores, essa etapa é chamada de seleção de modelos. Diversos métodos de seleção de modelos têm sido propostos nos últimos anos para o modelo linear clássico, os mais comumente utilizados são *forward*, *backward* e *stepwise*, para saber mais sobre esses métodos ver Draper e Smith (1981). Também são bastante conhecidos o critério de informação de Akaike (AIC) e o critério bayesiano de Schwarz (BIC), introduzidos por Akaike (1973) e Schwarz (1978), respectivamente. Além desses métodos é costume verificar a soma de quadrados de resíduos e algumas funções dessa quantidade, tais como os coeficientes R^2 e o R^2 ajustado, em que ambas medidas avaliam o quanto de variação da resposta pode se explicada pelo modelo ajustado. Buscando generalizar o coeficiente R^2 para diversas outras classes de modelos de regressão, alguns autores apresentam diferentes formulações para essa medida (geralmente chamadas de pseudos R^2), como Nagelkerke (1991), que se baseia na razão de verossimilhanças. No caso dos modelos de regressão beta também foram propostos alguns critérios de seleção, a princípio, Ferrari e Cribari-Neto (2004) utilizam uma nova formulação para o coeficiente R^2, com base na correlação entre medidas calculadas com os valores observados da resposta e com os valores preditos dessa variável, Bayer e Cribari-Neto (2017) apresentam para essa classe de modelos os critérios AIC, BIC e alguns pseudos R^2.

Outra estatística utilizada como critério para selecionar modelos é a PRESS (Predictive Residual Sum of Squares), proposta por Allen (1971) para os modelos lineares. Geralmente, essa estatística é aplicada para identificar o modelo com maior poder de predição da variável resposta. De forma análoga à definição do coeficiente R^2 baseado na soma de quadrados de resíduos, Mediavilla e Shah (2008) propõem a estatística P^2, cuja medida é utilizada para avaliar o poder de predição do modelo ajustado. Dessa forma, Espinheira et al. (2019) estendem as estatísticas PRESS e P^2 para a classe de modelos beta. Além disso, de modo similar a Bayer e Cribari-Neto (2017), eles também propõem uma versão corrigida de P^2.

É notável que a definição de alguns critérios de seleção é baseada em um determinado fator de interesse, seja no poder de explicação da variabilidade da resposta, ou no poder de predição dos valores desta. Neste trabalho, temos como objetivo principal, propor uma medida para selecionar modelos de regressão beta, considerando ambos os fatores, ou seja, buscamos apresentar um critério de balanceamento viés-variância, que chamamos de *BV* (Bias and Variability). Para isso, propomos uma constante α para calcular a combinação linear entre critérios dos tipos R^2 e P^2. Decidimos utilizar a metodologia *bootstrap*, introduzida por Efron (1979), para estimar α. O comportamento das estatísticas *BV* e $\hat{\alpha}$ foram avaliadas por meio de estudos de simulação. Estes estudos revelaram que tanto má especificação, quanto especificações corretas, foram identificadas a partir da avaliação conjunta de *BV* e $\hat{\alpha}$. Neste sentido, aqui estamos propondo um esquema de seleção de modelos. As aplicações a dados reais demonstram que o método proposto fornece ferramentas bastante úteis e eficazes na escolha de bons ajustes para a classe de modelos de regressão beta.

1.1 ORGANIZAÇÃO DA TESE

O presente trabalho, encontra-se dividido em sete capítulos. Nesse primeiro capítulo temos uma breve introdução a respeito de nossa motivação e os principais materiais de suporte computacional utilizados no estudo. No segundo capítulo introduzimos a distribuição beta e algumas propriedades, seguido pelo modelo de regressão beta. No segundo capítulo, temos a função escore, a matriz de informação de Fisher e os estimadores para os parâmetros do modelo. Também encontramos uma breve abordagem da obtenção dos resíduos ponderados padronizados para o modelo beta.

No terceiro capítulo, apresentamos algumas das principais medidas usadas na seleção de modelos beta, bem como os critérios aplicados para a análise da qualidade do ajuste, como também os critérios usados para a análise do poder de predição. Além disso, apresentamos um novo critério de seleção para esse modelo, tomando como base uma média ponderada das medidas tradicionais de adequabilidade para o modelo de regressão beta. Prosseguindo, apresentamos algumas características observadas para o novo critério de seleção. Também introduzimos uma metodologia para a obtenção dos pesos que compõem a nova medida.

No quarto capítulo, avaliamos por meio de estudos de simulação de Monte Carlo o desempenho dos critérios de seleção propostos, considerando o modelo de regressão beta linear e não linear, com dispersão variável. Para esses estudos, abordamos situações em que o modelo beta está corretamente especificado e também mal especificado. Ademais, comparamos os desempenhos das medidas propostas com as demais medidas de adequabilidade. Além disso, tanto para as situações com o modelo de regressão beta linear, como para o modelo beta não linear, foram realizadas breves análises dos valores obtidos para a estatística proposta como pesos no cálculo do novo critério de seleção de modelos. No quinto capítulo, são apresentados os estudos de simulação realizados em modelos de regressão beta linear com dispersão fixa.

No sexto capítulo, aplicamos o modelo de regressão beta em alguns conjuntos de dados reais e verificamos a adequação do ajuste por meio das medidas de adequabilidade já existentes e também pelo novo critério de balanceamento, junto com a medida estimada para os pesos. Também comparamos os resultados obtidos por essas estatísticas com os resultados segundo outros métodos de diagnósticos, bastante conhecidos na literatura. Por último, no sétimo capítulo, são apresentadas as conclusões gerais, obtidas ao longo desse trabalho.

1.2 SUPORTE COMPUTACIONAL

Todos os resultados numéricos apresentados no quarto, quinto e sexto capítulos, alcançados pelos estudos de simulações de Monte Carlo e aplicações a dados reais, foram obtidos por meio da linguagem de programação `Ox`, na sua versão 9.0, utilizando o sistema operacional *Windows*. Esta linguagem de programação foi desenvolvida por Jurgen Doornik em 1994, e encontra-se disponível em http://www.doornik.com. Para mais detalhes, ver Doornik (2001).

Todos os resultados gráficos, foram produzidos utilizando o ambiente computacional e estatístico R, na sua versão 4.1.2, para o sistema operacional *Windows*. O R foi criado por Ross

Ihaka e Robert Gentleman em 1996, tratando-se de uma implementação da linguagem S, porém com a vantagem de ser livre e possuir código-fonte aberto. Ele encontra-se disponível gratuitamente em http://www.r-project.org. Para mais detalhes, ver Peternelli e Mello (2011).

A parte escrita do presente trabalho, foi desenvolvida pelo uso do sistema de tipografia LaTeX, criado por Leslie Lamport em 1985. Trata-se de um conjunto de macros ou rotinas TeX. O TeX por sua vez é um sistema de editor de textos, usado para produzir livros, artigos, etc., com alta qualidade, desenvolvido por Donald E. Knuth. Para mais detalhes sobre o sistema de tipografia LaTeX, ver Knuth (1986) ou Lamport (1994).

2 MODELO DE REGRESSÃO BETA

2.1 DISTRIBUIÇÃO BETA

Comumente, muitos estudos nas mais diversas áreas buscam modelar dados na forma de taxas, frações ou proporções, ou seja, dados cujos valores são distribuídos restritamente no intervalo contínuo $(0,1)$. Nesse caso, a distribuição beta tem sido uma das principais distribuições contínuas utilizadas para modelar esses tipos de dados. Dizemos que uma variável aleatória y segue uma distribuição beta com parâmetros p e q se sua função densidade de probabilidade for dada por

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad 0<y<1, \tag{2.1}$$

em que $p>0$, $q>0$ e $\Gamma(.)$ é a função gama, definida por: $\Gamma(z)=\int_0^\infty x^{z-1}e^{-x}dx$. Neste caso, a média e a variância de y são dadas respectivamente por

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

Buscando desenvolver um modelo de regressão que permita modelar a média de uma variável resposta que assuma valores no intervalo $(0,1)$ e ainda envolver um parâmetro de precisão (ou dispersão), Ferrari e Cribari-Neto (2004) propuseram uma reparametrização para a distribuição beta, de modo que

$$\mu=\frac{p}{p+q} \qquad \text{e} \qquad \phi=p+q,$$

então,

$$p=\mu\phi \qquad \text{e} \qquad q=(1-\mu)\phi.$$

Dessa forma, a função densidade de probabilidade (2.1) pode ser reescrita em termos de μ e ϕ. Obtemos então

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0<y<1, \tag{2.2}$$

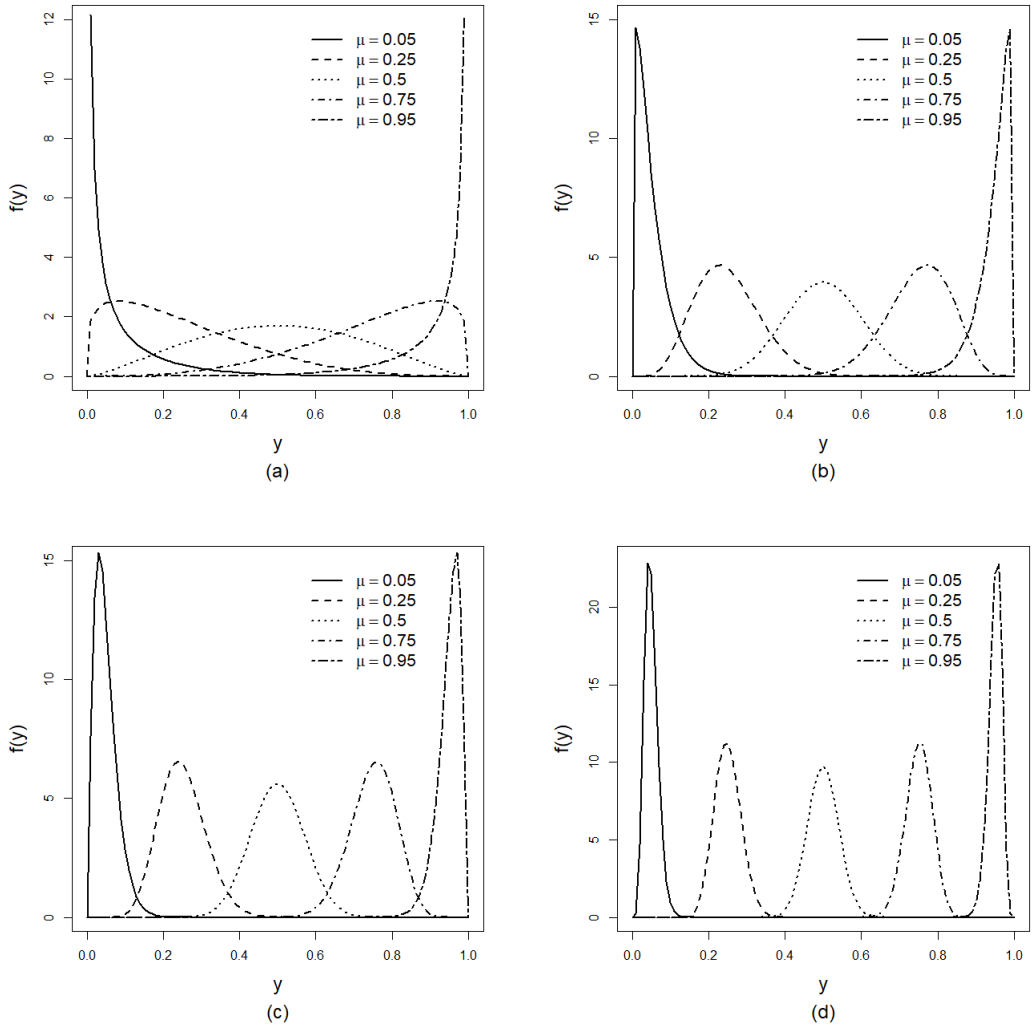
em que $0<\mu<1$ e $\phi>0$.

Nesse caso, a média e variância ficam dadas por

$$\mathbb{E}(y)=\mu \qquad \text{e} \qquad \text{Var}(y)=\frac{\mu(1-\mu)}{1+\phi}, \tag{2.3}$$

respectivamente. Podemos notar que para μ dado, quanto maior o valor de ϕ, menor será o valor de $\text{Var}(y)$. Por essa razão, ϕ é dito parâmetro de precisão.

Figura 1 – Gráficos da função densidade da distribuição beta para diferentes valores de μ, com $\phi = 5$ em (a), $\phi = 25$ em (b), $\phi = 50$ em (c) e $\phi = 150$ em (d).



Fonte: Autoria própria.

Na Figura 1 encontramos os gráficos das curvas referentes à densidade da distribuição beta, para diferentes valores de μ e ϕ. Observando inicialmente a Figura 1(a), vemos que para um valor de ϕ pequeno e $\mu = 0,5$, a distribuição apresenta uma curva simétrica e bastante aberta, já para valores de μ nos extremos do intervalo $(0,1)$, a curva da distribuição assume formatos assimétricos. No entanto, pelas Figuras 1(b)-(d) vemos que quando o valor de ϕ cresce, as curvas assumem formatos cada vez mais simétricos, mesmo para μ localizado nos extremos do intervalo unitário. Além disso, podemos notar que distribuição também torna-se mais afilada quando a precisão ϕ aumenta.

2.2 MODELO DE REGRESSÃO BETA COM DISPERSÃO VARIÁVEL

Sejam $y_1, \ldots, y_n$ variáveis aleatórias independentes, em que cada y_t, com $t = 1, \ldots, n$, segue uma distribuição beta, com função densidade de probabilidade dada por (2.2), com média μ_t

e precisão ϕ_t. O modelo de regressão beta não linear admite que a média e a precisão de y_t satisfaçam, respectivamente, as seguintes relações funcionais:

$$g(\mu_t) = f_1(\boldsymbol{x}_t^\top, \beta) = \eta_{1t} \qquad \text{e} \qquad h(\phi_t) = f_2(\boldsymbol{z}_t^\top, \gamma) = \eta_{2t}, \tag{2.4}$$

em que η_{1t} e η_{2t} são preditores não lineares, $g : (0,1) \to \mathbb{R}$ e $h : (0,\infty) \to \mathbb{R}$, são funções estritamente monótonas e duas vezes diferenciáveis, chamadas de funções de ligação, $\beta = (\beta_1, \ldots, \beta_k)^\top$ é o vetor de parâmetros do modelo para média, com $\beta \in \mathbb{R}^k$, $\gamma = (\gamma_1, \ldots, \gamma_r)^\top$ é o vetor de parâmetros do modelo para precisão, com $\gamma \in \mathbb{R}^r$, $k + r < n$, $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tk_1})^\top$ e $\boldsymbol{z}_t = (z_{t1}, \ldots, z_{tr_1})^\top$ são k_1 e r_1 valores conhecidos e fixados das variáveis explicativas dos modelos, respectivamente, com $t = 1, \ldots, n$, de modo que podem coincidir totalmente ou em partes com os números de parâmetros, tal que $k_1 \leq k$ e $r_1 \leq r$. Por fim, $f_1(\cdot)$ e $f_2(\cdot)$ são funções contínuas e diferenciáveis, tais que as matrizes de derivadas $J_1 = \partial\eta_1/\partial\beta$ e $J_2 = \partial\eta_2/\partial\gamma$ possuem postos k e r, respectivamente, sendo $\eta_1 = (\eta_{11}, \ldots, \eta_{1n})^\top$ e $\eta_2 = (\eta_{21}, \ldots, \eta_{2n})^\top$.

O modelo dado por (2.2) e (2.4) engloba o modelo de regressão beta linear com dispersão variável, quando os preditores η_{1t} e η_{2t} são funções lineares dos parâmetros, nesse caso, $g(\mu_t) = \eta_{1t} = \boldsymbol{x}_t^\top \beta$ e $h(\phi_t) = \eta_{2t} = \boldsymbol{z}_t^\top \gamma$.

Na literatura encontramos diversos tipos de funções de ligação $g(\cdot)$ e $h(\cdot)$, as funções mais conhecidas para $g(\cdot)$ são:

- probit: $g(\mu_t) = \Phi^{-1}(\mu_t)$, em que $\Phi(\cdot)$ corresponde a função de distribuição acumulada da normal padrão;
- log-log: $g(\mu_t) = \log[-\log(\mu_t)]$;
- complemento log-log: $g(\mu_t) = \log\{-\log(1-\mu_t)\}$;
- logit: $g(\mu_t) = \log(\mu_t/(1-\mu_t))$.

As funções mais comuns para $h(\cdot)$:

◇ logarítmica: $h(\phi_t) = \log(\phi_t)$;

◇ raiz quadrada: $h(\phi_t) = \sqrt{\phi_t}$;

◇ identidade: $h(\phi_t) = \phi_t$.

Em McCullagh e Nelder (1989) encontramos mais detalhes sobre as funções de ligações.

2.2.1 Vetor Escore, Matriz de Informação de Fisher e Estimadores de Máxima Verossimilhança

Com base na expressão (2.2), temos que o logaritmo da função de verossimilhança é dada por

$$\ell(\beta, \gamma) \;=\; \sum_{t=1}^{n} \ell_t(\mu_t, \phi_t), \tag{2.5}$$

em que

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

A partir da derivada do logaritmo da função de verossimilhança dada em (2.5), podemos obter a função escore para o vetor β calculando a função escore para β_i, com $i=1,\ldots,k$, que por sua vez é obtida a partir da escore de μ_t, dada por

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

em que $\psi(\cdot)$ é a função digama, ou seja, $\psi(z)=\partial\log\Gamma(z)/\partial z$, para $z>0$. Para simplificar, comumente, definimos

$$y_t^* = \log\left(\frac{y_t}{1-y_t}\right) \qquad \text{e} \qquad \mu_t^* = \psi(\mu_t\phi_t)-\psi[(1-\mu_t)\phi_t]. \tag{2.6}$$

Dessa forma, temos que

$$\begin{aligned}\frac{\partial\ell(\beta,\gamma)}{\partial\beta_i} &= \sum_{t=1}^{n}\frac{\partial\ell_t(\mu_t,\phi_t)}{\partial\beta_i}\\ &= \sum_{t=1}^{n}\left\{\frac{\partial\ell_t(\mu_t,\phi_t)}{\partial\mu_t}\frac{d\mu_t}{d\eta_{1t}}\frac{\partial\eta_{1t}}{\partial\beta_i}\right\}\\ &= \sum_{t=1}^{n}\left\{\phi_t(y_t^*-\mu_t^*)\frac{1}{g'(\mu_t)}\frac{\partial\eta_{1t}}{\partial\beta_i}\right\}.\end{aligned}$$

Vale notar que $d\mu_t/d\eta_{1t}=1/g'(\mu_t)$. Logo, podemos reescrever a função escore para β pela forma matricial

$$U_\beta(\beta,\gamma) = J_1^\top\Phi T(\mathbf{y}^*-\mu^*), \tag{2.7}$$

em que $J_1=\partial\eta_1/\partial\beta$ é uma matriz de ordem $n\times k$, $\Phi=\text{diag}\{\phi_1,\ldots,\phi_n\}$, $\mathbf{y}^*$ e μ^* vetores com t-ésimos elementos dados por y_t^* e μ_t^*, respectivamente, e

$$T = \text{diag}\left\{1/g'(\mu_1),\ldots,1/g'(\mu_n)\right\}. \tag{2.8}$$

Obteremos agora, a função escore para γ, a partir da derivada da expressão (2.5) em relação a γ_j, com $j=1,\ldots,r$, temos então

$$\begin{aligned}\frac{\partial\ell(\beta,\gamma)}{\partial\gamma_j} &= \sum_{t=1}^{n}\frac{\partial\ell_t(\mu_t,\phi_t)}{\partial\gamma_j}\\ &= \sum_{t=1}^{n}\left\{\frac{\partial\ell_t(\mu_t,\phi_t)}{\partial\phi_t}\frac{d\phi_t}{d\eta_{2t}}\frac{\partial\eta_{2t}}{\partial\gamma_j}\right\}\end{aligned}$$

$$= \sum_{t=1}^{n} a_t \frac{1}{h'(\phi_t)} \frac{\partial \eta_{2t}}{\partial \gamma_j},$$

em que $d\phi_t/d\eta_{2t} = 1/h'(\phi_t)$ e

$$a_t = \mu_t(y_t^* - \mu_t^*) + \log(1-y_t) - \psi[(1-\mu_t)\phi_t] + \psi(\phi_t). \tag{2.9}$$

Assim, podemos expressar a função escore para γ na seguinte forma matricial;

$$U_\gamma(\beta,\gamma) = J_2^\top H \mathbf{a}, \tag{2.10}$$

em que $J_2 = \partial\eta_2/\partial\gamma$ é uma matriz de ordem $n \times r$, $\mathbf{a} = (a_1,\ldots,a_n)^\top$, com a_t definido em (2.9), e

$$H = \operatorname{diag}\left\{1/h'(\phi_1),\ldots,1/h'(\phi_n)\right\}. \tag{2.11}$$

Seguiremos obtendo a matriz de informação de Fisher para os parâmetros β e γ, a partir das derivadas de segunda ordem do logaritmo da função de verossimilhança em (2.5). Essa matriz é dada por

$$K = K(\beta,\gamma) = \begin{bmatrix} K_{\beta\beta} & K_{\beta\gamma} \\ K_{\gamma\beta} & K_{\gamma\gamma} \end{bmatrix}, \tag{2.12}$$

sendo

$$K_{\beta\beta} = -\mathbb{E}\left[\frac{\partial^2 \ell(\beta,\gamma)}{\partial\beta\partial\beta^\top}\right], \tag{2.13}$$

em que, o (i,l)-ésimo elemento de (2.13) com $i,l = 1,\ldots,k$ é

$$-\mathbb{E}\left[\frac{\partial^2 \ell(\beta,\gamma)}{\partial\beta_i\partial\beta_l}\right] = \sum_{t=1}^{n} -\mathbb{E}\left[\frac{\partial^2 \ell_t(\mu_t,\phi_t)}{\partial\beta_i\partial\beta_l}\right].$$

Após alguns cálculos podemos facilmente ver que

$$\frac{\partial^2 \ell_t(\mu_t,\phi_t)}{\partial\beta_i\partial\beta_l} = \frac{\partial^2 \ell_t(\mu_t,\phi_t)}{\partial\mu_t^2}\frac{1}{[g'(\mu_t)]^2}\frac{\partial\eta_{1t}}{\partial\beta_i}\frac{\partial\eta_{1t}}{\partial\beta_l} + \frac{\partial}{\partial\mu_t}\left[\frac{1}{g'(\mu_t)}\right]\frac{\partial\ell_t(\mu_t,\phi_t)}{\partial\mu_t}\frac{1}{g'(\mu_t)}\frac{\partial\eta_{1t}}{\partial\beta_i}\frac{\partial\eta_{1t}}{\partial\beta_l}.$$

em que

$$\frac{\partial^2 \ell_t(\mu_t,\phi_t)}{\partial\mu_t^2} = -\phi_t^2\left\{\psi'(\mu_t\phi_t) + \psi'[(1-\mu_t)\phi_t]\right\} \qquad \text{e} \qquad \frac{\partial}{\partial\mu_t}\left[\frac{1}{g'(\mu_t)}\right] = -\frac{g''(\mu_t)}{[g'(\mu_t)]^2}.$$

Lembrando que $\partial \ell_t(\mu_t,\phi_t)/\partial \mu_t = \phi_t(y_t^* - \mu_t^*)$, temos que $\mathbb{E}[\partial \ell_t(\mu_t,\phi_t)/\partial \mu_t] = 0$, uma vez que é possível mostrar que $\mathbb{E}(y_t^*) = \mu_t^*$. Logo,

$$-\mathbb{E}\left[\frac{\partial^2 \ell_t(\mu_t,\phi_t)}{\partial \beta_i \partial \beta_l}\right] = \phi_t^2\left\{\psi'(\mu_t\phi_t) + \psi'[(1-\mu_t)\phi_t]\right\}\frac{1}{[g'(\mu_t)]^2}\frac{\partial \eta_{1t}}{\partial \beta_i}\frac{\partial \eta_{1t}}{\partial \beta_l}.$$

Assim, definindo,

$$w_t = \phi_t \upsilon_t \frac{1}{[g'(\mu_t)]^2}, \tag{2.14}$$

sendo

$$\upsilon_t = \{\psi'(\mu_t\phi_t) + \psi'[(1-\mu_t)\phi_t]\}, \tag{2.15}$$

temos que

$$-\mathbb{E}\left[\frac{\partial^2 \ell_t(\mu_t,\phi_t)}{\partial \beta_i \partial \beta_l}\right] = \phi_t w_t \frac{\partial \eta_{1t}}{\partial \beta_i}\frac{\partial \eta_{1t}}{\partial \beta_l}.$$

Então,

$$-\mathbb{E}\left[\frac{\partial^2 \ell(\beta,\gamma)}{\partial \beta_i \partial \beta_l}\right] = \sum_{t=1}^{n} \phi_t w_t \frac{\partial \eta_{1t}}{\partial \beta_i}\frac{\partial \eta_{1t}}{\partial \beta_l}.$$

Matricialmente temos que (2.13) fica dado por

$$K_{\beta\beta} = J_1^\top \Phi W J_1, \tag{2.16}$$

sendo $W = \text{diag}\{w_1,\ldots,w_n\}$, com w_t dado por (2.14). Seguiremos obtendo as segundas derivadas de $\ell(\beta,\gamma)$, com respeito a β_i e γ_j, para $i = 1,\ldots,k$ e $j = 1,\ldots,r$, que são dadas por

$$\begin{aligned}
\frac{\partial^2 \ell(\beta,\gamma)}{\partial \beta_i \partial \gamma_j} &= \sum_{i=1}^{n}\frac{\partial}{\partial \gamma_j}\left[\frac{\partial \ell_t(\mu_t,\phi_t)}{\partial \beta_i}\right] \\
&= \sum_{i=1}^{n}\frac{\partial}{\partial \phi_t}\left[\frac{\partial \ell_t(\mu_t,\phi_t)}{\partial \mu_t}\frac{d\mu_t}{d\eta_{1t}}\frac{\partial \eta_{1t}}{\partial \beta_i}\right]\frac{d\phi_t}{d\eta_{2t}}\frac{\partial \eta_{2t}}{\partial \gamma_j} \\
&= \sum_{i=1}^{n}\frac{\partial}{\partial \phi_t}\left[\phi_t(y_t^* - \mu_t^*)\frac{1}{g'(\mu_t)}\frac{\partial \eta_{1t}}{\partial \beta_i}\right]\frac{1}{h'(\phi_t)}\frac{\partial \eta_{2t}}{\partial \gamma_j}.
\end{aligned}$$

É fácil mostrar que

$$\frac{\partial}{\partial \phi_t}\left[\phi_t(y_t^* - \mu_t^*)\frac{1}{g'(\mu_t)}\frac{\partial \eta_{1t}}{\partial \beta_i}\right] = \left\{(y_t^* - \mu_t^*) - \phi_t[\psi'(\mu_t\phi_t)\mu_t - \psi'((1-\mu_t)\phi_t)(1-\mu_t)]\right\}\frac{1}{g'(\mu_t)}\frac{\partial \eta_{1t}}{\partial \beta_i}.$$

Logo,

$$\frac{\partial^2 \ell(\beta,\gamma)}{\partial\beta_i\partial\gamma_j} = \sum_{i=1}^{n}\{(y_t^*-\mu_t^*)-\phi_t[\psi'(\mu_t\phi_t)\mu_t-\psi'((1-\mu_t)\phi_t)(1-\mu_t)]\}\frac{1}{g'(\mu_t)}\frac{\partial\eta_{1t}}{\partial\beta_i}\frac{1}{h'(\phi_t)}\frac{\partial\eta_{2t}}{\partial\gamma_j},$$

como $\mathbb{E}(y_t^*-\mu_t^*)=0$, então, o (i,j)-ésimo elemento de $K_{\beta\gamma}$ com $i=1,\ldots,k$ e $j=1,\ldots,r$ fica dado por

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

em que

$$c_t = \phi\{\psi'(\mu_t\phi_t)\mu_t-\psi'[(1-\mu_t)\phi_t](1-\mu_t)\}. \tag{2.17}$$

Assim, temos matricialmente que

$$K_{\beta\gamma} = = J_1^\top CTHJ_2, \tag{2.18}$$

em que $C=\text{diag}\{c_1,\ldots,c_n\}$, com c_t dado por (2.17). Temos ainda da matriz (2.12) que $K_{\beta\gamma}=K_{\gamma\beta}{}^\top$. Por fim, considerando a derivada de segunda ordem de $\ell(\beta,\gamma)$ com respeito a γ_j e γ_s, para $j,s=1,\ldots,r$, dada por

$$\begin{aligned}
\frac{\partial^2\ell(\beta,\gamma)}{\partial\gamma_j\partial\gamma_s} &= \sum_{t=1}^{n}\frac{\partial^2\ell_t(\mu_t,\phi_t)}{\partial\gamma_j\partial\gamma_s}\\
&= \sum_{t=1}^{n}\frac{\partial}{\partial\phi_t}\left[\frac{\partial\ell_t(\mu_t,\phi_t)}{\partial\phi_t}\frac{d\phi_t}{d\eta_{2t}}\frac{\partial\eta_{2t}}{\partial\gamma_j}\right]\frac{d\phi_t}{d\eta_{2t}}\frac{\partial\eta_{2t}}{\partial\gamma_s}\\
&= \sum_{t=1}^{n}\frac{\partial}{\partial\phi_t}\left[\frac{\partial\ell_t(\mu_t,\phi_t)}{\partial\phi_t}\frac{1}{h'(\phi_t)}\right]\frac{1}{h'(\phi_t)}\frac{\partial\eta_{2t}}{\partial\gamma_j}\frac{\partial\eta_{2t}}{\partial\gamma_s}\\
&= \sum_{t=1}^{n}\left\{\frac{\partial^2\ell_t(\mu_t,\phi_t)}{\partial\phi_t^2}\frac{1}{h'(\phi_t)}+\frac{\partial}{\partial\phi_t}\left[\frac{1}{h'(\phi_t)}\right]\frac{\partial\ell_t(\mu_t,\phi_t)}{\partial\phi_t}\right\}\frac{1}{h'(\phi_t)}\frac{\partial\eta_{2t}}{\partial\gamma_j}\frac{\partial\eta_{2t}}{\partial\gamma_s},
\end{aligned}$$

como

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

e $\mathbb{E}[\partial\ell_t(\mu_t,\phi_t)/\partial\phi_t]=0$, definindo

$$d_t = \varsigma_t\frac{1}{[h'(\phi_t)]^2}, \tag{2.19}$$

sendo

$$\varsigma_t = \mu_t^2\psi'(\mu_t\phi_t)+(1-\mu_t)^2\psi'[(1-\mu_t)\phi_t]-\psi'(\phi_t), \tag{2.20}$$

obtemos que

$$-\mathbb{E}\left[\frac{\partial^2 \ell(\beta,\gamma)}{\partial\gamma_j\partial\gamma_s}\right] = \sum_{t=1}^{n} d_t \frac{\partial\eta_{2t}}{\partial\gamma_j}\frac{\partial\eta_{2t}}{\partial\gamma_s}$$

é o (j,s)-ésimo elemento da matriz

$$K_{\gamma\gamma} = -\mathbb{E}\left[\frac{\partial^2 \ell(\beta,\gamma)}{\partial\gamma\partial\gamma^\top}\right]. \tag{2.21}$$

Reescrevendo em forma matricial, temos

$$K_{\gamma\gamma} = J_2^\top D J_2, \tag{2.22}$$

em que $D = \text{diag}\{d_1,\ldots,d_n\}$, com d_t dado por (2.19). Para n suficientemente grande e considerando válidas as condições de regularidades, temos que a distribuição assintótica dos estimadores de máxima verossimilhança é tal que

$$\begin{pmatrix}\widehat{\beta}\\ \widehat{\gamma}\end{pmatrix} \sim N_{k+r}\left[\begin{pmatrix}\beta\\ \gamma\end{pmatrix}; K^{-1}\right],$$

aproximadamente, em que $\widehat{\beta}$ e $\widehat{\gamma}$ são os estimadores de máxima verossimilhança de β e γ, respectivamente, e K^{-1} é o inverso da matriz de informação de Fisher, dada por (2.12).

Como os estimadores de máxima verossimilhança para β e γ não possuem forma fechada, torna-se necessária a utilização de métodos numéricos que utilizam algoritmos de otimização não linear para maximizar o logaritmo da função de verossimilhança, e assim, obter valores plausíveis para β e γ. Dentre os métodos iterativo de otimização não linear, destacam-se o de Newton-Raphson, Scoring de Fisher e BFGS.

Para inicializar o processo iterativo os algoritmos de otimização requerem a especificação de valores iniciais para β e γ. Para o modelo beta linear com dispersão fixa, em que $\eta_{1t} = \mathbf{x}_t^\top\beta$, Ferrari e Cribari-Neto (2004) propõem usar como valor inicial para β, digamos $\beta^{(0)}$, o estimador de mínimos quadrados ordinários da regressão linear da resposta y_t transformada pela função $g(\cdot)$ em X, sendo X uma matriz de posto completo de ordem $n\times k$, com a t-ésima linha denotada por $\mathbf{x}_t^\top$. Assim, seja $\mathbf{z} = (z_1,\ldots,z_n)^\top$, com $z_t = g(y_t)$, então $\beta^{(0)} = (X^\top X)^{-1}X^\top\mathbf{z}$.

Para o parâmetro de precisão ϕ, o valor inicial se baseia no fato que $\text{Var}(y_t) = \mu_t(1-\mu_t)/(1+\phi)$, que implica $\phi = \mu_t(1-\mu_t)/\text{Var}(y_t)$ - 1. Dessa maneira, temos que

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

em que

$$\check{\mu}_t = g^{-1}(\mathbf{x}_t^\top\beta^{(0)}) \qquad \text{e} \qquad \widehat{\sigma}_t^2 = \frac{\check{\mathbf{e}}^\top\check{\mathbf{e}}}{(n-k)[g'(\check{\mu}_t)]^2} = \frac{\check{\sigma}^2}{[g'(\check{\mu}_t)]^2},$$

sendo $\check{\mathbf{e}}$ o vetor de resíduos de mínimos quadrados ordinários da regressão de $\mathbf{z}$ em X, ou seja, $\check{\mathbf{e}} = \mathbf{z} - X(X^\top X)^{-1}X^\top \mathbf{z}$. A definição de $\check{\sigma}_t^2$ está baseada no fato de que

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

logo, $\mathrm{Var}(y_t) \approx \mathrm{Var}[g(y_t)]/[g'(\mu_t)]^2$. Além disso, sabemos que $\widehat{\mathrm{Var}}[g(y_t)] = \check{\sigma}^2 = \check{\mathbf{e}}^\top\check{\mathbf{e}}/(n-k)$.

Para o modelo linear beta com dispersão variável, em que adicionalmente temos que $\eta_{2t} = \boldsymbol{z}_t^\top\gamma$, Ferrari et al. (2011) sugerem utilizar o valor inicial para β similar ao da dispersão fixa. No caso do valor inicial de γ é preciso obter inicialmente os valores iniciais para $\phi_1, \ldots, \phi_n$. Nesse caso, é sugerido utilizar

$$\phi_t^{(0)} = \frac{\check{\mu}_t(1-\check{\mu}_t)}{\widehat{\sigma}_t^2} - 1.$$

Assim, o valor inicial para γ é dado por $\gamma^{(0)} = (Z^\top Z)^{-1}Z^\top\mathbf{k}$, em que Z é uma matriz de posto completo de ordem $n \times r$, com a t-ésima linha dada por $\boldsymbol{z}_t^\top$ e $\mathbf{k} = (k_1, \ldots, k_n)^\top$, com $k_t = h(\phi_t^{(0)})$.

Para o modelo beta não linear, o processo para estabelecer valores iniciais para β e γ é um pouco mais extenso. De forma que, considerando a princípio o modelo não linear para μ_t e supondo que $k_1 = k$, seja $f_1(\mathbf{x}_t^\top, \beta)$ a função não linear, sua expansão em série de Taylor até a primeira ordem em torno de $\beta^{(0)}$ é dada por

$$f_1(\mathbf{x}_t^\top, \beta) \approx f_1(\mathbf{x}_t^\top, \beta^{(0)}) + \sum_{i=1}^{k}\left[\frac{\partial f_1(\mathbf{x}_t^\top, \beta)}{\partial \beta_i}\right]_{\beta=\beta^{(0)}}(\beta_i - \beta_i^{(0)}),$$

em que $\beta^{(0)} = (\beta_1^{(0)}, \ldots, \beta_k^{(0)})^\top$ é um valor inicial para β, que será definido mais adiante, sendo utilizado na obtenção do valor inicial do modelo não linear.

Vale notar que

$$\left[\frac{\partial f_1(\mathbf{x}_t^\top, \beta)}{\partial \beta_i}\right]_{\beta=\beta^{(0)}} = j_{1ti}^{(0)},$$

em que j_{1ti} é o elemento localizado na t-ésima linha e na i-ésima coluna da matriz J_1, então,

$$f_1(\mathbf{x}_t^\top, \beta) \approx f_1(\mathbf{x}_t^\top, \beta^{(0)}) + \sum_{i=1}^{k} j_{1ti}^{(0)}(\beta_i - \beta_i^{(0)}). \tag{2.23}$$

No caso do modelo normal não linear temos que $y_t = f_1(\mathbf{x}_t^\top, \beta) + \xi_t$, em que ξ_t é o t-ésimo erro aleatório, sobre o qual assume-se os pressupostos usuais, e como $\mathbb{E}(y_t|\mathbf{x}_t^\top) = \mu_t = f_1(\mathbf{x}_t^\top, \beta)$ e na Equação (2.23) $f_1(\mathbf{x}_t^\top, \beta)$ é substituido por $y_t \approx \mu_t$. Aqui, vamos considerar $g(y_t) \approx g(\mu_t) \approx f_1(\mathbf{x}_t^\top, \beta)$. Supondo ainda que $\theta_i^{(0)} = (\beta_i - \beta_i^{(0)})$ e $g(y_t) - f_1(\mathbf{x}_t^\top, \beta^{(0)}) = y_t^\dagger$ a

expressão (2.23) passa a ser

$$y_t^\dagger = \sum_{i=1}^{k} j1_{ti}^{(0)} \theta_i^{(0)}, \tag{2.24}$$

que pode ser visto como um modelo linear alternativo. Assim, pode-se deduzir que o estimador de mínimos quadrados ordinários de $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_k^{(0)})^\top$ fica então dado de modo que

$$\begin{aligned} \theta^{(0)} = (\beta - \beta^{(0)}) &= [J_1^{(0)\top} J_1^{(0)}]^{-1} J_1^{(0)\top} \mathbf{y}^\dagger \\ &= [J_1^{(0)\top} J_1^{(0)}]^{-1} J_1^{(0)\top} [g(\mathbf{y}) - f_1(X, \beta^{(0)})], \end{aligned}$$

em que $J_1^{(0)} = [\partial \eta_1 / \partial \beta]_{\beta=\beta^{(0)}}$, $\mathbf{y}^\dagger = (y_1^\dagger, \ldots, y_n^\dagger)^\top$, com $y_t^\dagger$ dado em (2.24), $\mathbf{y} = (y_1, \ldots, y_n)^\top$ e lembrando que X é a matriz de posto completo de ordem $n \times k$, com a t-ésima linha denotada por $\mathbf{x}_t^\top$. Podemos então apresentar o seguinte processo iterativo:

$$\beta^{(m)} = (J_1^{(m-1)\top} J_1^{(m-1)})^{-1} J_1^{(m-1)\top} [g(\mathbf{y}) - f_1(X, \beta^{(m-1)})] + \beta^{(m-1)}. \tag{2.25}$$

em que $m = 0, 1, 2, \ldots$ são os passos do processo. Dessa forma, a cada iteração é gerada uma nova aproximação $\beta^{(m)}$ para o vetor de parâmetros $\beta^{(m-1)}$ através do processo (2.25), que será repetido até que se obtenha uma convergência definida a partir de algum critério, como $||\beta^{(m)} - \beta^{(m-1)}|| < \varepsilon$, sendo ε um valor muito pequeno. Finalmente, a proposta para o valor inicial de β é dado por

$$\beta_{NL}^{(0)} = (J_1^{(0)\top} J_1^{(0)})^{-1} J_1^{(0)\top} [g(\mathbf{y}) - f_1(X, \beta_L^{(0)})] \qquad \text{e} \qquad \beta_L^{(0)} = (X^\top X)^{-1} X^\top g(\mathbf{y}). \tag{2.26}$$

Podemos notar que para o modelo não linear, a proposta do valor inicial para β utiliza um outro valor inicial, definido para o modelo linear. Para o modelo da precisão ϕ_t, supondo $r_1 = r$ e considerando $h(\phi_t) = f_2(\mathbf{z}_t^\top, \gamma)$, temos de forma análoga que

$$\gamma_{NL}^{(0)} = (J_2^{(0)\top} J_2^{(0)})^{-1} J_2^{(0)\top} [h(\phi_{NL}^{(0)}) - f_2(Z, \gamma_L^{(0)})] \qquad \text{e} \qquad \gamma_L^{(0)} = (Z^\top Z)^{-1} Z^\top h(\phi_L^{(0)}), \tag{2.27}$$

lembrando que Z é a matriz de posto completo de ordem $n \times r$, com a t-ésima linha dada por $\mathbf{z}_t^\top$, $J_2^{(0)} = [\partial \eta_2 / \partial \gamma]_{\gamma=\gamma^{(0)}}$ e $\phi_L^{(0)} = (\phi_{L1}^{(0)}, \ldots, \phi_{Ln}^{(0)})^\top$, com

$$\phi_{Lt}^{(0)} = \frac{\check{\mu}_{Lt}(1 - \check{\mu}_{Lt})}{\check{\sigma}_{Lt}^2} - 1,$$

em que $\check{\mu}_{Lt} = g^{-1}(\mathbf{x}_t^\top \beta_L^{(0)})$ e $\check{\sigma}_{Lt}^2 = \check{\mathbf{e}}_L^\top \check{\mathbf{e}}_L / \{(n-k)[g'(\check{\mu}_{Lt})]^2\}$, com $\check{\mathbf{e}}_L = g(\mathbf{y}) - \check{\mu}_L$, em que $\check{\mu}_L = (\check{\mu}_{L1}, \ldots, \check{\mu}_{Ln})^\top$. Finalizando, de (2.27) segue que $\phi_{NL}^{(0)} = (\phi_{NL1}^{(0)}, \ldots, \phi_{NLn}^{(0)})^\top$, com

$$\phi_{NLt}^{(0)} = \frac{\check{\mu}_{NLt}(1 - \check{\mu}_{NLt})}{\widehat{\sigma}_{NLt}^2} - 1,$$

em que $\check{\mu}_{NLt} = g^{-1}[f_1(\mathbf{x}_t^\top, \beta_{NL}^{(0)})]$ e $\widehat{\sigma}^2_{NLt} = \check{\mathbf{e}}_{NL}^\top \check{\mathbf{e}}_{NL} / \{(n-k)[g'(\check{\mu}_{NLt})]^2\}$, com $\check{\mathbf{e}}_{NL} = g(\mathbf{y}) - \check{\mu}_{NL}$, em que $\check{\mu}_{NL} = (\check{\mu}_{NL1}, \ldots, \check{\mu}_{NLn})^\top$. Possivelmente, existem situações em que para o modelo não linear o número de covariáveis é inferior ao número de parâmetros, para, pelo menos, um dos submodelos, ou seja, $k_1 < k$ e/ou $r_1 < r$, nesse caso, antes de obter os valores iniciais descritos anteriormente, é preciso fornecer certos valores para alguns dos parâmetros. Algumas técnicas são utilizadas para escolha desses valores iniciais nos modelos não lineares, dentre elas temos: a linearização, que consiste em realizar uma transformação no preditor a fim de torná-lo linear nos parâmetros, a partir de uma função, digamos $l(\cdot)$. Quando existe essa possibilidade, realiza-se então uma regressão linear de $l[g(y)]$ sobre os novos parâmetros e usa-se a transformação inversa das estimativas dessa regressão como valores iniciais para o modelo não linear.

Outra forma utilizada na determinação dos valores iniciais é estudando o comportamento do preditor, quando para certos valores atribuídos para um ou mais parâmetros, um determinado termo da expressão não converge. Dessa forma, podemos substitui tais parâmetros por valores iniciais que garantam a convergência. Além disso, é interessante estabelecer os valores iniciais dos parâmetros que não estejam associados num contexto linear no preditor. Uma última maneira de determinar valores iniciais para alguns parâmetros é pela análise gráfica dos dados. Nesse caso, é importante o conhecimento e familiaridade do pesquisador com o modelo utilizado.

2.2.2 Resíduos

Um procedimento bastante importante para validar o ajuste de qualquer modelo de regressão é a análise de diagnóstico, cujos principais objetivos são identificar pontos discrepantes ou mal ajustados e verificar possíveis afastamentos das suposições feitas sobre o modelo. Parte das técnicas de diagnóstico é chamada análise de resíduos, definidos como uma função que busca medir a discrepância entre o valor observado e o valor ajustado. O tipo de resíduo mais usual é chamado "resíduo ordinário", que serve de base para obtenção de outros tipos e pode ser definido como sendo $y_t - \widehat{\mathbb{E}(y_t)}$.

No caso do modelo de regressão linear beta, assumindo que o parâmetro de precisão ϕ é conhecido e constante, Espinheira et al. (2008) propõem um resíduo baseado no processo iterativo Scoring de Fisher para β. Ferrari et al. (2011) apresentam uma extensão desse resíduo para o modelo linear beta com dispersão variável. De modo que, considerando o processo iterativo Scoring de Fisher para estimar β e γ, dado por

$$\beta^{(m+1)} = \beta^{(m)} + [K_{\beta\beta}^{(m)}]^{-1} U_\beta^{(m)}(\beta) \qquad \text{e} \qquad \gamma^{(m+1)} = \gamma^{(m)} + [K_{\gamma\gamma}^{(m)}]^{-1} U_\gamma^{(m)}(\gamma), \tag{2.28}$$

em que $m = 0, 1, 2, \ldots$ são os passos do processo, (2.28) é repetido até que as diferenças $||\beta^{(m+1)} - \beta^{(m)}||$ e $||\gamma^{(m+1)} - \gamma^{(m)}||$ sejam menores que uma tolerância especificada. Aqui iremos abordar a estratégia apresentada por Ferrari et al. (2011), considerando o modelo de regressão beta não linear.

Dessa forma, trabalhando o lado direito da primeira expressão em (2.28), utilizando a função escore e a matriz de informação de Fisher para β, dadas por (2.7) e (2.16), respectivamente,

obtemos

$$\beta^{(m+1)} = \beta^{(m)} + [J_1^\top \Phi^{(m)} W^{(m)} J_1]^{-1} J_1^\top \Phi^{(m)} T^{(m)} [\mathbf{y}^* - \mu^{*(m)}], \tag{2.29}$$

vale lembrar que os t-ésimos elementos de $\mathbf{y}^*$ e de μ^* são definidos em (2.6), sendo $\mathbb{E}(\mathbf{y}^*) = \mu^*$, a matriz T é definida por (2.8), W é uma matriz diagonal com elementos dados por (2.14), $J_1 = \partial\eta_1/\partial\beta$ é uma matriz de ordem $n \times k$ e $\Phi = \text{diag}\{\phi_1, \ldots, \phi_n\}$. Vale observar que para o modelo beta com ϕ fixo, teríamos $\phi_1 = \phi_2 = \cdots = \phi_n = \phi$. Similarmente, utilizando a função escore e a matriz de informação de Fisher para γ, dadas por (2.10) e (2.22), respectivamente, pelo lado direito da segunda expressão em (2.28), obtemos

$$\gamma^{(m+1)} = \gamma^{(m)} + (J_2^\top D^{(m)} J_2)^{-1} J_2^\top H^{(m)} \mathbf{a}^{(m)}, \tag{2.30}$$

em que o t-ésimo elemento de $\mathbf{a}$ é definido em (2.9), a matriz H é definida por (2.11), D é uma matriz diagonal com elementos dados por (2.19) e $J_2 = \partial\eta_2/\partial\gamma$ é uma matriz de ordem $n \times r$.

É possível reescrever as expressões (2.29) e (2.30) como processos iterativos de mínimos quadrados reponderados, dados por

$$\beta^{(m+1)} = [J_1^\top \Phi^{(m)} W^{(m)} J_1]^{-1} J_1^\top \Phi^{(m)} W^{(m)} \mathbf{u}_1^{(m)} \qquad \text{e} \qquad \gamma^{(m+1)} = [J_2^\top D^{(m)} J_2]^{-1} J_2^\top D^{(m)} \mathbf{u}_2^{(m)},$$

em que $\mathbf{u}_1^{(m)} = J_1\beta^{(m)} + W^{-1(m)} T^{(m)} [\mathbf{y}^* - \mu^{*(m)}]$ e $\mathbf{u}_2^{(m)} = J_2\gamma^{(m)} + D^{-1(m)} H^{(m)} \mathbf{a}^{(m)}$.

Sob convergência dos processos, temos

$$\widehat{\beta} = [J_1^\top \widehat{\Phi}\widehat{W} J_1]^{-1} J_1^\top \widehat{\Phi}\widehat{W}\mathbf{u}_1, \qquad \text{e} \qquad \widehat{\gamma} = [J_2^\top \widehat{D} J_2]^{-1} J_2^\top \widehat{D}\mathbf{u}_2, \tag{2.31}$$

em que $\mathbf{u}_1 = J_1\widehat{\beta} + \widehat{W}^{-1}\widehat{T}(\mathbf{y}^* - \widehat{\mu}^*)$ e $\mathbf{u}_2 = J_2\widehat{\gamma} + \widehat{D}^{-1}\widehat{H}\widehat{\mathbf{a}}$. Aqui, as matrizes $\widehat{W}$, $\widehat{T}$, $\widehat{D}$ e $\widehat{H}$ correspondem às matrizes W, T, D e H, avaliadas nos estimadores de máxima verossimilhança. Podemos notar que $\widehat{\beta}$ e $\widehat{\gamma}$ em (2.31) podem serem vistos como soluções de mínimos quadrados das regressões lineares de $\widehat{\Phi}^{1/2}\widehat{W}^{1/2}u_1$ em $\widehat{\Phi}^{1/2}\widehat{W}^{1/2}J_1$ e de $\widehat{D}^{1/2}u_2$ em $\widehat{D}^{1/2}J_2$, respectivamente.

Então, os resíduos ordinários ficam definidos como $\mathbf{r}^\beta = \widehat{\Phi}^{1/2}\widehat{W}^{-1/2}\widehat{T}(\mathbf{y}^* - \widehat{\mu}^*)$ e $\mathbf{r}^\gamma = \widehat{D}^{-1/2}\widehat{H}\widehat{\mathbf{a}}$. Assim, é possível ver que os resíduos da t-ésima observação são dados por

$$r_t^\beta = \frac{y_t^* - \widehat{\mu}_t^*}{\sqrt{\widehat{\upsilon}_t}}, \qquad \text{com} \qquad r_t^\gamma = \frac{\hat{a}_t}{\sqrt{\hat{\varsigma}_t}}, \tag{2.32}$$

com υ_t e ς_t definidos em (2.15) e (2.20), respectivamente. A partir de Espinheira et al. (2008) é possível mostrar que, $\text{var}(y_t^*) = \upsilon_t$, e também que $\mathbb{E}(a_t) = 0$ e $\text{var}(a_t) = \varsigma_t$. Logo, os resíduos em (2.32) também são chamados de *resíduos ponderados padronizados*.

Como já mencionado, esses resíduos são bastante úteis para a análise de diagnóstico do modelo de regressão beta. De modo que um dos métodos mais conhecidos é análise gráfica dos resíduos *versus* os índices das observações, ou *versus* os valores ajustados, ou os valores

das covariáveis. Além disso, tais resíduos também são utilizados no gráfico de probabilidade com envelopes simulados, aplicado para avaliar a veracidade da hipótese sobre a distribuição assumida para a variável resposta.

Ademais, os resíduos ponderados padronizados, apresentados por (2.32) são importantes medidas usadas nas definições de alguns critérios de seleção de modelos de regressão beta, conforme veremos no próximo capítulo desse trabalho, e também vemos em Espinheira et al. (2019).

3 CRITÉRIOS DE SELEÇÃO DE MODELOS

3.1 INTRODUÇÃO

Diferentes estratégias têm surgido para modelar dados que assumem valores em um certo intervalo, havendo maior interesse em quando $y_t \in (0,1)$, com $t = 1,\ldots,n$, o intervalo unitário. Os modelos de regressão usuais normal linear não são adequados para modelar tais dados, devido à possibilidade dos valores ajustados estarem fora do intervalo e não permitir assimetria e heteroscedasticidade na distribuição da variável resposta. Alguns modelos foram então adaptados e introduzidos para tais respostas, por exemplo: os modelos de regressão simplex, por Barndorff-Nielsen e Jørgensen (1991), os modelos de regressão beta, por Ferrari e Cribari-Neto (2004), o modelo Kumaraswamy, por Mitnik e Baek (2013), o Johnson S_B, por Lemonte e Bazan (2015) e o modelo gama unitária, Mousa et al. (2016). Os mais citados na literatura são da classe de modelos de regressão beta e também da classe de modelos de regressão simplex. Pesquisadores podem utilizar o pacote `betareg` nas modelagens de regressões beta, disponível para o software estatístico `R`, e nas regressões simplex o pacote `simplexreg`, também disponível para o ambiente computacional `R`. Aqui, o nosso foco principal é a regressão beta, e o processo de selecionar modelos que produzam valores preditos plausíveis (então as estimativas da média da variável resposta, μ_t, apresentam baixos vieses) e a variabilidade da variável resposta seja bem explicada. Mais especificamente, estamos interessados no processo de escolher modelos beta bem ajustados e com alto poder preditivo.

Sabemos que, quando omitimos covariáveis importantes em um modelo de regressão, isso pode resultar em um viés sobre as estimativas dos coeficientes do modelo de regessão. Outro fator que provoca más estimativas dos parâmetros é quando uma covariável desnecessária é incluída erroneamente nos preditores. Existem vários outros casos de modelos mal especificados que podem produzir elevados vieses nas estimativas, por exemplo, modelo inadequado para a dispersão, negligência da não-linearidade nos preditores, especificação incorreta da distribuição de probabilidade da variável resposta, entre outros. Além disso, todos estes problemas de especificação também afetam a variabilidade na modelagem da regressão.

Uma ferramenta muito conhecida e utilizada para avaliar a qualidade do ajuste de um modelo de regressão é o coeficiente de explicação R^2, desenvolvido a princípio para o modelo normal linear clássico. Essa medida busca indicar a proporção da variabilidade da resposta explicada pelo modelo. Varias versões desse coeficiente foram propostas, possibilitando a aplicação dessa medida em diversas classes de modelos de regressão, todas mantendo a ideia de avaliar a variabilidade explicada pelo modelo estimado. Aqui consideraremos o R^2_{LR} apresentado por Nagelkerke (1991), sendo definido com base na razão das funções de máxima verossimilhança do modelo sem regressores e do modelo investigado. Desse modo, a razão das verossimilhanças aponta o nível de melhoria oferecido pelo modelo proposto em relação ao modelo nulo. Veremos mais detalhes sobre essa medida mais adiante.

Outra versão de R^2 que consideraremos é o R^2_{FC}, proposto de modo intuitivo por Ferrari e Cribari-Neto (2004). Essa medida é baseada no coeficiente de correlação amostral de Pearson

entre $g(\mathbf{y})$ e $\widehat{\eta}_1$, de forma que, em algumas situações, quanto maior for a correlação, maior será a covariância. Consequentemente, grandes mudanças ocorrem em $g(\mathbf{y})$ e $\widehat{\eta}_1$ na mesma direção e intensidade, e portanto, o mesmo deve ocorrer entre $\mathbf{y}$ e $\widehat{\mu}$. Vale lembrar que $g(\cdot)$ é a função de ligação do modelo para a média μ, e $\widehat{\eta}_1$ é o preditor linear desse modelo. Mais adiante veremos mais detalhes sobre a medida R^2_{FC}. Como essas medidas R^2 estão relacionadas diretamente com a qualidade do ajuste do modelo de regressão ao conjunto de dados, temos que elas são ótimas ferramentas para identificar possíveis casos de "*underfitting*", que justamente correspondem às situações em que o modelo utilizado não se ajusta bem aos dados. Logo, valores baixos dessas medidas indicam que o modelo não apresenta um bom ajuste devido a sua variância elevada.

Contudo, essas versões (chamadas de pseudo R^2's ou critérios R^2) não conseguem identificar se as predições são tendenciosas. Tal informação é geralmente obtida através da inspeção visual dos gráficos de resíduos. Entretanto, Allen (1974) propôs algumas medidas que podem ser consideravelmente úteis para indicar estimativas de μ com altos vieses, uma vez que tais medidas são baseadas nos resíduos, como é o caso da estatística PRESS. Uma outra estatística mais fácil de ser interpretada, baseada na PRESS e tipicamente conhecida como a estatística R^2 de predição, é a medida P^2, proposta por Quan (1988) como critério Q^2. Essa medida é capaz de validar o poder de predição do modelo, sem selecionar amostras adicionais, ou considerar subconjuntos dos dados em conjuntos de formação e validação, tal como é exigido num esquema de validação cruzada. Dessa forma, a análise da estatística P^2 pode ser bastante útil para evitar o que chamamos de "*overfitting*", que ocorre quando o modelo aparenta apresentar um bom ajuste para um conjunto de dados em particular, mas fornece más predições para novas observações. Além disso, tal como demonstrado por Espinheira et al. (2019), a estatística P^2 também fornece informação sobre a falta de robustez empírica do esquema de estimação por máxima verosimilhança, devido à ocorrência de observações influentes nos dados.

Resulta que para escolher um modelo adequado devemos considerar conjuntamente as informações contidas nos critérios P^2, R^2_{FC} e R^2_{LR} e suas versões corrigidas. Nesse capítulo, serão melhor apresentados os critérios de seleção citados anteriormente na escolha dos modelos de regressão beta. Propomos também um novo critério de seleção de modelos beta, que corresponde a uma função dos critérios P^2, R^2_{FC} e R^2_{LR}, e de suas versões corrigidas segundo o número de covariáveis inseridas no modelo. Então, esse novo critério de seleção é construído no sentido de resumir a informação contida nessas medidas, sendo assim um critério de balanceamento viés-variância.

3.2 CRITÉRIOS DE SELEÇÃO R^2

Sabemos que durante a análise de regressão em um modelo linear clássico, além estimar os coeficientes do modelo, os erros-padrão e avaliar algumas propriedades, também temos o interesse que avaliar a "qualidade do ajuste" da linha de regressão ao conjunto de dados. Ou seja, buscamos descobrir o quão bem a linha de regressão estimada adequa-se aos dados. Sabemos que se todas as observações disponíveis estivessem situadas sobre a linha de regressão, obteríamos um "ajuste perfeito", no entanto, raramente isso pode ocorrer, pois geralmente existem

resíduos em torno dessa linha de regressão, porém, vale observar que quanto menor forem os resíduos, melhor será o ajuste.

Dessa forma, considerando o modelo linear clássico $\mathbf{y} = X\beta + \varepsilon$ (com intercepto), em que $\mathbf{y}$ é o vetor das variáveis respostas $n \times 1$, X é a matriz de covariáveis de ordem $n \times k$, β é o vetor de parâmetros $k \times 1$ e ε o vetor de erros $n \times 1$, temos que a variação total dos valores observados de $\mathbf{y}$ em torno de sua média amostral $\bar{y}$ é dada por $SST = \sum_{t=1}^{n}(y_t - \bar{y})^2$, comumente chamada de **Soma de Quadrados Total**. É fácil mostrar que

$$\sum_{t=1}^{n}(y_t - \bar{y})^2 \quad = \quad \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2 + \sum_{t=1}^{n} r_t^2,$$

em que $r_t = y_t - \hat{y}_t$ corresponde ao t-ésimo resíduo, sendo $\hat{y}_t$ o valor predito de y_t. De modo que $\sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2 = SSR$ corresponde à variação dos valores preditos de y_t em torno de $\bar{y}$, comumente chamado de **Soma de Quadrados da Regressão**, enquanto $\sum_{t=1}^{n} r_t^2 = SSE$ corresponde à variação residual, isto é, a variação não explicada dos valores de y_t em relação à linha de regressão, chamado de **Soma dos Quadrados dos Resíduos**. Dessa forma,

$$SST \quad = \quad SSR + SSE.$$

Logo, vemos que a variação total dos valores observados da variável resposta em torno de sua média pode ser dividida em duas partes, uma parte atribuível à regressão, ou seja, que é explicada pelo modelo de regressão ajustado, e a outra parte atribuível à erros não observáveis (não sendo explicada pelo modelo de regressão). Dessa forma, a medida

$$R^2 \quad = \quad \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{3.1}$$

é um indicador usado para medir a qualidade do ajustamento na linha de regressão estimada. Em outras palavras, ela mede a proporção da variabilidade total de y_t explicada pelo modelo de regressão. A medida (3.1) é chamada de coeficiente de determinação ou explicação. Vale destacar algumas propriedades importantes sobre essa medida R^2:

- o coeficiente de determinação R^2 varia no intervalo $[0, 1]$;
- R^2 corresponde ao quadrado do coeficiente de correlação entre y e $\hat{y}$;
- quanto mais próximo de 1 for o valor de R^2, maior o poder de explicação do modelo;
- o valor de R^2 cresce com o número de regressores.

Essa última observação é bastante importante, uma vez que R^2 é uma ferramenta comumente usada para comparar diferentes modelos para uma mesma variável resposta. Logo, aquele que apresentar maior R^2 é tido como o melhor modelo. No entanto, devemos ter cuidado quando os modelos comparados apresentam diferentes números de regressores, pois nesses casos é sugerido considerar o coeficiente R^2 junto a um termo penalizador, quando novas covariáveis

são incluídas no modelo. Chamamos essa medida penalizada de R^2 *ajustado*, sendo denotado por

$$\overline{R^2} = 1 - (1 - R^2)\left(\frac{n-1}{n-p}\right), \tag{3.2}$$

em que p corresponde ao número de coeficientes do modelo de regressão. Buscando inserir o coeficiente R^2 em outras classes de modelos de regressão, diferentes formulações foram propostas para essa medida, como encontramos em Ferrari e Cribari-Neto (2004) e Nagelkerke (1991). Tais formulações são geralmente conhecidas como pseudo R^2's ou critério R^2.

3.2.1 Principais Critérios R^2 Para Modelos Beta

Para o modelo de regressão beta linear, Ferrari e Cribari-Neto (2004) definem uma medida de ajuste pseudo R^2 como sendo o quadrado do coeficiente de correlação amostral de Pearson entre $g(\mathbf{y})$ e $\widehat{\eta}_1$, de modo que aqui representaremos esse pseudo R^2 por R^2_{FC}. Dessa forma, temos que Ferrari e Cribari-Neto (2004) propõem

$$R^2_{FC} \;=\; [\mathrm{r}(g(\mathbf{y}), \widehat{\eta}_1)]^2, \tag{3.3}$$

em que $\mathrm{r}(\cdot)$ corresponde ao coeficiente de correlação amostral de Pearson e $\widehat{\eta}_1 = X\widehat{\beta} = g(\widehat{\mu})$, onde $\widehat{\beta}$ é o estimador de máxima verossimilhança de β. Para o caso do modelo não linear temos que $\widehat{\eta}_1 = f_1(x, \widehat{\beta}) = g(\widehat{\mu})$. Vale observar que altos valores da correlação geralmente implicam alta covariância, logo, as medidas $g(\mathbf{y})$ e $\widehat{\eta}_1$ mudam conjuntamente na mesma direção e magnitude.

Outra observação importante é que a correlação entre $g(\mathbf{y})$ e $\widehat{\eta}_1$ depende das suas respectivas variâncias, então, se a variância das estimativas do preditor da média é pequena, esperamos que a medida R^2_{FC} apresente um valor alto. Embora seja importante considerar a possibilidade da variância de $g(\mathbf{y})$ ser alta, o que irá penalizar o R^2_{FC}, isso faz com que esse critério assuma valores mais baixos do que outras versões apresentadas na literatura para pseudo R^2. Além disso, para um valor baixo de R^2_{FC}, temos que a correlação entre $g(\mathbf{y})$ e $\widehat{\eta}_1$ é baixa, e uma vez que $\widehat{\mu}$ é o preditor natural para $\mathbf{y}$, temos nesse caso que $\widehat{\mu}$ não consegue predizer bem $\mathbf{y}$. Então, valores baixos de R^2_{FC} possivelmente indicam má especificação no submodelo da média. Isso pode ocorrer devido a falta de covariáveis importantes, ou função de ligação $g(\cdot)$ inapropriada, ou má especificação sobre a expressão matemática que definem o modelo da média. Vale notar também que $R^2_{FC} = 1$ representa uma relação perfeita entre $g(\mathbf{y})$ e $\widehat{\eta}_1$, e portanto entre $\mathbf{y}$ e $\widehat{\mu}$.

Adicionalmente, Bayer e Cribari-Neto (2017) propuseram uma versão corrigida dessa medida, inserindo um termo penalizador quando são incluídas novas covariáveis no modelo, similar ao R^2 ajustado para o modelo de regressão linear clássico, dado em (3.2). Assim, a versão corrigida do critério R^2_{FC} fica dado por:

$$R^2_{FC_c} \;=\; 1 - (1 - R^2_{FC})\left(\frac{n-1}{n-p}\right), \tag{3.4}$$

em que $p = k + r$ é o número de parâmetros do modelo, sendo k e r o número de parâmetros do modelo para a média e para a precisão, respectivamente.

Outra proposta de pseudo R^2, foi apresentada por Nagelkerke (1991), tratando-se de uma generalização do coeficiente R^2 para uma classe maior de modelos, com base na razão de verossimilhanças, sendo definido por

$$R^2_{LR} = 1 - \left(\frac{L_{null}}{L_{fit}}\right)^{2/n}, \tag{3.5}$$

em que L_{null} é a função de verossimilhança maximizada do modelo sem regressores e L_{fit} é a função de verossimilhança maximizada do modelo ajustado. Podemos notar que, quanto maior for L_{fit} em relação a L_{null}, maior será a evidência de que o modelo ajustado é melhor do que o modelo nulo e, assim, menor será a razão das verossimilhanças, consequentemente, mais próximo de um será R^2_{LR}. Além disso, dado que a comparação do modelo proposto é feita sob o modelo nulo, R^2_{LR} tende a ser tipicamente superior a outras versões de R^2.

Quanto à variabilidade, Nagelkerke (1991) cita que R^2 pode ser interpretada como a proporção de 'variação explicada', ou melhor, $1 - R^2$ interpreta a proporção de 'variação não explicada'. Nagelkerke (1991) também diz que, muito geralmente, a variação deve ser explicada como qualquer medida estendida para qual uma distribuição não é degenerada, o que nos parece uma interpretação bastante superficial da medida R^2, em termos de variabilidade. Ademais, por utilizar funções de verossimilhança em sua definição, a medida R^2_{LR} expressa a plausividade do modelo estimado ser o verdadeiro processo de geração dos dados. Então, em termos de variabilidade, R^2_{FC} é uma estatística mais rigorosa na escolha de um melhor modelo. Neste caso, quando o valor de R^2_{FC} é bem menor do que R^2_{LR}, temos uma forte evidência de erro de especificação no submodelo da média, uma vez que a correlação entre $g(\mathbf{y})$ e $\widehat{\eta}_1$ é baixa, ou seja, $\widehat{\mu}$ não é bom preditor de $\mathbf{y}$.

Como tipicamente o R^2_{LR} assume valores superiores a R^2_{FC}, temos de forma intuitiva que, quando o oposto ocorre, ou seja, $R^2_{LR} < R^2_{FC}$, existe grande possibilidade do modelo estimado estar mal especificado, ainda que ambas medidas apresentem valores elevados ou razoavelmente altos. No entanto, se os valores são elevados, especialmente de R^2_{FC}, então, os $\widehat{\mu}$ e $\mathbf{y}$ estão próximos e provavelmente o problema de especificação então não está sobre o modelo da média e sim no modelo postulado para a precisão.

Baseado na medida R^2_{LR}, podemos ainda considerar o critério de seleção para o modelo beta, proposto por Bayer e Cribari-Neto (2017), dado por

$$R^2_{LR_c} = 1 - (1 - R^2_{LR})\left(\frac{n-1}{n-p}\right), \tag{3.6}$$

para o modelo com dispersão constante e

$$R^2_{LR_c} = 1 - (1 - R^2_{LR})\left[\frac{n-1}{n-(1+\alpha)k-(1-\alpha)r}\right]^{\delta}, \tag{3.7}$$

com dispersão variável, em que $\alpha \in (0,1)$ e $\delta > 0$. Com base em estudos de simulações, Bayer e Cribari-Neto (2017) sugerem que $\alpha = 0,4$ e $\delta = 1$. Além disso, o critério R^2_{LR} tem ligação com critérios de informação tais como o Critério de Informação de Akaike (AIC), proposto por Akaike (1973), e o Critério Baysiano de Schwarz (BIC), apresentado por Schwarz (1978). Bayer e Cribari-Neto (2017) fornecem um extenso estudo, no qual são avaliados vários critérios de seleção e informação para modelos de regressão beta. Adicionalmente, os autores propuseram um esquema de seleção rápida de modelos em duas fases, tendo em consideração tanto os submodelos da média como os da dispersão.

3.3 ESTATÍSTICAS PRESS e P^2

A estatística PRESS foi apresentada inicialmente por Allen (1971) para os modelos lineares e busca avaliar a habilidade do modelo em predizer valores para a variável resposta. Além disso, essa medida é bastante útil para indicar altos vieses em $\hat{\mu}_t$, uma vez que sua definição está baseada nos resíduos, como veremos a seguir.

Considerando então o modelo $\mathbf{y} = X\beta + \varepsilon$, em que $\mathbf{y}$ é o vetor das variáveis respostas $n \times 1$, X é a matriz de covariáveis de ordem $n \times k$, β é o vetor de parâmetros $k \times 1$ e ε o vetor de erros $n \times 1$. Sabemos que o estimador de mínimos quadrados de β é definido por $\widehat{\beta} = (X^\top X)^{-1} X^\top \mathbf{y}$. Dessa forma, o t-ésimo valor predito da resposta fica dado por $\widehat{y}_t = \mathbf{x}_t^\top \widehat{\beta}$, com $t = 1, \ldots, n$, e a diferença entre os valores observados e os valores preditos da variável resposta é chamada de resíduo ordínário, sendo $r_t = y_t - \widehat{y}_t$.

Sejam $\widehat{\beta}_{(t)}$ o estimador de β sem a t-ésima observação e $\widehat{y}_{(t)} = \mathbf{x}_t^\top \widehat{\beta}_{(t)}$ o valor predito do caso deletado quando a covariável assume valor $\mathbf{x}_t$, então, $e_{(t)} = y_t - \widehat{y}_{(t)}$ é dito ser o erro predito. Dessa forma, a estatística PRESS para um modelo de regressão linear é dada por

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

em que h_{tt} é o t-ésimo elemento da diagonal principal da matriz de projeção $X(X^\top X)^{-1} X^\top$. Logo, para valores baixos dessa medida, temos que o modelo ajustado possui alto poder preditivo. Para mais detalhes, ver Allen (1971), Hocking (1972) e Geisser e Eddy (1979).

Para o modelo de regressão beta não linear, temos, segundo Espinheira et al. (2019), que $\widehat{\beta}$ dado em (2.31) pode ser visto como solução de mínimos quadrados da regressão linear de

$$\mathbf{y}^\dagger = \widehat{\Phi}^{1/2} \widehat{W}^{1/2} \mathbf{u}_1 \qquad \text{em} \qquad J_1^\dagger = \widehat{\Phi}^{1/2} \widehat{W}^{1/2} J_1,$$

em que $\Phi = \text{diag}\{\phi_1, \ldots, \phi_n\}$, $W = \text{diag}\{w_1, \ldots, w_n\}$, com w_t dado em (2.14) e $\mathbf{u}_1$ é dado a partir de (2.31). Seja $\widehat{\beta}_{(t)}$ o estimador de β sem a t-ésima observação. Definimos $\widehat{y^\dagger}_{(t)} = \widehat{\phi}_t^{1/2} \widehat{w}_t^{1/2} \boldsymbol{j}_{1t}^\top \widehat{\beta}_{(t)}$, em que $\boldsymbol{j}_{1t}^\top$ é a t-ésima linha da matriz J_1. Dessa forma, o erro predito fica dado por

$$e^\dagger_{(t)} \quad = \quad y_t^\dagger - \widehat{y^\dagger}_{(t)}$$

$$= \widehat{\phi}_t^{1/2}\widehat{w}_t^{1/2}u_{1,t} - \widehat{\phi}_t^{1/2}\widehat{w}_t^{1/2}\boldsymbol{j}_{1t}^{\top}\widehat{\beta}_{(t)}. \tag{3.8}$$

Entretanto, de acordo com Pregibon (1981),

$$\widehat{\beta}_{(t)} = \widehat{\beta} - \frac{(J_1^{\top}\widehat{\Phi}\widehat{W}J_1)^{-1}\boldsymbol{j}_{1t}\widehat{\phi}_t^{1/2}\widehat{w}_t^{1/2}r_t^{\beta}}{1-h_{tt}^*}, \tag{3.9}$$

em que r_t^{β} é o resíduo ponderado padronizado da t-ésima observação, apresentado em (2.32) e h_{tt}^* é o t-ésimo elemento da diagonal principal de

$$H^* = (\widehat{W}\widehat{\Phi})^{1/2}J_1(J_1\widehat{\Phi}\widehat{W}J_1)^{-1}J_1^{\top}(\widehat{\Phi}\widehat{W})^{1/2}.$$

Então, a partir de (3.9) temos que o erro predito em (3.8) pode ser reescrito como

$$e_{(t)}^{\dagger} = \frac{r_t^{\beta}}{1-h_{tt}^*}.$$

Logo, de acordo com Espinheira et al. (2019), a estatística PRESS para o modelo de regressão beta não linear pode ser definida como

$$\text{PRESS} = \sum_{t=1}^{n}[e_{(t)}^{\dagger}]^2 = \sum_{t=1}^{n}\left(\frac{r_t^{\beta}}{1-h_{tt}^*}\right)^2. \tag{3.10}$$

Podemos notar que a medida PRESS apresentada em (3.10) é definida como sendo a soma dos quadrados dos resíduos preditos (digamos $SSE_{(t)}$). Logo, para valores baixos dessa medida, concluímos que o modelo ajustado possui alto poder preditivo. De modo similar ao cálculo do coeficiente de determinação R^2, Mediavilla e Shah (2008) seguindo as sugestões de Quan (1988), propõem o seguinte coeficiente de predição:

$$P^2 = 1 - \frac{\text{PRESS}}{SST_{(t)}}, \tag{3.11}$$

em que $SST_{(t)} = \sum_{t=1}^{n}[y_t^{\dagger} - \overline{y^{\dagger}}_{(t)}]^2$, sendo $\overline{y^{\dagger}}_{(t)}$ a média aritmética dos valores de $\mathbf{y}^{\dagger}$ sem a t-ésima observação. É possível mostrar que $SST_{(t)} = [n/(n-p)]^2 SST$, em que $SST = \sum_{t=1}^{n}[y_t^{\dagger} - \overline{y^{\dagger}}]^2$, com $\overline{y^{\dagger}}$ sendo a média aritmética dos valores de $\mathbf{y}^{\dagger}$ e p é o número de parâmetros do modelo.

Embora P^2 seja definido de modo similar ao coeficiente R^2, ambos possuem funções distintas, uma vez que o R^2 busca avaliar a qualidade do ajuste, ou seja, ele mede a proporção da variabilidade total da variável resposta explicada pelo modelo enquanto o coeficiente P^2 procura medir o poder de predição do modelo ajustado. Similarmente ao $R^2_{FC_c}$, Espinheira et al. (2019) propõem a versão corrigida de P^2, adicionando um termo penalizador ao incluir novas covariáveis no modelo. Assim, a versão corrigida fica dada por

$$P_c^2 = 1 - (1-P^2)\left(\frac{n-1}{n-p}\right). \tag{3.12}$$

Após ajustar um certo modelo de regressão, é possível haver casos em que aparentemente esse modelo apresente um bom ajuste ao conjunto de dados de treinamento, no entanto, ele fornece más predições para novas observações. Nesse caso, ocorre o que chamamos de "*overfitting*". Com base nos critérios de ajuste apresentados até o presente momento para a classe de modelos beta, um cenário típico de *overfitting* pode ser identificado quando uma versão de R^2 tem valor significativamente elevado enquanto os valores das estatísticas do tipo P^2 são consideravelmente baixos (até próximos de zero).

Outra situação pode ocorrer quando as estatísticas P^2 apresentam valores elevados, enquanto as estatísticas do tipo R^2 (especialmente R^2_{FC}) apresentam valores significativamente baixos. Nesse caso, como P^2 é alto, temos então que $\hat{\mu}_t$, com $t = 1, \ldots, n$, aparentemente apresenta vieses consideravelmente baixos. Porém, $\hat{\mu}_t$ corresponde apenas a um estimador pontual de μ_t. Então, se o modelo ajustado apresenta dificuldades em estimar adequadamente a variância da resposta, isto poderá afetar alguns resultados inferenciais, como testes de hipóteses e intervalos de confiança. Quando isso ocorre, as conclusões inferenciais sobre o submodelo da média, e sobre os parâmetros β's, e consequentemente sobre μ_t, ficam comprometidas, além de existir a possibilidade da inclusão de covariáveis não tão importantes para a explicação da resposta em uma futura amostra analisada. Até onde sabemos, a questão da modelagem de regressão envolve a seleção de modelos parcimoniosos, a saber, poucos parâmetros e muita informação.

Além disso, sempre que as medidas R^2_{FC} e R^2_{LR} forem baixas, temos fortes indícios de possíveis casos de *underfitting*, ou seja, o modelo não está bem ajustado aos dados de treinamento. Adicionalmente, de modo intuitivo, quando as estatísticas do tipo R^2 são baixas, temos que: se R^2_{FC} for bem menor do que R^2_{LR}, provavelmente o modelo não se adequa bem ao conjunto de dados devido à variabilidade estimada ser elevada. Já quando as medidas do tipo P^2 são altas, o modelo não se ajusta bem ao conjunto de dados devido à distribuição imposta para a resposta não ser adequada. Dessa forma, um critério de balanceamento viés-variância seria bastante útil na escolha de um bom modelo.

3.4 NOVO CRITÉRIO DE SELEÇÃO *BV* (*BIAS AND VARIABILITY*)

As definições dos critérios de seleção de modelos usuais são geralmente baseadas no poder de explicação da variabilidade da resposta ou no poder de predição dos valores desta, ou seja, esses critérios são formulados tendo como objetivo controlar apenas um desses dois fatores. Em muitos casos, quando ajustamos um modelo de regressão e calculamos esses critérios usuais, um dos fatores se sobressai, isto é, ou o modelo apresenta um alto poder de explicação da variabilidade e um poder inferior de predição, ou o contrário. Então, nessa situação, precisamos escolher qual dos fatores pesará mais na escolha do modelo. Nosso interesse consiste em apresentar um critério de seleção que busque avaliar conjuntamente a proporção de variabilidade explicada pelo modelo e o seu poder de predição.

Dispomos aqui das estatísticas P^2, R^2_{FC} e R^2_{LR}, dadas por (3.11), (3.3) e (3.5), respectivamente, e de suas versões corrigidas, P^2_c, dado por (3.12), $R^2_{FC_c}$, dado por (3.4) e $R^2_{LR_c}$, dado por (3.6) ou (3.7), como critérios de seleção. Essas estatísticas possuem em comum o fato de que,

quanto mais próximos seus valores estiverem de um, mais adequado será considerado o modelo ajustado, e quanto menores forem seus valores, menos adequado será o modelo. Vale observar que algumas destas medidas podem assumir valores negativos.

A proposta para o novo critério de seleção, que chamaremos de BV (Bias and Variability), refere-se à média ponderada entre o máximo e o mínimo dessas seis estatísticas. Dessa forma, sejam $T = \min\{P^2, P_c^2, R_{FC}^2, R_{FC_c}^2, R_{LR}^2, R_{LR_c}^2\}$ e $U = \max\{P^2, P_c^2, R_{FC}^2, R_{FC_c}^2, R_{LR}^2, R_{LR_c}^2\}$, o novo critério de seleção proposto é denotado por

$$BV \quad = \quad (1-\alpha)T + \alpha U, \quad \text{com } \alpha \in (0,1). \tag{3.13}$$

É possível observar que os valores de BV variam entre os valores de T e U, independentemente do valor de α. Entretanto, podemos notar que:

- se $\alpha < 0.5$, o mínimo das estatísticas terá um peso maior sobre o valor de BV;
- se $\alpha > 0.5$, o máximo das estatísticas terá um peso maior sobre o valor de BV;
- se $\alpha = 0.5$, o de mínimo e o máximo exercem o mesmo peso sobre o valor de BV, ou seja, BV consiste na média aritmética dessas duas medidas.

Agora, sejam as estatísticas $T = \min\{P^2, P_c^2, R_{FC}^2, R_{FC_c}^2, R_{LR}^2, R_{LR_c}^2\}$ e $U = \max\{P^2, P_c^2, R_{FC}^2, R_{FC_c}^2, R_{LR}^2, R_{LR_c}^2\}$ variáveis aleatórias tais que $t \in (-\infty, 1.0)$ e $u \in (0, 1.0)$, respectivamente. Podemos dizer que se o modelo está perfeitamente especificado, então, $t \to 1 \Rightarrow u \to 1$, poderíamos escolher $\alpha = 1$ e, consequentemente, $BV \to 1$. Agora suponha outro caso extremo. Quando o erro de predição do modelo tende a infinito, teríamos $t \to -\infty$ e, mantendo uma escolha equivalente para a constante de combinação entre o máximo e o mínimo, teríamos a escolha de $\alpha = 0$, $(1-\alpha) = 1$ e $BV \to -\infty$. Note então que podemos definir α considerando a distribuição de probabilidades do mínimo.

Por definição, temos que a função de distribuição acumulada de T é dada por

$$F(t) \quad = \quad P(T \leq t), \quad \text{com } t \in (-\infty, 1). \tag{3.14}$$

Temos também que a função $F(t)$ satisfaz as seguintes propriedades:

- $0 \leq F(t) \leq 1$, para todo $t \in (-\infty, 1)$;
- $F(t)$ é não decrescente;
- $\lim_{t \to -\infty} F(t) = 0$;
- $\lim_{t \to 1} F(t) = 1$.

Assim, nossa proposta é considerar que em (3.13), α seja obtido a partir de

$$\alpha \quad = \quad P(T \leq t). \tag{3.15}$$

Podemos notar que BV torna-se um critério de seleção rigoroso, cujo valor é penalizado com pesos maiores para o mínimo das estatísticas, sempre que este decresce. No entanto, como não conhecemos a distribuição de probabilidades de T, não dispomos de uma expressão para $F(t) = P(T \leq t)$. Uma solução para isso é estimarmos $F(t)$ a partir da distribuição empírica, utilizando o método *bootstrap* paramétrico.

3.4.1 Esquema Bootstrap

O método *Bootstrap*, introduzido por Efron (1979), é uma técnica estatística computacionalmente intensiva de reamostragem que tem como finalidade fazer inferências sobre distribuições de variáveis aleatórias. A ideia básica do método consiste em obter réplicas dos dados, realizando reamostragens de tamanhos iguais e com reposição, ou gerando os dados com base em um modelo paramétrico estimado a partir da amostra original.

Assim, o uso desse método nos permite obter a distribuição de probabilidade empírica de uma determinada estatística, quando dispomos ou não do modelo gerador dos dados, e também é possível estimar o viés e o erro padrão dessa estatística. Outra vantagem do uso desse método é a obtenção de intervalos de confiança e realização de testes de hipóteses. Para mais detalhes ver Efron e Tibishiran (1993). Existem diversos tipos de métodos *bootstrap*, dentre eles, o "não-paramétrico", geralmente utilizado quando não dispomos da informação necessária sobre a distribuição dos dados, e o método "paramétrico", usado sempre que dispomos de toda informação necessária sobre a distribuição dos dados. Nesse trabalho utilizamos o método *bootstrap* paramétrico, cuja metodologia é dada a seguir.

Considere a amostra $y_1, \ldots, y_n$ de variáveis aleatórias independentes, em que cada y_t, $t = 1, \ldots, n$, segue distribuição beta com função densidade de probabilidade dada por (2.2), com média μ_t e precisão ϕ_t, ou seja, $y_t \sim B(\mu_t, \phi_t)$. As amostras *bootstrap* são geradas a partir de $B(\widehat{\mu}_t, \widehat{\phi}_t)$, em que $\widehat{\mu}_t$ e $\widehat{\phi}_t$ são os estimadores de máxima verossimilhança de μ_t e ϕ_t, respectivamente, obtidos através da amostra original.

Dessa forma, seja T nossa estatística de interesse e t_{obs} o valor que ela assume para a amostra $y_1, \ldots, y_n$. Podemos obter a estimativa *bootstrap* do valor da função de distribuição acumulada de T, $F(t_{obs})$, realizando o seguinte procedimento:

1. Geramos B amostras de tamanho n da distribuição $B(\widehat{\mu}_t, \widehat{\phi}_t)$.

2. Para cada amostra *bootstrap*, calculamos o valor da estatística de interesse, obtendo, respectivamente, $T_1^* = t_1^*, T_2^* = t_2^*, \ldots, T_B^* = t_B^*$.

3. Estimamos $F(t_{obs})$ a partir da função de distribuição empírica dos t_b^*, com $b = 1, \ldots, B$, digamos $\widehat{F}(t_{obs})$, de modo que

$$\widehat{F}(t_{obs}) = \frac{1}{B} \sum_{b=1}^{B} I(t_b^* \leq t_{obs}),$$

em que $I(\cdot)$ corresponde à função indicadora.

Para o nosso caso, dado que $T = \min\{P^2, P_c^2, R_{FC}^2, R_{FC_c}^2, R_{LR}^2, R_{LR_c}^2\}$, propomos que $\widehat{\alpha} = \widehat{F}(t_{obs})$. Vale notar que para cada amostra $y_1, \ldots, y_n$ obtemos diferentes valores de $\hat{\alpha}$, logo, dispomos de uma variável aleatória, cuja distribuição aproximada pode ser obtida, por exemplo, pelo método *bootstrap* duplo, implicando na possibilidade de construções de intervalos de confiança e testes de hipóteses. Para detalhes sobre o método *bootstrap* duplo, ver Efron (1983), Hall e Martin (1988) ou Beran (1988).

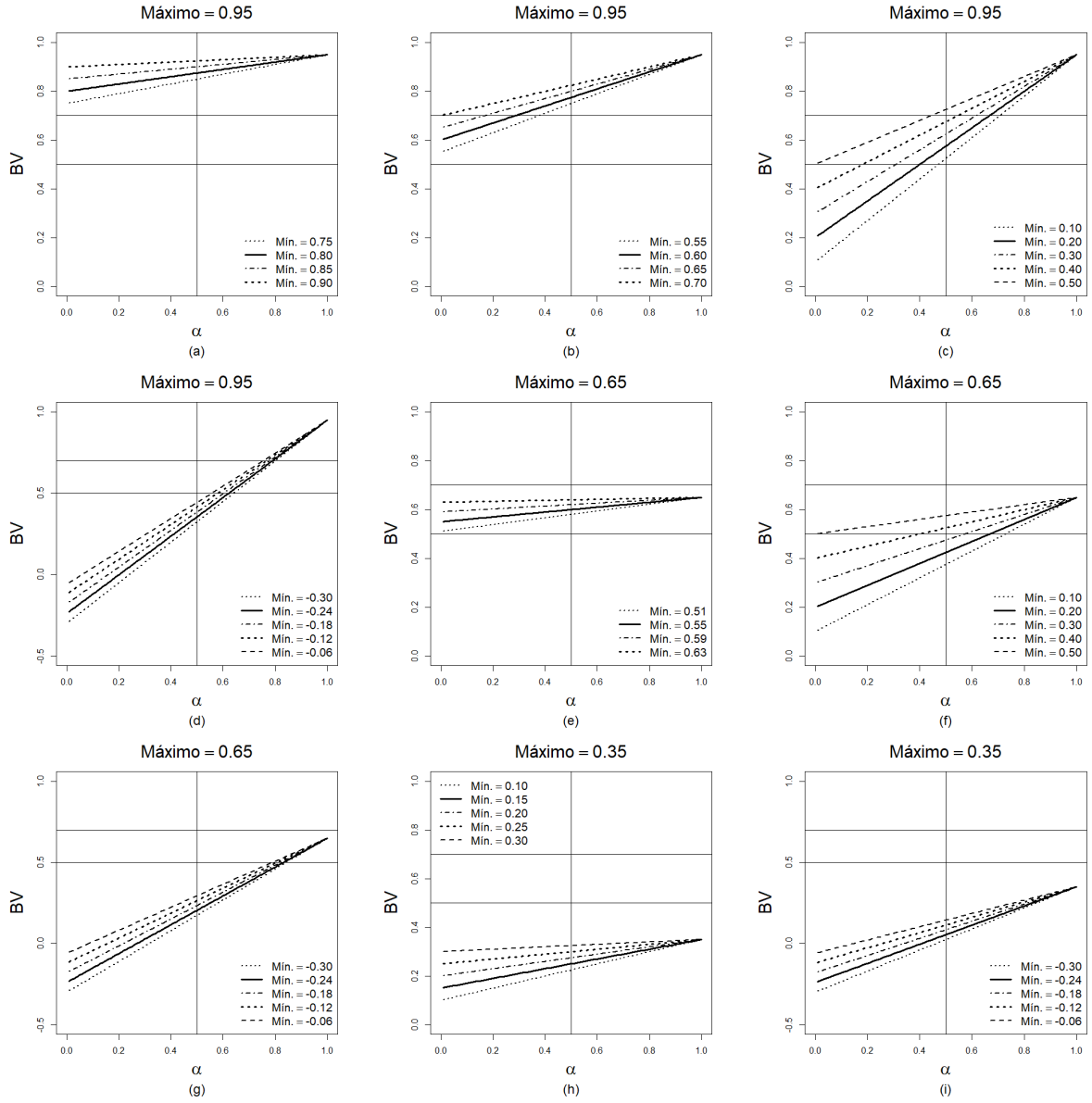
Nossa proposta é com base nos dados e no método *bootstrap*, descrito anteriormente, estimar α, e assim calcular a estatística BV. Os diferentes valores que $\hat{\alpha}$ pode assumir, para modelos bem e mal especificados, serão estudados no próximo capítulo a partir das simulações de Monte Carlo. No entanto, buscando verificar o comportamento da estatística BV para possíveis valores das estatísticas P^2, P_c^2, R_{FC}^2, $R_{FC_c}^2$, R_{LR}^2 e $R_{LR_c}^2$ e de α, foram fixados alguns valores para o mínimo T e para o máximo U, e consideramos valores de α variando no intervalo $(0,1)$, com incremento de 0.01.

Na Figura 2 encontramos os gráficos com os valores de BV *versus* α para possíveis cenários de T e U. Nesses gráficos temos que a primeira linha horizontal corresponde a $BV = 0.5$ e a segunda, a $BV = 0.7$. Iremos considerar aqui que valores de T, U e BV entre 0.5 e 0.7 serão classificados como medianos, uma vez que valores das estatísticas nesse intervalo geralmente indicam ajustes razoáveis dos modelos, havendo no entanto algum problema de especificação. Para valores das estatísticas acima de 0.7, classificamos como altos e menores que 0.5 como baixos. De modo que obtemos indícios de modelos bem especificados no primeiro caso e mal especificados no segundo. Além disso, a linha na vertical corresponde a $\alpha = 0.5$. Para a Figura 2a fixamos o valor do máximo U em 0.95 e consideramos altos valores para o mínimo T. Neste contexto, temos que qualquer valor de α resulta em altos valores para BV. No entanto, uma vez que para esse caso, o mínimo claramente não tende para menos infinito, e também não está tão próximo de um, não faz sentido tomar $\alpha \approx 0$ ou $\alpha \approx 1$ (até maior que o máximo), e atribuir pesos muito desiguais ao mínimo e ao máximo para obter o valor de BV. Então, o valor de α mais aceitável é $\alpha \approx 0.5$. Logo, quando $\alpha = 0.5$ temos que $BV \geq 0.8$, compatível com todos os valores admitidos para o mínimo em combinação com o máximo igual a 0.95. Essa figura gera uma primeira evidência que $\alpha \approx 1$ e $BV \approx 1$ não parece plausível, mesmo para modelos adequadamente ajustados.

Já pela Figura 2b, em que mantivemos o valor do máximo e atribuímos valores entre 0.55 e 0.7 ao mínimo, vemos que α próximo de 0.5 eleva consideravelmente os valores de BV, como por exemplo, para o mínimo igual a 0.55, notamos que nesse caso BV é aproximadamente 0.75, o que não parece plausível, devido o mínimo ser consideravelmente menor que 0.75. Aqui é importante ressaltar que os cenários abordados pela Figura 2b representam modelos com algum problema de especificação. De modo que nestes casos, os modelos apresentam altos poderes de predição e proporções de variabilidade explicada consideravelmente menores, ou o contrário. Ainda no cenário com mínimo igual a 0.55, considerando α entre 0.3 e 0.4, obtemos para essa situação valores para BV em torno de 0.65, o que parece aceitável, de acordo com o cenário abordado. Então, $0.3 < \alpha < 0.4$ resulta em valores de BV bem menores que o máximo, tornando então BV capaz de indicar que o modelo ajustado apresenta diferença em seus poderes

de predição e explicação da variância da resposta.

Figura 2 – Gráficos *BV versus* α, para diferentes valores do mínimo T e do maximo U.



Fonte: Autoria própria.

Observando as Figuras 2c e 2d vemos que, quando os valores atribuídos ao mínimo diminuem, então, os valores de α que proporcionam *BV*'s que assimilam significativamente os baixos valores dos mínimos, sem desconsiderar o valor do máximo, são cada vez menores. Ou seja, os valores de α que fazem sentido são tipicamente menores que 0.2. Dessa forma, os valores para *BV* são fortemente influenciados pelo valor do mínimo das estatísticas de adequabilidade de ajuste. Adicionalmente, notamos que $0.10 < \alpha < 0.20$ é indicação de má especificação do modelo ou valores de *BV* obtidos com pesos bem maiores para o mínimo, quando este indica má adequabilidade do modelo.

Aqui é necessária uma observação muito importante. Notemos que na Figura 2d existe uma forte discrepância entre o máximo e o mínimo, representando então cenários que podem ser

overfitting ou *undefitting*. Aqui, vemos que à medida de α se aproxima de um, BV também se aproxima de um, mesmo quando o mínimo, por exemplo, é igual a -0.30. Isso demonstra que $\alpha \approx 1$ pode indicar boa adequação de um modelo mal especificado. Logo, esse gráfico demonstra enfaticamente que a combinação de valores de $BV \approx 1$ e $\alpha \approx 1$ simultaneamente, pode resultar em má seleção de modelos. A Figura 2c corrobora com essa assertiva. Como, já observamos em cenários de modelos bem especificados (Figura 2a) que $\alpha = 0.5$ resulta valores de BV que indicam adequabilidade dos modelos e ainda representam bem a combinação entre o mínimo e máximo. Uma vez, $\alpha \approx 1 \Rightarrow BV \approx 1$, acreditamos que o padrão de α e BV altos simultaneamente não deva ser considerado como indicação de bons ajustes de modelos.

Nas Figuras 2e, 2f e 2g foi estabelecido o valor 0.65 para o máximo. Vemos pela Figura 2e que, quando o mínimo está próximo desse valor, obtemos sempre valores medianos para BV, sob qualquer α. Entretanto, como nesses cenários os valores do máximo e mínimo são bastante próximos, provavelmente, o poder de predição e de explicação da variabilidade são similarmente baixos. Desse modo, não faz sentido considerar $\alpha \approx 0$ ou $\alpha \approx 1$, atribuindo pesos desiguais ao mínimo e ao máximo para obter o valor de BV. Então, o valor para α mais aceitável é $\alpha \approx 0.5$. Observando as Figuras 2f e 2g e sabendo que valores de α abaixo de 0.5 fornecem pesos maiores ao mínimo, vemos que para tais cenários isso resulta baixos valores para BV, indicando assim que o modelo ajustado não estaria adequado, o que parece aceitável, considerando os cenários em questão, em que, além do mínimo ser baixo, o máximo também não assume um valor tão elevado.

Por fim, avaliando as Figuras 2h e 2i, observamos que quando o máximo das estatísticas for inferior a 0.5, então, BV apresentará valores baixos ou até negativos, para qualquer que seja o valor de α. No entanto, definitivamente $\alpha \approx 0$ é o que conduziria a valores de BV compatíveis com os cenários de má adequabilidade dos modelos, em termos de predição e explicação da variabilidade da variável resposta.

Logo, sempre que todas as estatísticas indicaram um mal ajuste, BV conduziu à mesma conclusão, para $\alpha \approx 0$ quando o mínimo for muito distante do máximo. Ou seja, quando o poder de explicação da variabilidade for bastante inferior ao poder de predição, ou caso contrário. E também para $\alpha \approx 0.5$ quando o modelo apresenta poderes similares e baixos. Adicionalmente, observamos que $\alpha \approx 1$ não deve ser considerado para indicação de modelos bem especificados, pois podem produzir valores de BV que não representam bem os cenários do máximo e mínimo. Logo, valores de $\alpha < 0.4$ ou $\alpha \approx 1$ indicam falhas na especificação do modelo ajustado. Por fim, $\alpha = 0.5$ ou $\alpha \approx 0.5$ apontam similaridade do modelo para predizer valores da variável resposta e explicar a varição desta. Portanto, esse valor de α associado a um alto valor de BV seria um possível padrão para modelos bem especificados, com altos poderes de predição e elevadas proporções de variabilidade sendo explicada.

4 AVALIAÇÃO NUMÉRICA NO MODELO BETA COM DISPERSÃO VARIÁVEL

4.1 INTRODUÇÃO

O objetivo desse capítulo é investigar o comportamento das estatísticas BV e $\widehat{\alpha}$ em avaliar corretamente o desempenho do modelo postulado e estimado com base nos dados. Adicionalmente, serão estudados os comportamentos das estatísticas do tipo P^2 e do tipo R^2. Para isso, foram realizadas simulações de Monte Carlo com $MC = 5000$ réplicas, considerando alguns cenários para o modelo de regressão beta, como o modelo linear com dispersão variável e também o modelo não linear, tanto no submodelo da média quanto no submodelo da precisão. Adicionalmente, diferentes tamanhos de amostra, graus de heteroscedasticidade, valores da precisão da resposta e intervalos de valores para μ também foram considerados. Finalmente, o método *bootstrap*, utilizado para estimação de α, baseou-se em $B = 500$ e $B = 250$ réplicas.

4.2 AVALIAÇÃO NUMÉRICA NO MODELO BETA LINEAR

Consideramos inicialmente o modelo de regressão beta com dispersão variável. De modo que, inicialmente, buscamos observar o desempenho das estatísticas de adequabilidade de ajuste sob um modelo corretamente especificado. Dessa forma, seja $y_1, y_2, \ldots, y_n$ uma amostra de n variáveis aleatórias independentes, cuja média e precisão de y_t, com $t = 1, \ldots, n$, satisfazem respectivamente as seguintes relações funcionais:

$$g(\mu_t) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} \qquad \text{e} \qquad \log(\phi_t) = \gamma_1 + \gamma_2 z_{t2} + \gamma_3 z_{t3}. \tag{4.1}$$

É conhecido o fato que, para modelos lineares clássicos, o grau de heterocedasticidade pode ser medido por $\lambda = \max \sigma_t^2 / \min \sigma_t^2$; com $t = 1, \ldots, n$ e sendo σ_t^2 a variância da resposta. De modo que sob homocedasticidade, $\lambda = 1$, e sob heteroscedasticidade $\lambda > 1$. Dessa forma, para os modelos de regressão beta consideraremos em nosso estudo

$$\lambda \;=\; \frac{\max\limits_{t=1,\ldots,n} \mathrm{Var}(y_t)}{\min\limits_{t=1,\ldots,n} \mathrm{Var}(y_t)}, \tag{4.2}$$

sendo $\mathrm{Var}(y_t)$ denotado por (2.3). Vale notar que inicialmente consideramos aqui três covariáveis no submodelo da média e duas no submodelo da precisão, de modo que após terem sido geradas, permaneceram fixas durante toda a simulação.

Os tamanhos amostrais aqui considerados foram $n = 20$, 40, 80 e 120, em que foram geradas 20 observações das covariáveis x_{ti} e z_{tj}, com $i = 2, 3, 4$ e $j = 2, 3$, e em seguida, cada uma das amostras são replicadas até obter-se os valores amostrais desejados. Esse procedimento de replicação garante que tanto a média da variável resposta se mantenha em uma mesma faixa de valores, quanto o grau de heteroscedasticidade se mantenha constante, independentemente do tamanho amostral.

Além disso, buscando ainda avaliar o desempenho das medidas sob diferentes cenários para μ_t, definimos os valores dos parâmetros do modelo para média, ou seja, de β_1, β_2, β_3 e β_4 e também a distribuição das covariáveis x_{ti}, com $i = 2,\ldots,4$, de modo que os valores de μ_t estejam distribuídos em três diferentes intervalos: próximos do zero, próximos de um e centrais (ou seja, valores em torno de 0.5). A partir de estudos de simulações, Espinheira et. al (2019) mostra que para cada cenário de μ_t existem diferentes funções de ligação com melhor adequabilidade. Dessa forma, para cada cenário da média foram assumidas diferentes funções $g(\cdot)$, de maneira que consideramos a função log-log para μ_t perto do zero, complemento log-log para μ_t perto de um e logit para valores de μ_t centrais, com $t = 1,\ldots,n$. Todas essas funções estão definidas no capítulo 2. Também foram combinados os valores dos parâmetros γ's e a distribuição de z_{tj}, com $j = 2,3$, de maneira que foram obtidos os graus de heteroscedasticidade 30, 50 e 100.

Desenvolvemos então a simulação de Monte Carlo para calcular os valores de $\widehat{\alpha}$ e BV, quando o modelo ajustado está corretamente especificado, tendo como base a seguinte metodologia:

1. fixamos os valores dos parâmetros β_i e γ_j, com $i = 1,\ldots,4$ e $j = 1,2,3$;

2. geramos x_{ti} e z_{tj} com $t = 1,\ldots,20$, $i = 2,3,4$ e $j = 2,3$, (replicamos os valores das covariáveis de acordo com o tamanho amostral considerado);

3. obtemos $\mu_t = g^{-1}(\eta_{1t})$ e $\phi_t = h^{-1}(\eta_{2t})$, em que $\eta_{1t} = \sum_{i=1}^{4} x_{ti}\beta_i$ e $\eta_{2t} = \sum_{j=1}^{3} z_{tj}\gamma_j$;

4. iniciamos o laço de Monte Carlo, de modo que para cada réplica executamos as seguintes tarefas:

   - geramos uma amostra aleatória $y_1, y_2, \ldots, y_n$ de n variáveis aleatórias independentes, em que cada $y_t \sim B(\mu_t, \phi_t)$ e obtemos os estimadores $\widehat{\mu}_t$ e $\widehat{\phi}_t$ pelo método BFGS;
   - calculamos as estatísticas P^2, P_c^2, R_{FC}^2, $R_{FC_c}^2$, R_{LR}^2 e $R_{LR_c}^2$;
   - obtemos os valores de $T = \min\{P^2, P_c^2, R_{FC}^2, R_{FC_c}^2, R_{LR}^2, R_{LR_c}^2\}$ e $U = \max\{P^2, P_c^2, R_{FC}^2, R_{FC_c}^2, R_{LR}^2, R_{LR_c}^2\}$;
   - iniciamos o laço *bootstrap*, em que para cada réplica ($b = 1,\ldots,B$, sendo $B = 500$) realizamos as seguintes tarefas:

     (i) geramos uma amostra *bootstrap* $(y_{b1}^*, y_{b2}^*, \ldots, y_{bn}^*)$ de n variáveis aleatórias independentes, em que cada $y_{bt}^* \sim B(\widehat{\mu}_t, \widehat{\phi}_t)$;

     (ii) calculamos as estatísticas P_b^{2*}, P_{cb}^{2*}, R_{FCb}^{2*}, $R_{FC_cb}^{2*}$, R_{LRb}^{2*} e $R_{LR_cb}^{2*}$;

     (iii) obtemos $T_b^* = \min\{P_b^{2*}, P_{cb}^{2*}, R_{FCb}^{2*}, R_{FC_cb}^{2*}, R_{LRb}^{2*}, R_{LR_cb}^{2*}\}$;

   - sejam t_{obs} e t_b^* os valores obtidos para T e T_b^*, respectivamente, tal que $\widehat{\alpha}$ é obtido a partir de

$$\widehat{\alpha} = \frac{1}{B}\sum_{b=1}^{B} I(t_b^* \leq t_{obs});$$

   - calculamos $BV = (1-\widehat{\alpha})T + \widehat{\alpha}U$;

5. calculamos a média dos valores de BV e das demais estatísticas P^2, P_c^2, R_{FC}^2, $R_{FC_c}^2$, R_{LR}^2 e $R_{LR_c}^2$, obtidos pelo laço Monte Carlo e também a média dos valores de $\widehat{\alpha}$.

Na Tabela 1 são apresentados os valores fixados para os parâmetros dos submodelos da média e da precisão, temos também os intervalos obtidos para μ_t, a função de ligação usada em cada cenário de μ_t, as distribuições assumidas para gerar os valores das covariáveis do modelo da precisão, e os respectivos graus de heteroscedasticidade. Além disso, para gerar os valores das covariáveis x_{ti}, com $t = 1,\ldots,n$ e $i = 2,3,4$, foi considerado que $x_{ti} \sim U(0,1)$.

Tabela 1 – Valores dos parâmetros β's e γ's, fixados para a análise no modelo beta linear com dispersão variável corretamente especificado, com os respectivos intervalos dos valores de μ_t, as funções de ligação para o modelo da média, as distribuições das covariáveis do modelo da precisão e os graus de heteroscedasticidade.

β_1	β_2	β_3	β_4	μ	$g(\cdot)$	γ_1	γ_2	γ_3	z_{t2}	z_{t3}	λ
						2.4	0.58	2.66	$U(0,1)$	$U(0,1)$	30
-1.95	1.0	1.2	0.5	$\mu_t \in (0.0193, 0.2517)$	Log-log	3.68	-1.0	2.85	$U(0,1)$	$U(-0.1, 0.9)$	50
						3.5	2.38	-2.0	$U(-0.13, 0.87)$	$U(0,1)$	100
						5.1	3.3	-2.6	$U(-0.25, 0.75)$	$U(0.15, 1.15)$	30
-2.48	1.5	1.7	3.1	$\mu_t \in (0.2650, 0.8702)$	Logit	3.1	1.3	3.67	$U(-0.1, 0.9)$	$U(-0.3, 0.7)$	50
						5.2	-4.3	0.36	$U(-0.4, 0.6)$	$U(-0.5, 0.5)$	100
						4.18	1.1	3.3	$U(-0.6, 0.6)$	$U(-0.5, 0.5)$	30
1.2	-0.6	-1.23	0.63	$\mu \in (0.6404, 0.9881)$	C-Log-Log	3.9	-1.1	3.22	$U(0,1)$	$U(-0.1, 0.9)$	50
						4.68	-2.55	1.52	$U(-0.4, 0.6)$	$U(-0.6, 0.4)$	100

Fonte: Autoria própria.

Na Tabela 2 temos os valores médios das estatísticas de adequabilidade de ajuste para o modelo linear beta com dispersão variável corretamente especificado. Com base nesses valores vemos que as estatísticas apresentam melhores resultados para $\mu_t \in (0.2650, 0.8702)$, isso provavelmente ocorre devido ao fato dos estimadores de máxima verossimilhança serem mais viesados nas fronteiras do espaço paramétrico. Além disso, o modelo beta apresentar melhor adequabilidade para dados centrais. Notamos que para esse cenário da média, tanto as estatísticas de predição, quanto as estatísticas de qualidade de ajuste, como também a estatística BV, os quais apresentam valores mais elevados, indicam que o modelo está bem especificado. Além disso, para este cenário de μ_t, vemos que a estatística $\hat{\alpha}$ apresenta sempre valores mais próximos de 0.5, indicando que o máximo e o mínimo das estatísticas de ajuste exercem pesos similares na obtenção de BV. Logo, acreditamos que para um modelo linear beta bem ajustado, com habilidades similares e elevadas em explicar a variabilidade da resposta e predizer bons valores, é geralmente esperado que o valor de BV seja elevado (próximo de um) e $\hat{\alpha}$ esteja em torno de 0.5. Podemos observar também que para $n = 20$ e $n = 40$, de modo geral, quando o grau de heteroscedasticidade aumenta as medidas de ajuste tendem a diminuir, especialmente as medidas do tipo R^2, o que é compreensível, uma vez que as estatísticas R^2 estão associadas com a quantidade de variação da resposta explicada pelo modelo, e altos valores de heteroscedasticidade implicam alta dispersão da variância, nesse caso, existe uma maior dificuldade em estimar corretamente os parâmetros. Esse fato é bem representado por BV e $\hat{\alpha}$, pois observamos que os valores dessas medidas tendem a diminuir quando a heteroscedasticidade aumenta, e no caso

de $\hat{\alpha}$, vemos que seus valores tornam-se mais distante de 0.5. Quando o tamanho da amostra é maior, ou seja, $n = 80$ e $n = 120$, vemos que geralmente, os valores de $\hat{\alpha}$ se aproximam de 0.5 quando o grau de heteroscedasticidade é baixo, indicando então que o modelo está bem ajustado. Além disso, notamos que as medidas de ajuste do tipo P^2 e R^2 tornam-se mais próximas, o que aponta um equilíbrio entre o poder de predição e a proporção de variabilidade explicada (qualidade do ajuste).

Na Tabela 3 vemos os intervalos dos valores estabelecidos para ϕ_t, $\text{Var}(y_t)$ e λ, junto com suas respectivas médias das estimativas, além das médias das estimativas para μ_t, para cada tamanho amostral e cenário de μ_t mencionados anteriormente, obtidas a partir das simulações de Monte Carlo. Observando então tais valores vemos que eles complementam as conclusões feitas anteriormente, uma vez que tanto a precisão quanto a variância e também o grau de heteroscedasticidade são mal estimados (estão superestimados) quando o tamanho amostral é baixo e o grau de heteroscedasticidade verdadeiro é alto. Logo, embora o modelo ajustado esteja corretamente ajustado, para $n = 20$ e $n = 40$ notamos uma maior dificuldade em estimar corretamente o modelo, e esse fato é bem representado pelas medidas de ajuste BV e $\hat{\alpha}$, em que BV tende a assumir valores não tão elevados quanto a maioria das demais medidas de ajuste, e $\hat{\alpha}$ não atinge exatamente o valor 0.5. Ainda de acordo com a Tabela 3, vemos a partir de $n = 80$, que as estimativas da precisão, da variância e do grau de heteroscedasticidade melhoram quando o tamanho amostral aumenta, esse fato é bem representado por $\hat{\alpha}$ na Tabela 2, cujos valores também se aproximam de 0.5 quando o tamanho da amostra cresce.

Avaliando os gráficos de *Boxplot* das medidas de adequabilidade de ajuste P^2, P_c^2, R_{FC}^2, $R_{FC_c}^2$, R_{LR}^2, $R_{LR_c}^2$ e BV, apresentados pelas Figuras 3, 4 e 5, vemos que, quando o tamanho amostral aumenta as estatísticas corrigidas tendem a apresentar valores mais próximos de suas versões não corrigidas e mais distantes das demais medidas. Além disso, vemos que geralmente, as estatísticas P^2, P_c^2, R_{LR}^2 e $R_{LR_c}^2$ apresentam menores dispersões. Outra observação, é que estatística BV apresenta valores distribuídos sempre de maneira centralizada entre as distribuições das demais estatísticas de ajuste. Vemos também que para $n = 20$ e $\lambda = 100$, as distribuições de $R_{LR_c}^2$ e de BV se assemelham. Já quando comparada com R_{FC}^2 e/ou $R_{FC_c}^2$ notamos que BV apresenta valores sempre maiores que essas estatísticas.

Tabela 2 – Valores médios das estatísticas de avaliação de desempenho no modelo beta linear com dispersão variável corretamente especificado, com $g(\cdot)$ adequada à cada cenário de μ_t e $h(\cdot)$ log.

	$g(\cdot)$: Log-Log $\mu_t \in (0.0193, 0.2517)$			$g(\cdot)$: Logit $\mu_t \in (0.2650, 0.8702)$			$g(\cdot)$: C-Log-Log $\mu_t \in (0.6404, 0.9881)$			
λ	30	50	100	30	50	100	30	50	100	n
P^2	0.9539	0.9577	0.9346	0.9776	0.9867	0.9851	0.9765	0.9778	0.9650	20
P_c^2	0.9327	0.9381	0.9044	0.9672	0.9806	0.9782	0.9656	0.9676	0.9488	
R_{FC}^2	0.7743	0.7729	0.7317	0.9046	0.8795	0.8689	0.8222	0.8281	0.7825	
$R_{FC_c}^2$	0.6701	0.6681	0.6079	0.8606	0.8238	0.8084	0.7402	0.7487	0.6821	
R_{LR}^2	0.8772	0.8892	0.8547	0.9654	0.9497	0.9401	0.9298	0.9299	0.8953	
$R_{LR_c}^2$	0.7800	0.8015	0.7396	0.9381	0.9098	0.8927	0.8742	0.8743	0.8123	
BV	0.7919	0.7939	0.7397	0.9139	0.9003	0.9006	0.8463	0.8536	0.8056	
$\hat{\alpha}$	0.4413	0.4460	0.4141	0.4857	0.4761	0.5122	0.4739	0.4728	0.4504	
P^2	0.9200	0.9207	0.8807	0.9628	0.9776	0.9747	0.9603	0.9573	0.9330	40
P_c^2	0.9054	0.9062	0.8590	0.9560	0.9735	0.9700	0.9530	0.9495	0.9208	
R_{FC}^2	0.7614	0.7564	0.7223	0.8966	0.8740	0.8627	0.8093	0.8165	0.7672	
$R_{FC_c}^2$	0.7180	0.7121	0.6718	0.8778	0.8511	0.8377	0.7747	0.7832	0.7248	
R_{LR}^2	0.8517	0.8634	0.8234	0.9572	0.9367	0.9229	0.9158	0.9148	0.8739	
$R_{LR_c}^2$	0.8110	0.8259	0.7749	0.9454	0.9193	0.9018	0.8927	0.8914	0.8392	
BV	0.7990	0.7971	0.7541	0.9134	0.9045	0.8971	0.8539	0.8578	0.8104	
$\hat{\alpha}$	0.4178	0.4250	0.4053	0.4594	0.4382	0.4407	0.4555	0.4486	0.4327	
P^2	0.9015	0.9006	0.8552	0.9542	0.9724	0.9675	0.9510	0.9450	0.9159	80
P_c^2	0.8934	0.8924	0.8432	0.9505	0.9701	0.9649	0.9470	0.9404	0.9090	
R_{FC}^2	0.7516	0.7483	0.7117	0.8901	0.8680	0.8583	0.8031	0.8092	0.7578	
$R_{FC_c}^2$	0.7312	0.7277	0.6880	0.8811	0.8572	0.8467	0.7870	0.7935	0.7379	
R_{LR}^2	0.8401	0.8543	0.8107	0.9535	0.9309	0.9158	0.9107	0.9088	0.8672	
$R_{LR_c}^2$	0.8211	0.8370	0.7882	0.9480	0.9227	0.9058	0.9000	0.8979	0.8513	
BV	0.8031	0.8010	0.7573	0.9124	0.9067	0.8998	0.8591	0.8597	0.8139	
$\hat{\alpha}$	0.4374	0.4397	0.4281	0.4660	0.4455	0.4468	0.4652	0.4572	0.4489	
P^2	0.8958	0.8929	0.8450	0.9509	0.9708	0.9648	0.9474	0.9406	0.9093	120
P_c^2	0.8903	0.8872	0.8367	0.9482	0.9692	0.9629	0.9446	0.9375	0.9044	
R_{FC}^2	0.7504	0.7450	0.7087	0.8888	0.8661	0.8565	0.7997	0.8059	0.7531	
$R_{FC_c}^2$	0.7371	0.7314	0.6932	0.8829	0.8590	0.8489	0.7890	0.7956	0.7400	
R_{LR}^2	0.8381	0.8512	0.8062	0.9523	0.9294	0.9133	0.9088	0.9069	0.8649	
$R_{LR_c}^2$	0.8258	0.8399	0.7915	0.9487	0.9240	0.9067	0.9019	0.8998	0.8546	
BV	0.8056	0.8013	0.7580	0.9134	0.9079	0.9007	0.8600	0.8605	0.8127	
$\hat{\alpha}$	0.4456	0.4477	0.4389	0.4738	0.4531	0.4546	0.4689	0.4671	0.4497	

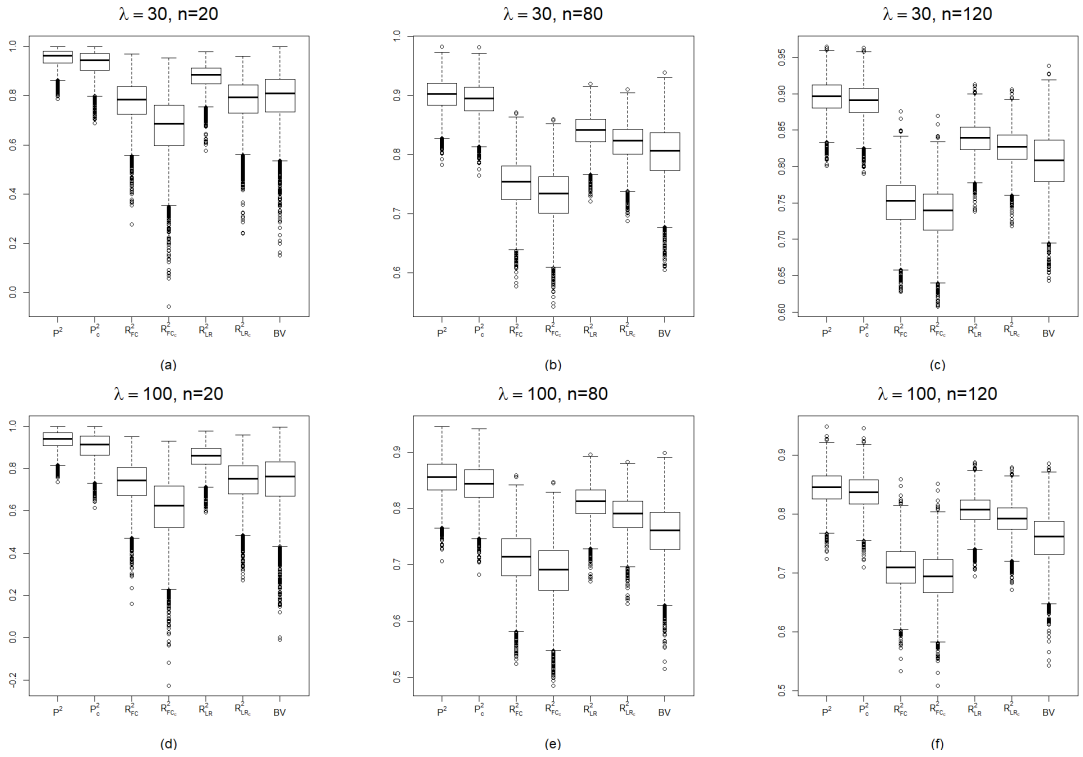
Fonte: Autoria própria.

Tabela 3 – Intervalos de valores de ϕ_t e da $\text{Var}(y_t)$, e valores de λ, juntamente com suas respectivas médias das estimativas, e também as médias das estimativas dos intervalos de μ_t, no modelo beta linear com dispersão variável corretamente especificado, usando $g(\cdot)$ log-log para $\mu_t \in (0.0193, 0.2517)$, $g(\cdot)$ logit para $\mu_t \in (0.2650, 0.8702)$, $g(\cdot)$ c-log-log para $\mu_t \in (0.6404, 0.9881)$ e $h(.)$ log.

$\hat{\mu}_t$	ϕ_t	$\hat{\phi}_t$	$\text{Var}(y_t)$	$\widehat{\text{Var}}(y_t)$	λ	$\hat{\lambda}$	n
$\mu_t \in (0.0193, 0.2517)$							
(0.0116, 0.2958)	(17.082, 250.09)	(41.879, 2928.6)	(0.0002, 0.0076)	(0.0002, 0.0092)	30	56.1	20
(0.0073, 0.2541)	(12.959, 246.32)	(36.803, 2770.7)	(0.0002, 0.0122)	(0.0002, 0.0154)	50	82.8	
(0.0250, 0.2785)	(9.838, 215.61)	(37.323, 1892.3)	(0.0002, 0.0152)	(0.0001, 0.0208)	100	164.2	
(0.0213, 0.2596)	(17.082, 250.09)	(24.351, 412.43)	(0.0002, 0.0076)	(0.0002, 0.0077)	30	40.1	40
(0.0195, 0.2579)	(12.959, 246.32)	(15.588, 393.57)	(0.0002, 0.0122)	(0.0002, 0.0127)	50	61.6	
(0.0194, 0.2525)	(9.838, 215.61)	(13.716, 352.02)	(0.0002, 0.0152)	(0.0001, 0.0157)	100	119.8	
(0.0165, 0.2694)	(17.082, 250.09)	(19.574, 303.96)	(0.0002, 0.0076)	(0.0002, 0.0076)	30	36.3	80
(0.0189, 0.2586)	(12.959, 246.32)	(14.150, 301.46)	(0.0002, 0.0122)	(0.0002, 0.0123)	50	55.2	
(0.0183, 0.2485)	(9.838, 215.61)	(11.260, 266.37)	(0.0002, 0.0152)	(0.0001, 0.0153)	100	109.7	
(0.0206, 0.2582)	(17.082, 250.09)	(18.866, 284.99)	(0.0002, 0.0076)	(0.0002, 0.0076)	30	35.2	120
(0.0260, 0.2460)	(12.959, 246.32)	(13.691, 279.57)	(0.0002, 0.0122)	(0.0002, 0.0123)	50	53.3	
(0.0224, 0.2895)	(9.838, 215.61)	(10.763, 245.11)	(0.0002, 0.0152)	(0.0001, 0.0152)	100	105.7	
$\mu_t \in (0.2650, 0.8702)$							
(0.2791, 0.8655)	(15.272, 1011.19)	(66.235, 9221.8)	(0.0002, 0.0072)	(0.0002, 0.0114)	30	52.6	20
(0.2369, 0.8469)	(12.882, 732.38)	(41.131, 11706.0)	(0.0002, 0.0124)	(0.0002, 0.0183)	50	111.4	
(0.2884, 0.8843)	(14.299, 727.37)	(25.460, 78062.0)	(0.0002, 0.0158)	(0.0002, 0.0240)	100	146.9	
(0.2750, 0.8659)	(15.272, 1011.19)	(19.850, 1777.6)	(0.0002, 0.0072)	(0.0002, 0.0077)	30	39.7	40
(0.2600, 0.8677)	(12.882, 732.38)	(17.232, 1280.5)	(0.0002, 0.0124)	(0.0002, 0.0130)	50	65.5	
(0.2825, 0.8720)	(14.299, 727.37)	(17.181, 1622.1)	(0.0002, 0.0158)	(0.0001, 0.0169)	100	145.3	
(0.2575, 0.8707)	(15.272, 1011.19)	(16.843, 1294.6)	(0.0002, 0.0072)	(0.0002, 0.0072)	30	34.1	80
(0.2622, 0.8736)	(12.882, 732.38)	(14.534, 915.34)	(0.0002, 0.0124)	(0.0002, 0.0126)	50	56.2	
(0.2685, 0.8706)	(14.299, 727.37)	(15.479, 994.55)	(0.0002, 0.0158)	(0.0001, 0.0161)	100	119.6	
(0.2717, 0.8599)	(15.272, 1011.19)	(16.434, 1177.5)	(0.0002, 0.0072)	(0.0002, 0.0071)	30	31.8	120
(0.2599, 0.8769)	(12.882, 732.38)	(13.901, 852.60)	(0.0002, 0.0124)	(0.0002, 0.0125)	50	54.7	
(0.2549, 0.8642)	(14.299, 727.37)	(15.072, 883.37)	(0.0002, 0.0158)	(0.0001, 0.0160)	100	112.54	
$\mu_t \in (0.6404, 0.9881)$							
(0.6665, 0.9787)	(14.824, 535.26)	(42.978, 7570.7)	(0.0002, 0.0073)	(0.0002, 0.0085)	30	53.2	20
(0.5835, 0.9924)	(14.369, 397.15)	(25.092, 5197.0)	(0.0002, 0.0125)	(0.0002, 0.0159)	50	74.6	
(0.6046, 0.9687)	(11.282, 224.82)	(19.878, 10098.0)	(0.0002, 0.0156)	(0.0002, 0.0203)	100	109.7	
(0.6290, 0.9876)	(14.824, 535.26)	(19.336, 927.27)	(0.0002, 0.0073)	(0.0002, 0.0072)	30	37.4	40
(0.6588, 0.9828)	(14.369, 397.15)	(17.215, 657.57)	(0.0002, 0.0125)	(0.0002, 0.0131)	50	58.1	
(0.6626, 0.9801)	(11.282, 224.82)	(13.528, 429.28)	(0.0002, 0.0156)	(0.0001, 0.0163)	100	115.0	
(0.6251, 0.9914)	(14.824, 535.26)	(16.564, 676.49)	(0.0002, 0.0073)	(0.0002, 0.0070)	30	32.7	80
(0.6381, 0.9900)	(14.369, 397.15)	(15.774, 486.88)	(0.0002, 0.0125)	(0.0002, 0.0126)	50	53.6	
(0.6413, 0.9867)	(11.282, 224.82)	(12.288, 294.95)	(0.0002, 0.0156)	(0.0001, 0.0158)	100	107.3	
(0.6403, 0.9813)	(14.824, 535.26)	(15.928, 621.32)	(0.0002, 0.0073)	(0.0002, 0.0071)	30	31.9	120
(0.6492, 0.9869)	(14.369, 397.15)	(15.164, 451.98)	(0.0002, 0.0125)	(0.0002, 0.0126)	50	52.8	
(0.6204, 0.9836)	(11.282, 224.82)	(11.875, 265.84)	(0.0002, 0.0156)	(0.0001, 0.0158)	100	105.3	

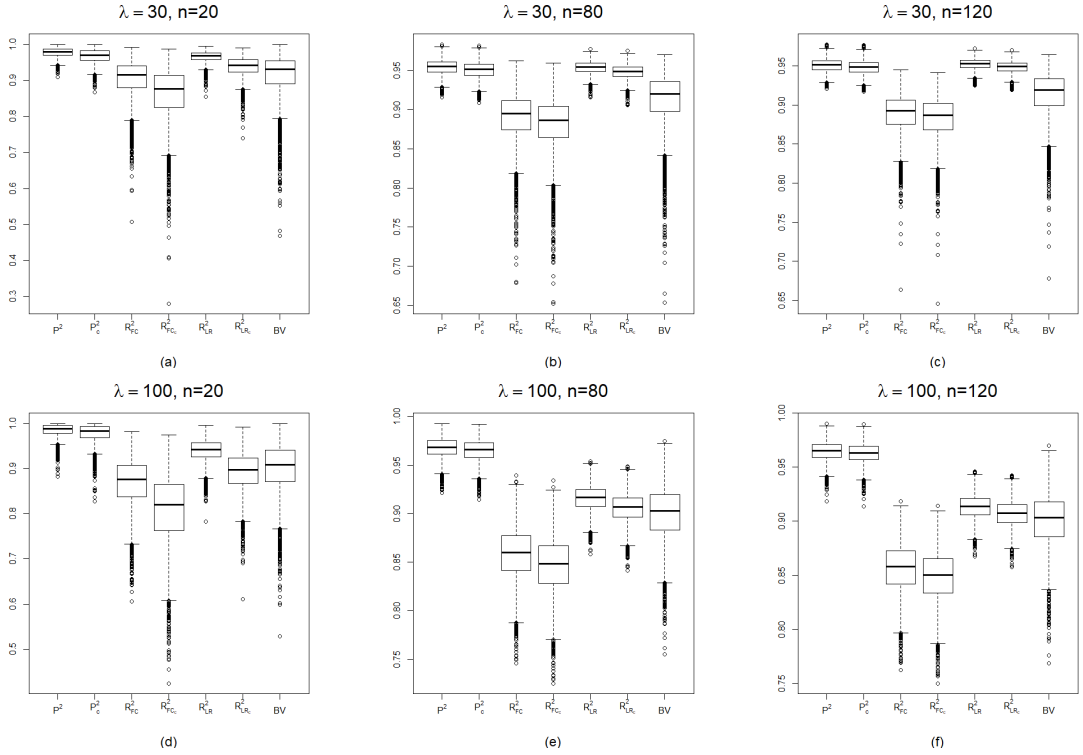
Fonte: Autoria própria.

Figura 3 – Gráficos de *boxplot* dos valores das estatísticas de adequabilidade de ajuste, no modelo beta linear com dispersão variável corretamente especificado, para $\mu_t \in (0.0193, 0.2517)$, com $g(\cdot)$ log-log e $h(\cdot)$ log.



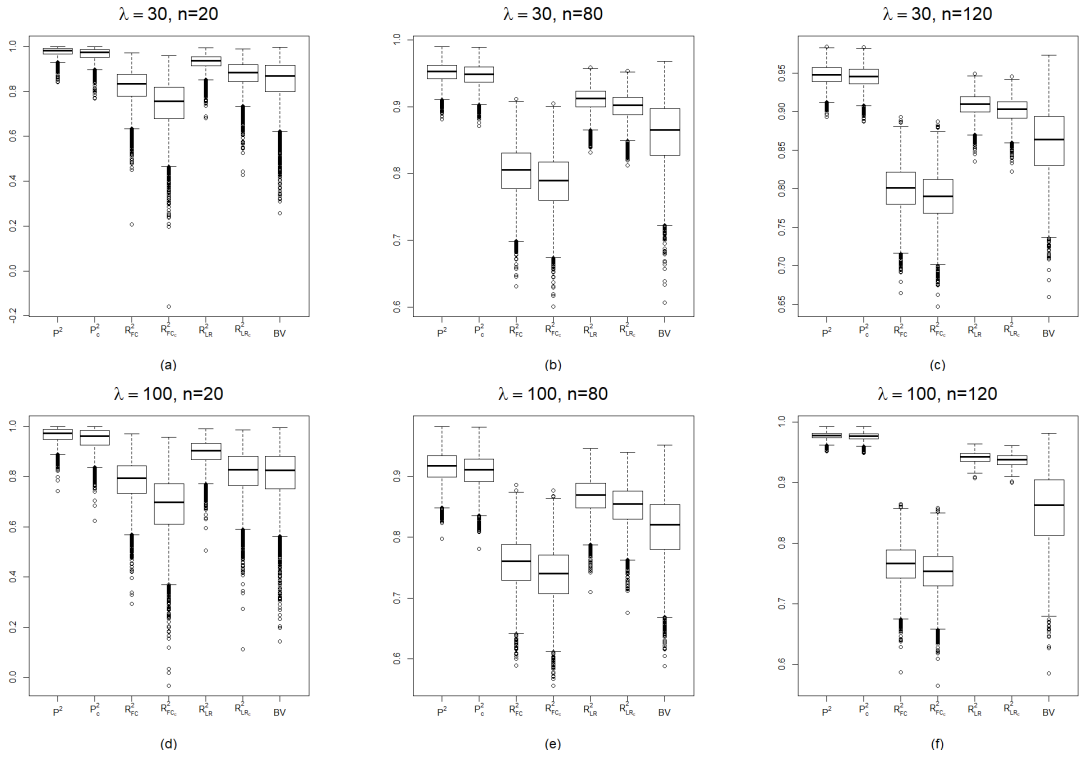
Fonte: Autoria própria.

Figura 4 – Gráficos de *Boxplot* dos valores das estatísticas de adequabilidade de ajuste, no modelo beta linear com dispersão variável corretamente especificado, para $\mu_t \in (0.2650, 0.8702)$, com $g(\cdot)$ logit e $h(\cdot)$ log.



Fonte: Autoria própria.

Figura 5 – Gráficos de *Boxplot* dos valores das estatísticas de adequabilidade de ajuste, no modelo beta linear com dispersão variávelcorretamente especificado, para $\mu_t \in (0.6404, 0.9881)$, com $g(\cdot)$ c-log-log e $h(\cdot)$ log.



Fonte: Autoria própria.

4.2.1 Estudo Descritivo da Estatística $\hat{\alpha}$ - Caso Linear

Buscando realizar uma avaliação sobre o comportamento dos valores de $\hat{\alpha}$, para o modelo beta linear corretamente especificado, com dispersão variável, foram calculadas algumas medidas descritivas dessa quantidade, baseando-se nos 5000 valores obtidos em cada um dos cenários considerados anteriormente. Na Tabela 3 apresentamos essas medidas descritivas; podemos notar que as médias de $\hat{\alpha}$ geralmente estão entre 0.4 e 0.5. No entanto, essas médias tornam-se levemente mais próximas quando aumentamos o tamanho amostral. Além disso, como já mencionado anteriormente, para μ_t localizado nos extremos do intervalo $(0,1)$ verificamos que as médias de $\hat{\alpha}$ em geral, são menores do que para $\mu_t \in (0.2650, 0.8702)$.

Na Figura 6 temos ilustrado o comportamento da média de $\hat{\alpha}$ em cada um dos cenários de μ, para os diferentes tamanhos amostrais e valores de λ. Dessa forma, podemos observar mais facilmente que existe um determinado padrão para os valores esperados de $\hat{\alpha}$, quando o modelo está corretamente especificado, de modo que tais valores tornam-se próximos de 0.5 quando o tamanho amostral e λ aumentam.

Ainda pela Tabela 3, avaliando o desvio-padrão de $\hat{\alpha}$, observamos que em todos os cenários, para $n = 20$, foram apresentados valores bem próximos de 0.2. Entretanto, quando o tamanho amostral cresce, os desvios de $\hat{\alpha}$ quando $\mu_t \in (0.2650, 0.8702)$ e $\lambda \geq 50$, diminuem mais rapidamente do que para os demais cenários de μ_t. Ou seja, para μ_t nos extremos do intervalo unitário, notamos que existe menos variação nos valores dos desvios, quando λ e n aumentam. Com relação à curtose de $\hat{\alpha}$, notamos valores sempre abaixo de 3.0. Dessa forma, a curva da

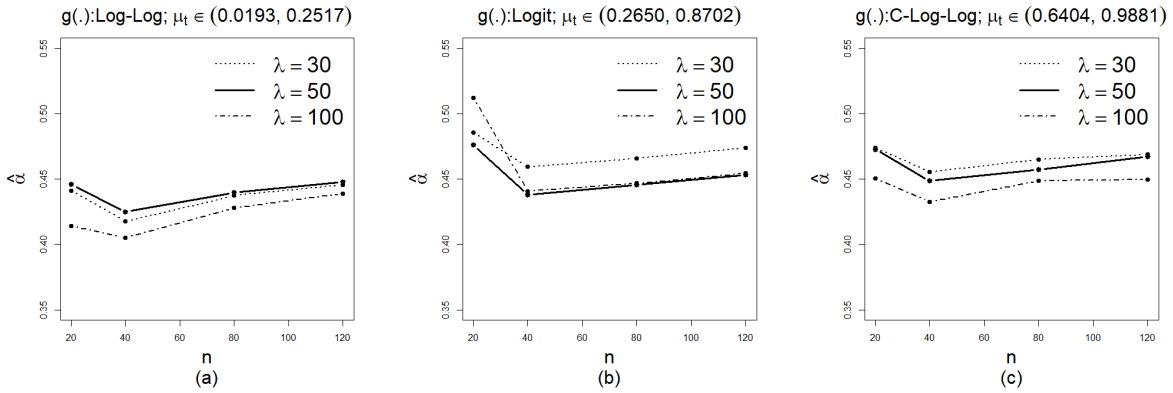
distribuição de $\hat{\alpha}$ possui uma forma mais aberta do que a da distribuição normal padrão, principalmente para μ_t localizado nos extremos do intervalo $(0,1)$, com $\lambda \geq 50$ e $n \geq 40$. No que se refere ao coeficiente de assimetria da distribuição de $\hat{\alpha}$, acreditamos que essa medida tende a se aproximar de zero quando o tamanho da amostra aumenta.

Tabela 4 – Medidas descritivas de $\hat{\alpha}$ para diferentes valores de n e λ, em diferentes cenários da média μ_t, no modelo beta linear com dispersão variável corretamente especificado, usando $g(\cdot)$ log-log para $\mu_t \in (0.0193, 0.2517)$, $g(\cdot)$ logit para $\mu_t \in (0.2650, 0.8702)$, $g(\cdot)$ c-log-log para $\mu_t \in (0.6404, 0.9881)$ e $h(.)$ log.

	Média			Desvio-Padrão			
λ \ $\mu_t \in$	$(0.0193, 0.2517)$	$(0.2650, 0.8702)$	$(0.6404, 0.9881)$	$(0.0193, 0.2517)$	$(0.2650, 0.8702)$	$(0.6404, 0.9881)$	n
30	0.4413	0.4857	0.4739	0.1908	0.2100	0.2105	20
50	0.4460	0.4761	0.4728	0.1928	0.1919	0.1958	
100	0.4141	0.5122	0.4504	0.1854	0.1966	0.1982	
30	0.4178	0.4594	0.4555	0.1596	0.1980	0.1952	40
50	0.4250	0.4382	0.4486	0.1686	0.1640	0.1763	
100	0.4053	0.4407	0.4327	0.1542	0.1439	0.1796	
30	0.4374	0.4660	0.4652	0.1615	0.2025	0.1980	80
50	0.4397	0.4455	0.4572	0.1642	0.1599	0.1799	
100	0.4281	0.4468	0.4489	0.1499	0.1354	0.1868	
30	0.4456	0.4738	0.4689	0.1644	0.2064	0.1984	120
50	0.4477	0.4531	0.4671	0.1665	0.1648	0.1840	
100	0.4389	0.4546	0.4497	0.1487	0.1354	0.1908	
	Curtose			Assimetria			
30	2.6718	2.3020	2.4118	0.3950	0.1930	0.1868	20
50	2.6867	2.6694	2.5070	0.4454	0.4688	0.2218	
100	2.9351	2.4458	2.5554	0.5877	0.4434	0.2848	
30	2.5492	2.2625	2.2934	0.1730	0.0883	0.0260	40
50	2.4777	2.6844	2.5097	0.1624	0.2555	0.0188	
100	2.9250	3.0543	2.4438	0.4012	0.4355	0.0278	
30	2.5485	2.3041	2.2902	0.1586	0.0093	0.0659	80
50	2.4705	2.6352	2.3631	0.0659	0.0322	−0.0379	
100	2.6774	2.8347	2.3159	0.1958	0.2290	0.0116	
30	2.4814	2.2403	2.2823	0.0993	−0.0345	0.1061	120
50	2.4830	2.6974	2.4017	0.0409	0.0112	−0.0720	
100	2.5262	2.7085	2.2722	0.1415	0.1376	0.0355	

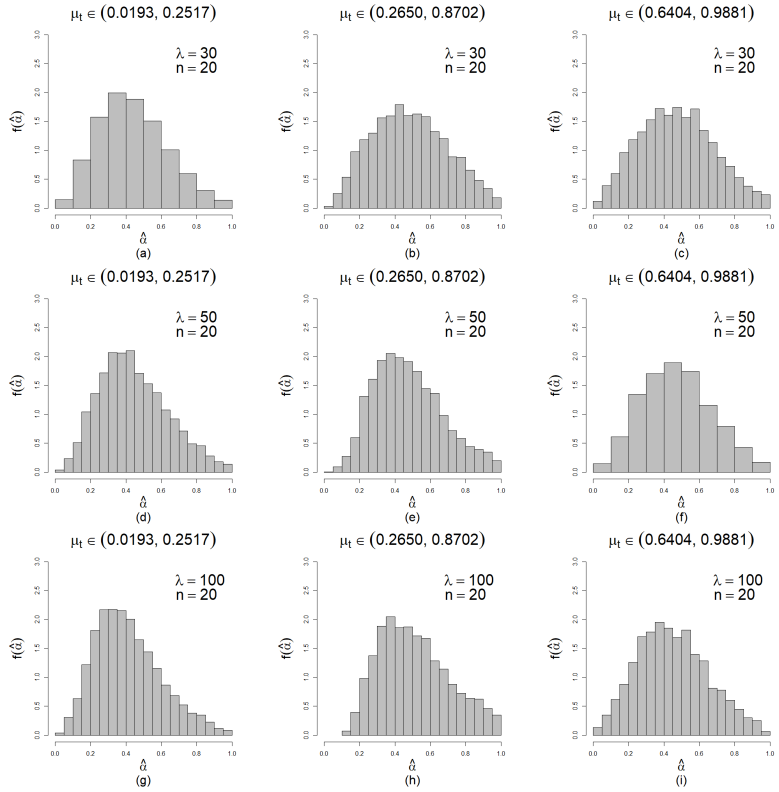
Fonte: Autoria própria.

Figura 6 – Valores médios de $\hat{\alpha}$ *versus* tamanhos amostrais para diferentes valores de λ, no modelo beta linear com dispersão variável corretamente especificado, sendo $h(.)$ log.



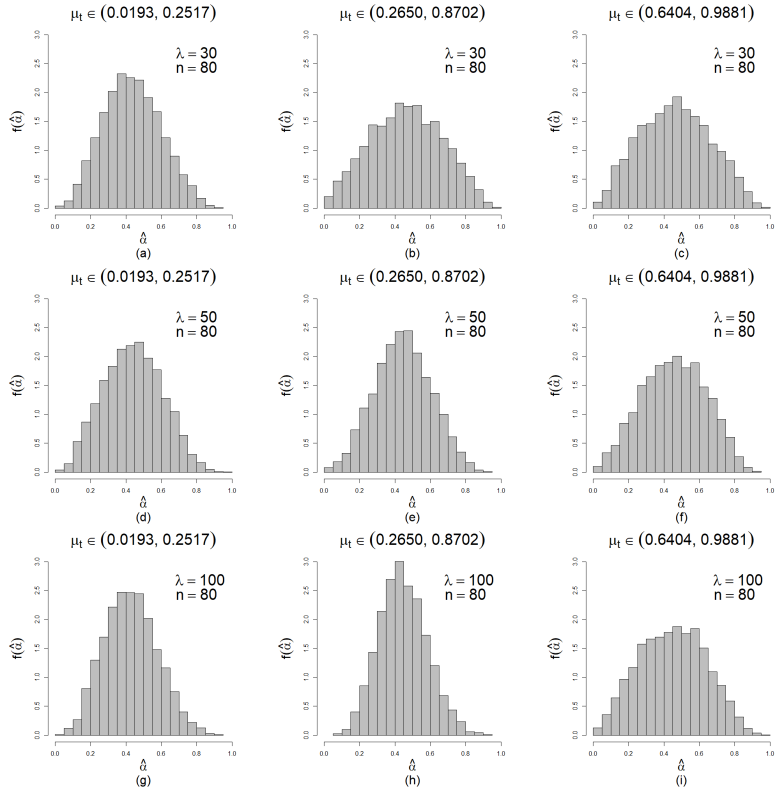
Fonte: Autoria própria.

Figura 7 – Histogramas de $\hat{\alpha}$ no modelo beta linear corretamente especificado, com ϕ variável e $n = 20$.



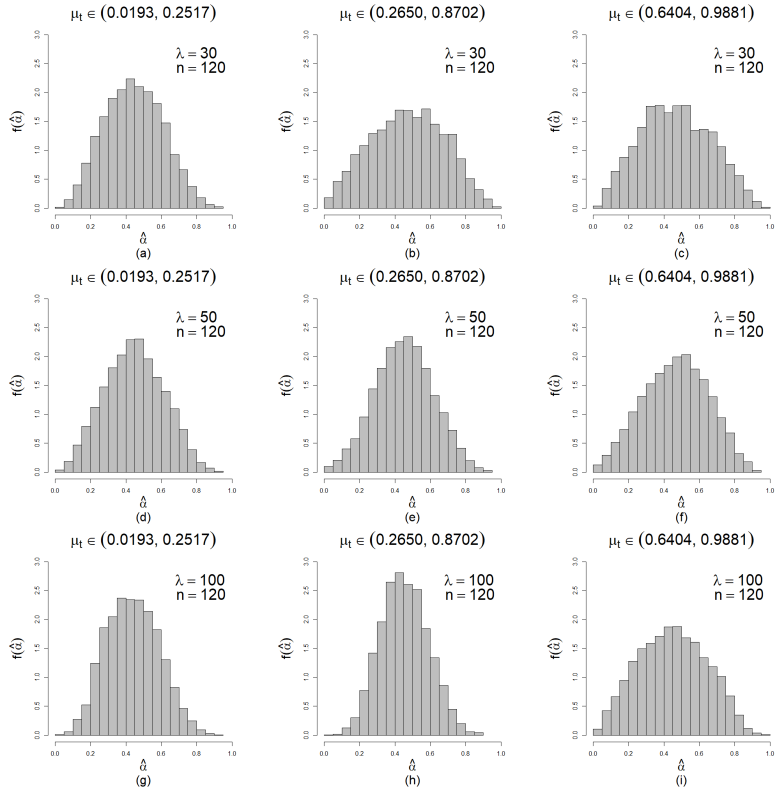
Fonte: Autoria própria.

Figura 8 – Histogramas de $\hat{\alpha}$ no modelo beta linear corretamente especificado, com ϕ variável e $n = 80$.



Fonte: Autoria própria.

Figura 9 – Histogramas de $\hat{\alpha}$ no modelo beta linear corretamente especificado, com ϕ variável e $n = 120$.



Fonte: Autoria própria.

Nas Figuras 7, 8 e 9 temos os histogramas dos valores de $\hat{\alpha}$, para o modelo linear beta corretamente especificado e com dispersão variável, para $n = 20$, 80 e 120, respectivamente, para cada um dos cenários de μ_t, utilizando a função de ligação adequada e os graus de heteroscedasticidade 30, 50 e 100. Por essas figuras podemos observar desde já que, a partir do tamanho amostral $n = 80$ notamos a presença de uma aparente simetria na distribuição de $\hat{\alpha}$ em torno de suas respectivas médias, para todos os cenários de μ_t e para os três graus de heteroscedasticidade considerados. Dessa forma, para $n = 20$, observamos distribuições assimétricas positiva, para qualquer intervalo de μ_t ou valor de λ. Entretanto, essas distribuições tornam-se consideravelmente mais simétricas quando o tamanho amostral aumenta. Também notamos que para todo n, os valores de $\hat{\alpha}$ estão sempre distribuídos ao logo de todo o intervalo $(0,1)$, porém, para $\mu_t \in (0.2650, 0.8702)$ e $\lambda \geq 50$, observamos valores mais concentrados em torno de $0.45 \approx 0.5$.

É importante notar que quando o modelo beta linear está corretamente ajustado, incluído o uso da função de ligação e dados centrais, teríamos o que podemos chamar de cenário ideal para a estimação por máxima verossimilhança. Pois bem, neste cenário com altos valores para a precisão dos dados $(\lambda = 30)$ e tamanho amostral grande $(n = 120)$, notamos com base na Figura 9(h) que $\hat{\alpha}$ apresenta uma distribuição simétrica e valor médio concentrado em torno de 0.5. Em outras palavras, $\hat{\alpha} \approx 0.5$ é de fato indicação de um modelo bem ajustado em um cenário ideal.

4.2.2 Avaliação no Modelo Beta com Dispersão Variável Mal Especificado: Omitindo uma Covariável Sob Diferentes Distribuições

Nosso objetivo agora será avaliar o desempenho da estatística BV para um modelo beta com parâmetro de precisão variável, quando ajustado um modelo omitindo uma determinada covariável. Para isso, consideramos o seguinte modelo linear:

$$\log\left(\frac{\mu_t}{1-\mu_t}\right) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} \qquad \text{e} \qquad \log(\phi_t) = \gamma_1 + \gamma_2 z_{t2},$$

de modo que geramos $x_{t2} \sim U(0,1)$ e $x_{t3} \sim U(0.3,1.3)$, com $t = 1,\ldots,20$ e para gerar os valores da covariável x_{t4} consideramos três diferentes distribuições: $x_{t4} \sim U(0.5,1.5)$, $x_{t4} \sim N(0,4)$ e $x_{t4} \sim Exp(2)$.

Nosso interesse aqui, consiste em verificar o desempenho das estatísticas, em especial BV e $\hat{\alpha}$, quando uma covariável que em geral apresenta valores maiores que os valores da distribuição $U(0,1)$ ou valores atípicos devido a assimetria, é omitida da modelagem da média da variável resposta. Além disso, determinamos os valores dos parâmetros β's apenas para o caso em que $\mu_t \in (0.25,0.80)$, ou seja, apenas para valores de μ_t centrais. Temos assim, três diferentes cenários para o modelo da média, cujos valores estão restritos em intervalos similares e considerando ainda a mesma função de ligação $g(\cdot)$, nesse caso a logit. Além disso, para o modelo da precisão, tomamos γ_1 e γ_2 de modo que, em cada grau de heteroscedasticidade (λ) temos intervalos similares para a variância da resposta, em cada cenário do modelo da média. Vale observar que consideramos nesse estudo $\lambda = 30$, 50 e 100. Os valores fixados para os parâmetros β's e γ's, junto com as distribuições de x_{t4} e de z_{t2}, com os intervalos das variâncias e os graus de heteroscedasticidade estão apresentados na Tabela 5.

Tabela 5 – Valores dos parâmetros β's e γ's, fixados para a análise no modelo beta linear com dispersão variável, omitindo uma covariável do modelo da média sob diferentes distribuições, com os intervalos dos valores de μ_t, as distribuições de x_{t4} e de z_{t2}, os intervalos das variâncias e os graus de heteroscedasticidade.

β_1	β_2	β_3	β_4	x_{t4}	μ_t	γ_1	γ_2	z_{t2}	$\text{Var}(y_t)$	λ
						2.4	4.26	$U(0,1)$	$(0.0003,0.0103)$	30
-2.85	2.4	0.8	1.2	$U(0.5,1.5)$	$(0.2529,0.8014)$	1.0	4.94	$U(0.15,1.15)$	$(0.0003,0.0174)$	50
						0.9	5.9	$U(0,1)$	$(0.0003,0.0323)$	100
						2.8	3.7	$U(0,1)$	$(0.0003,0.0106)$	30
-0.7	0.7	0.53	0.31	$N(0,4.0)$	$(0.2500,0.8063)$	1.56	4.32	$U(0.15,1.15)$	$(0.0003,0.0177)$	50
						1.4	5.2	$U(0,1)$	$(0.0003,0.0342)$	100
						2.5	4.23	$U(0,1)$	$(0.0003,0.0105)$	30
1.0	-1.7	-1.4	1.5	$Exp(2.0)$	$(0.2520,0.8064)$	1.2	4.82	$U(0.15,1.15)$	$(0.0003,0.0171)$	50
						1.12	5.74	$U(0,1)$	$(0.0003,0.0324)$	100

Fonte: Autoria própria.

Os valores obtidos para as médias das estatísticas de avaliação de desempenho são mostradas na Tabela 6. Avaliando tais resultados vemos que muitas das estatísticas apresentam dificuldade em assimilar a omissão da covariável de distribuição $U(0.5,1.5)$. Provavelmente, isso ocorre devido o fato de que, embora apresente valores maiores do que as demais covariáveis

presentes no modelo, a covariável Uniforme omitida é bastante similar a essas outras variáveis. Logo, é compreensível muitas estatísticas apresentarem valores elevados, não conseguindo captar a falta dessa variável no modelo, com exceção da estatística $R^2_{FC_c}$, cujos valores são baixos, em especial quando o grau de heteroscedasticidade aumenta, conseguindo então identificar a má adequabilidade do modelo ajustado. Vale notar que consequentemente, a nossa medida BV também apresenta valores menores do que as demais medidas de ajuste, e seus valores diminuem com o aumento do grau de heteroscedasticidade. Além disso, observando os valores de $\hat{\alpha}$ para $n = 20$, vemos que estes estão distantes de 0.5 e também dos valores médios apresentados pela Tabela 4, para valores de μ_t centrais. Dessa forma, embora algumas estatísticas, incluindo a BV, possam indicar que o modelo ajustado esteja aceitável, $\hat{\alpha}$ indica que o modelo não está adequado. Logo, reforçamos a ideia de que devemos avaliar conjuntamente as medidas BV e $\hat{\alpha}$.

Já para o caso em que é omitida uma covariável de distribuição normal, vemos que tanto os valores de $\hat{\alpha}$ quanto das demais medidas são consideravelmente baixos, quando $n \leq 40$, logo, conseguem detectar a má especificação do modelo. Dessa forma, estamos em uma situação em que a variável omitida é bastante relevante para o modelo. De modo similar, vemos os cenários em que a distribuição da covariável omitida trata-se de uma exponencial, cujas estatísticas também apresentam valores baixos. Porém, algumas observações importantes podem ser feitas para esse caso, como o fato dos valores das estatísticas R^2_{LR} e $R^2_{LR_c}$ serem bastante distantes, o que reforça a suposição de má adequabilidade do modelo e também mostra a importância de também utilizarmos as versões corrigidas das estatísticas usuais na obtenção da BV.

Quando o tamanho da amostra aumenta, observamos que para os cenários em que são omitidas as covariáveis Normal ou Exponencial, os valores de BV permanecem consideravelmente baixos. Já no caso de $\hat{\alpha}$, notamos que apenas quando $x_{t4} \sim Exp(2)$ os valores dessa estatística tendem a diminuir quando n aumenta, entretanto, em todas as situações ainda dispomos de valores de $\hat{\alpha}$ distantes de 0.5. Além disso, em todos os cenários, vemos que a questão da variabilidade não ser bem explicada pelo modelo se dilui quando a amostra cresce e as versões corrigidas das medidas do tipo R^2 aumentam levemente. Por outro lado, podemos observar que para esse cenário de omissão de uma covariável relevante, os valores de P^2 diminuem quando n aumenta. Isso provavelmente indica aumento no viés de $\hat{\mu}_t$, resultando em baixo poder de predição da variável resposta. Dessa forma, uma vez que a BV considera a combinação linear dos dois tipos de medidas, seu valor é o que sempre capta uma das duas situações.

Tabela 6 – Valores médios das estatísticas de avaliação de desempenho no modelo beta linear com dispersão variável, omitindo uma covariáveis com diferentes distribuições, com $g(.)$ logit e $h(.)$ log, para $\mu_t \in (0.25, 0.80)$.

$x_{t4} \sim$	$U(0.5, 1.5)$			$N(0, 4.0)$			Exp(2.0)			
λ	30	50	100	30	50	100	30	50	100	n
P^2	0.8482	0.8421	0.8361	0.2942	0.3057	0.3255	0.5250	0.5265	0.5331	20
P_c^2	0.8078	0.8000	0.7923	0.1060	0.1206	0.1457	0.3983	0.4003	0.4086	
R_{FC}^2	0.7218	0.6742	0.5927	0.1063	0.1038	0.0997	0.2152	0.2091	0.1907	
$R_{FC_c}^2$	0.6476	0.5874	0.4841	−0.1320	−0.1352	−0.1404	0.0060	−0.0019	−0.0251	
R_{LR}^2	0.8154	0.8114	0.8125	0.1245	0.1262	0.1382	0.3342	0.3516	0.3878	
$R_{LR_c}^2$	0.7216	0.7156	0.7173	−0.3203	−0.3177	−0.2995	−0.0040	0.0222	0.0768	
BV	0.7087	0.6675	0.5973	−0.1997	−0.1976	−0.1758	0.1075	0.1116	0.1114	
$\hat{\alpha}$	0.3299	0.3359	0.3434	0.2042	0.2016	0.2107	0.2703	0.2814	0.2941	
P^2	0.8250	0.8197	0.8159	0.2024	0.2093	0.2198	0.4449	0.4496	0.4590	40
P_c^2	0.8050	0.7991	0.7948	0.1113	0.1190	0.1306	0.3815	0.3867	0.3971	
R_{FC}^2	0.7100	0.6596	0.5623	0.1029	0.1011	0.0950	0.2098	0.2008	0.1776	
$R_{FC_c}^2$	0.6768	0.6207	0.5123	0.0004	−0.0016	−0.0085	0.1195	0.1095	0.0836	
R_{LR}^2	0.8100	0.8059	0.8132	0.1109	0.1111	0.1219	0.3320	0.3508	0.3956	
$R_{LR_c}^2$	0.7727	0.7678	0.7765	−0.0637	−0.0634	−0.0505	0.2008	0.2234	0.2770	
BV	0.7297	0.6898	0.6154	0.0164	0.0152	0.0245	0.2127	0.2084	0.1963	
$\hat{\alpha}$	0.3661	0.3614	0.3559	0.3085	0.2949	0.2958	0.2947	0.3008	0.3082	
P^2	0.8122	0.8067	0.8027	0.1517	0.1559	0.1625	0.4047	0.4090	0.4197	80
P_c^2	0.8022	0.7964	0.7921	0.1065	0.1108	0.1179	0.3729	0.3774	0.3888	
R_{FC}^2	0.7052	0.6485	0.5406	0.1021	0.0988	0.0915	0.2075	0.1968	0.1728	
$R_{FC_c}^2$	0.6894	0.6298	0.5161	0.0542	0.0507	0.0431	0.1652	0.1540	0.1287	
R_{LR}^2	0.8070	0.8039	0.8127	0.1057	0.1039	0.1141	0.3321	0.3512	0.3976	
$R_{LR_c}^2$	0.7900	0.7866	0.7962	0.0268	0.0250	0.0360	0.2732	0.2940	0.3445	
BV	0.7359	0.6926	0.6101	0.0731	0.0701	0.0732	0.2307	0.2253	0.2146	
$\hat{\alpha}$	0.3839	0.3659	0.3398	0.3773	0.3500	0.3354	0.2777	0.2855	0.2952	
P^2	0.8079	0.8020	0.7979	0.1343	0.1374	0.1429	0.3900	0.3947	0.4065	120
P_c^2	0.8012	0.7951	0.7909	0.1042	0.1074	0.1131	0.3688	0.3736	0.3858	
R_{FC}^2	0.7037	0.6458	0.5340	0.1016	0.0979	0.0901	0.2063	0.1946	0.1696	
$R_{FC_c}^2$	0.6934	0.6335	0.5178	0.0704	0.0666	0.0584	0.1787	0.1665	0.1407	
R_{LR}^2	0.8058	0.8034	0.8128	0.1038	0.1014	0.1110	0.3305	0.3503	0.3997	
$R_{LR_c}^2$	0.7948	0.7923	0.8022	0.0528	0.0503	0.0605	0.2924	0.3133	0.3656	
BV	0.7374	0.6941	0.6070	0.0853	0.0818	0.0823	0.2326	0.2261	0.2153	
$\hat{\alpha}$	0.3881	0.3662	0.3264	0.4050	0.3672	0.3386	0.2584	0.2655	0.2761	

Fonte: Autoria própria.

Tabela 7 – Intervalos de valores de ϕ_t e de $\text{Var}(y_t)$, e valores de λ, juntamente com suas respectivas médias das estimativas, e também as médias das estimativas dos intervalos de μ, no modelo beta linear com dispersão variável, omitindo uma covariável do modelo da média sob diferentes distribuições, com $g(.)$ logit e $h(.)$ log, para $\mu_t \in (0.25, 0.80)$.

$\hat{\mu}_t$	ϕ_t	$\hat{\phi}_t$	$\text{Var}(y_t)$	$\widehat{\text{Var}}(y_t)$	λ	$\hat{\lambda}$	n
			$x_{t4} \sim U(0.5, 1.5)$				
(0.2315, 0.7232)	(15.561, 636.99)	(18.762, 121.50)	(0.0003, 0.0103)	(0.0020, 0.0229)	30	11.5	20
(0.2693, 0.7201)	(8.506, 629.75)	(13.733, 127.66)	(0.0003, 0.0174)	(0.0019, 0.0312)	50	16.2	
(0.2663, 0.7332)	(3.965, 677.57)	(9.0705, 139.00)	(0.0003, 0.0323)	(0.0018, 0.0463)	100	26.0	
(0.2423, 0.7247)	(15.561, 636.99)	(10.225, 115.99)	(0.0003, 0.0103)	(0.0019, 0.0238)	30	12.4	40
(0.2447, 0.7383)	(8.506, 629.75)	(7.527, 120.85)	(0.0003, 0.0174)	(0.0018, 0.0318)	50	17.1	
(0.2387, 0.7321)	(3.965, 677.57)	(4.438, 133.85)	(0.0003, 0.0323)	(0.0017, 0.0488)	100	29.0	
(0.2493, 0.7316)	(15.561, 636.99)	(9.0569, 113.12)	(0.0003, 0.0103)	(0.0019, 0.0239)	30	12.5	80
(0.2535, 0.7223)	(8.506, 629.75)	(6.4108, 118.69)	(0.0003, 0.0174)	(0.0018, 0.0324)	50	17.7	
(0.2587, 0.7339)	(3.965, 677.57)	(3.7813, 131.19)	(0.0003, 0.0323)	(0.0017, 0.0501)	100	30.2	
(0.2517, 0.7297)	(15.561, 636.99)	(8.712, 112.11)	(0.0003, 0.0103)	(0.0019, 0.0239)	30	12.5	120
(0.2469, 0.7316)	(8.506, 629.75)	(6.147, 117.58)	(0.0003, 0.0174)	(0.0018, 0.0326)	50	17.8	
(0.2618, 0.7175)	(3.965, 677.57)	(3.590, 130.24)	(0.0003, 0.0323)	(0.0017, 0.0505)	100	30.5	
			$x_{t4} \sim N(0, 4.0)$				
(0.4290, 0.6361)	(22.185, 557.51)	(8.964, 18.24)	(0.0003, 0.0106)	(0.0206, 0.0260)	30	1.26	20
(0.4252, 0.6134)	(12.905, 556.63)	(9.408, 17.25)	(0.0003, 0.0177)	(0.0223, 0.0252)	50	1.13	
(0.4169, 0.5815)	(6.177, 573.59)	(9.335, 21.03)	(0.0003, 0.0342)	(0.0229, 0.0335)	100	1.46	
(0.4365, 0.6280)	(22.185, 557.51)	(9.159, 12.47)	(0.0003, 0.0106)	(0.0214, 0.0249)	30	1.16	40
(0.4377, 0.6368)	(12.905, 556.63)	(9.531, 10.59)	(0.0003, 0.0177)	(0.0229, 0.0256)	50	1.12	
(0.4288, 0.5912)	(6.177, 573.59)	(7.569, 10.57)	(0.0003, 0.0342)	(0.0218, 0.0359)	100	1.65	
(0.4295, 0.5926)	(22.185, 557.51)	(9.252, 11.21)	(0.0003, 0.0106)	(0.0217, 0.0245)	30	1.13	80
(0.4001, 0.5758)	(12.905, 556.63)	(9.151, 9.77)	(0.0003, 0.0177)	(0.0228, 0.0265)	50	1.16	
(0.4188, 0.6224)	(6.177, 573.59)	(6.355, 10.74)	(0.0003, 0.0342)	(0.0212, 0.0370)	100	1.75	
(0.4279, 0.5950)	(22.185, 557.51)	(9.292, 10.86)	(0.0003, 0.0106)	(0.0218, 0.0244)	30	1.11	120
(0.4034, 0.6253)	(12.905, 556.63)	(8.844, 9.80)	(0.0003, 0.0177)	(0.0227, 0.0266)	50	1.17	
(0.4046, 0.5965)	(6.177, 573.59)	(6.059, 10.76)	(0.0003, 0.0342)	(0.0211, 0.0373)	100	1.77	
			$x_{t4} \sim \text{Exp}(2.0)$				
(0.3631, 0.6837)	(17.156, 684.16)	(3.320, 27.69)	(0.0003, 0.0105)	(0.0087, 0.0644)	30	7.40	20
(0.3572, 0.7075)	(10.106, 673.87)	(3.169, 28.50)	(0.0003, 0.0171)	(0.0086, 0.0695)	50	8.09	
(0.3823, 0.6626)	(4.877, 724.98)	(2.859, 30.91)	(0.0003, 0.0324)	(0.0080, 0.0814)	100	10.11	
(0.3821, 0.6927)	(17.156, 684.16)	(2.975, 27.52)	(0.0003, 0.0105)	(0.0085, 0.0657)	30	7.73	40
(0.3646, 0.6993)	(10.106, 673.87)	(2.666, 28.67)	(0.0003, 0.0171)	(0.0082, 0.0718)	50	8.75	
(0.3665, 0.6971)	(4.877, 724.98)	(2.101, 31.42)	(0.0003, 0.0324)	(0.0075, 0.0857)	100	11.35	
(0.3700, 0.6878)	(17.156, 684.16)	(2.816, 27.54)	(0.0003, 0.0105)	(0.0084, 0.0666)	30	7.94	80
(0.3613, 0.6654)	(10.106, 673.87)	(2.496, 28.67)	(0.0003, 0.0171)	(0.0081, 0.0731)	50	9.05	
(0.3562, 0.6959)	(4.877, 724.98)	(1.941, 31.41)	(0.0003, 0.0324)	(0.0074, 0.0871)	100	11.76	
(0.3688, 0.6827)	(17.156, 684.16)	(2.785, 27.43)	(0.0003, 0.0105)	(0.0084, 0.0667)	30	7.95	120
(0.3679, 0.7041)	(10.106, 673.87)	(2.443, 28.59)	(0.0003, 0.0171)	(0.0081, 0.0735)	50	9.12	
(0.3633, 0.6841)	(4.877, 724.98)	(1.876, 31.42)	(0.0003, 0.0324)	(0.0073, 0.0881)	100	11.98	

Fonte: Autoria própria.

Com base agora nos valores da Tabela 7, onde encontramos as médias das estimativas dos intervalos de μ_t, ϕ e da variância de y_t, podemos observar que as estimativas para μ_t são menos afetadas quando omitimos a covariável de distribuição Uniforme. Logo, embora o modelo não esteja bem especificado, o viés de $\hat{\mu}_t$, com $t = 1,\ldots,n$, não é tão elevado quanto nos casos em que omitimos $x_{t4} \sim N(0,4.0)$ ou $x_{t4} \sim Exp(2)$. Isso explica o fato dos valores das estatísticas P^2 e P_c^2 serem altos, como mostrado na Tabela 6. Além disso, vemos ainda pela Tabela 7 que independentemente da distribuição de x_{t4}, as variâncias e as precisões são mal estimadas, de modo que, as variâncias são sempre superestimadas e as precisões são subestimadas, principalmente quando n cresce. Por isso as estatísticas R^2_{FC} e $R^2_{FC_c}$ são baixas, pois apontam para esses problemas nos modelos. Especialmente quando $x_{t4} \sim N(0,4.0)$, em que vemos que as estimativas das variâncias de y_t, com $t = 1,\ldots,n$, são elevadas e praticamente contante. No entanto, como os valores de $\hat{\alpha}$ estão sempre distantes do que é esperado quando dispomos de bons ajustes, podemos dizer que tal medida demostra aqui melhor eficácia em apontar os problemas relacionados as estimativas das variâncias e das precisões, e a adequabilidade do modelo.

4.2.3 Avaliação no Modelo Beta com Dispersão Variável Mal Especificado: Má Especificação da Heteroscedasticidade

Antes de prosseguir com a apresentação do nosso próximo estudo de simulação é importante fazer uma observação: diferentemente dos modelos de regressão normal linear, a variância da resposta y_t nos modelos de regressão betas depende da média μ_t, com $t = 1,\ldots,n$, como vemos pela definição 2.3, então, mesmo nos casos em que ϕ é constante, existe ainda uma variabilidade na variância (ainda que muito pequena). Dessa forma, o grau de heteroscedasticidade λ, definido em 4.2, é maior que 1.0 quando a precisão ϕ varia de acordo com as observações, e também quando ϕ é constante. No entanto, o valor de λ torna-se consideravelmente elevado quando a precisão dos dados é variável, ou seja, a heteroscedasticidade é mais grave.

O nosso objetivo aqui consiste em investigar o desempenho das estatísticas de adequabilidade, quando ajustado um modelo assumindo precisão constante, sob dados gerados a partir de um modelo com média e precisão variável. Dessa forma, para esse estudo os valores da variável resposta são gerados com base no seguinte modelo:

$$g(\mu_t) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} \qquad \text{e} \qquad \log(\phi_t) = \gamma_1 + \gamma_2 z_{t2} + \gamma_3 z_{t3}. \tag{4.3}$$

Ou seja, tanto a média quanto a precisão estão associados a uma estrutura de regressão. No entanto, foram calculadas as estatísticas referentes ao ajuste do modelo dado por:

$$g(\mu_t) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} \qquad \text{e} \qquad \log(\phi_t) = \gamma_1. \tag{4.4}$$

Na Tabela 8 dispomos dos valores fixados para os parâmetros dos submodelos da média e da precisão em nosso estudo de simulação, também temos os intervalos de μ_t e as respectivas funções de ligação usada em cada cenário da média, as distribuições assumidas para gerar os valores das covariáveis do modelo da precisão, e por fim os graus de heteroscedasticidade.

Além disso, vale informar que os valores das covariáveis x_{ti}, com $t = 1,\ldots,n$ e $i = 1,2,3$, foram geradas a partir da distribuição $U(0,1)$.

Tabela 8 – Valores dos parâmetros β's e γ's, fixados para a análise no modelo beta linear, omitindo a precisão variável, com os respectivos intervalos dos valores de μ, as funções de ligação para o modelo da média, as distribuições das covariáveis do modelo da precisão e os graus de heteroscedasticidade.

β_1	β_2	β_3	β_4	μ_t	$g(\cdot)$	γ_1	γ_2	γ_3	z_{t2}	z_{t3}	λ
						5.0	-1.9	-1.0	$U(0.1,1.1)$	$U(0.2,1.2)$	30
-1.95	1.0	1.2	0.5	$\mu_t \in (0.0193,0.2517)$	Log-log	3.8	-0.58	-2.35	$U(-0.1,0.9)$	$U(-0.3,0.7)$	50
						4.3	-0.4	-3.35	$U(0.35,1.35)$	$U(-0.15,0.85)$	100
						4.2	3.9	1.08			30
-2.48	1.5	1.7	3.1	$\mu_t \in (0.2650,0.8702)$	Logit	3.6	1.1	3.88	$U(-0.5,0.5)$	$U(-0.5,0.5)$	50
						4.0	-1.5	4.18			100
						3.3	1.15	3.4	$U(-0.4,0.6)$	$U(-0.5,0.5)$	30
1.2	-0.6	-1.23	0.63	$\mu_t \in (0.6404,0.9881)$	C-Log-Log	3.0	1.5	4.0	$U(-0.4,0.6)$	$U(-0.5,0.5)$	50
						0.25	0.6	4.81	$U(0,1)$	$U(0,1)$	100

Fonte: Autoria própria.

Na Tabela 9 encontramos os valores médios das estatísticas de adequabilidade de ajuste, quando omitimos o fato da precisão ser variável. Em todos os cenários de μ_t observamos que os valores das estatísticas P^2, P_c^2, R^2_{FC}, $R^2_{FC_c}$, R^2_{LR}, $R^2_{LR_c}$ e BV diminuem quando o grau de heteroscedasticidade aumenta. Logo, as medidas apresentam maior facilidade em detectar a omissão do modelo da precisão quando a heteroscedasticidade existente for elevada. Outra observação importante é que os valores das estatísticas do tipo P^2 diminuem quando aumenta-se o tamanho da amostra. Dessa forma, tais medidas indicam que o poder de predição da variável resposta para o modelo ajustado diminui com o aumento da amostra, ou seja, existe um aumento no viés.

Podemos observar ainda que, para $\mu_t \in (0.2650,0.8702)$, as medidas de adequabilidade P^2, R^2 e BV apresentam dificuldades em captar a falta de ajuste de um modelo para a precisão, principalmente quando o grau de heteroscedasticidade é baixo. No entanto, os valores da estatística $\hat{\alpha}$ diferem consideravelmente de 0.5, para todos os valores de λ. Logo, para esse cenário de μ_t, tal medida foi a única capaz de indicar que o modelo ajustado não está bem especificado. Para valores de μ_t localizados próximos de zero ou próximos de um, notamos um melhor desempenho das estatísticas R^2_{FC} e $R^2_{FC_c}$ em indicar má adequabilidade do modelo ajustado, principalmente o R^2_{FC}, cujo valor tende a diminuir quando o tamanho amostral aumenta. Dessa forma, fica claro que o modelo não explica bem a variabilidade da resposta.Isso é comprovado quando observamos a Tabela 10, em que os intervalos referentes às estimativas das variâncias diferem consideravelmente dos intervalos verdadeiros. Consequentemente, os valores de BV também conseguem refletir essa dificuldade. Observando ainda μ_t nos extremos do intervalo $(0,1)$, vemos que os valores de $\hat{\alpha}$ para $\mu_t \in (0.6404,0.9881)$ são bem menores do que $\mu_t \in (0.0193,0.2517)$, isso ocorre devido às medidas R^2 e P^2 serem próximas, quando μ_t é perto de zero e n grande. Logo, $\hat{\alpha}$ indica que quando n aumenta o modelo não apresenta bom ajuste e possui baixo poder de predição da resposta.

Tabela 9 – Valores médios das estatísticas de avaliação de desempenho no modelo beta linear, omitindo a precisão variável, com $g(.)$ logit adequada ao cenário de μ_t e $h(.)$ log.

	$g(\cdot)$: Log-Log $\mu_t \in (0.0193, 0.2517)$			$g(\cdot)$: Logit $\mu_t \in (0.2650, 0.8702)$			$g(\cdot)$: C-Log-Log $\mu_t \in (0.6404, 0.9881)$			
λ	30	50	100	30	50	100	30	50	100	n
P^2	0.7419	0.7250	0.6898	0.8996	0.7890	0.7692	0.8138	0.8139	0.7780	20
P_c^2	0.6730	0.6517	0.6070	0.8729	0.7327	0.7076	0.7642	0.7643	0.7188	
R_{FC}^2	0.6495	0.6459	0.5490	0.8959	0.7844	0.7581	0.5890	0.5219	0.4797	
$R_{FC_c}^2$	0.5560	0.5514	0.4288	0.8681	0.7269	0.6936	0.4794	0.3944	0.3410	
R_{LR}^2	0.6898	0.6760	0.5971	0.9111	0.7938	0.7622	0.7868	0.7524	0.6460	
$R_{LR_c}^2$	0.5005	0.4783	0.3512	0.8568	0.6680	0.6171	0.6567	0.6013	0.4300	
BV	0.5631	0.5422	0.4355	0.8630	0.6977	0.6547	0.4922	0.4123	0.3606	
$\hat{\alpha}$	0.3044	0.2821	0.2875	0.2152	0.2359	0.2337	0.1187	0.1161	0.1881	
P^2	0.7050	0.6785	0.6245	0.8793	0.7465	0.7199	0.7861	0.7860	0.7486	40
P_c^2	0.6713	0.6417	0.5816	0.8655	0.7175	0.6879	0.7617	0.7615	0.7199	
R_{FC}^2	0.6169	0.6107	0.5042	0.8782	0.7583	0.7347	0.5201	0.4540	0.4253	
$R_{FC_c}^2$	0.5732	0.5662	0.4475	0.8643	0.7307	0.7044	0.4653	0.3916	0.3596	
R_{LR}^2	0.6654	0.6459	0.5583	0.8997	0.7695	0.7376	0.7807	0.7511	0.6565	
$R_{LR_c}^2$	0.5897	0.5657	0.4582	0.8769	0.7173	0.6781	0.7310	0.6947	0.5787	
BV	0.6105	0.5932	0.4924	0.8656	0.7261	0.6954	0.4729	0.4015	0.3901	
$\hat{\alpha}$	0.3588	0.3487	0.3441	0.2190	0.2959	0.3089	0.0519	0.0443	0.1180	
P^2	0.6842	0.6534	0.5891	0.8694	0.7228	0.6890	0.7718	0.7701	0.7300	80
P_c^2	0.6674	0.6349	0.5671	0.8624	0.7080	0.6724	0.7596	0.7578	0.7156	
R_{FC}^2	0.5988	0.5939	0.4840	0.8701	0.7438	0.7177	0.4801	0.4187	0.4011	
$R_{FC_c}^2$	0.5774	0.5722	0.4565	0.8632	0.7301	0.7027	0.4524	0.3877	0.3691	
R_{LR}^2	0.6529	0.6320	0.5417	0.8945	0.7567	0.7206	0.7800	0.7566	0.6639	
$R_{LR_c}^2$	0.6181	0.5951	0.4957	0.8839	0.7323	0.6926	0.7580	0.7322	0.6302	
BV	0.6143	0.6009	0.5017	0.8648	0.7221	0.6884	0.4537	0.3891	0.3819	
$\hat{\alpha}$	0.3762	0.3877	0.3799	0.2439	0.3153	0.3229	0.0091	0.0062	0.0469	
P^2	0.6800	0.6430	0.5764	0.8655	0.7150	0.6799	0.7661	0.7638	0.7211	120
P_c^2	0.6688	0.6306	0.5617	0.8608	0.7051	0.6688	0.7580	0.7556	0.7115	
R_{FC}^2	0.5957	0.5871	0.4799	0.8663	0.7394	0.7138	0.4657	0.4075	0.3927	
$R_{FC_c}^2$	0.5816	0.5728	0.4618	0.8617	0.7303	0.7039	0.4471	0.3869	0.3716	
R_{LR}^2	0.6515	0.6256	0.5376	0.8926	0.7525	0.7166	0.7780	0.7577	0.6664	
$R_{LR_c}^2$	0.6290	0.6014	0.5079	0.8857	0.7366	0.6984	0.7637	0.7421	0.6449	
BV	0.6167	0.6001	0.5037	0.8638	0.7192	0.6841	0.4474	0.3871	0.3774	
$\hat{\alpha}$	0.3814	0.4077	0.3968	0.2438	0.3181	0.3130	0.0017	0.0011	0.0218	

Fonte: Autoria própria.

Devemos ressaltar a habilidade de $\hat{\alpha}$ em identificar a omissão da modelagem da precisão. Tipicamente, esse erro de especificação é muito difícil de ser identificado. Mesmo para $\mu_t \approx 0$, os valores de $\hat{\alpha}$ ainda podem ser considerados distantes de 0.5, o que indicaria um bom ajuste.

Tabela 10 – Intervalos de valores de ϕ_t e de $\mathrm{Var}(y_t)$, e valores de λ, juntamente com suas respectivas médias das estimativas, e também as médias das estimativas dos intervalos de μ_t, no modelo beta linear omitindo a precisão variável, usando $g(\cdot)$ log-log para $\mu_t \in (0.0193, 0.2517)$, $g(\cdot)$ logit para $\mu_t \in (0.2650, 0.8702)$, $g(\cdot)$ c-log-log para $\mu_t \in (0.6404, 0.9881)$ e $h(.)$ log.

$\hat{\mu}_t$	ϕ_t	$\hat{\phi}_t$	$\mathrm{Var}(y_t)$	$\widehat{\mathrm{Var}}(y_t)$	λ	$\hat{\lambda}$	n
$\mu_t \in (0.0193, 0.2517)$							
(0.0067, 0.2765)	(7.557, 64.56)	23.584	(0.0005, 0.0170)	(0.0011, 0.0089)	30	8.07	20
(0.0056, 0.3108)	(5.744, 64.38)	24.964	(0.0005, 0.0255)	(0.0012, 0.0090)	50	7.50	
(0.0137, 0.2448)	(2.742, 64.42)	15.176	(0.0005, 0.0472)	(0.0025, 0.0154)	100	6.08	
(0.0193, 0.2392)	(7.557, 64.56)	19.154	(0.0005, 0.0170)	(0.00116, 0.0101)	30	8.67	40
(0.0181, 0.2060)	(5.744, 64.38)	19.581	(0.0005, 0.0255)	(0.0013, 0.0103)	50	7.82	
(0.0335, 0.3195)	(2.742, 64.42)	11.622	(0.0005, 0.0472)	(0.0028, 0.0178)	100	6.44	
(0.0135, 0.2367)	(7.557, 64.56)	17.394	(0.0005, 0.0170)	(0.0012, 0.0107)	30	8.95	80
(0.0181, 0.2679)	(5.744, 64.38)	17.680	(0.0005, 0.0255)	(0.0014, 0.0109)	50	8.05	
(0.0370, 0.2653)	(2.742, 64.42)	10.212	(0.0005, 0.0472)	(0.0028, 0.0193)	100	6.84	
(0.0315, 0.2947)	(7.557, 64.56)	16.889	(0.0005, 0.0170)	(0.0012, 0.0109)	30	9.16	120
(0.0189, 0.2654)	(5.744, 64.38)	16.995	(0.0005, 0.0255)	(0.0014, 0.0113)	50	8.01	
(0.0335, 0.3168)	(2.742, 64.42)	9.8811	(0.0005, 0.0472)	(0.0028, 0.0196)	100	6.99	
$\mu_t \in (0.2650, 0.8702)$							
(0.2885, 0.8855)	(12.098, 427.56)	88.875	(0.0005, 0.0170)	(0.0016, 0.0037)	30	2.25	20
(0.2188, 0.8397)	(5.995, 356.79)	30.213	(0.0005, 0.0246)	(0.0044, 0.0096)	50	2.18	
(0.2149, 0.8770)	(4.080, 309.23)	26.359	(0.0005, 0.0474)	(0.0054, 0.0113)	100	2.10	
(0.2697, 0.8946)	(12.098, 427.56)	67.611	(0.0005, 0.0170)	(0.0018, 0.0042)	30	2.29	40
(0.2946, 0.9036)	(5.995, 356.79)	24.046	(0.0005, 0.0246)	(0.0049, 0.0108)	50	2.19	
(0.2833, 0.8795)	(4.080, 309.23)	21.134	(0.0005, 0.0474)	(0.0059, 0.0125)	100	2.09	
(0.2538, 0.8706)	(12.098, 427.56)	59.916	(0.0005, 0.0170)	(0.0019, 0.0044)	30	2.30	80
(0.2809, 0.8639)	(5.995, 356.79)	21.570	(0.0005, 0.0246)	(0.0052, 0.0115)	50	2.20	
(0.2489, 0.8787)	(4.080, 309.23)	18.615	(0.0005, 0.0474)	(0.0064, 0.0134)	100	2.08	
(0.2691, 0.8659)	(12.098, 427.56)	57.301	(0.0005, 0.0170)	(0.0019, 0.0045)	30	2.30	120
(0.2577, 0.8713)	(5.995, 356.79)	20.888	(0.0005, 0.0246)	(0.0053, 0.0117)	50	2.19	
(0.2699, 0.8432)	(4.080, 309.23)	18.053	(0.0005, 0.0474)	(0.0065, 0.0135)	100	2.09	
$\mu_t \in (0.6404, 0.9881)$							
(0.5944, 0.9863)	(5.855, 238.37)	28.637	(0.0005, 0.0170)	(0.0005, 0.0092)	30	19.56	20
(0.7204, 0.9832)	(3.197, 276.51)	20.182	(0.0005, 0.0251)	(0.0006, 0.0130)	50	20.23	
(0.6620, 0.9842)	(2.632, 249.09)	11.784	(0.0005, 0.0508)	(0.0015, 0.0210)	100	14.26	
(0.5987, 0.9949)	(5.855, 238.37)	23.387	(0.0005, 0.0170)	(0.0004, 0.0102)	30	23.18	40
(0.5885, 0.9931)	(3.197, 276.51)	15.993	(0.0005, 0.0251)	(0.0006, 0.0147)	50	24.64	
(0.6737, 0.9810)	(2.632, 249.09)	9.382	(0.0005, 0.0508)	(0.0013, 0.0236)	100	17.99	
(0.6232, 0.9945)	(5.855, 238.37)	21.316	(0.0005, 0.0170)	(0.0004, 0.0108)	30	25.62	80
(0.6318, 0.9925)	(3.197, 276.51)	14.771	(0.0005, 0.0251)	(0.0006, 0.0152)	50	27.46	
(0.6321, 0.9812)	(2.632, 249.09)	8.5241	(0.0005, 0.0508)	(0.0012, 0.0247)	100	19.91	
(0.6505, 0.9912)	(5.855, 238.37)	20.567	(0.0005, 0.0170)	(0.0004, 0.0110)	30	26.32	120
(0.6288, 0.9912)	(3.197, 276.51)	14.201	(0.0005, 0.0251)	(0.0005, 0.0156)	50	28.44	
(0.6568, 0.9908)	(2.632, 249.09)	8.208	(0.0005, 0.0508)	(0.0012, 0.0252)	100	20.43	

Fonte: Autoria própria.

4.3 AVALIAÇÃO NUMÉRICA NO MODELO BETA NÃO LINEAR

Nosso interesse nessa seção consiste em avaliar a estatísticas considerando agora um modelo de regressão beta não linear, com média e precisão variáveis. Dessa forma, seja $y_1, y_2, \ldots, y_n$ uma amostra de n variáveis aleatórias independentes, cuja média e precisão de cada y_t, com $t = 1, \ldots, n$, satisfazem, respectivamente, as seguintes relações funcionais:

$$g(\mu_t) = \beta_1 + x_{t2}{}^{\beta_2} \qquad \text{e} \qquad \log(\phi_t) = \gamma_1 + z_{t2}{}^{\gamma_2}. \tag{4.5}$$

Inicialmente, vamos realizar a análise das medidas considerando situações em que o modelo ajustado está corretamente especificado. Nesse estudo de simulação, seguimos a metodologia similar a descrita para o modelo linear. Na Tabela 11 temos os valores fixados para os parâmetros, as distribuições assumidas para gerar os valores das covariáveis para o modelo da média e da precisão, os intervalos obtidos para μ_t e os graus de heteroscedasticidade considerados.

Tabela 11 – Valores dos parâmetros β's e γ's, fixados para a análise no modelo beta não linear com precisão variável corretamente especificado, com os intervalos dos valores de μ_t, as funções de ligação para o modelo da média, as distribuições das covariáveis do modelo da média e da precisão e os graus de heteroscedasticidade.

β_1	β_2	x_{t2}	μ_t	$g(\cdot)$	γ_1	γ_2	z_{t2}	λ
					3.0	2.6		30
-1.37	4.3	$U(0,1)$	$\mu_t \in (0.0195, 0.2300)$	Log-log	2.5	3.67	$U(0.3,1.3)$	50
					1.5	5.0		100
					3.5	3.25		30
-1.2	2.8	$U(0.5,1.5)$	$\mu_t \in (0.2726, 0.8690)$	Logit	3.0	3.61	$U(0.5,1.5)$	50
					1.95	4.1		100
					3.65	7.1		30
-0.35	1.0	$U(0.2,1.2)$	$\mu_t \in (0.6067, 0.9028)$	C-Log-Log	3.13	8.3	$U(0.2,1.2)$	50
					2.1	9.8		100

Fonte: Autoria própria.

Na Tabela 12 são apresentados os valores médios das medidas de diagnóstico calculadas nas replicações de Monte Carlo. Inicialmente avaliaremos os resultados relacionados a $\mu_t \approx 0$ e $\mu_t \approx 1$, extremos do intervalo unitário, de modo que pelos resultados da Tabela 12 notamos um comportamento similar ao ocorrido no caso linear com precisão variável, em que a medida P^2 aponta para um alto poder preditivo do modelo quanto a dados futuros, o que para nós é plausível, uma vez que o modelo está corretamente especificado. No entanto, no caso dos critérios do tipo R^2, em especial R^2_{FC} e $R^2_{FC_c}$, observamos por meio dos valores apresentados que tais medidas indicam haver dificuldade em ajustar o conjunto dos dados original. Esse padrão é muito mais evidente quando $\mu_t \approx 0$.

Para compreendermos a causa deste comportamento das medidas R^2 em um modelo bem especificado, devemos considerar inicialmente que esse pode ser um problema referente à amostra, associada a algumas características da distribuição beta e ao método de estimação. Dessa forma, os valores mais reduzidos dos critérios R^2 para $\mu_t \approx 0$ e $\mu_t \approx 1$ estão relacionados com as amostras geradas durante o estudo de simulação. É importante observar que se y_t segue distri-

buição beta com parâmetros μ_t e ϕ_t, então $(1-y_t)$ também segue distribuição beta com média $(1-\mu_t)$ e precisão ϕ_t. Assim, espera-se comportamentos equivalentes da variável resposta quando sua média está próxima aos extremos do intervalo unitário.

Uma vez que os pseudos R^2 abordados ao longo desse trabalho estão associados ao desempenho da estimação dos parâmetros da distribuição de probabilidades de y_t. De fato, os baixos valores das medidas R^2 se deve à associação entre o método de estimação (no caso o método de máxima verossimilhança) e as características da função de densidade conjunta do vetor aleatório $(y_1,\ldots,y_n)^\top$ a ser maximizada. Os critérios do tipo R^2, em especial o critério R^2_{FC}, captam a dificuldade do método de estimação por máxima verossimilhança em estimar os parâmetros do modelo de regressão, baseado na distribuição beta, quando o domínio da variável resposta está próxima aos extremos do intervalo unitário.

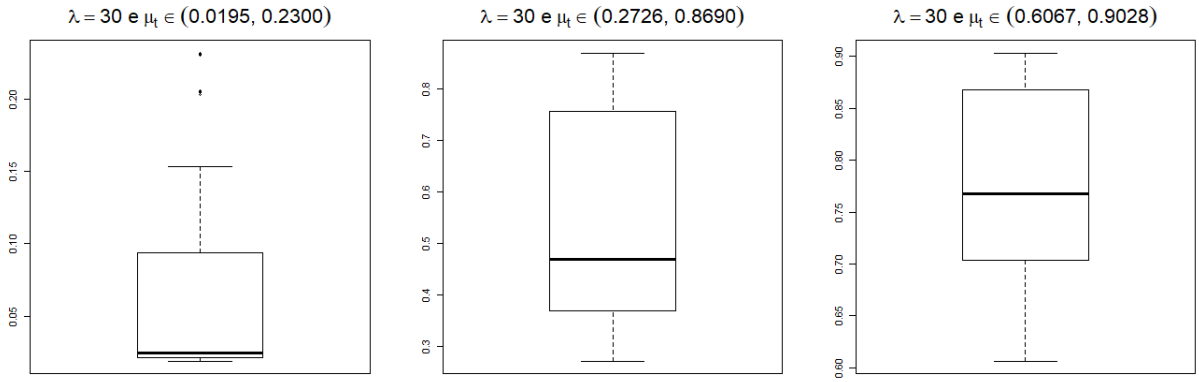
No caso do modelo de regressão beta essa dificuldade em estimar os parâmetros é ainda maior para o submodelo da precisão. Como mostrado por Ospina et al. (2006), $\hat{\phi}$ é extremamente viesado, em especial quando a média da variável resposta está próxima de zero ou de um. Essa é uma razão adicional para que os critérios do tipo R^2 apresentem valores mais reduzidos que os critérios do tipo P^2. O desempenho da estimação por máxima verossimilhança dos parâmetros do modelo de regressão beta difere bastante quando $\mu_t \in (0.2726, 0.8690)$, em que 50% ou mais das observações da resposta apresentam $\mu_t \approx 0.5$. Ou seja, são valores da variável respostas distribuídos em torno de 0.5, ou mais distantes dos extremos. Nesse caso, independentemente do tamanho da amostra e do grau de heteroscedasticidade, os valores das estatísticas são consideravelmente elevados e alcançam até valores muito próximos de um (como vemos para $\lambda = 30$). Vale observar que esse fato ocorre mesmo com o cenário em que ambos os preditores do submodelo da média e da precisão apresentam relações não lineares dos parâmetros.

Entretanto, podemos observar que para $\mu_t \approx 0$ os valores do R^2_{FC} são bem menores que para $\mu_t \approx 1$. Buscando compreender o motivo desse comportamento, apresentamos na Figura 10 os gráficos de *boxplot* dos valores gerados das respostas. De modo que para cada cenário de μ_t o valor da i-ésima resposta no gráfico corresponde à média das i-ésimas respostas das réplicas de Monte Carlo, com $n = 120$ e $\lambda = 30$. Podemos observar que as respostas geradas próximas de zero apresentam distribuição muito mais assimétrica, com pontos atípicos/influentes, os quais afetam fortemente o processo de estimação por máxima verossimilhança. Embora a função de ligação log-log amenize o efeito desses pontos atípicos, produzindo valores preditos menos afetados, segundo Espinheira, et al. (2019), o desempenho do processo de estimação nesse cenário da média ainda é inferior aos demais casos, em que $\mu_t \approx 0.5$ ou $\mu_t \approx 1$.

Outro fator referente à amostra que pode estar interferindo nos valores do R^2_{FC} em todos os cenários de μ_t com $\lambda = 100$ são as variâncias elevadas. Uma vez que os intervalos das variâncias são similares para todos os cenários de μ_t, conforme podemos observar pela Tabela 13. É possível que as variâncias sejam consideravelmente maiores do que deveriam, por exemplo, quando comparadas com as variâncias da distribuição gama unitária. Rocha et al.(2020) apresentam um gráfico na Figura 10, em que são comparadas as variâncias das distribuições beta e gamma unitária, quando $\mu = 0.1$, $\mu = 0.5$ e $\mu = 0.9$, para valores fixos e equivalentes de ϕ. Com base nesta figura nota-se que, apesar das variâncias nos extremos serem menores que as

variâncias no centro do intervalo $(0,1)$, é notável como essas variâncias são substancialmente ainda menores para a distribuição gama unitária que para a distribuição beta.

Figura 10 – Gráficos de *boxplot* dos valores de y_t com o modelo beta não linear corretamente especificado.



Fonte: Autoria própria.

O critério de seleção *BV* considera todas as questões acima discutidas. De modo que para $\mu_t \in (0.2726, 0.8690)$ ele apresenta sempre valores elevados, indicando a boa adequabilidade do modelo, tanto na questão de predição quanto no ajuste. Já para $\mu_t \in (0.0195, 0.2300)$ ou $\mu_t \in (0.6067, 0.9028)$, os menores valores do critério *BV* são próximos de 0.6, quando $\lambda = 100$. Além disso, em todos os casos temos $\hat{\alpha} \approx 0.5$, especialmente quando $n = 120$, isto é, o mínimo e o máximo tem aproximadamente o mesmo peso em todos os cenários considerados, inclusive para $\mu_t \approx 0$. Assim, para um modelo bem ajustado, esperamos que $\hat{\alpha}$ deva ser o mais próximo de 0.5 possível.

Nas Figuras 11, 12 e 13 temos os gráficos de *boxplot* dos critérios de seleção para o modelo beta não linear corretamente especificado, para cada cenário de μ_t, com os graus de heteroscedasticidade 30 e 100, e os tamanhos amostrais 20, 80 e 120. Observamos então por esses gráficos, comportamentos similares ao caso linear. De modo que as estatísticas P^2 geralmente apresentam valores mais próximos de um, e vemos também que existe a possibilidade de valores negativos para os critérios R^2 corrigidos, quando o tamanho da amostra é pequeno e μ_t localizado nos extremos de $(0,1)$. Diante da diferença entre os comportamentos das estatísticas P^2 e R^2, o critério *BV* apresenta dispersão consideravelmente mais elevada do que as demais estatísticas.

Observando ainda os gráficos das Figuras 11, 12 e 13, devemos ressaltar que, nos extremos do intervalo unitário, em especial para $\mu_t \approx 0$, os valores dos critérios do tipo R^2, em especial o R^2_{FC}, apresentam valores ainda menores que no caso linear, particularmente quando $\lambda = 100$. Este fato se deve a questões similares em ambos os casos, a saber: o baixo desempenho da estimação por máxima verossimilhança (principalmente referente à precisão), associado aos fatores como proximidade da resposta dos extremos do intervalo unitário, assimetria da distribuição da resposta, entre outros. Sendo que, agora existe uma dificuldade adicional, a não linearidade da relação entre γ_2 e z_2, o que torna $\hat{\gamma}_2$ mais viesado que para o modelo linear (Simas, et al. 2010). A estatística *BV* resume muito bem as informações contidas nos três tipos de critério.

Tabela 12 – Valores médios das estatísticas de avaliação de desempenho no modelo beta não linear com dispersão variável corretamente especificado, com $g(\cdot)$ adequada a cada cenário de μ_t e $h(\cdot)$ log.

	$g(\cdot)$: Log-Log $\mu_t \in (0.0195, 0.2300)$			$g(\cdot)$: Logit $\mu_t \in (0.2726, 0.8690)$			$g(\cdot)$: C-Log-Log $\mu_t \in (0.6067, 0.9028)$			
λ	30	50	100	30	50	100	30	50	100	n
P^2	0.8859	0.8919	0.8886	0.9840	0.9807	0.9681	0.9503	0.9400	0.9133	20
P_c^2	0.8645	0.8716	0.8678	0.9810	0.9771	0.9621	0.9410	0.9287	0.8971	
R_{FC}^2	0.6993	0.5946	0.4090	0.9580	0.9337	0.8332	0.7935	0.6949	0.4791	
$R_{FC_c}^2$	0.6429	0.5185	0.2982	0.9501	0.9212	0.8019	0.7548	0.6377	0.3815	
R_{LR}^2	0.7737	0.7285	0.6858	0.9774	0.9692	0.9406	0.8609	0.8079	0.7107	
$R_{LR_c}^2$	0.6928	0.6315	0.5736	0.9693	0.9582	0.9193	0.8113	0.7394	0.6074	
BV	0.7365	0.6747	0.5628	0.9656	0.9491	0.8797	0.8338	0.7583	0.5942	
$\hat{\alpha}$	0.4440	0.4772	0.4984	0.4643	0.4880	0.5120	0.4290	0.4375	0.4540	
P^2	0.8449	0.8538	0.8519	0.9792	0.9748	0.9597	0.9300	0.9272	0.9052	40
P_c^2	0.8320	0.8416	0.8396	0.9774	0.9727	0.9564	0.9241	0.9212	0.8973	
R_{FC}^2	0.6832	0.5774	0.3967	0.9553	0.9290	0.8246	0.7840	0.6839	0.4618	
$R_{FC_c}^2$	0.6568	0.5421	0.3465	0.9515	0.9231	0.8100	0.7660	0.6576	0.4170	
R_{LR}^2	0.7535	0.7086	0.6704	0.9752	0.9661	0.9353	0.8418	0.7871	0.6790	
$R_{LR_c}^2$	0.7172	0.6657	0.6219	0.9716	0.9611	0.9258	0.8186	0.7558	0.6318	
BV	0.7340	0.6726	0.5757	0.9635	0.9457	0.8774	0.8346	0.7720	0.6239	
$\hat{\alpha}$	0.4533	0.4605	0.4886	0.4571	0.4704	0.4958	0.4396	0.4553	0.4665	
P^2	0.8248	0.8347	0.8287	0.9762	0.9715	0.9537	0.9196	0.9174	0.8973	80
P_c^2	0.8179	0.8282	0.8220	0.9752	0.9704	0.9519	0.9164	0.9142	0.8932	
R_{FC}^2	0.6753	0.5676	0.3879	0.9537	0.9272	0.8165	0.7783	0.6788	0.4510	
$R_{FC_c}^2$	0.6624	0.5505	0.3638	0.9518	0.9243	0.8092	0.7696	0.6661	0.4293	
R_{LR}^2	0.7449	0.7023	0.6703	0.9741	0.9650	0.9331	0.8356	0.7805	0.6733	
$R_{LR_c}^2$	0.7276	0.6822	0.6480	0.9724	0.9626	0.9286	0.8245	0.7656	0.6512	
BV	0.7316	0.6764	0.5803	0.9625	0.9454	0.8742	0.8356	0.7770	0.6341	
$\hat{\alpha}$	0.4600	0.4740	0.4929	0.4668	0.4790	0.4895	0.4623	0.4694	0.4711	
P^2	0.8186	0.8280	0.8192	0.9753	0.9702	0.9514	0.9166	0.9128	0.8915	120
P_c^2	0.8139	0.8235	0.8146	0.9747	0.9694	0.9502	0.9144	0.9106	0.8886	
R_{FC}^2	0.6719	0.5647	0.3828	0.9533	0.9264	0.8129	0.7767	0.6777	0.4508	
$R_{FC_c}^2$	0.6635	0.5534	0.3669	0.9521	0.9245	0.8081	0.7709	0.6694	0.4366	
R_{LR}^2	0.7425	0.7007	0.6690	0.9738	0.9646	0.9322	0.8344	0.7792	0.6730	
$R_{LR_c}^2$	0.7312	0.6876	0.6545	0.9727	0.9630	0.9293	0.8272	0.7695	0.6587	
BV	0.7319	0.6786	0.5756	0.9627	0.9451	0.8727	0.8367	0.7786	0.6448	
$\hat{\alpha}$	0.4673	0.4819	0.4843	0.4772	0.4793	0.4855	0.4714	0.4728	0.4860	

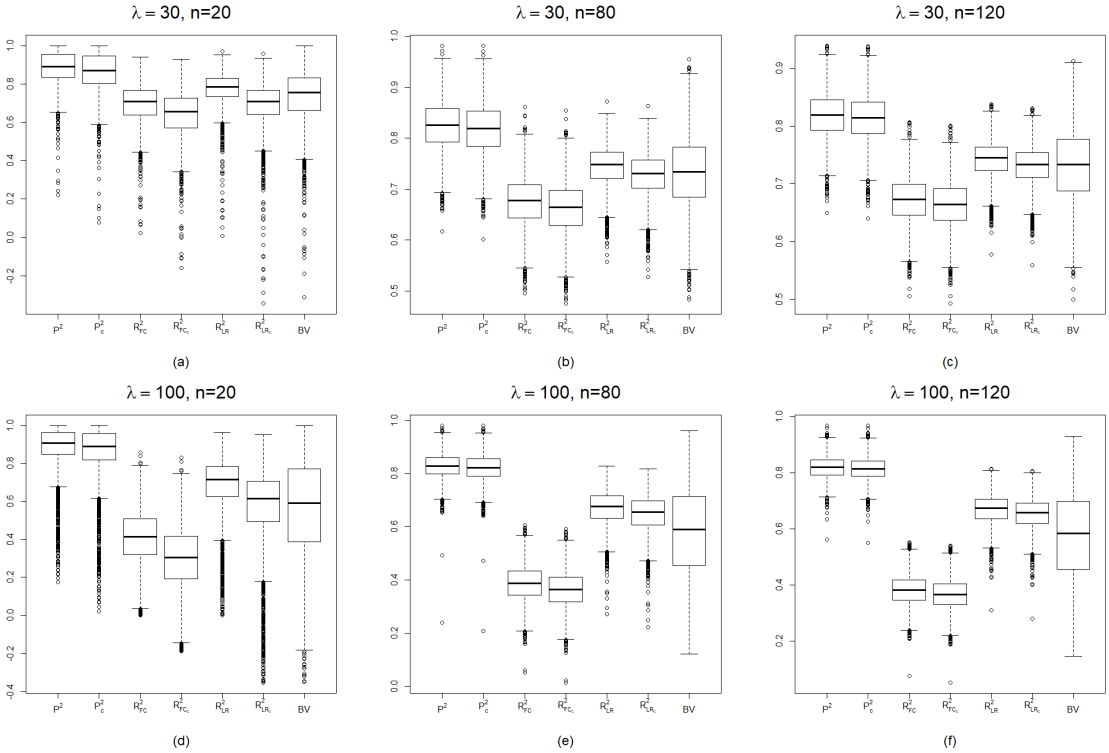
Fonte: Autoria própria.

Tabela 13 – Intervalos de valores de ϕ_t e de $\mathrm{Var}(y_t)$, e valores de λ, juntamente com suas respectivas médias das estimativas, e também as médias das estimativas dos intervalos de μ_t, no modelo beta não linear com dispersão variável corretamente especificado, usando $g(\cdot)$ log-log para $\mu_t \in (0.0195, 0.2300)$, $g(\cdot)$ logit para $\mu_t \in (0.2726, 0.8690)$, $g(\cdot)$ c-log-log para $\mu_t \in (0.6067, 0.9028)$ e $h(.)$ log.

$\hat{\mu}_t$	ϕ_t	$\hat{\phi}_t$	$\mathrm{Var}(y_t)$	$\widehat{\mathrm{Var}}(y_t)$	λ	$\hat{\lambda}$	n
$\mu_t \in (0.0195, 0.2300)$							
(0.0236, 0.2456)	(21.787, 120.89)	(40.175, 8773.9)	(0.0002, 0.0057)	(0.0002, 0.0051)	30	26.5	20
(0.0157, 0.2126)	(12.540, 119.51)	(20.616, 11059)	(0.0002, 0.0095)	(0.0002, 0.0090)	50	40.0	
(0.0205, 0.2331)	(4.518, 97.50)	(6.414, 18280)	(0.0002, 0.0243)	(0.0003, 0.0235)	100	70.8	
(0.0161, 0.2142)	(21.787, 120.89)	(27.424, 203.59)	(0.0002, 0.0057)	(0.0002, 0.0053)	30	29.2	40
(0.0158, 0.2133)	(12.540, 119.51)	(14.840, 213.45)	(0.0002, 0.0095)	(0.0002, 0.0089)	50	47.1	
(0.0190, 0.2289)	(4.518, 97.50)	(5.008, 192.54)	(0.0002, 0.0243)	(0.0002, 0.0235)	100	93.3	
(0.0183, 0.2253)	(21.787, 120.89)	(23.811, 145.70)	(0.0002, 0.0057)	(0.0002, 0.0054)	30	29.8	80
(0.0262, 0.2539)	(12.540, 119.51)	(13.260, 148.06)	(0.0002, 0.0095)	(0.0002, 0.0093)	50	49.9	
(0.0160, 0.2142)	(4.518, 97.50)	(4.727, 125.20)	(0.0002, 0.0243)	(0.0002, 0.0240)	100	104.5	
(0.0197, 0.2301)	(21.787, 120.89)	(22.897, 135.97)	(0.0002, 0.0057)	(0.0002, 0.0055)	30	30.3	120
(0.0198, 0.2318)	(12.540, 119.51)	(12.985, 137.50)	(0.0002, 0.0095)	(0.0002, 0.0094)	50	50.4	
(0.0178, 0.2217)	(4.518, 97.50)	(4.676, 113.60)	(0.0002, 0.0243)	(0.0002, 0.0240)	100	104.7	
$\mu_t \in (0.2726, 0.8690)$							
(0.2747, 0.8801)	(39.298, 955.84)	(79.556, 7550.3)	(0.0002, 0.0056)	(0.0002, 0.0067)	30	40.7	20
(0.2650, 0.8943)	(23.121, 940.15)	(45.475, 10360)	(0.0002, 0.0095)	(0.0001, 0.0116)	50	73.2	
(0.2943, 0.8712)	(7.829, 711.69)	(14.470, 7606.8)	(0.0002, 0.0262)	(0.0002, 0.0309)	100	149.3	
(0.2741, 0.8616)	(39.298, 955.84)	(45.227, 1512.1)	(0.0002, 0.0056)	(0.0002, 0.0058)	30	34.2	40
(0.2819, 0.8548)	(23.121, 940.15)	(25.949, 1582.7)	(0.0002, 0.0095)	(0.0002, 0.0098)	50	58.0	
(0.2694, 0.8642)	(7.829, 711.69)	(8.689, 1239.4)	(0.0002, 0.0262)	(0.0002, 0.0271)	100	125.1	
(0.2732, 0.8678)	(39.298, 955.84)	(41.529, 1141.6)	(0.0002, 0.0056)	(0.0002, 0.0057)	30	31.9	80
(0.2723, 0.8623)	(23.121, 940.15)	(24.214, 1151.4)	(0.0002, 0.0095)	(0.0002, 0.0096)	50	53.2	
(0.2684, 0.8664)	(7.829, 711.69)	(8.125, 894.6)	(0.0002, 0.0262)	(0.0002, 0.0265)	100	114.7	
(0.2706, 0.8693)	(39.298, 955.84)	(40.533, 1082.1)	(0.0002, 0.0056)	(0.0002, 0.0057)	30	31.7	120
(0.2711, 0.8663)	(23.121, 940.15)	(23.953, 1064.2)	(0.0002, 0.0095)	(0.0002, 0.0095)	50	51.9	
(0.2677, 0.8768)	(7.829, 711.69)	(7.991, 819.8)	(0.0002, 0.0262)	(0.0002, 0.0265)	100	111.4	
$\mu_t \in (0.6067, 0.9028)$							
(0.6130, 0.9464)	(38.479, 593.39)	(60.475, 22580)	(0.0002, 0.0057)	(0.0004, 0.0049)	30	10.9	20
(0.5989, 0.8834)	(22.875, 585.89)	(34.262, 25365)	(0.0002, 0.0095)	(0.0007, 0.0082)	50	12.4	
(0.5910, 0.9018)	(8.166, 451.03)	(11.527, 55282)	(0.0002, 0.0247)	(0.0014, 0.0217)	100	15.4	
(0.6057, 0.8966)	(38.479, 593.39)	(46.233, 1669.4)	(0.0002, 0.0057)	(0.0002, 0.0054)	30	23.4	40
(0.6017, 0.9078)	(22.875, 585.89)	(26.438, 1864.4)	(0.0002, 0.0095)	(0.0002, 0.0091)	50	35.7	
(0.6190, 0.9003)	(8.166, 451.03)	(9.295, 1413.9)	(0.0002, 0.0247)	(0.0004, 0.0236)	100	52.3	
(0.6148, 0.9010)	(38.479, 593.39)	(40.663, 843.30)	(0.0002, 0.0057)	(0.0002, 0.0056)	30	30.7	80
(0.6086, 0.9045)	(22.875, 585.89)	(24.109, 861.94)	(0.0002, 0.0095)	(0.0002, 0.0093)	50	50.4	
(0.6042, 0.8908)	(8.166, 451.03)	(8.490, 687.98)	(0.0002, 0.0247)	(0.0002, 0.0244)	100	101.7	
(0.6014, 0.9004)	(38.479, 593.39)	(39.682, 733.62)	(0.0002, 0.0057)	(0.0002, 0.0056)	30	31.5	120
(0.6117, 0.9035)	(22.875, 585.89)	(23.633, 736.52)	(0.0002, 0.0095)	(0.0002, 0.0093)	50	51.9	
(0.6163, 0.9104)	(8.166, 451.03)	(8.353, 583.00)	(0.0002, 0.0247)	(0.0002, 0.0245)	100	108.2	

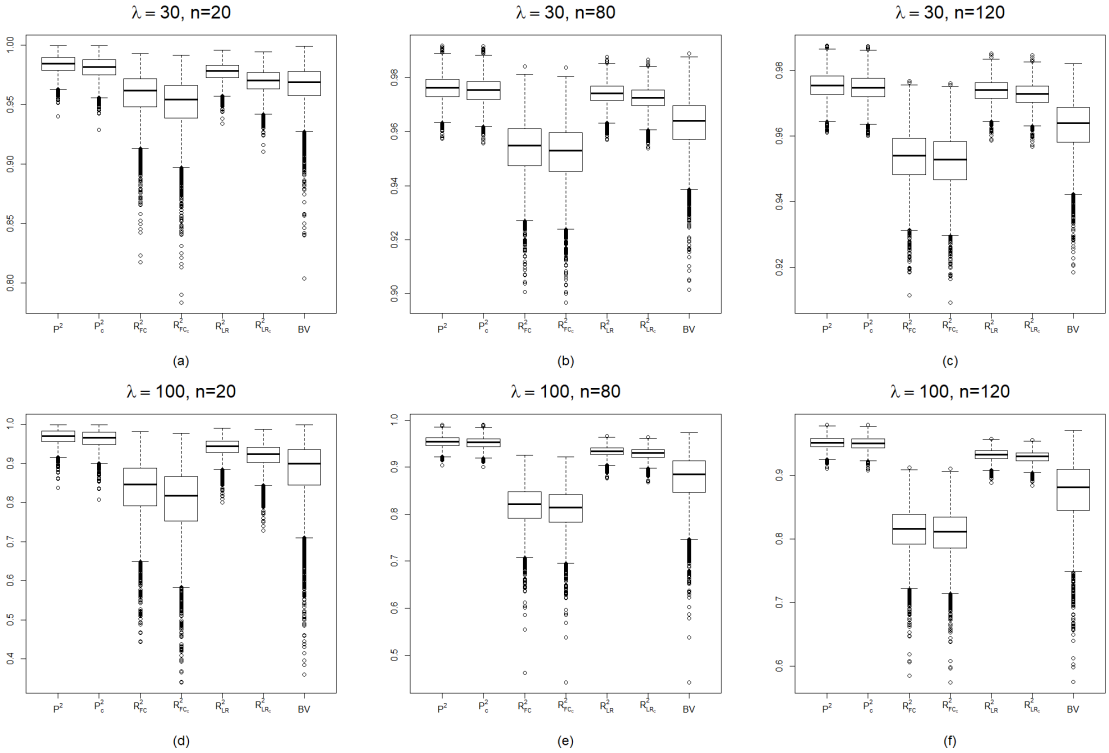
Fonte: Autoria própria.

Figura 11 – Gráficos de *boxplot* dos valores das estatísticas de adequabilidade de ajuste, no modelo beta não linear com dispersão variável corretamente especificado, para $\mu_t \in (0.0195, 0.2300)$, com $g(\cdot)$ log-log e $h(\cdot)$ log.



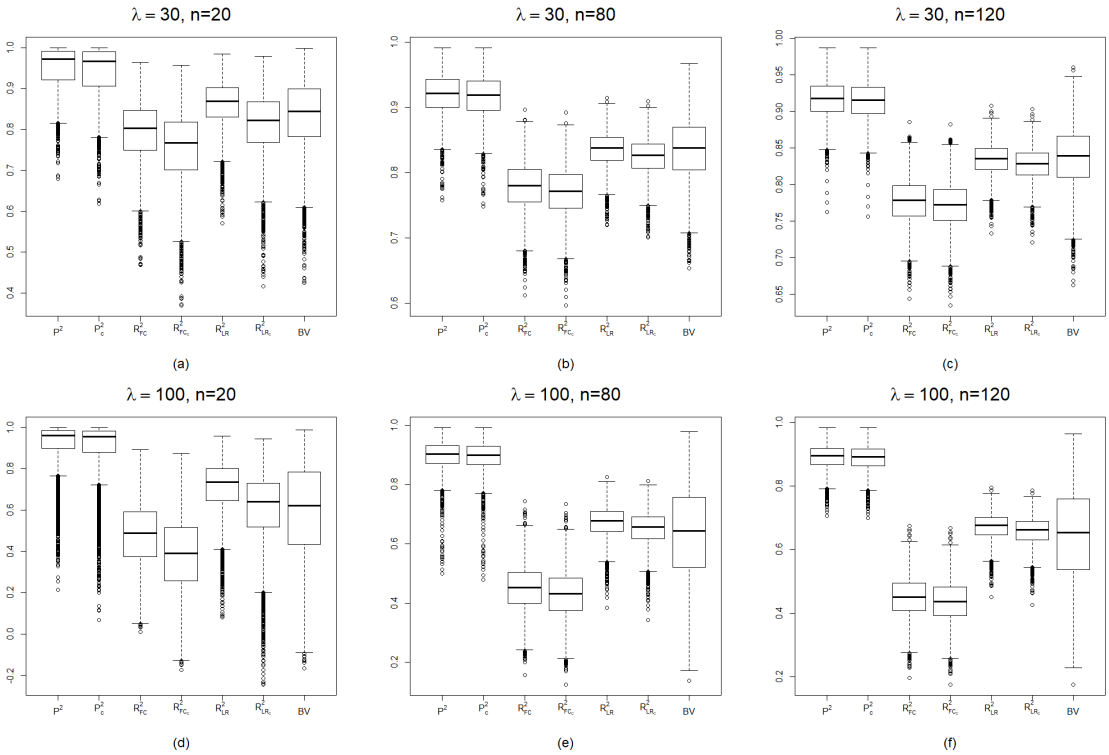
Fonte: Autoria própria.

Figura 12 – Gráficos de *boxplot* dos valores das estatísticas de adequabilidade de ajuste, no modelo beta não linear com dispersão variável corretamente especificado, para $\mu_t \in (0.2726, 0.8690)$, com $g(\cdot)$ logit e $h(\cdot)$ log.



Fonte: Autoria própria.

Figura 13 – Gráficos de *boxplot* dos valores das estatísticas de adequabilidade de ajuste, no modelo beta não linear com dispersão variável corretamente especificado, para $\mu_t \in (0.6067, 0.9028)$, com $g(\cdot)$ c-log-log e $h(\cdot)$ log.



Fonte: Autoria própria.

4.3.1 Estudo Descritivo da Estatística $\hat{\alpha}$ - Caso Não Linear

Com o objetivo de avaliar o comportamento de $\hat{\alpha}$ para o modelo beta não linear, corretamente especificado e com dispersão variável, encontramos na Tabela 14 as medidas descritivas dessa variável aleatória, com base nos valores obtidos pela simulação descrita anteriormente. Podemos então verrificar que, geralmente, as médias de $\hat{\alpha}$ encontram-se em torno de 0.48, principalmente para $\mu_t \in (0.0195, 0.2300)$ e $\mu_t \in (0.2726, 0.8690)$ com $\lambda \geq 50$, essa afirmação torna-se mais evidente quando aumenta-se o tamanho amostral. Observando o desvio-padrão encontramos valores próximos de 0.2, no entanto, vemos que para valores de μ_t próximos de zero ou de um, os desvios são maiores do que para valores de μ_t centrais. Ademais, em todos os cenários de μ_t notamos que os desvios de $\hat{\alpha}$ tendem a aumentar quando λ aumenta, porém, mantêm-se constante quando o tamanho amostral cresce. Dessa forma, podemos supor que a média e o desvio-padrão de $\hat{\alpha}$ podem ser considerados fixos, quando ajustamos corretamente um modelo beta não linear.

Pelos valores apresentados para curtose vemos que estes são sempre próximos de 2.0, e que essa medida torna-se ainda mais próxima desse valor quando λ ou o tamanho amostral aumentam. Vale notar também que para $\mu_t \in (0.0195, 0.2300)$ a distribuição de $\hat{\alpha}$ apresenta menores curtoses do que para $\mu_t \in (0.2726, 0.8690)$ e $\mu_t \in (0.6067, 0.9028)$. Dessa forma, a curva da distribuição normal padrão apresenta uma forma mais fechada, com valores mais concentrados em torno da média, que o gráfico da distribuição dos valores de $\hat{\alpha}$, principalmente para valores de μ_t próximos de zero. Notamos também que quando o tamanho amostral ou o

grau de heteroscedasticidade aumentam o coeficiente de assimetria diminui. Logo, a curva da distribuição de $\hat{\alpha}$ torna-se mais simétrica.

Tabela 14 – Medidas descritivas de $\hat{\alpha}$ para diferentes valores de n e λ, em diferentes cenários da média μ_t, no modelo beta não linear com dispersão variável corretamente especificado, usando $g(\cdot)$ log-log para $\mu_t \in (0.0195, 0.2300)$, $g(\cdot)$ logit para $\mu_t \in (0.2726, 0.8690)$, $g(\cdot)$ c-log-log para $\mu_t \in (0.6067, 0.9028)$ e $h(.)$ log.

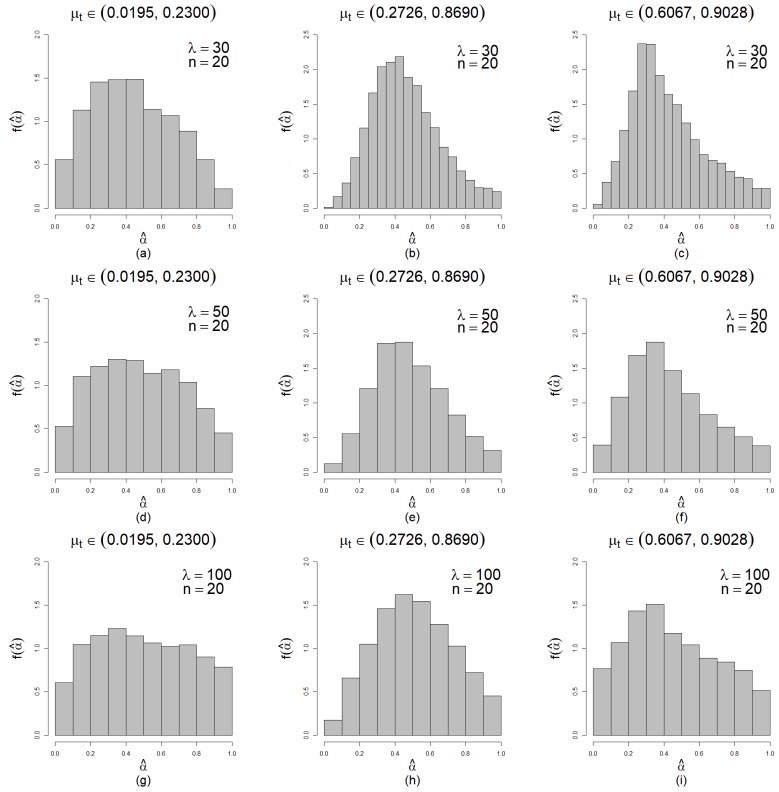
	Média			Desvio-Padrão			
λ \ $\mu_t \in$	$(0.0195, 0.2300)$	$(0.2726, 0.8690)$	$(0.6067, 0.9028)$	$(0.0195, 0.2300)$	$(0.2726, 0.8690)$	$(0.6067, 0.9028)$	n
20	0.4440	0.4643	0.4290	0.2322	0.1915	0.2101	20
50	0.4772	0.4880	0.4375	0.2485	0.2036	0.2274	
100	0.4984	0.5120	0.4540	0.2673	0.2213	0.2562	
20	0.4533	0.4571	0.4396	0.2327	0.1737	0.1900	40
50	0.4605	0.4704	0.4553	0.2530	0.1879	0.2175	
100	0.4886	0.4958	0.4665	0.2635	0.2134	0.2547	
20	0.4600	0.4668	0.4623	0.2362	0.1745	0.1910	80
50	0.4740	0.4790	0.4694	0.2525	0.1883	0.2160	
100	0.4929	0.4895	0.4711	0.2695	0.2160	0.2525	
20	0.4673	0.4772	0.4714	0.2364	0.1760	0.1893	120
50	0.4819	0.4793	0.4728	0.2522	0.1920	0.2173	
100	0.4843	0.4855	0.4860	0.2693	0.2179	0.2534	
	Curtose			Assimetria			
20	2.1504	2.8506	2.8128	0.2290	0.4929	0.6786	20
50	2.0001	2.5007	2.4690	0.1225	0.3090	0.4966	
100	1.8939	2.2649	2.0766	0.0824	0.1152	0.2782	
20	2.1432	2.6269	2.6701	0.2167	0.1220	0.3852	40
50	1.9872	2.4617	2.3115	0.1828	0.0713	0.2231	
100	1.9531	2.2428	2.0139	0.0964	-0.0169	0.1429	
20	2.1105	2.5026	2.3720	0.1987	-0.0391	0.1414	80
50	1.9961	2.3517	2.1830	0.1025	0.0329	0.1389	
100	1.8953	2.2617	2.0278	0.0405	-0.0717	0.1202	
20	2.0658	2.5063	2.3512	0.1619	0.0073	0.1213	120
50	1.9996	2.3795	2.1793	0.0938	-0.0157	0.1053	
100	1.9066	2.2350	1.9892	0.0522	-0.0677	0.0621	

Fonte: Autoria própria.

Os histogramas dos valores de $\hat{\alpha}$ para o modelo beta não linear corretamente especificado são apresentados pelas Figuras 14, 15 e 16, para os tamanhos amostrais 20, 80 e 120, respectivamente, e para cada cenário de μ_t, com os graus de heteroscedasticidade 30, 50 e 100. Observando essas figuras vemos que para valores de μ_t próximos de zero, temos uma maior frequência de valores de $\hat{\alpha}$ localizados nos extremos do intervalo $(0,1)$, do que para os demais cenários de μ_t, isso ocorre mesmo para tamanhos amostrais altos.

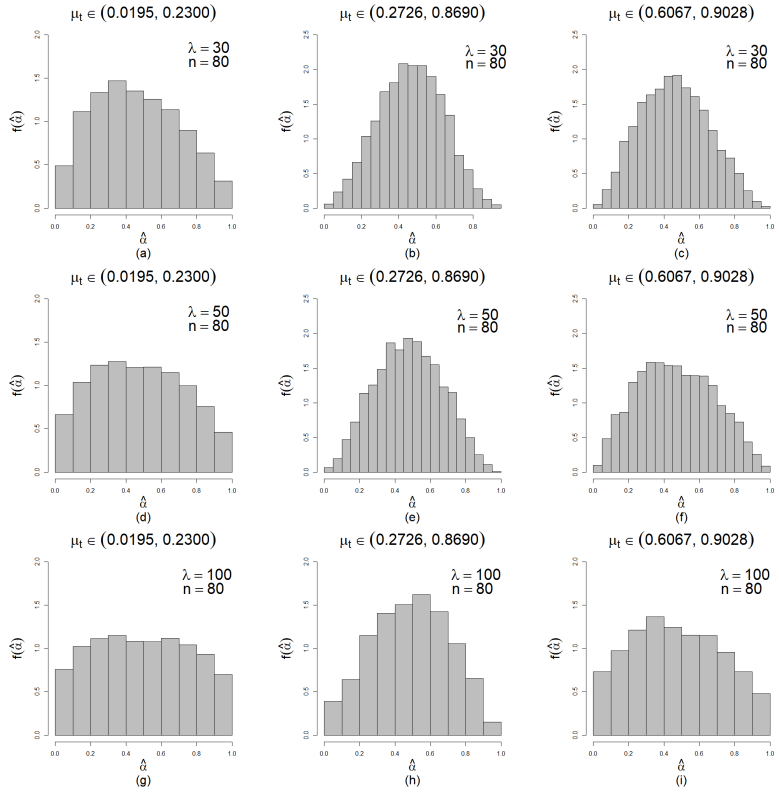
Prosseguindo, podemos observar que tanto para $\mu_t \in (0.2726, 0.8690)$ quanto para $\mu_t \in (0.6067, 0.9028)$, a distribuição dos valores de $\hat{\alpha}$ apresenta forma mais simétrica quando o tamanho amostral aumenta. Dessa forma, assim como para o modelo beta linear, temos que, quando ajustamos corretamente o modelo beta não linear com parâmetro dispersão variável e o tamanho amostral é consideravelmente alto, $\hat{\alpha}$ segue distribuição aproximadamente simétrica, cuja média, desvio-padrão e curtose são aproximadamente 0.48, 0.2 e 2.0, respectivamente.

Figura 14 – Histogramas de $\hat{\alpha}$ no modelo beta não linear corretamente especificado, com ϕ variável e $n = 20$.



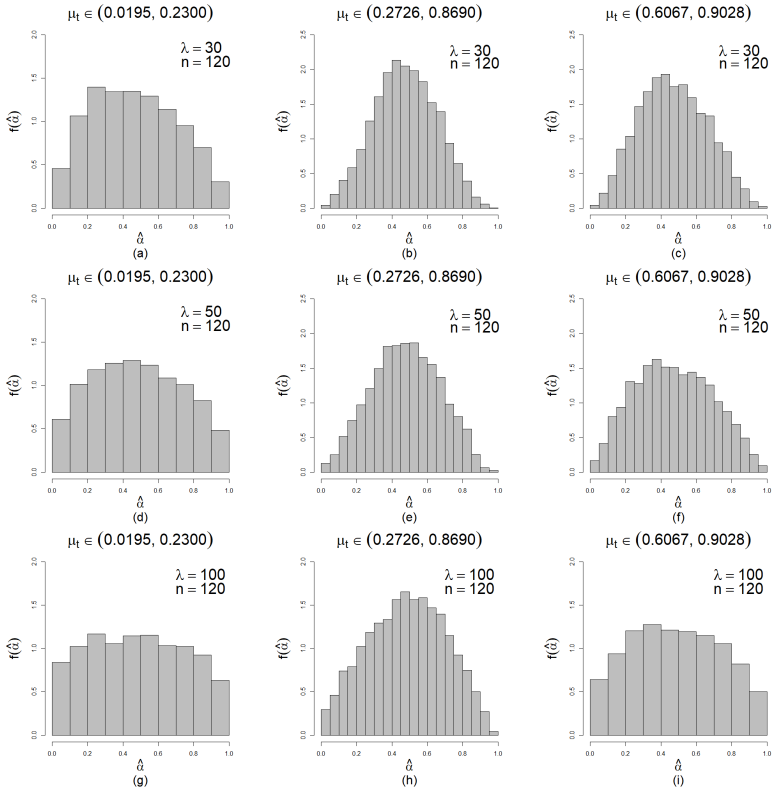
Fonte: Autoria própria.

Figura 15 – Histogramas de $\hat{\alpha}$ no modelo beta não linear corretamente especificado, com ϕ variável e $n = 80$.



Fonte: Autoria própria.

Figura 16 – Histogramas de $\hat{\alpha}$ no modelo beta não linear corretamente especificado, com ϕ variável e $n = 120$.



Fonte: Autoria própria.

4.3.2 Avaliação Para Tamanhos de Amostras Maiores

Prosseguimos com nosso estudo tendo o interesse de avaliar o comportamento e desempenho das estatísticas de adequabilidade para um tamanho amostral consideravelmente elevado, $n = 500$, sob um modelo beta não linear corretamente especificado e também sob um modelo beta mal especificado, de modo que nesse segundo caso foi desconsiderado no ajuste o fator de não linearidade. Nesse estudo foram utilizadas $MC = 5000$ réplicas de Monte Carlo e $B = 250$ réplicas *bootstrap*. Além disso, consideramos apenas o cenário de $\mu_t \in (0.25830.8695)$, com $g(\cdot)$ logit e $\lambda = 50$.

Dessa forma, seja $y_1, y_2, \ldots, y_n$ uma amostra de n variáveis aleatórias independentes, cuja média e precisão de cada y_t, com $t = 1, \ldots, n$, satisfazem, respectivamente, as seguintes relações funcionais:

$$\log\left(\frac{\mu_t}{1-\mu_t}\right) = \beta_1 + x_{t2}{}^{\beta_2} \qquad \text{e} \qquad \log(\phi_t) = \gamma_1 + \gamma_2 z_{t2}. \tag{4.6}$$

Podemos notar que foi assumida uma estrutura não linear apenas no submodelo da média. Os valores assumidos para os parâmetros foram: $(\beta_1, \beta_2) = (-1.2, 2.8)$ e $(\gamma_1, \gamma_2) = (1.7, 3.3)$, e os valores das covariáveis x_{t2} e z_{t2} foram gerados a partir da distribuição $U(0.5, 1.5)$. Na Tabela 15 encontramos os valores médios das estatísticas de avaliação de desempenho, com o modelo beta não linear corretamente especificado e quando é omitida a não linearidade da média.

Tabela 15 – Valores médios das estatísticas de avaliação de desempenho no modelo beta não linear corretamente especificado e quando é omitida a não linearidade da média, para $\mu_t \in (0.2583, 0.8695)$, $g(\cdot)$ logit, $\lambda = 50$ e $n = 500$.

Modelo Verdadeiro	$\log\left(\frac{\mu_t}{1-\mu_t}\right) = \beta_1 + x_{t2}{}^{\beta_2}$ e $\log(\phi_t) = \gamma_1 + \gamma_2 z_{t2}$		
Modelo Ajustado	$\log\left(\frac{\mu_t}{1-\mu_t}\right) = \beta_1 + x_{t2}{}^{\beta_2}$ e $\log(\phi_t) = \gamma_1 + \gamma_2 z_{t2}$	$\log\left(\frac{\mu_t}{1-\mu_t}\right) = \beta_1 + \beta_2 x_{t2}$ e $\log(\phi_t) = \gamma_1 + \gamma_2 z_{t2}$	n
P^2	0.9677	0.9096	
P_c^2	0.9675	0.9091	
R_{FC}^2	0.9371	0.8936	
$R_{FC_c}^2$	0.9367	0.8929	500
R_{LR}^2	0.9999	0.9993	
$R_{LR_c}^2$	0.9999	0.9993	
BV	0.9672	0.9154	
$\hat{\alpha}$	0.4856	0.2133	

Fonte: Autoria própria.

Tabela 16 – Valores verdadeiros e estimados dos parâmetros β's e γ's, dos intervalos de μ, ϕ e dE $\text{Var}(y_t)$, e do grau de heteroscedasticidade λ, para o modelo corretamente especificado: $\log[(\mu_t)/(1-\mu_t)] = \beta_1 + x_{t2}{}^{\beta_2}$ e $\log(\phi_t) = \gamma_1 + \gamma_2 z_{t2}$, e para o modelo omitindo a não linearidade da média: $\log[(\mu_t)/(1-\mu_t)] = \beta_1 + \beta_2 x_{t2}$ e $\log(\phi_t) = \gamma_1 + \gamma_2 z_{t2}$, com $n = 500$.

Quantidades Estimadas	Valores Verdadeiros	Modelo Corretamente Especificado	Modelo Omitindo a Não Linearidade da Média	n
β_1	-1.2	-1.2000	-2.9412	
β_2	2.8	2.8003	2.9193	
γ_1	1.7	1.6934	2.5516	
γ_2	3.3	3.3138	1.6677	500
μ_t	$(0.2583, 0.8695)$	$(0.2586, 0.8705)$	$(0.1825, 0.8066)$	
ϕ_t	$(28.696, 769.88)$	$(28.941, 787.33)$	$(29.842, 156.84)$	
$\text{Var}(y_t)$	$(0.00017, 0.00841)$	$(0.00016, 0.00847)$	$(0.00105, 0.00819)$	
λ	50	51.1	7.8	

Fonte: Autoria própria.

Observando os valores Na Tabela 15 vemos que, apesar das estatísticas do tipo P^2 e R^2, e também BV (quando avaliado isoladamente), apresentarem valores um pouco mais elevados e próximos a um, quando o modelo ajustado está corretamente especificado, os valores dessas medidas quando desconsiderado o fator de não linearidade da média ainda são bastante elevados, indicando então que o modelo linear ajustado está aceitável. No entanto, se observarmos agora as médias das estimativas de μ_t, ϕ_t, $\text{Var}(y_t)$ e λ, apresentados na Tabela 16 para ambos os casos, podemos verificar que a omissão da não linearidade da média durante o ajuste afeta con-

sideravelmente tais estimativas, especialmente para os valores da precisão. Sabendo que o valor mediano verdadeiro de ϕ_t é 150.78, então, para cada réplica de Monte Carlo é provável que 50% das estimativas de ϕ_t, com $t = 1,\ldots,500$ estejam altamente viesadas, mesmo que o submodelo da precisão esteja bem especificado. Consequentemente, as estimativas dos mínimos das variâncias também são bastante afetados e assim também o grau de heteroscedasticidade.

Dessa forma, vemos que mesmo para um tamanho amostral bastante alto, as estatísticas de adequabilidade P^2, P_c^2, R_{FC}^2, $R_{FC_c}^2$, R_{LR}^2 e $R_{LR_c}^2$, e também BV sozinho, não conseguem captar os problemas destacados. Entretanto, podemos notar que $\hat{\alpha}$ foi a única estatística que apresentou alteração drástica em seu valor, de modo que foi apresentado para essa medida um valor médio bem menor do esperado, quando o modelo está correto. Muito provavelmente esse fato se deve aos altos vieses de $\hat{\phi}_t$, pois como já indicado anteriormente, $\hat{\alpha}$ é um forte indicador da qualidade das estimativas da precisão. Logo, reforçamos a importância de avaliar conjuntamente os valores de BV e $\hat{\alpha}$.

É notável o valor de $\hat{\alpha} = 0.21$ indicando a má especificação do modelo. Geralmente, em estudos de análise de adequabilidade de modelos de regressão, não é comum dispor de métodos que apontam problemas no ajuste do modelo linear devido a omissão da não linearidade no submodelo da média. Assim, ressaltamos o mérito da estatística $\hat{\alpha}$. De fato, e finalmente, o método que propomos após o estudo das estatísticas BV e $\hat{\alpha}$ é inicialmente avaliar $\hat{\alpha}$. Caso seu valor esteja muito distante de 0.5, como por exemplo, 0.3 ou 0.7, isto é um sinal que o modelo está mal especificado e deve ser descartado. Já para valores de $\hat{\alpha} \approx 0.5$, a seleção do modelo envolve a escolha de BV com valores mais próximos de um.

5 AVALIAÇÃO NUMÉRICA NO MODELO BETA COM DISPERSÃO FIXA

5.1 INTRODUÇÃO

Neste capítulo buscamos avaliar o desempenho das estatísticas BV e $\widehat{\alpha}$ considerando o modelo de regressão beta linear com o parâmetro de precisão ϕ fixo ao longo das observações. Iremos observar a partir dos estudos de simulação que esse caso merece um certo cuidado e destaque ao usarmos as nossas estatísticas de adequabilidade de ajuste, especialmente para pequenos tamanhos de amostras.

5.2 AVALIAÇÃO NO MODELO CORRETAMENTE ESPECIFICADO

Consideramos agora o modelo de regressão beta linear, com ϕ fixo. Dessa forma, seja $y_1, y_2, \ldots, y_n$ uma amostra de n variáveis aleatórias independentes, em que cada y_t, $t = 1, \ldots, n$ segue distribuição beta, com função densidade dada por (2.2), com média μ_t e precisão constante ϕ, cuja média de y_t satisfaz a seguinte relação:

$$g(\mu_t) \quad = \quad \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5}. \tag{5.1}$$

Logo, consideramos quatro covariáveis nesse estudo, que após geradas permaneceram fixas durante todo o processo de simulação. Foram utilizados a princípio os tamanhos amostrais $n = 20$, 40, 60, 80, 100 e 120, de modo que são geradas 20 observações das covariáveis $x_{ti} \sim U(0,1)$, com $i = 2, \ldots, 5$, em seguida, para garantir que a média da resposta se mantenha em um mesmo conjunto de valores para todo n, as amostras são replicadas até obter os valores amostrais desejados.

Buscando novamente avaliar as estatísticas BV e $\hat{\alpha}$ em diferentes cenários para μ_t, foram definidos os valores dos parâmetros do modelo para média e também a distribuição das covariáveis x_{ti}, com $i = 2, \ldots, 5$, de modo que os valores de μ_t estejam distribuídos em três diferentes cenários (próximos do zero, próximos de um e centrais). Além disso, consideramos para $g(\cdot)$ quatro diferentes funções de ligação: logit, probit, log-log e complemento log-log, todas definidas no capítulo 2.

Tabela 17 – Valores dos parâmetros β's fixados para a análise no modelo beta linear corretamente especificado com parâmetro ϕ fixo, junto com os respectivos intervalos produzidos para μ_t e a função de ligação $g(\cdot)$ utilizada.

β_1	β_2	β_3	β_4	β_5	μ	$g(\cdot)$
-1.6	1.0	0.7	1.0	1.48	$\mu_t \in (0.2925, 0.8252)$	Probit
-2.3	1.5	1.8	1.2	1.7	$\mu_t \in (0.2933, 0.8239)$	Logit
-2.45	1.1	1.2	1.0	0.28	$\mu_t \in (0.0100, 0.2678)$	Log-log
-0.8	1.12	1.25	1.0	0.4	$\mu_t \in (0.6923, 0.9860)$	C-Log-log

Fonte: Autoria própria.

Inicialmente, desenvolvemos a simulação de Monte Carlo para calcular os valores de $\widehat{\alpha}$ e BV, quando o modelo ajustado está corretamente especificado. Tal simulação está baseada em

metodologia similar à apresentada para o modelo beta com ϕ variável. Os valores fixados para os parâmetros β's e os respectivos intervalos obtidos para μ_t são apresentados na Tabela 17, além da função $g(\cdot)$ utilizada em cada cenário.

Podemos ver na Tabela 18 os valores médios das estatísticas de avaliação de desempenho para o modelo beta linear com precisão fixa. Analisando essa tabela é possível ver que BV possui um comportamento semelhante às demais estatísticas de qualidade de ajuste, assumindo valores próximo de um, o que implica em um bom desempenho dessas medidas, uma vez que o modelo está corretamente especificado. Notadamente, os valores das estatísticas crescem à medida que ϕ aumenta e para intervalos de μ_t centrais, em que tipicamente o processo de estimação dos parâmetros do modelo de regressão beta tem melhor desempenho. Entretanto, para todos os intervalos de μ_t as estatísticas apresentam valores aproximadamente entre 0.5 e 0.7, quando $\phi = 20$ e $n \leq 40$, mesmo com a função de ligação considerada apropriada ao intervalo da média da variável resposta, segundo Espinheira, et al. (2019). De fato, apesar do modelo estar corretamente especificado, quando a precisão e o tamanho da amostra são baixos as medidas refletem a dificuldade em ajustar os dados, em especial as estatísticas do tipo R^2 e mais enfaticamente suas versões corrigidas.

É interessante enfatizar que no modelo de regressão beta este fato está diretamente relacionado com $\hat{\phi}$, o qual é tipicamente viesado quando a precisão verdadeira dos dados é pequena e o tamanho da amostra é reduzido, um comportamento comumente relacionado à estimação por máxima verossimilhança. De fato, uma vez que critérios do tipo R^2 avaliam a qualidade do modelo ajustado em explicar a variabilidade da variável resposta, seus valores são mais afetados pela precisão baixa $\phi = 20$ que o critério P^2. No entanto, o valor reduzido da precisão do modelo também torna $\hat{\mu}_t$ mais viesado, o que é refletido pelos valores de P^2 abaixo de 0.79. Podemos confirmar essas informações a partir dos resultados apresentados pela Tabela 19, em que dispomos das médias das estimativas de ϕ e dos vieses de $\hat{\phi}$, e também dos máximos e mínimos das estimativas das variâncias e de μ_t. De fato, observamos que existem diferenças mais sgnificativas entre as estimativas e os valores verdadeiros dos parâmetros e da variância quando n é pequeno e $\phi = 20$, em todos os cenários de μ_t. Temos ainda que à medida que o tamanho da amostra aumenta, já a partir de $n = 40$, os valores das estatísticas tornam-se mais próximos, indicando equilíbrio entre os poderes de predição e de explicação da variabilidade da resposta. Esse cenário também é válido quando $\phi = 20$, ou seja, a precisão do modelo é baixa e consequentemente a variabilidade da variável resposta é grande. Neste caso, os valores das estatísticas se aproximam de 0.7, que é um valor baixo, considerando que o modelo está corretamente especificado. **No entanto, é importante enfatizar que este valor está diretamente relacionado à precisão dos dados**.

Um fato importante observado nesse estudo de simulação com o modelo corretamente especificado e ϕ constante é que em média $\hat{\alpha}$ apresenta valores bem menores que 0.5, de modo que quanto menor o tamanho da amostra menor a média de $\hat{\alpha}$. Tal comportamento está de acordo com a definição 3.15, uma vez que, quando o tamanho da amostra diminui o processo de estimação por máxima verossimilhança apresenta maior dificuldade em estimar corretamente os parâmetros. Logo, as estimativas são consideravelmente viesadas quando o tamanho amostral

é baixo, especialmente para ϕ, conforme observamos na Tabela 19. Dessa forma, é esperado que o valor do mínimo das estatísticas diminua quando n decresce, consequentemente, a probabilidade dada por 3.15 converge para zero, uma vez que a área correspondente ao gráfico da distribuição do mínimo diminui. Vale ainda destacar que mesmo quando $\phi = 400$, em que as estatísticas do tipo P^2, R^2 e também BV assumem valores bastante elevados e bem próximos a um, os valores de $\hat{\alpha}$ são menores que para $\phi = 20$ ou $\phi = 100$. Isso provavelmente ocorre devido ao viés de $\hat{\phi}$ ser mais elevado, conforme notamos pela Tabela 19. No entanto, como o viés diminui quando o tamanho da amostra aumenta, vemos que os valores de $\hat{\alpha}$ aumentam e também tornam-se próximos para os três valores de ϕ, em cada um dos cenários de μ_t. Adicionalmente, vemos também que os valores de $\hat{\alpha}$ são maiores para μ_t localizado nos extremos do intervalo unitário, isso possivelmente ocorre devido as variâncias serem menores nessas situações, o que contribui para o ajuste. Vemos então que quanto menor o tamanho da amostra, menor o valor de $\hat{\alpha}$, e maior é o peso atribuído ao valor mínimo das estatísticas no cálculo de BV. Consequentemente, os valores mais reduzidos da BV refletem a dificuldade em estimar os parâmetros do modelo de regressão beta, quando dispomos de pequenas amostras. Neste sentido, podemos dizer que BV possui um desempenho superior às demais estatísticas, quando avaliadas isoladamente. De fato, estamos dispondo de uma medida de qualidade de ajuste rigorosa.

Com base no que foi discutido nos parágrafos anteriores obtemos algumas conclusões associadas ao modelo de regressão beta com precisão fixa. A primeira é que o valor de $\hat{\alpha}$ aumenta juntamente com o tamanho da amostra, porém, mesmo para $n = 120$ temos $0.36 \leq \hat{\alpha} \leq 0.43$ aproximadamente, indicando que o ajuste ainda não está bom, e isso se deve a $\hat{\phi}$ ainda ser bastante viesado. Logo, torna-se mais apropriado utilizar amostras maiores. A Segunda observação é que quando a precisão do modelo é pequena tanto $\hat{\mu}$ quanto $\hat{\phi}$ são consideravelmente viesados, o que é evidenciado por valores das estatísticas P^2, R^2 e BV em torno de 0.7. E a terceira e última observação é que quanto menor o tamanho da amostra mais rigorosa é a medida BV, e quanto maior o tamanho da amostra mais próximos são os valores das medidas de adequabilidade P^2, R^2 e também BV.

Tabela 18 – Valores médios das estatísticas de avaliação de desempenho no modelo beta linear com precisão fixa corretamente especificado, com $g(\cdot)$ adequada à cada cenário de μ_t e $h(\cdot)$ log.

	$g(\cdot)$: Probit $\mu_t \in (0.2925, 0.8252)$			$g(\cdot)$: Logit $\mu_t \in (0.2933, 0.8239)$			$g(\cdot)$: Log-Log $\mu_t \in (0.0100, 0.2678)$			$g(\cdot)$: C-Log-Log $\mu_t \in (0.6923, 0.9860)$			
ϕ	20	100	400	20	100	400	20	100	400	20	100	400	n
P^2	0.7907	0.9435	0.9851	0.7313	0.9216	0.9791	0.8173	0.8799	0.9581	0.8197	0.8930	0.9646	20
P_c^2	0.7160	0.9233	0.9797	0.6354	0.8936	0.9717	0.7521	0.8370	0.9432	0.7552	0.8547	0.9519	
R_{FC}^2	0.7723	0.9398	0.9842	0.7194	0.9204	0.9789	0.7359	0.9104	0.9752	0.7214	0.9090	0.9747	
$R_{FC_c}^2$	0.6910	0.9182	0.9785	0.6193	0.8919	0.9714	0.6415	0.8784	0.9664	0.6219	0.8765	0.9656	
R_{LR}^2	0.7741	0.9404	0.9844	0.7275	0.9234	0.9798	0.7762	0.9194	0.9773	0.7532	0.9151	0.9761	
$R_{LR_c}^2$	0.5873	0.8911	0.9714	0.5022	0.8601	0.9631	0.5912	0.8528	0.9585	0.5492	0.8450	0.9563	
BV	0.6273	0.9007	0.9739	0.5494	0.8720	0.9661	0.6332	0.8498	0.9494	0.6022	0.8540	0.9559	
$\hat{\alpha}$	0.1929	0.1795	0.1771	0.1969	0.1817	0.1795	0.2724	0.2099	0.1885	0.2361	0.1948	0.1879	
P^2	0.7453	0.9300	0.9816	0.6765	0.9039	0.9740	0.7888	0.8552	0.9484	0.7894	0.8702	0.9568	40
P_c^2	0.7078	0.9197	0.9789	0.6290	0.8898	0.9701	0.7578	0.8339	0.9408	0.7584	0.8511	0.9504	
R_{FC}^2	0.7388	0.9299	0.9816	0.6795	0.9081	0.9754	0.6968	0.8963	0.9710	0.6857	0.8946	0.9710	
$R_{FC_c}^2$	0.7003	0.9195	0.9789	0.6324	0.8945	0.9718	0.6522	0.8810	0.9667	0.6395	0.8791	0.9667	
R_{LR}^2	0.7415	0.9307	0.9819	0.6894	0.9118	0.9764	0.7497	0.9073	0.9736	0.7229	0.9015	0.9725	
$R_{LR_c}^2$	0.6684	0.9112	0.9768	0.6016	0.8868	0.9698	0.6789	0.8810	0.9661	0.6445	0.8737	0.9647	
BV	0.6904	0.9165	0.9782	0.6264	0.8933	0.9714	0.6898	0.8543	0.9494	0.6744	0.8650	0.9563	
$\hat{\alpha}$	0.2785	0.2656	0.2678	0.2839	0.2694	0.2645	0.3575	0.2866	0.2658	0.3259	0.2881	0.2660	
P^2	0.7299	0.9252	0.9804	0.6566	0.8970	0.9721	0.7811	0.8480	0.9452	0.7774	0.8626	0.9536	60
P_c^2	0.7049	0.9183	0.9786	0.6248	0.8874	0.9695	0.7608	0.8339	0.9401	0.7568	0.8499	0.9493	
R_{FC}^2	0.7275	0.9263	0.9808	0.6646	0.9031	0.9740	0.6871	0.8922	0.9698	0.6717	0.8903	0.9694	
$R_{FC_c}^2$	0.7023	0.9195	0.9790	0.6336	0.8942	0.9716	0.6581	0.8822	0.9670	0.6413	0.8801	0.9666	
R_{LR}^2	0.7305	0.9274	0.9811	0.6757	0.9071	0.9752	0.7433	0.9037	0.9724	0.7117	0.8975	0.9710	
$R_{LR_c}^2$	0.6845	0.9150	0.9778	0.6203	0.8913	0.9709	0.6995	0.8872	0.9677	0.6625	0.8800	0.9660	
BV	0.6999	0.9188	0.9788	0.6362	0.8935	0.9712	0.7004	0.8560	0.9499	0.6839	0.8649	0.9560	
$\hat{\alpha}$	0.3186	0.3062	0.3086	0.3241	0.3112	0.3045	0.3959	0.3257	0.3039	0.3627	0.3236	0.3062	
P^2	0.7227	0.9231	0.9796	0.6480	0.8938	0.9711	0.7751	0.8434	0.9434	0.7728	0.8586	0.9525	80
P_c^2	0.7039	0.9179	0.9782	0.6242	0.8866	0.9692	0.7600	0.8329	0.9396	0.7575	0.8490	0.9493	
R_{FC}^2	0.7224	0.9250	0.9802	0.6588	0.9011	0.9734	0.6795	0.8897	0.9691	0.6665	0.8876	0.9690	
$R_{FC_c}^2$	0.7037	0.9199	0.9789	0.6358	0.8944	0.9716	0.6578	0.8823	0.9670	0.6440	0.8800	0.9669	
R_{LR}^2	0.7257	0.9260	0.9805	0.6702	0.9051	0.9746	0.7392	0.9016	0.9718	0.7082	0.8950	0.9706	
$R_{LR_c}^2$	0.6922	0.9169	0.9781	0.6299	0.8935	0.9715	0.7074	0.8896	0.9683	0.6726	0.8822	0.9670	
BV	0.7039	0.9199	0.9789	0.6391	0.8928	0.9710	0.7008	0.8562	0.9502	0.6892	0.8645	0.9563	
$\hat{\alpha}$	0.3418	0.3330	0.3356	0.3493	0.3357	0.3307	0.4071	0.3483	0.3294	0.3850	0.3479	0.3314	
P^2	0.7176	0.9220	0.9793	0.6423	0.8918	0.9708	0.7724	0.8407	0.9428	0.7694	0.8567	0.9515	100
P_c^2	0.7026	0.9178	0.9782	0.6233	0.8860	0.9692	0.7603	0.8323	0.9397	0.7571	0.8491	0.9489	
R_{FC}^2	0.7186	0.9243	0.9800	0.6550	0.8997	0.9733	0.6764	0.8880	0.9689	0.6625	0.8867	0.9684	
$R_{FC_c}^2$	0.7036	0.9203	0.9790	0.6367	0.8944	0.9718	0.6592	0.8820	0.9672	0.6446	0.8807	0.9668	
R_{LR}^2	0.7218	0.9253	0.9803	0.6665	0.9038	0.9744	0.7369	0.9002	0.9716	0.7046	0.8940	0.9700	
$R_{LR_c}^2$	0.6953	0.9182	0.9785	0.6347	0.8947	0.9720	0.7118	0.8907	0.9689	0.6765	0.8840	0.9672	
BV	0.7049	0.9204	0.9790	0.6386	0.8923	0.9711	0.7027	0.8565	0.9507	0.6907	0.8649	0.9562	
$\hat{\alpha}$	0.3594	0.3514	0.3559	0.3643	0.3533	0.3488	0.4206	0.3651	0.3471	0.3992	0.3628	0.3494	
P^2	0.7142	0.9208	0.9790	0.6380	0.8913	0.9704	0.7710	0.8394	0.9418	0.7663	0.8552	0.9509	120
P_c^2	0.7017	0.9173	0.9781	0.6221	0.8865	0.9691	0.7610	0.8323	0.9393	0.7561	0.8488	0.9487	
R_{FC}^2	0.7161	0.9234	0.9798	0.6514	0.8996	0.9729	0.6744	0.8875	0.9684	0.6596	0.8857	0.9682	
$R_{FC_c}^2$	0.7036	0.9200	0.9789	0.6361	0.8952	0.9718	0.6601	0.8826	0.9671	0.6447	0.8807	0.9668	
R_{LR}^2	0.7194	0.9244	0.9801	0.6632	0.9037	0.9741	0.7359	0.8998	0.9712	0.7015	0.8932	0.9697	
$R_{LR_c}^2$	0.6975	0.9186	0.9786	0.6370	0.8962	0.9721	0.7154	0.8919	0.9690	0.6783	0.8849	0.9674	
BV	0.7055	0.9199	0.9788	0.6374	0.8928	0.9709	0.7039	0.8571	0.9507	0.6921	0.8651	0.9563	
$\hat{\alpha}$	0.3726	0.3632	0.3703	0.3767	0.3660	0.3629	0.4277	0.3752	0.3588	0.4142	0.3767	0.3615	

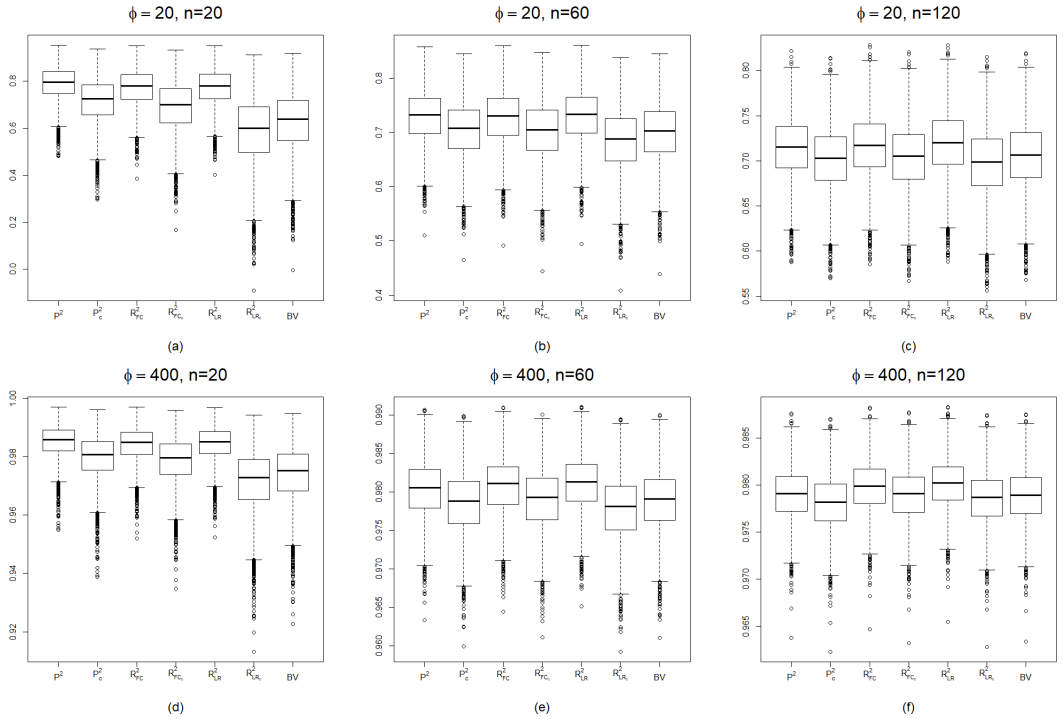
Fonte: Autoria própria.

Tabela 19 – Valores médios de $\hat{\phi}$ e de seus respectivos vieses, e também dos intervalos de $\hat{\mu}_t$ e da $\widehat{\text{Var}}(y_t)$, para cada tamanho amostral, no modelo beta linear corretamente especificado com precisão fixa, utilizando a função $g(\cdot)$ adequada ao cenário da média.

	$g(\cdot)$: Probit			$g(\cdot)$: Logit			
	$\mu_t \in (0.2925, 0.8252)$			$\mu_t \in (0.2933, 0.8239)$			
ϕ	20	100	400	20	100	400	n
$\text{Var}(y_t)$	(0.0069, 0.0119)	(0.0014, 0.0025)	(0.0004, 0.0006)	(0.0069, 0.0118)	(0.0014, 0.0024)	(0.0004, 0.0006)	
$\hat{\phi}$ (Viés)	30.6 (10.6521)	154.6 (54.0841)	628.7 (219.1603)	30.6 (10.6071)	154.6 (54.0764)	631.0 (221.5168)	20
$\widehat{\text{Var}}(y_t)$	(0.0052, 0.0089)	(0.0011, 0.0018)	(0.0003, 0.0004)	(0.0052, 0.0088)	(0.0011, 0.0018)	(0.0002, 0.0004)	
$\hat{\mu}_t$	(0.2639, 0.8207)	(0.2789, 0.8189)	(0.2799, 0.8251)	(0.3264, 0.8256)	(0.2460, 0.7929)	(0.2897, 0.8374)	
$\hat{\phi}$ (Viés)	24.3 (4.2909)	122.6 (22.1581)	497.8 (88.3034)	24.2 (4.2205)	121.4 (20.9171)	496.43 (86.9024)	40
$\widehat{\text{Var}}(y_t)$	(0.0059, 0.0104)	(0.0012, 0.0021)	(0.0003, 0.0005)	(0.0060, 0.0103)	(0.0012, 0.0021)	(0.0003, 0.0005)	
$\hat{\mu}_t$	(0.2305, 0.7935)	(0.2590, 0.8130)	(0.2835, 0.8320)	(0.3214, 0.8483)	(0.2814, 0.7973)	(0.2794, 0.8225)	
$\hat{\phi}$ (Viés)	22.6 (2.6134)	113.3 (12.8097)	463.0 (53.4928)	22.7 (2.6795)	113.5 (13.0299)	464.2 (54.6423)	60
$\widehat{\text{Var}}(y_t)$	(0.0063, 0.0109)	(0.0013, 0.0023)	(0.0003, 0.0005)	(0.0063, 0.0108)	(0.0013, 0.0022)	(0.0003, 0.0005)	
$\hat{\mu}_t$	(0.2753, 0.8198)	(0.2994, 0.8311)	(0.2904, 0.8164)	(0.2289, 0.8796)	(0.2994, 0.8124)	(0.2975, 0.8187)	
$\hat{\phi}$ (Viés)	21.8 (1.8336)	110.2 (9.6944)	450.5 (40.9364)	21.9 (1.9035)	110.1 (9.5912)	448.9 (39.3442)	80
$\widehat{\text{Var}}(y_t)$	(0.0064, 0.0112)	(0.0013, 0.0023)	(0.0003, 0.0006)	(0.0065, 0.0110)	(0.0013, 0.0023)	(0.0003, 0.0005)	
$\hat{\mu}_t$	(0.2861, 0.8607)	(0.2992, 0.8119)	(0.2882, 0.8289)	(0.3382, 0.8469)	(0.2925, 0.8456)	(0.2866, 0.8114)	
$\hat{\phi}$ (Viés)	21.6 (1.5573)	108.0 (7.5694)	439.5 (30.0012)	21.5 (1.4553)	107.96 (7.4772)	440.8 (31.2916)	100
$\widehat{\text{Var}}(y_t)$	(0.0065, 0.0112)	(0.0013, 0.0023)	(0.0003, 0.0006)	(0.0066, 0.0112)	(0.0013, 0.0023)	(0.0003, 0.0006)	
$\hat{\mu}_t$	(0.2501, 0.8152)	(0.2947, 0.8275)	(0.2862, 0.8299)	(0.2966, 0.8506)	(0.2924, 0.8366)	(0.2967, 0.8277)	
$\hat{\phi}$ (Viés)	21.2 (1.2497)	106.8 (6.2827)	434.3 (24.7341)	21.2 (1.2374)	106.7 (6.2353)	433.8 (24.2298)	120
$\widehat{\text{Var}}(y_t)$	(0.0066, 0.0114)	(0.0013, 0.0023)	(0.0003, 0.0006)	(0.0066, 0.0113)	(0.0014, 0.0023)	(0.0003, 0.0006)	
$\hat{\mu}_t$	(0.2992, 0.8383)	(0.2796, 0.8266)	(0.2961, 0.8241)	(0.2989, 0.8079)	(0.2837, 0.8197)	(0.3038, 0.8241)	
	$g(\cdot)$: Log-Log			$g(\cdot)$: C-Log-Log			
	$\mu_t \in (0.0100, 0.2678)$			$\mu_t \in (0.6923, 0.9860)$			
ϕ	20	100	400	20	100	400	n
$\text{Var}(y_t)$	(0.0005, 0.0093)	(0.0001, 0.0019)	(0.0000, 0.0005)	(0.0007, 0.0101)	(0.0001, 0.0021)	(0.0000, 0.0005)	
$\hat{\phi}$ (Viés)	31.4 (11.4023)	154.8 (54.3731)	630.0 (220.5187)	32.4 (12.4211)	155.9 (55.4412)	629.5 (219.9987)	20
$\widehat{\text{Var}}(y_t)$	(0.0004, 0.0069)	(0.0000, 0.0014)	(0.0000, 0.0003)	(0.0005, 0.0074)	(0.0001, 0.0016)	(0.0000, 0.0004)	
$\hat{\mu}_t$	(0.0169, 0.3081)	(0.0087, 0.3001)	(0.0118, 0.2621)	(0.6110, 0.9954)	(0.6863, 0.9904)	(0.6957, 0.9837)	
$\hat{\phi}$ (Viés)	24.5 (4.4544)	122.6 (22.0915)	495.2 (85.6433)	24.5 (4.5181)	122.6 (22.0877)	495.2 (85.7089)	40
$\widehat{\text{Var}}(y_t)$	(0.0004, 0.0081)	(0.0000, 0.0017)	(0.0000, 0.0004)	(0.0006, 0.0088)	(0.0001, 0.0018)	(0.0000, 0.0004)	
$\hat{\mu}_t$	(0.0059, 0.2893)	(0.0096, 0.2862)	(0.0122, 0.2534)	(0.7203, 0.9864)	(0.7033, 0.9823)	(0.6854, 0.9877)	
$\hat{\phi}$ (Viés)	22.8 (2.8149)	114.1 (13.5866)	463.2 (53.6604)	22.9 (2.9818)	114.6 (14.1432)	462.8 (53.2568)	60
$\widehat{\text{Var}}(y_t)$	(0.0004, 0.0085)	(0.0000, 0.0018)	(0.0000, 0.0004)	(0.0006, 0.0092)	(0.0001, 0.0019)	(0.0000, 0.0005)	
$\hat{\mu}_t$	(0.0075, 0.2968)	(0.0088, 0.2566)	(0.0097, 0.2711)	(0.6934, 0.9929)	(0.6875, 0.9928)	(0.6952, 0.9875)	
$\hat{\phi}$ (Viés)	21.969 (1.9641)	110.1 (9.6391)	448.6 (39.1113)	22.0 (2.0253)	110.4 (9.9578)	448.7 (39.1952)	80
$\widehat{\text{Var}}(y_t)$	(0.0005, 0.0087)	(0.0000, 0.0018)	(0.0000, 0.0004)	(0.0006, 0.0095)	(0.0001, 0.0019)	(0.0000, 0.0005)	
$\hat{\mu}_t$	(0.0074, 0.3055)	(0.0110, 0.2581)	(0.0107, 0.2594)	(0.7051, 0.9878)	(0.6654, 0.9897)	(0.6911, 0.9866)	
$\hat{\phi}$ (Viés)	21.6 (1.6185)	108.3 (7.8343)	438.9 (29.4012)	21.7 (1.6630)	108.0 (7.5595)	440.4 (30.9095)	100
$\widehat{\text{Var}}(y_t)$	(0.0004, 0.0088)	(0.0000, 0.0018)	(0.0000, 0.0004)	(0.0006, 0.0096)	(0.0001, 0.0020)	(0.0000, 0.0005)	
$\hat{\mu}_t$	(0.0110, 0.2568)	(0.0093, 0.2680)	(0.0089, 0.2643)	(0.6798, 0.9921)	(0.6919, 0.9855)	(0.6917, 0.9890)	
$\hat{\phi}$ (Viés)	21.3 (1.2874)	106.7 (6.2253)	434.7 (25.2275)	21.237 (1.2320)	107.1 (6.6432)	434.61 (25.0795)	120
$\widehat{\text{Var}}(y_t)$	(0.0005, 0.0089)	(0.0000, 0.0018)	(0.0000, 0.0004)	(0.0006, 0.0097)	(0.0001, 0.0020)	(0.0000, 0.0005)	
$\hat{\mu}_t$	(0.0129, 0.2685)	(0.0117, 0.2685)	(0.0107, 0.2621)	(0.6828, 0.9867)	(0.7030, 0.9833)	(0.6961, 0.9856)	

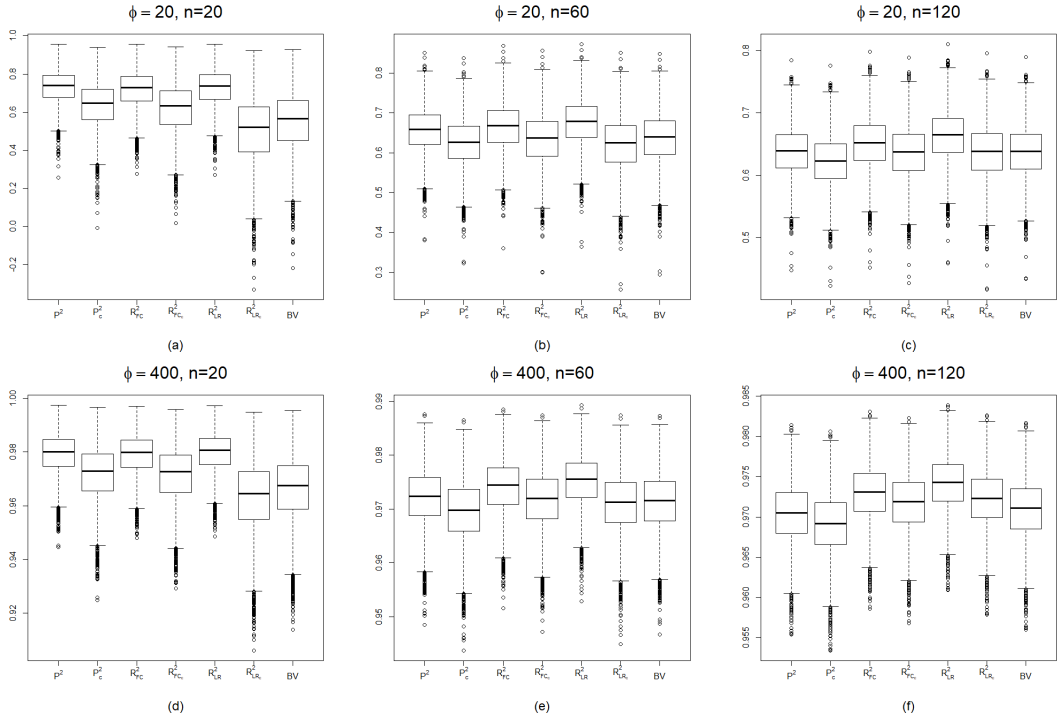
Fonte: Autoria própria.

Figura 17 – Gráficos de *boxplot* dos valores estatísticas de avaliação de desempenho no modelo beta linear corretamente especificado, com precisão fixa e função de ligação probit para $\mu_t \in (0.2925, 0.8252)$.



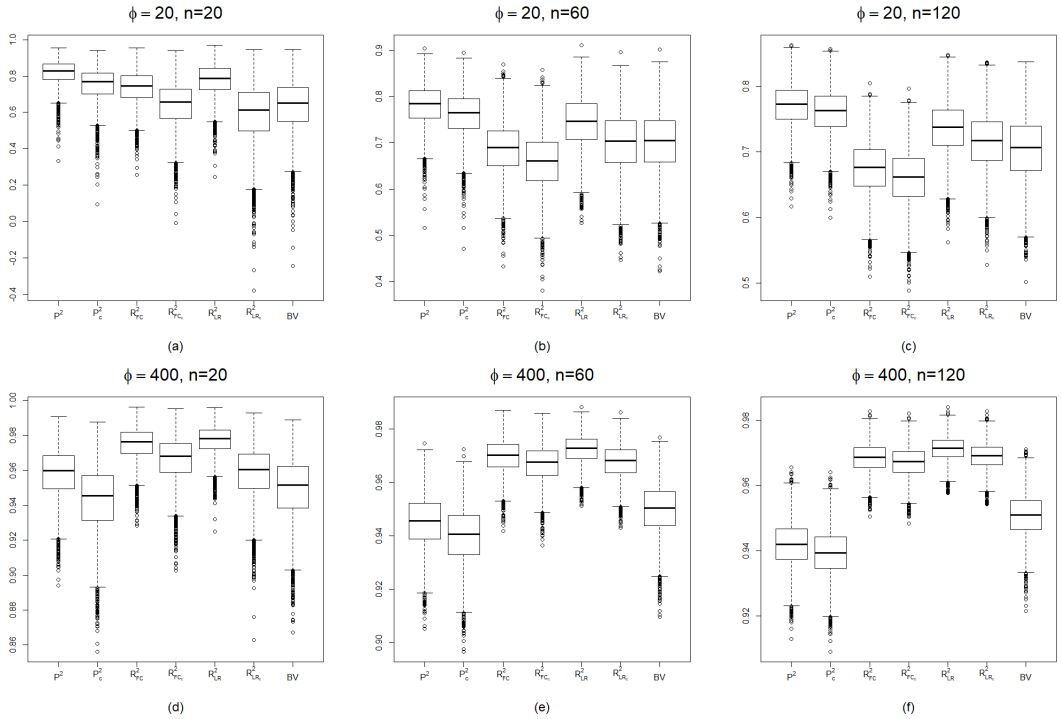
Fonte: Autoria própria.

Figura 18 – Gráficos de *boxplot* dos valores estatísticas de avaliação de desempenho no modelo beta linear corretamente especificado, com precisão fixa e função de ligação logit para $\mu_t \in (0.2933, 0.8239)$.



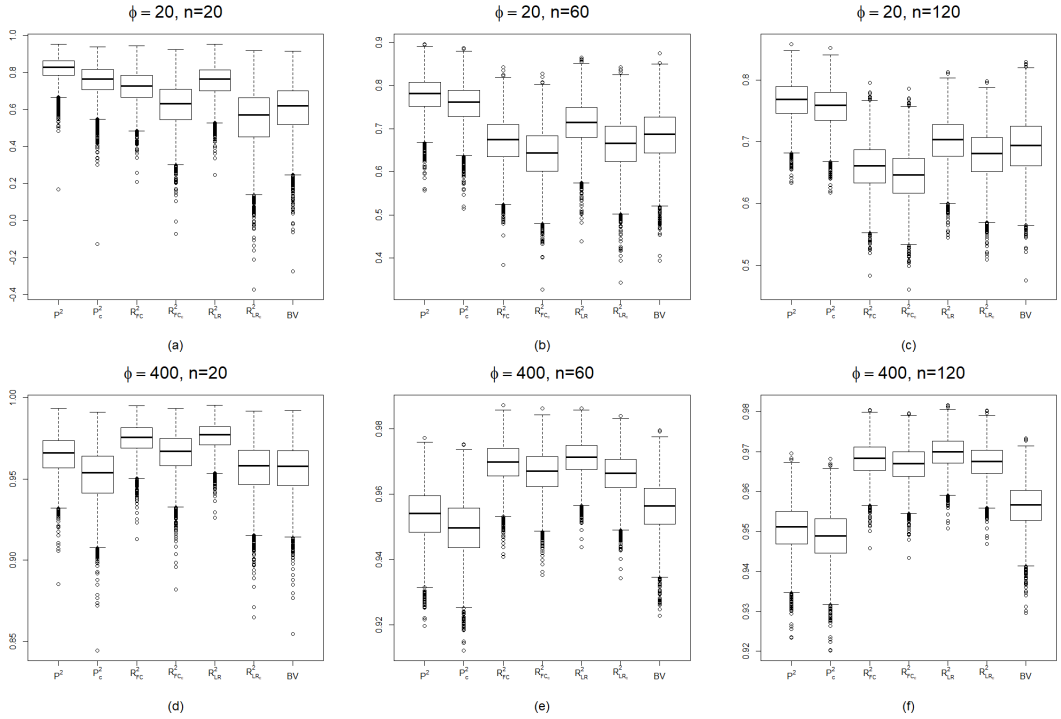
Fonte: Autoria própria.

Figura 19 – Gráficos de *boxplot* dos valores das estatísticas de avaliação de desempenho no modelo beta linear corretamente especificado, com precisão fixa e função de ligação log-log para $\mu_t \in (0.0100, 0.2678)$.



Fonte: Autoria própria.

Figura 20 – Gráficos de *boxplot* dos valores das estatísticas de avaliação de desempenho no modelo beta linear corretamente especificado, com precisão fixa e função de ligação c-log-log para $\mu_t \in (0.6923, 0.9860)$.



Fonte: Autoria própria.

Nas Figuras 17 e 18 são apresentados os gráficos de *boxplot* dos 5000 valores das estatísticas de avaliação de desempenho do modelo beta, para $\mu_t \in (0.30, 0.83)$ aproximadamente, utilizando as funções probit e logit, respectivamente. Com base nestas figuras percebemos que as distribuições das estatísticas tornam-se semelhantes entre si à medida que o tamanho da amostra cresce, independentemente do valor de ϕ. No entanto, os valores das estatísticas concentram-se próximos de um quando o valor da precisão aumenta. Quando $\phi = 20$ e $n = 20$ a dispersão das distribuições empíricas das estatísticas é consideravelmente alta, em especial de $R^2_{LR_c}$ que apresenta valores até negativos, em especial quando a função de ligação é logit (Figura 18a). Consequentemente, enquanto para o modelo probit os valores de BV estão dentro do intervalo $(0,1)$, para o modelo logit esses valores estão em $(-0.2,1)$. De forma geral, a distribuição dos valores da BV apontam para melhor adequabilidade modelo probit, quando a média da variável resposta encontra-se bem distribuída em torno do valor central do intervalo unitário.

As distribuições empíricas das estatísticas analisadas com base nos gráficos de *boxplot* das 5000 réplicas são muito úteis para evidenciar o desempenho da regressão beta em modelar respostas concentradas nos extremos do intervalo $(0,1)$, considerando a Figura 19, para $\mu_t \approx 0$, com função de ligação log-log, e a Figura 20, para $\mu_t \approx 1$, com ligação c-log-log.

De forma geral, podemos notar que quando a precisão do modelo é baixa, o critério R^2_{FC} aponta para o baixo poder do modelo de regressão beta em explicar a variabilidade da variável resposta. De fato, neste cenário $\hat{\phi}$ geralmente é consideravelmente viesado e os valores mais baixos do critério R^2_{FC} em comparação ao critério R^2_{LR}, em especial, quando o tamanho da amostra aumenta, evidenciam que o primeiro critério apresenta melhor desempenho em captar esse alto viés. Por outro lado, para tamanhos de amostras reduzidos, o critério R^2_{LR} e em especial sua versão corrigida apresentam melhores desempenhos, uma vez que o $R^2_{LR_c} \in (-0.4, 1.0)$ e o $R^2_{FC_c} \in (0.0, 1.0)$. Ressaltando que o modelo está corretamente especificado e esses valores negativos do critério do tipo R^2 estão estritamente relacionados ao viés de $\hat{\phi}$. Um ponto importante da medida BV é que em ambas situações ela absorve as informações de $R^2_{LR_c}$ e de $R^2_{FC_c}$, dado que seus valores estão distribuídos sempre de forma centralizada entre as estatísticas com altos e baixos valores. Por outro lado, quando a precisão do modelo aumenta, notadamente o viés de $\hat{\phi}$ diminui, como mencionado por Ospina, Cribari-Neto e Vasconcellos, (2006) e os valores das estatísticas do tipo R^2 aumentam.

Com relação a predição de novas observações, fica evidente que o poder de predição do modelo de regressão beta é menor quando as observações estão próximas de zero ou próximas de um. À medida que a precisão fixa das observações aumenta fica evidente que o modelo consegue ajustar-se bem ao conjunto de dados original. No entanto, seu poder de predição não aumenta. Esse fato pode ser notado claramente a partir das Figuras 19f e e 20f, em que $\mu_t \approx 0$ e $\mu_t \approx 1$, respectivamente, $n = 120$ e $\phi = 400$. Notadamente, os valores das medidas do tipo R^2 concentram-se próximos a um, enquanto as medidas P^2 são mais dispersas e chegam a assumir valores menores, próximos de 0.90. Assim, nessas situações, o modelo explica muito bem o comportamento aleatório da resposta para a amostra disponível, mas produz viese elevados de $\hat{\mu}_t$ no processo de estimação. Essa é então uma sua situação típica de *overfitting*. Ressaltamos mais uma vez a importância e o bom desempenho do critério BV, em que a distribuição de seus

valores representa a informação transmitida pelos critérios do tipo P^2, como pode ser notado nas Figuras 19e-f e 20e-f, para $\phi = 400$, com $n = 60$ e $n = 120$, em que os gráficos de *boxplot* referentes a BV ilustram valores intermediários aos das medidas P^2 e R^2.

Como mencionado anteriormente, com base nos estudos realizados acreditamos que, para o modelo beta linear com precisão constante, $\hat{\alpha}$ está fortemente associado com o viés de $\hat{\phi}$, e sabemos que este viés tende diminuir quando aumenta-se o tamanho amostral. Então, realizamos mais um estudo de simulação de Monte Carlo para o modelo corretamente especificado, considerando tamanhos amostrais maiores e apenas o cenário de μ_t central. Além disso, tomamos $\phi = 15$, 50 e 100, e consideramos o modelo da média com três covariáveis, com valores gerados a partir da distribuição $U(0,1)$, utilizamos também $g(\cdot)$ logit, com $\mu_t \in (0.1758, 0.8471)$, e os valores assumidos para os parâmetros β's foram $(\beta_1, \beta_2, \beta_3, \beta_4) = (-2.3, 1.5, 1.8, 1.6)$. Na Tabela 20 dispomos dos valores médios das estatísticas de avaliação de desempenho e do viés de $\hat{\phi}$ para cada tamanho amostral e precisão ϕ.

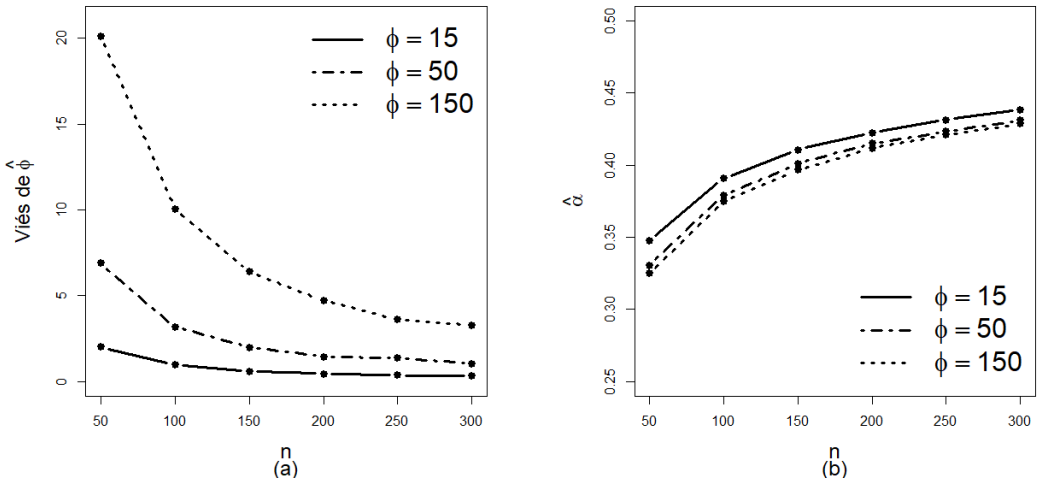
Tabela 20 – Valores médios das estatísticas de avaliação de desempenho e do viés de $\hat{\phi}$ no modelo beta linear com precisão fixa corretamente especificado, com $g(\cdot)$ logit e $\mu_t \in (0.1758, 0.8471)$ e tamanhos amostrais maiores.

	$g(\cdot)$: Logit $\mu_t \in (0.1758, 0.8471)$										
ϕ	P^2	P_c^2	R^2_{FC}	$R^2_{FC_c}$	R^2_{LR}	$R^2_{LR_c}$	BV	$\hat{\alpha}$	Viés de $\hat{\phi}$	Viés Rel. de $\hat{\phi}$	n
15	0.6706	0.6413	0.6789	0.6504	0.6855	0.6314	0.6492	0.3475	2.0028	0.1333	50
50	0.8611	0.8488	0.8702	0.8587	0.8733	0.8514	0.8565	0.3307	6.8933	0.1368	
150	0.9471	0.9424	0.9514	0.9470	0.9526	0.9444	0.9457	0.3250	20.0904	0.1338	
15	0.6539	0.6393	0.6692	0.6552	0.6765	0.6511	0.6537	0.3909	0.9604	0.0639	100
50	0.8523	0.8461	0.8647	0.8590	0.8679	0.8576	0.8544	0.3792	3.1821	0.0631	
150	0.9439	0.9416	0.9496	0.9475	0.9508	0.9470	0.9451	0.3748	10.0381	0.0668	
15	0.6475	0.6378	0.6651	0.6558	0.6726	0.6560	0.6520	0.4108	0.5914	0.0394	150
50	0.8492	0.8450	0.8628	0.8590	0.8661	0.8593	0.8535	0.4011	1.9845	0.0394	
150	0.9427	0.9411	0.9488	0.9474	0.9501	0.9476	0.9447	0.3967	6.3960	0.0426	
15	0.6447	0.6374	0.6635	0.6566	0.6711	0.6588	0.6516	0.4225	0.4520	0.0301	200
50	0.8479	0.8448	0.8620	0.8592	0.8654	0.8604	0.8534	0.4150	1.4464	0.0287	
150	0.9421	0.9409	0.9484	0.9473	0.9497	0.9478	0.9445	0.4116	4.7111	0.0314	
15	0.6426	0.6367	0.6620	0.6565	0.6698	0.6599	0.6509	0.4314	0.3582	0.0238	250
50	0.8477	0.8452	0.8620	0.8598	0.8655	0.8615	0.8538	0.4235	1.3688	0.0272	
150	0.9417	0.9407	0.9482	0.9473	0.9495	0.9480	0.9444	0.4211	3.5958	0.0239	
15	0.6419	0.6371	0.6619	0.6573	0.6696	0.6614	0.6512	0.4386	0.3093	0.0206	300
50	0.8467	0.8446	0.8613	0.8594	0.8648	0.8614	0.8533	0.4310	1.0378	0.0206	
150	0.9416	0.9408	0.9481	0.9474	0.9495	0.9482	0.9445	0.4287	3.2597	0.0217	

Fonte: Autoria própria.

Observamos na Tabela 20 que como esperado, quando o tamanho da amostra aumenta o viés de $\hat{\phi}$ e também o viés relativo diminuem. No entanto, isso ocorre de maneira lenta, especialmente, a partir de $n = 150$. Ou seja, apesar do tamanho da amostra aumentar consideravelmente

Figura 21 – Gráficos dos valores médios de $\hat{\alpha}$ e do viés de $\hat{\phi}$ *versus* o tamanho amostral, no modelo beta linear com precisão fixa corretamente especificado, com $g(\cdot)$ logit e $\mu_t \in (0.1758, 0.8471)$.



Fonte: Autoria própria.

de 150 para 300, não verificamos desempenhos tão bons quanto gostaríamos nas estimativas de ϕ. Observando os valores de $\hat{\alpha}$, vemos que, embora essa estatística possua a tendência de aumentar com n, a partir de também $n = 150$, ela não apresenta valores significativamente maiores e próximos de 0.5, quando elevamos o tamanho amostral, de modo que $n > 150$. Pela Figura 21 temos ilustrado graficamente o comportamento do viés de $\hat{\phi}$ e de $\hat{\alpha}$ em relação ao tamanho amostral, para os cenários abordados nesse estudo. Logo, reforçamos a ideia de que para modelos beta cujos valores de BV indicam bons ajustes, $\hat{\alpha}$ traz informação a respeito da precisão das estimativas de ϕ. Então, $\hat{\alpha}$ junto com BV tornam-se uma ferramenta eficaz na seleção de bons modelos beta.

5.2.1 Estudo Descritivo da Estatística $\hat{\alpha}$

Buscando mais uma vez realizar uma breve avaliação sobre o comportamento da distribuição dos valores de $\hat{\alpha}$, agora para o modelo beta linear com parâmetro de precisão ϕ constante, foram calculadas algumas medidas descritivas dessa quantidade, baseando-se nos 5000 valores obtidos em nosso estudo inicial para o modelo com ϕ fixo. Assim, na Tabela 21 encontramos os valores dessas medidas. Podemos então verificar que os valores médios para $\hat{\alpha}$ tendem a aumentar de acordo com o tamanho amostral, entretanto, a diferença entre os valores torna-se cada vez menor a partir de $n = 80$. Além disso, não existe muita variação nesses valores quando altera-se o valor de ϕ ou o intervalo para μ_t, usando a função de ligação adequada em cada caso. Vale observar também que para $n = 20$ os valores médios de $\hat{\alpha}$ são consideravelmente baixos, principalmente quando o valor de ϕ aumenta. Desta forma, como já mencionado, os valores para BV foram obtidos atribuindo pesos substancialmente maiores ao mínimo das estatísticas, o que então torna BV uma medida rigorosa, principalmente quando a precisão dos dados aumenta. Para $n = 40$, verificamos que as médias de $\hat{\alpha}$, embora sejam maiores do que para $n = 20$, ainda estão em geral abaixo de 0.3, indicando que quanto menor for o tamanho da amostra, mais rigorosa será

a estatística BV. Além disso, para o modelo beta com precisão constante e n pequeno, o valor de $\hat{\alpha}$ em média é baixo, e está geralmente associado ao alto viés de $\hat{\phi}$, que pode ser amenizado quando aumenta-se o tamanho da amostra.

Tabela 21 – Medidas descritivas referentes a distribuição de $\hat{\alpha}$ para diferentes tamanhos amostrais e valores de ϕ, no modelo beta linear corretamente especificado, usando a função de ligação probit para $\mu_t \in (0.2925, 0.8252)$, logit para $\mu_t \in (0.2933, 0.8239)$, log-log para $\mu_t \in (0.0100, 0.2678)$ e c-log-log para $\mu_t \in (0.6923, 0.9860)$.

	Média				Desvio-Padrão				
ϕ \ $g(\cdot)$	Probit	Logit	Log-Log	C-Log-Log	Probit	Logit	Log-Log	C-Log-Log	n
20	0.1929	0.1969	0.2724	0.2361	0.0243	0.0271	0.1242	0.0994	20
100	0.1795	0.1817	0.2099	0.1946	0.0183	0.0188	0.0811	0.0590	
400	0.1771	0.1795	0.1885	0.1879	0.0177	0.0178	0.0707	0.0603	
20	0.2785	0.2839	0.3575	0.3259	0.0264	0.0304	0.1543	0.1215	40
100	0.2656	0.2694	0.2866	0.2881	0.0211	0.0250	0.0902	0.0878	
400	0.2678	0.2645	0.2658	0.2660	0.0205	0.0257	0.0575	0.0589	
20	0.3186	0.3241	0.3959	0.3627	0.0265	0.0351	0.1736	0.1395	60
100	0.3062	0.3112	0.3257	0.3236	0.0227	0.0356	0.0986	0.0959	
400	0.3086	0.3045	0.3039	0.3062	0.0216	0.0325	0.0554	0.0559	
20	0.3418	0.3493	0.4071	0.3850	0.0281	0.0405	0.1857	0.1525	80
100	0.3330	0.3357	0.3483	0.3479	0.0238	0.0326	0.1018	0.1010	
400	0.3356	0.3307	0.3294	0.3314	0.0233	0.0293	0.0556	0.0562	
20	0.3594	0.3643	0.4206	0.3992	0.0291	0.0432	0.1877	0.1633	100
100	0.3514	0.3533	0.3651	0.3628	0.0280	0.0314	0.1042	0.1064	
400	0.3559	0.3488	0.3471	0.3494	0.0277	0.0268	0.0569	0.0576	
20	0.3726	0.3767	0.4277	0.4142	0.0301	0.0437	0.1927	0.1674	120
100	0.3632	0.3660	0.3752	0.3767	0.0303	0.0308	0.1063	0.1093	
400	0.3703	0.3629	0.3588	0.3615	0.0277	0.0270	0.0582	0.0569	
	Curtose				Assimetria				
20	3.1651	3.6890	3.6097	3.3443	0.1634	0.2087	0.6645	0.5971	20
100	3.0391	2.9811	3.0162	3.1208	0.0979	0.1268	0.2445	0.2472	
400	2.9102	3.0714	2.9973	2.5918	0.0193	0.0984	0.3230	−0.0392	
20	3.1420	3.6669	2.6987	2.9825	−0.0271	−0.1185	0.1824	0.2117	40
100	3.0613	3.7172	3.2780	3.2532	0.0586	−0.2638	0.1283	0.1697	
400	3.0259	3.9458	3.4029	3.5566	0.0137	−0.5102	0.2090	0.2660	
20	3.1244	4.3184	2.4812	2.6915	0.0012	−0.4976	0.0960	0.0660	60
100	3.0047	3.0008	3.5570	3.1514	0.0584	−0.0214	0.1311	0.0849	
400	2.9782	3.1440	3.5757	3.5512	0.0515	−0.0576	0.0775	0.1369	
20	3.4697	4.2074	2.3038	2.5270	−0.0909	−0.4493	0.0958	0.0207	80
100	3.2469	3.2205	3.6459	3.3465	−0.1032	0.0753	0.0149	−0.0262	
400	3.0998	3.0396	3.6728	3.3717	−0.0753	−0.0051	−0.1284	0.0776	
20	3.8904	4.2762	2.2837	2.3883	−0.2204	−0.4707	0.0387	0.0291	100
100	3.2765	3.1480	3.2554	3.1392	−0.2032	0.0119	−0.1430	−0.0529	
400	2.9297	3.0758	3.7434	3.4187	−0.0965	0.0121	−0.0510	0.0551	
20	4.0279	4.3727	2.2995	2.3495	−0.3548	−0.3868	0.0935	0.0118	120
100	3.0754	3.0137	3.2869	3.0845	−0.1084	−0.0039	−0.1703	−0.1462	
400	2.9558	2.9409	3.6869	3.3898	−0.0303	0.0224	−0.1489	−0.0549	

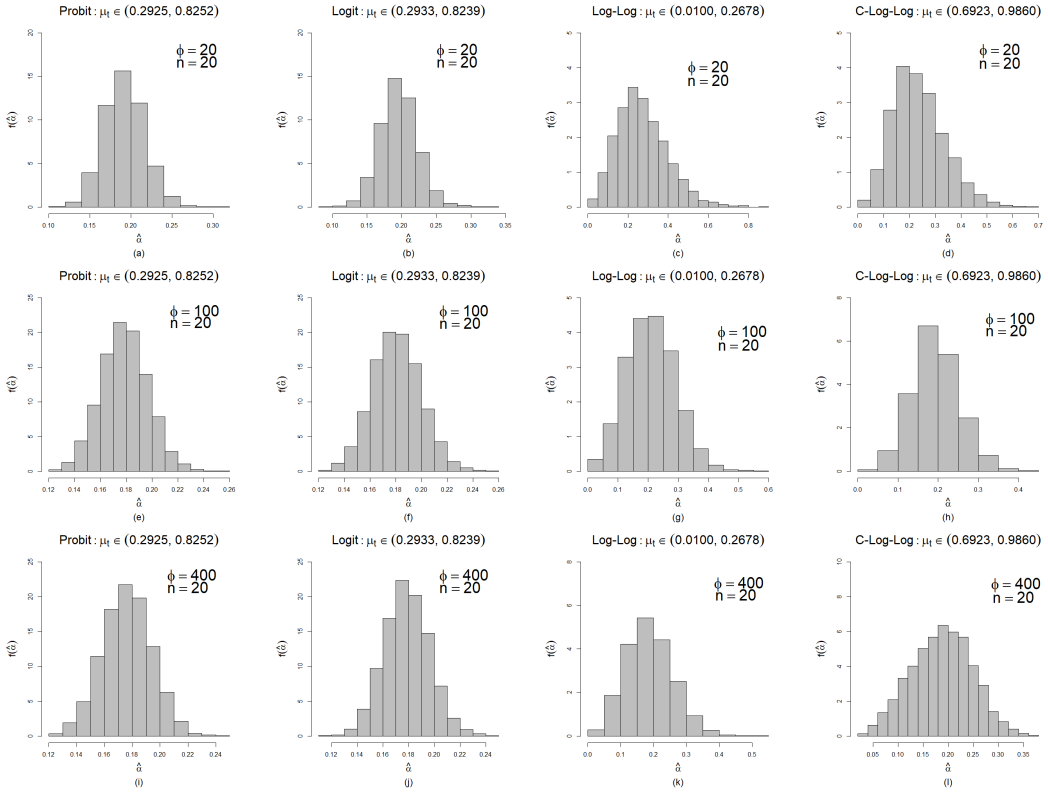
Fonte: Autoria própria.

Os desvios-padrão de $\hat{\alpha}$ estão bastante próximos de zero, então, não existe muita variação nos valores dessa medida, independentemente do cenário considerado. Além disso, temos também que os desvios-padrão de $\hat{\alpha}$ para valores de μ_t centrais são menores do que para valores de μ_t nos extremos do intervalo unitário e diminuem quando o valor de ϕ cresce. Esses valores também tendem aumentar com o tamanho amostral, entretanto, isso ocorre lentamenta. Quanto

aos coeficientes de assimetria de $\hat{\alpha}$ vemos que seus valores em geral são próximos de zero, principalmente quando ϕ é alto e os valores para μ estão próximos de 0.5, utilizando as funções de ligação logit e em especial a função probit. Para μ_t nos extremos do intervalo $(0,1)$ temos que, mesmo usando a função log-log para os valores próximos de zero e c-log-log para próximos de um, os coeficientes de assimetria encontram-se mais distantes de zero. No entanto, eles tendem diminuir quando o tamanho amostral e também o valor de ϕ aumentam.

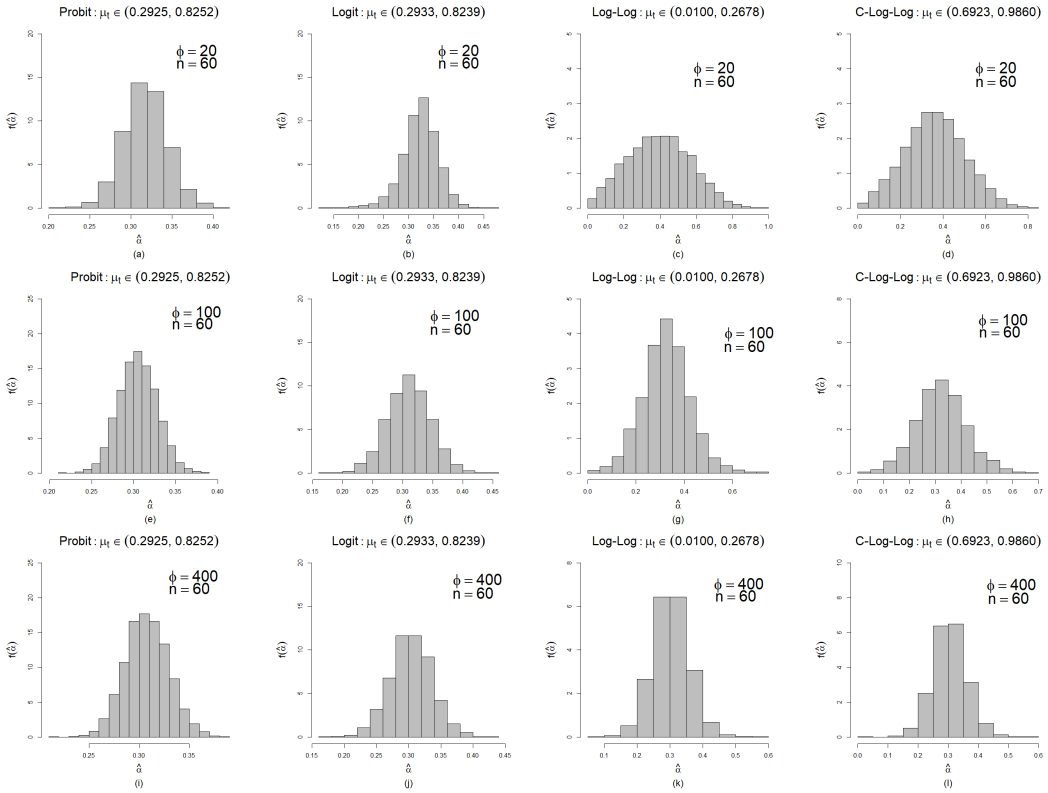
Avaliando a curtose, temos valores próximos de três na maioria dos casos, mas principalmente para valores de μ_t distribuídos longe dos extremos do intervalo $(0,1)$ e altos valores de ϕ. Dessa forma, nesses cenários, o gráfico da distribuição de $\hat{\alpha}$ ajusta-se bem à curva da distribuição normal padrão. Quando μ_t está perto de zero ou um, os valores da curtose mais próximos à três são para $\phi = 100$. Além disso, para μ_t nos extremos de $(0,1)$, com $\phi = 20$ e $n \geq 80$, os valores da curtose se aproximam de dois, ou seja, o gráfico da distribuição de $\hat{\alpha}$ possui um formato mais aberto do que a curva da distribuição normal padrão, consequentemente, apresenta maior dispersão. Dessa forma, mesmo em um modelo corretamente especificado, as chances de obtermos um valor de $\hat{\alpha}$ distante da sua média para esses cenários são maiores. Entretanto, em alguns casos vemos que as curtoses de $\hat{\alpha}$ são substancialmente maiores que três, consequentemente a dispersão dos dados são menores e os valores de $\hat{\alpha}$ estão muito próximos do seu valor médio.

Figura 22 – Histogramas dos valores $\hat{\alpha}$ no modelo beta linear corretamente especificado, com ϕ fixo e $n = 20$.



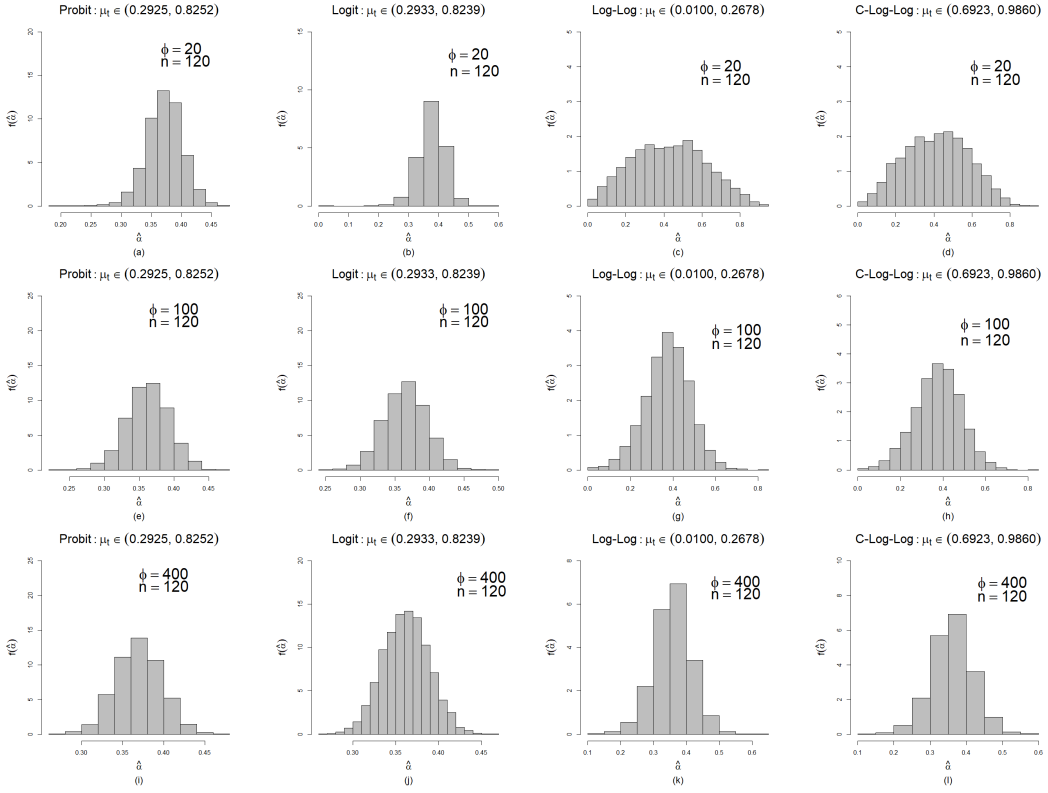
Fonte: Autoria própria.

Figura 23 – Histogramas dos valores $\hat{\alpha}$ no modelo beta linear corretamente especificado, com ϕ fixo e $n = 60$.



Fonte: Autoria própria.

Figura 24 – Histogramas dos valores $\hat{\alpha}$ no modelo beta linear corretamente especificado, com ϕ fixo e $n = 120$.



Fonte: Autoria própria.

Nas Figuras 22, 23 e 24 encontramos os histogramas dos valores de $\hat{\alpha}$, para os tamanhos amostrais 20, 60 e 120, respectivamente, para cada cenário de μ_t com as respectivas funções de ligação utilizadas e os valores de $\phi = 20$, 100 e 400. Pelos histogramas apresentados na Figura 22 observamos que quando $n = 20$ e $\mu_t \in (0.2925, 0.8252)$ a distribuição empírica de $\hat{\alpha}$ é aproximadamente simétrica em torno de 0.18 com dispersão aproximadamente constante, independentemente do valor de ϕ. Para os valores de μ_t concentrados nos extremos do intervalo unitário, notamos que distribuição empírica de $\hat{\alpha}$ é levemente assimétrica positiva apenas quando $\phi = 20$. Essa assimetria diminui quando o valor para ϕ aumenta. Nas Figuras 23 e 24, com $n = 60$ e $n = 120$, identificamos que a distribuição de $\hat{\alpha}$ é aproximadamente simétrica para todos os intervalos da média da resposta. Vale observar ainda que para $\mu_t \approx 0$ e $\mu_t \approx 1$ a dispersão dos dados é maior quando $\phi = 20$.

Com base no que foi visto sobre $\hat{\alpha}$, concluímos que essa estatística é bastante sensível às dificuldades encontradas em ajustes de modelos beta com precisão fixa. Como alto viés de $\hat{\phi}$, especialmente quando n é pequeno e alto viés também para μ_t quando ϕ é baixo e as variâncias elevadas. Ou seja, a medida $\hat{\alpha}$ capta o baixo desempenho do processo de estimação dos parâmetros. Logo, mesmo para dados gerados a partir de um modelo com precisão fixa, quando o tamanho amostral é pequeno, o alto viés de $\hat{\phi}$ faz com que $\hat{\alpha}$ capte um problema no ajuste, possivelmente, “a falta de um modelo para a precisão”.

6 APLICAÇÕES

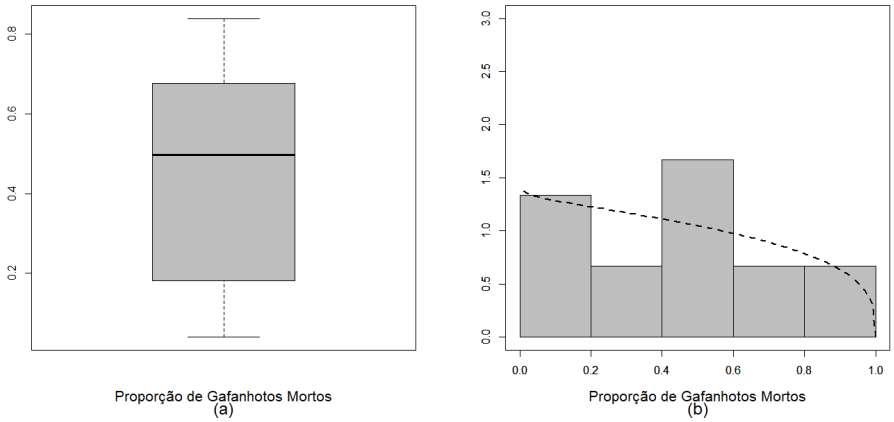
6.1 INTRODUÇÃO

No presente capítulo ajustamos o modelo de regressão beta a alguns conjuntos de dados e avaliamos a adequabilidade dos ajustes por meio das estatísticas BV e $\hat{\alpha}$, e das demais estatísticas presentes no decorrer desse trabalho. Além disso, foram realizadas algumas análises de diagnóstico tradicionais, como análises de resíduos e de pontos influentes. Dessa forma, podemos comparar as conclusões a respeito da adequação do modelo, obtidas por meio da análise tradicional e pelos resultados das estatísticas BV e $\hat{\alpha}$, para cada conjunto de dados.

6.2 APLICAÇÃO I: DADOS DO INSETICIDA

Iremos considerar nessa aplicação os dados disponíveis em McCullagh e Nelder (1989) de um estudo sobre a minimização de gastos com inseticidas, baseando-se na mistura de sinérgico piperonyl butoxide (PB) com o inseticida carbofuran, uma vez que o sinérgico torna o inseticida mais tóxico. O trabalho considera gafanhotos da espécie Melanopus sanguinipes e observa a proporção desses insetos mortos após a aplicação da mistura. Dispomos de um conjunto de 15 observações, com as variáveis explicativas sendo: as doses de inseticida (x_2) e as doses de sinérgico (x_3), utilizadas no preparo.

Figura 25 – Gráfico de *boxplot* e histograma da proporção de gafanhotos mortos.



Fonte: Autoria própria.

Na Figura 25 temos o gráfico de *boxplot* da proporção de gafanhotos mortos, e também o histograma dessa variável, junto à densidade beta, estimada para o modelo sem covariáveis. Podemos observar que os dados dessa variável apresentam valores distribuídos ao longo de todo intervalo $(0, 1)$, com maior concentração de observações em torno de 0.5 e também próximos de zero. Assim, observamos inicialmente uma distribuição assimétrica dos dados. Além disso, observando o *boxplot*, identificamos também uma maior dispersão dos dados para os valores

abaixo de 0.5. Temos também que os valores da média, mediana, curtose e assimetria, para a variável resposta desta aplicação, são dados respectivamente por: 0.450, 0.497, 1.701 e -0.208.

Ajustamos ao conjunto de dados:

$$\text{M1} \quad : \quad g(\mu_t) = \beta_1 + \beta_2 \log(x_{t2}) + \beta_3 x_{t3} \quad \text{e} \quad \sqrt{\phi_t} = \gamma_2 \log(x_{t2}),$$

$$\text{M2} \quad : \quad g(\mu_t) = \beta_1 + \beta_2 \log(x_{t2}) + \beta_3 x_{t3} \quad \text{e} \quad \phi_t = \gamma_2 \log(x_{t2}),$$

$$\text{M3} \quad : \quad g(\mu_t) = \beta_1 + \beta_2 \log(x_{t2} - \beta_3) + \beta_4 \frac{x_{t3}}{x_{t3} + \beta_5} \quad \text{e} \quad \sqrt{\phi_t} = \gamma_2 x_{t2} + \gamma_3 x_{t2} x_{t3},$$

$$\text{M4} \quad : \quad g(\mu_t) = \beta_1 + \beta_2 \log(x_{t2} - \beta_3) + \beta_4 \frac{x_{t3}}{x_{t3} + \beta_5} \quad \text{e} \quad \phi_t = \gamma_2 x_{t2} + \gamma_3 x_{t2} x_{t3},$$

com $t = 1, \ldots, 15$. Além disso, foram consideradas diferentes funções de ligação $g(\mu_t)$. Nas Tabelas 22 e 23 temos as estimativas dos parâmetros para os modelos M1 e M2, e para os modelos M3 e M4, respectivamente, junto com os erros-padrão e com os respectivos p-valores. Podemos então observar que para todos os modelos ajustados, os seus respectivos parâmetros podem ser considerados relevantes, ao nível de significância de 10%. No entanto, sendo um pouco mais rigoroso, e assumindo um nível de significância de 5%, vemos que, para o modelo não linear M4, com $h(\phi_t) = \phi_t$, o parâmetro γ_2 não é significativo para nenhum modelo com as quatro funções de ligação testadas, conforme observamos na Tabela 23.

Com base nos valores presentes na Tabela 24, observamos que as médias estimadas estão distribuídas ao longo do intervalo $(0,1)$, tanto para os modelos lineares como para os não lineares. Já para as estimativas de ϕ_t, temos valores extremamente maiores para os modelos não lineares, consequentemente, as variâncias e os graus de heteroscedasticidade são menores para essa classe.

Tabela 22 – Estimativas dos parâmetros, erros-padrão (s.e.) e respectivos p-valores, para os modelos lineares M1: $g(\mu_t) = \beta_1 + \beta_2 \log(x_{t2}) + \beta_3 x_{t3}$ e $\sqrt{\phi_t} = \gamma_2 \log(x_{t2})$ e M2: $g(\mu_t) = \beta_1 + \beta_2 \log(x_{t2}) + \beta_3 x_{t3}$ e $\phi_t = \gamma_2 \log(x_{t2})$, com diferentes funções de ligação para os modelos de μ_t. Dados do inseticida.

	M1											
Parâmetros	$g(\mu_t)$: Logit			$g(\mu_t)$: Probit			$g(\mu_t)$: Log-Log			$g(\mu_t)$: C-Log-Log		
	Estim.	s.e.	p-valor	Estim.	s.e.	p-valor	Estim.	s.e.	p-valor	Estim.	s.e.	p-valor
β_1	-3.4238	0.5763	0.0000	-2.0745	0.3332	0.0000	-1.9875	0.3070	0.0000	-2.6339	0.4426	0.0000
β_2	1.5574	0.2398	0.0000	0.9448	0.1385	0.0000	1.0958	0.1382	0.0000	1.0188	0.1746	0.0000
β_3	0.0410	0.0097	0.0000	0.0246	0.0056	0.0000	0.0300	0.0071	0.0000	0.0255	0.0059	0.0000
γ_2	2.3824	0.4193	0.0000	2.3785	0.4199	0.0000	2.4689	0.4440	0.0000	2.2164	0.3860	0.0000
	M2											
β_1	-3.8724	0.5772	0.0000	-2.3355	0.3227	0.0000	-2.1795	0.2820	0.0000	-3.0187	0.4590	0.0000
β_2	1.7644	0.2506	0.0000	1.0645	0.1404	0.0000	1.1968	0.1361	0.0000	1.1786	0.1862	0.0000
β_3	0.0383	0.0100	0.0001	0.0230	0.0058	0.0001	0.0256	0.0066	0.0001	0.0254	0.0065	0.0001
γ_2	9.8446	3.5248	0.0052	9.9080	3.5590	0.0054	10.9339	3.9892	0.0061	8.1449	2.8903	0.0048

Fonte: Autoria própria.

Tabela 23 – Estimativas dos parâmetros, erros-padrão (s.e.) e respectivos p-valores, para os modelos não lineares M3: $g(\mu_t) = \beta_1 + \beta_2 \log(x_{t2} - \beta_3) + \beta_4 \frac{x_{t3}}{x_{t3}+\beta_5}$ e $\sqrt{\phi_t} = \gamma_2 x_{t2} + \gamma_3 x_{t2} x_{t3}$ e M4: $g(\mu_t) = \beta_1 + \beta_2 \log(x_{t2} - \beta_3) + \beta_4 \frac{x_{t3}}{x_{t3}+\beta_5}$ e $\phi_t = \gamma_2 x_{t2} + \gamma_3 x_{t2} x_{t3}$, com diferentes funções de ligação para os modelos de μ_t. Dados do inseticida.

	M3											
Parâmetros	$g(\mu_t)$: Logit			$g(\mu_t)$: Probit			$g(\mu_t)$: Log-Log			$g(\mu_t)$: C-Log-Log		
	Estim.	s.e.	p-valor	Estim.	s.e.	p-valor	Estim.	s.e.	p-valor	Estim.	s.e.	p-valor
β_1	−3.3476	0.0994	0.0000	−2.0141	0.0598	0.0000	−3.3125	0.1400	0.0000	−2.0569	0.0612	0.0000
β_2	1.5179	0.0263	0.0000	0.9152	0.0159	0.0000	1.5760	0.0395	0.0000	0.8053	0.0126	0.0000
β_3	1.5556	0.0380	0.0000	1.4898	0.0373	0.0000	0.3560	0.0999	0.0004	1.8891	0.0138	0.0000
β_4	1.8446	0.0715	0.0000	1.0941	0.0424	0.0000	1.5463	0.0852	0.0000	1.0179	0.0520	0.0000
β_5	1.7234	0.2736	0.0000	1.6748	0.2606	0.0000	1.6600	0.2889	0.0000	1.5946	0.2855	0.0000
γ_2	1.0944	0.3058	0.0003	1.1079	0.3098	0.0003	0.6844	0.1846	0.0002	0.8940	0.2504	0.0004
γ_3	0.6236	0.1714	0.0003	0.6421	0.1761	0.0003	0.6479	0.1669	0.0001	0.6066	0.1632	0.0002
	M4											
β_1	−3.3017	0.1410	0.0000	−1.9996	0.0885	0.0000	−3.0400	0.2261	0.0000	−2.1092	0.0934	0.0000
β_2	1.5045	0.0431	0.0000	0.9107	0.0281	0.0000	1.5148	0.0772	0.0000	0.8123	0.0235	0.0000
β_3	1.5756	0.0455	0.0000	1.5028	0.0514	0.0000	0.4722	0.1673	0.0048	1.8901	0.0185	0.0000
β_4	1.8390	0.0949	0.0000	1.0963	0.0558	0.0000	1.4174	0.1053	0.0000	1.0666	0.0743	0.0000
β_5	1.8841	0.3232	0.0000	1.8291	0.3262	0.0000	1.8419	0.4073	0.0000	1.7471	0.3733	0.0000
γ_2	9.2980	5.2992	0.0793	9.8177	5.5984	0.0795	3.6730	2.0434	0.0723	5.9146	3.3624	0.0786
γ_3	26.9312	13.1505	0.0406	24.1774	11.8753	0.0418	17.2357	8.2727	0.0372	19.8559	9.6471	0.0396

Fonte: Autoria própria.

Tabela 24 – Estimativas dos valores mínimos e máximos de μ_t, ϕ_t e de $\text{Var}(y_t)$, e estimativas do grau de heteroscedasticidade λ, para os modelos lineares M1: $g(\mu_t) = \beta_1 + \beta_2 \log(x_{t2}) + \beta_3 x_{t3}$ e $\sqrt{\phi_t} = \gamma_2 \log(x_{t2})$ e M2: $g(\mu_t) = \beta_1 + \beta_2 \log(x_{t2}) + \beta_3 x_{t3}$ e $\phi_t = \gamma_2 \log(x_{t2})$ e para os modelos não lineares M3: $g(\mu_t) = \beta_1 + \beta_2 \log(x_{t2} - \beta_3) + \beta_4 \frac{x_{t3}}{x_{t3}+\beta_5}$ e $\sqrt{\phi_t} = \gamma_2 x_{t2} + \gamma_3 x_{t2} x_{t3}$ e M4: $g(\mu_t) = \beta_1 + \beta_2 \log(x_{t2} - \beta_3) + \beta_4 \frac{x_{t3}}{x_{t3}+\beta_5}$ e $\phi_t = \gamma_2 x_{t2} + \gamma_3 x_{t2} x_{t3}$, com diferentes funções de ligação para os modelos de μ_t. Dados do inseticida.

	Modelos Lineares				Modelos Não Lineares			
	M1				M3			
Estimadores	Logit	Probit	Log-Log	C-Log-Log	Logit	Probit	Log-Log	C-Log-Log
$\hat{\mu}_t$	(0.101, 0.853)	(0.093, 0.856)	(0.048, 0.834)	(0.148, 0.868)	(0.036, 0.839)	(0.031, 0.840)	(0.014, 0.839)	(0.044, 0.840)
$\hat{\phi}_t$	(2.73, 50.94)	(2.72, 50.77)	(2.93, 54.70)	(2.36, 44.08)	(19.16, 64584)	(19.64, 68386)	(7.49, 67363)	(12.79, 60286)
$\widehat{\text{Var}}(y_t)$	(0.003, 0.059)	(0.003, 0.059)	(0.003, 0.058)	(0.004, 0.065)	(0.000, 0.005)	(0.000, 0.005)	(< .000, 0.007)	(0.0000, 0.012)
$\hat{\lambda}$	17.493	17.476	17.617	17.210	2518.6	2616.1	8988.7	5369.9
	M2				M4			
$\hat{\mu}_t$	(0.076, 0.844)	(0.066, 0.844)	(0.030, 0.813)	(0.115, 0.863)	(0.034, 0.840)	(0.029, 0.840)	(0.015, 0.837)	(0.041, 0.842)
$\hat{\phi}_t$	(6.82, 29.49)	(6.87, 29.68)	(7.58, 32.75)	(5.65, 24.39)	(37.19, 10596)	(39.271, 9527)	(14.692, 6758)	(23.658, 7802)
$\widehat{\text{Var}}(y_t)$	(0.005, 0.023)	(0.005, 0.023)	(0.003, 0.021)	(0.006, 0.029)	(0.000, 0.003)	(0.0000, 0.003)	(0.000, 0.008)	(0.000, 0.006)
$\hat{\lambda}$	4.511	4.525	6.194	4.797	255.658	222.7	382.522	379.5

Fonte: Autoria própria.

Na Tabela 25 estão os valores das estatística de adequabilidade de ajuste, para cada um dos modelos ajustados. Podemos observar que, os modelos não lineares apresentam melhores ajustes, especialmente quando usada a função de ligação raiz quadrada no submodelo da precisão. O modelo com a função de ligação logit no submodelo da média fornece um maior valor para

a estatística BV e, além disso, o valor correspondente de $\hat{\alpha}$ é próximo de 0.5. Logo, é bastante provável que esse modelo apresente melhor ajuste ao conjunto de dados, além de melhor poder de predição.

Tabela 25 – Valores das estatísticas de adequabilidade para os modelos lineares M1: $g(\mu_t) = \beta_1 + \beta_2 \log(x_{t2}) + \beta_3 x_{t3}$ e $\sqrt{\phi_t} = \gamma_2 \log(x_{t2})$ e M2: $g(\mu_t) = \beta_1 + \beta_2 \log(x_{t2}) + \beta_3 x_{t3}$ e $\phi_t = \gamma_2 \log(x_{t2})$ e para os modelos não lineares M3: $g(\mu_t) = \beta_1 + \beta_2 \log(x_{t2} - \beta_3) + \beta_4 \frac{x_{t3}}{x_{t3}+\beta_5}$ e $\sqrt{\phi_t} = \gamma_2 x_{t2} + \gamma_3 x_{t2} x_{t3}$ e M4: $g(\mu_t) = \beta_1 + \beta_2 \log(x_{t2} - \beta_3) + \beta_4 \frac{x_{t3}}{x_{t3}+\beta_5}$ e $\phi_t = \gamma_2 x_{t2} + \gamma_3 x_{t2} x_{t3}$, com diferentes funções de ligação para os modelos de μ_t. Dados do inseticida.

	Modelos Lineares				Modelos Não Lineares			
	M1				M3			
Estatísticas	Logit	Probit	Log-Log	C-Log-Log	Logit	Probit	Log-Log	C-Log-Log
P^2	0.8358	0.8512	0.8725	0.7608	0.9994	0.9996	0.9995	0.9993
P_c^2	0.7911	0.8106	0.8377	0.6956	0.9990	0.9993	0.9991	0.9988
R^2_{FC}	0.8203	0.8292	0.8357	0.7977	0.9744	0.9775	0.9660	0.9467
$R^2_{FC_c}$	0.7713	0.7827	0.7909	0.7425	0.9553	0.9606	0.9404	0.9068
R^2_{LR}	0.8471	0.8454	0.8465	0.8307	0.9960	0.9961	0.9917	0.9953
$R^2_{LR_c}$	0.7389	0.7361	0.7380	0.7110	0.9882	0.9886	0.9759	0.9863
BV	0.8226	0.8100	0.7958	0.8039	0.9780	0.9766	0.9704	0.9347
$\hat{\alpha}$	0.7740	0.6420	0.4300	0.8020	0.5140	0.4100	0.5080	0.3020
	M2				M4			
P^2	0.8373	0.8564	0.8745	0.7708	0.9986	0.9989	0.9986	0.9974
P_c^2	0.7930	0.8173	0.8403	0.7083	0.9975	0.9981	0.9975	0.9955
R^2_{FC}	0.8487	0.8545	0.8595	0.8219	0.9739	0.9771	0.9701	0.9476
$R^2_{FC_c}$	0.8075	0.8148	0.8211	0.7734	0.9544	0.9598	0.9476	0.9083
R^2_{LR}	0.8547	0.8540	0.8579	0.8319	0.9951	0.9949	0.9885	0.9933
$R^2_{LR_c}$	0.7519	0.7508	0.7573	0.7130	0.9857	0.9850	0.9665	0.9805
BV	0.8105	0.8002	0.7866	0.7862	0.9604	0.9652	0.9606	0.9139
$\hat{\alpha}$	0.5700	0.4680	0.2500	0.6300	0.1360	0.1360	0.2540	0.0620

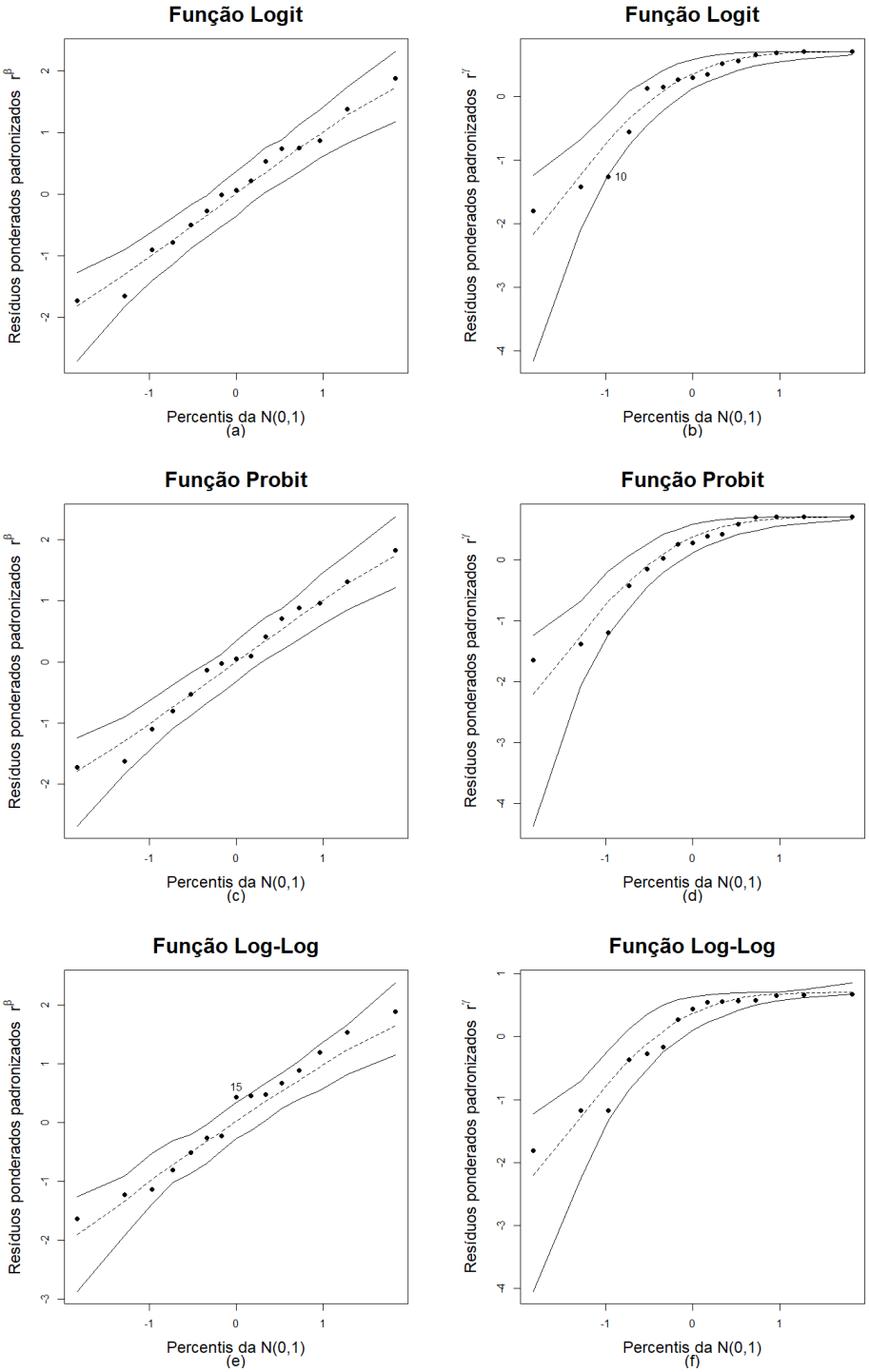
Fonte: Autoria própria.

A Tabela 25 contém informações muito importantes. Aparentemente, os modelos lineares apresentam valores razoáveis para as estatísticas de avaliação do modelo. No entanto, os valores de $\hat{\alpha}$ estão tipicamente distantes de 0.5. Um ponto interessante à ser observado são os valores consideravelmente reduzidos de $\hat{\alpha}$ para o modelo não linear M4, apesar dos altos valores dos critérios de seleção de modelos. Ou seja, a estatística $\hat{\alpha}$ é um critério poderoso para avaliar a especificação correta do modelo, incluindo a função de ligação do submodelo da precisão.

Vamos agora avaliar os modelos não lineares com função de ligação $\sqrt{\phi_t}$. Podemos observar então que, o modelo com a função log-log apresenta $\hat{\alpha}$ mais próximo de 0.5, logo, acreditamos que os vieses de $\hat{\phi}_t$, com $t = 1, \ldots, 15$, são menores para esse modelo, consequentemente, o ajuste do modelo da precisão nesse caso é melhor, como podemos confirmar pelo gráfico de envelopes simulados dado a seguir, pela Figura 26f. No entanto, a qualidade do ajuste do

modelo da média é inferior, conforme vemos pelo gráfico de envelopes da Figura 26e.

Figura 26 – Gráficos de envelopes dos resíduos ponderados padronizados r^{β} e r^{γ}, para o modelo não linear com as funções logit, probit e log-log no modelo da média e $h(\phi_t) = \sqrt{\phi_t}$, para os dados do inseticida.



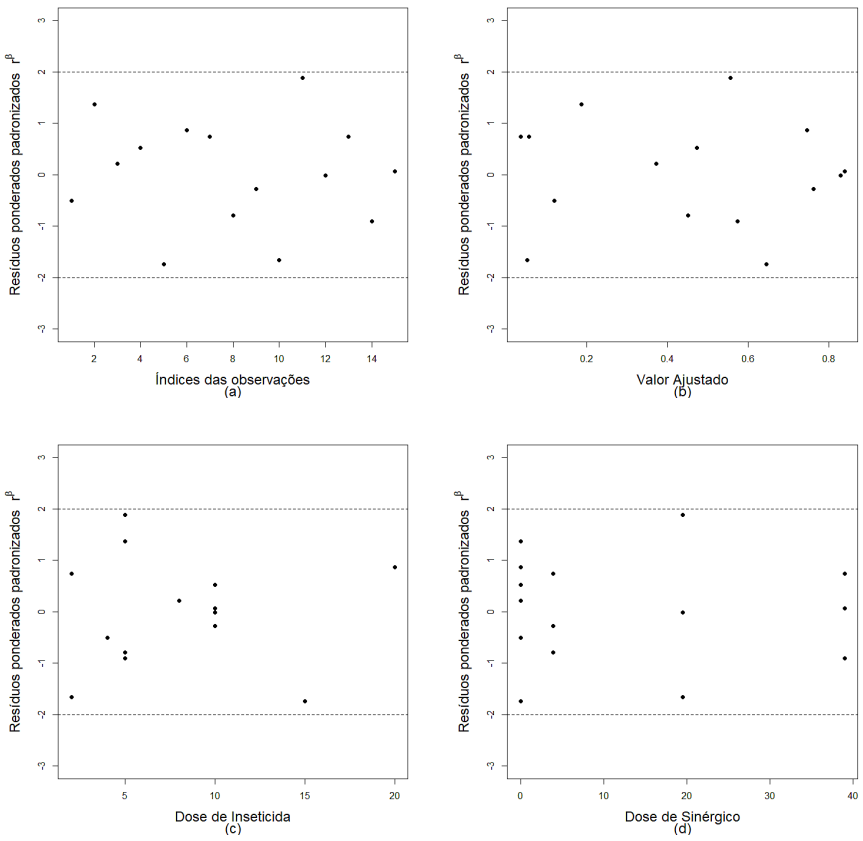
Fonte: Autoria própria.

Outra observação é para a função probit, pois segundo as medidas P^2 e R^2, esse seria o modelo com melhor ajuste, e de fato, o modelo possui uma boa adequação, conforme o gráfico da Figura 26c. Entretanto, o valor de $\hat{\alpha}$ para esse caso indica que os valores estimados da precisão são mais viesados. Esse fato vai de acordo com o gráfico na Figura 26d, uma vez que

para esse modelo notamos um número maior de pontos localizados nos limites do gráfico de envelopes simulados com os resíduos do submodelo da precisão.

Portanto, dentre os modelos avaliados para os dados de inseticida, notamos inicialmente que os valores $\hat{\alpha} = 0.514$ e $\hat{\alpha} = 0.508$, para os modelos não lineares, cujas funções de ligação para o submodelo da média são logit e log-log, respectivamente, apontam melhor adequabilidade desses modelos, dado que suas constantes de linearização deram próximas de 0.5. No entanto, é importante fazer a análise conjunta entre BV e $\hat{\alpha}$; e nesse sentido, o modelo com função logit mostra-se mais adequado.

Figura 27 – Gráficos dos resíduos ponderados padronizados r^{β}, para o modelo não linear com a função de ligação logit no modelo da média e $h(\phi_t) = \sqrt{\phi_t}$, para os dados da proporção de gafanhotos mortos.

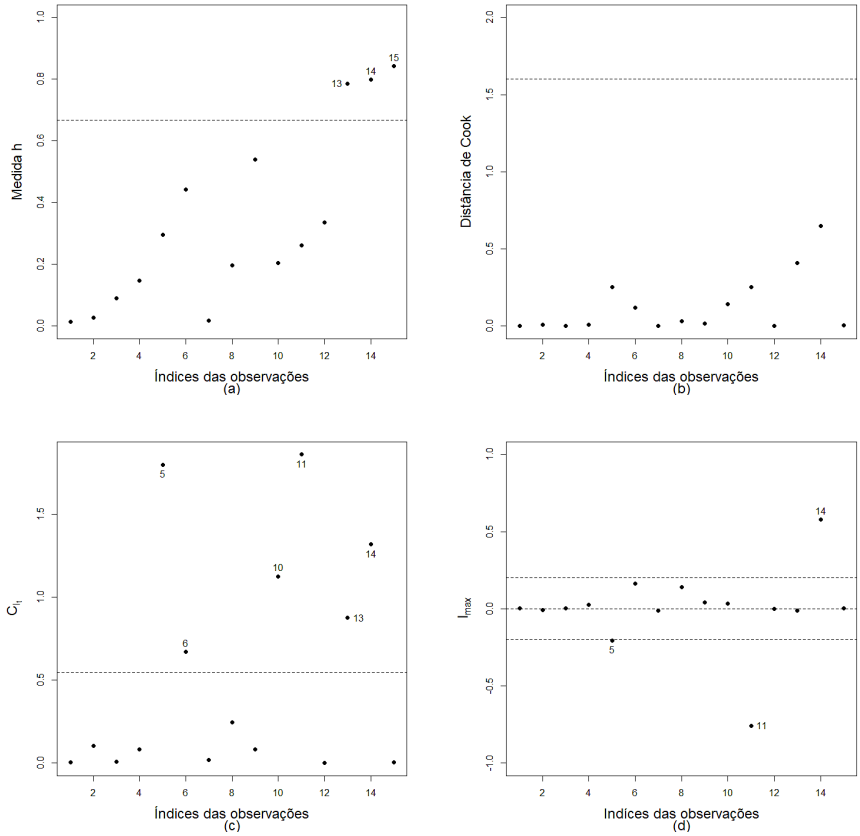


Fonte: Autoria própria.

Adicionalmente, realizamos uma análise de resíduos para validar as conclusões obtidas. Na Figura 27 estão os gráficos dos resíduos ponderados padronizados r^{β} *versus* os indíces das observações, os valores ajustados e as covariáveis do modelo, para o caso não linear, com as funções de ligação logit e raiz quadrada nos submodelos de μ_t e ϕ_t, respectivamente. Podemos observar que os resíduos encontram-se distribuídos aleatoriamente em torno de zero e dentro do intervalo $(-2, 2)$. Na Figura 28 temos alguns gráficos de diagnóstico para o modelo não linear com as funções logit e raiz quadrada. Podemos notar pela Figura 28a que as observações 13, 14 e 15 foram indicadas como sendo pontos com altos graus de alavancagem, podendo influenciar demasiadamente nas estimativas dos parâmetros. No entanto, pela distância de Cook não foram detectadas observações influentes, conforme vemos na Figura 28b. Com base nos gráficos de

influência local, vemos que as curvaturas normais $(C_{I_{max}})$, apresentadas pela Figura 28c, indicam os pontos 5, 6, 10, 11, 13 e 14 como prováveis observaçoes influentes. Além disso, pela Figura 28d, com o gráfico dos vetores I_{max}, verificamos que os pontos 5 e 11 são indicadas como sendo provavelmente observações conjuntamente influentes, também vemos a observação 14 novamente indicada como um possível ponto influente.

Figura 28 – Gráficos dos possíveis pontos de alavanca e influentes, para o modelo não linear com a função de ligação logit no modelo da média e $h(\phi_t) = \sqrt{\phi_t}$, para os dados da proporção de gafanhotos mortos.



Fonte: Autoria própria.

Dessa forma, decidimos avaliar o impacto que essas observações provocam nas estimativas dos parâmetros. Para isso, buscamos reajustar o modelo sem essas observações, em seguida calculamos a mudança relativa em porcentagem das estimativas dos parâmetros do modelo. Essa mudança é definida por

$$\text{Mudança Relativa em Porcentagem} \quad = \quad \left| \frac{\widehat{\theta} - \widehat{\theta}_{(i)}}{\widehat{\theta}} \right| \times 100\%,$$

em que $\widehat{\theta}_{(i)}$ é a estimativa do parâmetro $\widehat{\theta}$ sem a i-ésima observação. Na Tabela 26 encontramos mudança relativa provocada nas estimativas de cada um dos parâmetros, quando retirada cada uma das observações apontadas anteriormente como possíveis pontos influentes. Além disso, também temos os p-valores referentes aos testes de nulidade de cada parâmetro estimado.

Pela Tabela 26, vemos que a observação 5 exerce influência maior na estimativa do pa-

râmetro γ_2 do que nas estimativas dos demais parâmetros, enquanto que as observações 11 e 14 exercem de forma individual uma influência maior na estimativa do parâmetro γ_3. Porém, vale notar que as estimativas para esses parâmetros com todas as observações, são baixas, logo, qualquer pequena mudança torna-se muito significativa, em termos percentuais. Além disso, não observamos mudanças drásticas nos p-valores, de modo que os parâmetros permanecem significativos ao nível de 1%. Logo, essas observações não são influentes, então, o modelo de fato pode ser considerado adequado ao conjunto de dados.

Tabela 26 – Mudanças relativas (M. R.) em porcentagens das estimativas dos parâmetros e respectivos p-valores, verificados com a retirada de observações, para os dados da proporção de gafanhotos mortos, com $g(\cdot)$ logit e $h(\phi_t)$ raiz quadrada.

Pontos Retirados	Medidas	β_1	β_2	β_3	β_4	β_5	γ_2	γ_3
5	M. R. (%)	2.8288	0.0263	0.0514	4.8791	9.8061	57.2185	8.1622
	p-valor	0.0000	0.0000	0.0000	0.0000	0.0000	0.0010	0.0009
6	M. R. (%)	1.9327	0.2239	0.2314	3.8599	4.9495	4.4956	0.1924
	p-valor	0.0000	0.0000	0.0000	0.0000	0.0000	0.0010	0.0003
10	M. R. (%)	0.8812	0.5995	1.9413	0.1355	2.4834	1.3066	24.9358
	p-valor	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0005
11	M. R. (%)	1.8998	1.6470	1.4013	0.8402	8.6805	1.1421	49.4708
	p-valor	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0004
13	M. R. (%)	2.2613	1.6865	5.7405	0.3794	1.7233	1.8366	19.8203
	p-valor	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0005
14	M. R. (%)	6.6495	5.4417	4.7120	1.1113	15.5042	2.9148	36.3213
	p-valor	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0004
5 e 11	M. R. (%)	0.8603	1.6667	1.3885	4.0876	17.3378	54.7331	41.7896
	p-valor	0.0000	0.0000	0.0000	0.0000	0.0000	0.0012	0.0008

Fonte: Autoria própria.

6.3 APLICAÇÃO II: FATOR DE SIMULTANEIDADE

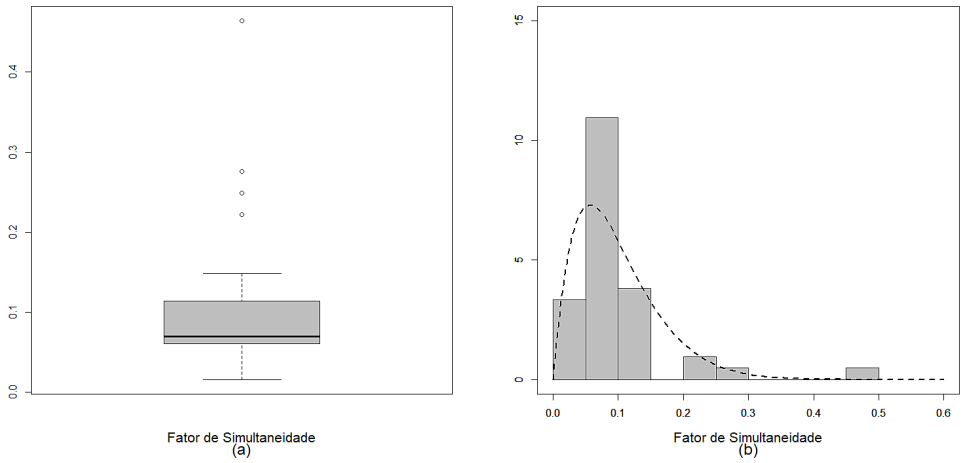
O conjunto de dados dessa aplicação possui 42 observações e corresponde a um estudo de medição desenvolvido por Zerbinatti (2008) e pelo Instituto de Pesquisas Tecnológicas (IPT) e a Companhia de Gás de São Paulo (COMGAS). O trabalho é referente à distribuição de gás natural para uso doméstico em São Paulo.

A distribuição de gás é baseada em dois fatores: o fator de simultaneidade e a capacidade máxima de consumo. O termo "fator de simultaneidade" corresponde a uma medida estimativa, que leva em consideração o fato de que todos os dispositivos de uma determinada instalação, nunca estão ligados simultaneamente, ou com força máxima. Esse fator permite indicar a quantidade máxima de gás que escoa em um trecho da tubulação por unidade de tempo. Esse cálculo é feito a partir da equação $Q_P = F \times Q_{max}$, em que Q_P é a vazão máxima que provavelmente escoa em um trecho da tubulação (a vazão adotada ou praticada), F é o fator de simultaneidade e Q_{max} é vazão máxima possível.

De acordo com Zerbinatti (2008), o fator de simultaneidade é um número adimensional, cujo valor está situado no intervalo no $(0,1)$, então, Q_P pode ser interpretado como uma proporção da vazão máxima possível Q_{max}. Logo, F representa a proporção da Q_{max} que provavelmente é utilizada em determinado trecho da tubulação. Zerbinatti (2008) ainda diz que os projetistas concordam que não é preciso que a distribuição de gás combustível atenda à capacidade máxima de consumo, bastando atender à demanda máxima praticada. Dessa forma, geralmente mensuram-se a capacidade máxima de consumo e a demanda máxima praticada por unidades de potências, sendo denominadas, respectivamente, como potência computada (que corresponde a Q_{max}) e potência adotada (que corresponde a Q_P).

Na Figura 29 vemos o *boxplot* e o histograma, respectivamente, para os valores dos fatores de simultaneidade disponíveis, junto ao histograma vemos a densidade beta, estimada para o modelo sem covariáveis. Podemos então observar que tais dados apresentam muitos valores perto de zero, além disso, vemos uma distribuição assimétrica dos dados. Pelo *boxplot*, identificamos maior dispersão dos dados para os valores acima da mediana, sendo esta igual a 0.0695. Ademais, os valores da média, curtose e assimetria, são dados respectivamente por 0.0974, 12.2891 e 2.8087.

Figura 29 – Gráficos de *boxplot* e histograma do fator de simultaneidade.



Fonte: Autoria própria.

Zerbinatti (2008) e Espinheira et al. (2014) consideram um modelo de regressão beta com precisão constante para modelar o fator de simultaneidade em relação da covariável potência computada, denotada por x_{t2}. Eles propõem utilizar a função de ligação logit no modelo da média e aplicar uma transformação logarítmica na covariável x_{t2}, esse modelo é apresentado a seguir por M1. Lima (2017) propõe considerar um modelo com precisão variável, dado por M2, utilizando a função de ligação complemento log-log no submodelo da média e a função log no submodelo da precisão. Espinheira et al. (2019) indicam utilizar o modelo proposto por Lima (2017), no entanto, aplicando a função de ligação log-log no submodelo de μ_t, conforme vemos em M3. Por fim, um novo modelo para esse conjunto de dados será proposto nesse trabalho, dado por M4. De modo que, similar ao modelo de Espinheira et al. (2019), consideramos a

função de ligação log-log no submodelo de μ_t e a função log no submodelo de ϕ_t. Entretanto, foi realizada uma transformação diferente na covariável do submodelo da precisão, sendo aplicanda a raiz quadrada na variável potência computada. Temos

$$\text{M1} \quad : \quad \log(\mu_t/(1-\mu_t)) = \beta_1 + \beta_2 \log(x_{t2}) \quad \text{e} \quad \phi \text{ constante},$$

$$\text{M2} \quad : \quad \log\{-\log(1-\mu_t)\} = \beta_1 + \beta_2 \log(x_{t2}) \quad \text{e} \quad \log(\phi_t) = \gamma_1 + \gamma_2 \log(x_{t2}),$$

$$\text{M3} \quad : \quad \log\{-\log(\mu_t)\} = \beta_1 + \beta_2 \log(x_{t2}) \quad \text{e} \quad \log(\phi_t) = \gamma_1 + \gamma_2 \log(x_{t2}),$$

$$\text{M4} \quad : \quad \log\{-\log(\mu_t)\} = \beta_1 + \beta_2 \log(x_{t2}) \quad \text{e} \quad \log(\phi_t) = \gamma_1 + \gamma_2 \sqrt{x_{t2}},$$

com $t = 1, \ldots, 42$. Nosso objetivo é avaliar cada um desses modelos a partir das nossas medidas de análise de adequabilidade, BV e $\hat{\alpha}$, e assim, selecionar o melhor modelo considerado adequado aos dados do fator de simultaneidade.

Na Tabela 27 encontramos as estimativas dos parâmetros para os quatros modelos ajustados ao conjunto de dados do gás, juntamente com os erros-padrão e com respectivos p-valores. Verificamos que para todos os modelos os seus os parâmetros são considerados relevantes, ao nível de significância de 10%. Entretato, sendo mais rigosos e considerando 5% de significância, temos que o parâmetro γ_2 em M2 não é significativo para esse modelo.

Tabela 27 – Estimativas dos parâmetros, erro-padrão (s.e.) e respectivos p-valores, para o modelo M1: $\log(\mu_t/(1-\mu_t)) = \beta_1 + \beta_2 \log(x_{t2})$ e ϕ constante, M2: $\log\{-\log(1-\mu_t)\} = \beta_1 + \beta_2 \log(x_{t2})$ e $\log(\phi_t) = \gamma_1 + \gamma_2 \log(x_{t2})$, M3: $\log\{-\log(\mu_t)\} = \beta_1 + \beta_2 \log(x_{t2})$ e $\log(\phi_t) = \gamma_1 + \gamma_2 \log(x_{t2})$ e M4: $\log\{-\log(\mu_t)\} = \beta_1 + \beta_2 \log(x_{t2})$ e $\log(\phi_t) = \gamma_1 + \gamma_2 \sqrt{x_{t2}}$. Dados do gás.

Parâmetros	Modelos Ajustados											
	M1			M2			M3			M4		
	Estim.	s.e.	p-valor	Estim.	s.e.	p-valor	Estim.	s.e.	p-valor	Estim.	s.e.	p-valor
β_1	−1.7122	0.0672	0.0000	−1.8172	0.0769	0.0000	−0.6270	0.0459	0.0000	−0.6073	0.0446	0.0000
β_2	−0.7935	0.0665	0.0000	−0.7399	0.0723	0.0000	−0.3135	0.0381	0.0000	−0.3323	0.0354	0.0000
γ_1	-	-	-	4.0451	0.3261	0.0000	3.8117	0.3252	0.0000	2.5978	0.7320	0.0004
γ_2	-	-	-	0.4919	0.2953	0.0957	0.7725	0.2948	0.0088	1.1548	0.4372	0.0083

Fonte: Autoria própria.

Pela Tabela 28 observamos os intervalos dos valores estimados para as médias, as precisões e as variâncias, também encontramos as estimativas dos graus de heteroscedasticidade e a precisão, quando aplicado o modelo com dispersão fixa. Notamos então que os valores estimados para as médias estão sempre próximos de zero. Já para as estimativas de ϕ_t, encontramos valores maiores para o modelo M4. Comparando os intervalos das variâncias estimadas para os modelos com dispersão variável, vemos que o modelo M1 apresenta variâncias mais próximas, consequentemente, esse modelo indica menor grau de heteroscedasticidade, diferentemente dos modelos M3 e M4, que estimam graus de heteroscedasticidade maiores e similares.

Tabela 28 – Estimativas dos valores mínimos e máximos de μ_t, ϕ_t e de $\mathrm{Var}(y_t)$, e estimativas do grau de heteroscedasticidade λ, para o modelo M1: $\log(\mu_t/(1-\mu_t)) = \beta_1 + \beta_2\log(x_{t2})$ e ϕ constante, M2: $\log\{-\log(1-\mu_t)\} = \beta_1 + \beta_2\log(x_{t2})$ e $\log(\phi_t) = \gamma_1 + \gamma_2\log(x_{t2})$, M3: $\log\{-\log(\mu_t)\} = \beta_1 + \beta_2\log(x_{t2})$ e $\log(\phi_t) = \gamma_1 + \gamma_2\log(x_{t2})$ e M4: $\log\{-\log(\mu_t)\} = \beta_1 + \beta_2\log(x_{t2})$ e $\log(\phi_t) = \gamma_1 + \gamma_2\sqrt{x_{t2}}$. Dados do gás.

Estimadores	Modelos Ajustados			
	M1	M2	M3	M4
$\hat{\mu}_t$	(0.0312, 0.4252)	(0.0320, 0.4542)	(0.0247, 0.3423)	(0.0228, 0.3618)
$\hat{\phi}_t$	79.345	(23.823, 166.347)	(11.453, 242.389)	(21.598, 411.952)
$\widehat{\mathrm{Var}}(y_t)$	(0.00037, 0.00304)	(0.00018, 0.00998)	(0.00000, 0.018078)	(0.00000, 0.010218)
$\hat{\lambda}$	8.094	53.920	182.43	189.01

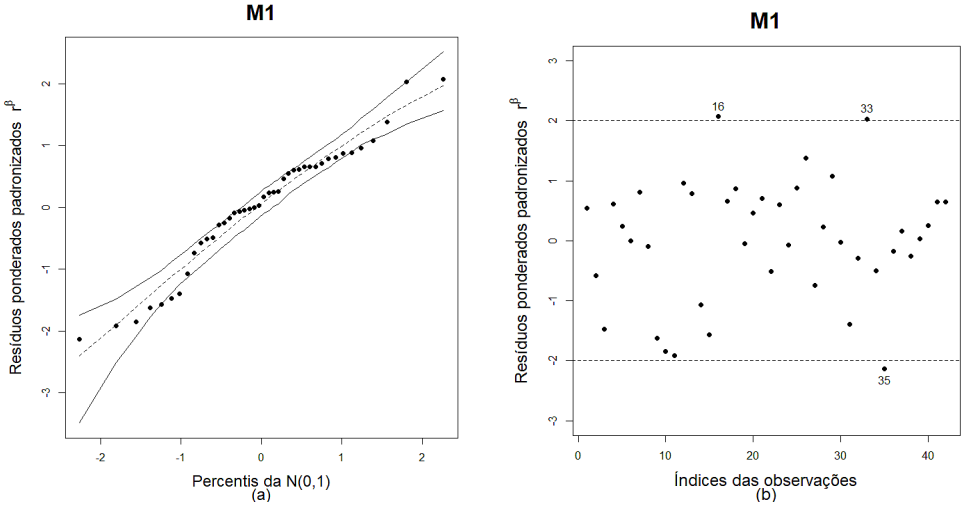
Fonte: Autoria própria.

Tabela 29 – Valores das medidas de qualidade de ajuste para o modelo M1: $\log(\mu_t/(1-\mu_t)) = \beta_1 + \beta_2\log(x_{t2})$ e ϕ constante, M2: $\log\{-\log(1-\mu_t)\} = \beta_1 + \beta_2\log(x_{t2})$ e $\log(\phi_t) = \gamma_1 + \gamma_2\log(x_{t2})$, M3: $\log\{-\log(\mu_t)\} = \beta_1 + \beta_2\log(x_{t2})$ e $\log(\phi_t) = \gamma_1 + \gamma_2\log(x_{t2})$ e M4: $\log\{-\log(\mu_t)\} = \beta_1 + \beta_2\log(x_{t2})$ e $\log(\phi_t) = \gamma_1 + \gamma_2\sqrt{x_{t2}}$. Dados do gás.

Estatísticas	M1	M2	M3	M4
P^2	0.4237	0.6251	0.8813	0.9073
P_c^2	0.3941	0.5955	0.8719	0.9000
R_{FC}^2	0.6905	0.6774	0.7241	0.7241
$R_{FC_c}^2$	0.6746	0.6520	0.7023	0.7023
R_{LR}^2	0.7219	0.7432	0.7427	0.7426
$R_{LR_c}^2$	0.7046	0.7229	0.7223	0.7223
BV	0.5626	0.6820	0.8143	0.7974
$\hat{\alpha}$	0.5140	0.5860	0.6260	0.4640

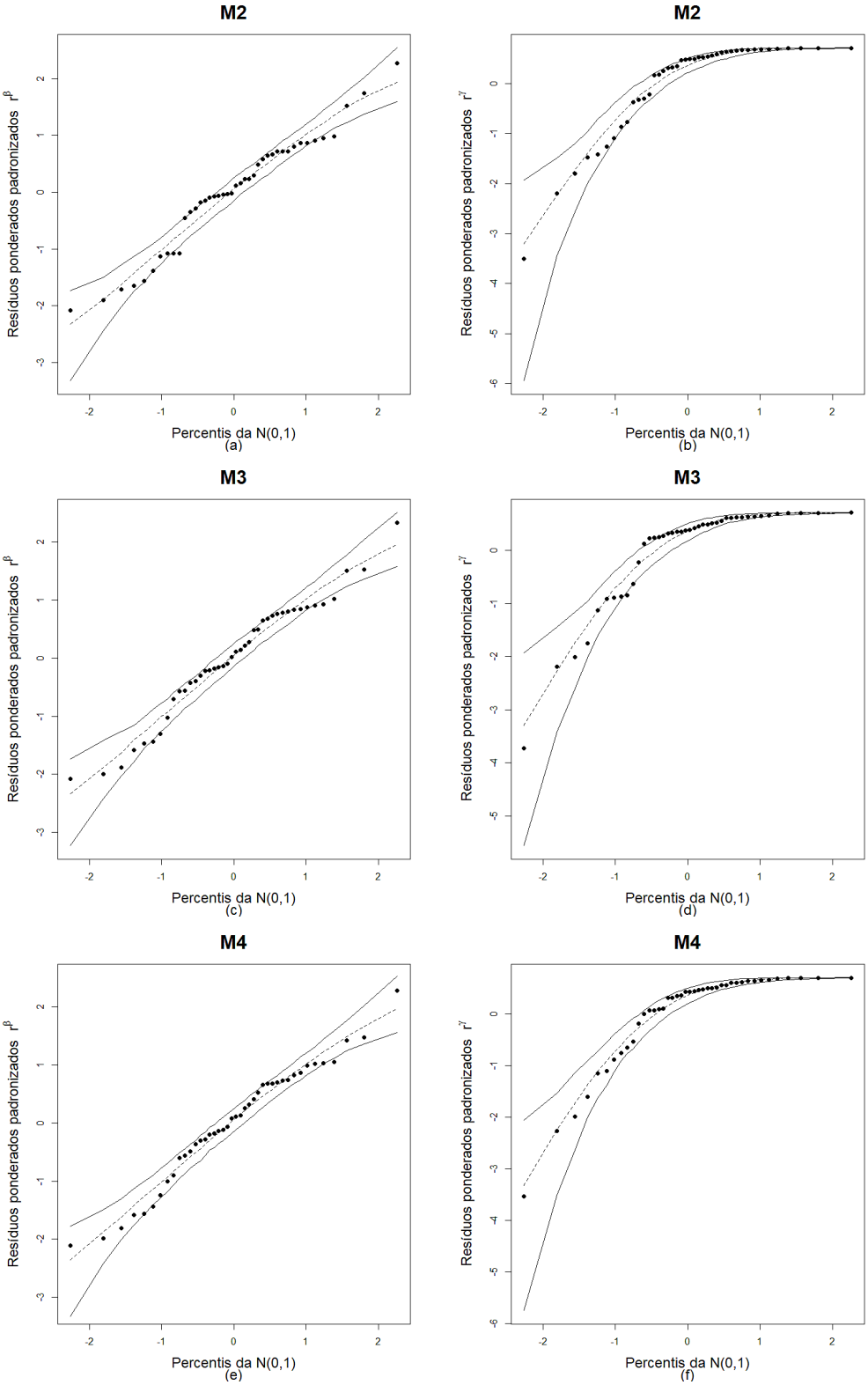
Fonte: Autoria própria.

Figura 30 – Gráficos de envelopes e dos resíduos ponderados padronizados r^β, para o modelo M1: $\log(\mu_t/(1-\mu_t)) = \beta_1 + \beta_2\log(x_{t2})$ e ϕ constante, para os dados do gás.



Fonte: Autoria própria.

Figura 31 – Gráficos de envelopes dos resíduos ponderados padronizados r^{β} e r^{γ}, para os modelos M2: $\log\{-\log(1-\mu_t)\} = \beta_1 + \beta_2 \log(x_{t2})$ e $\log(\phi_t) = \gamma_1 + \gamma_2 \log(x_{t2})$, M3: $\log\{-\log(\mu_t)\} = \beta_1 + \beta_2 \log(x_{t2})$ e $\log(\phi_t) = \gamma_1 + \gamma_2 \log(x_{t2})$ e M4: $\log\{-\log(\mu_t)\} = \beta_1 + \beta_2 \log(x_{t2})$ e $\log(\phi_t) = \gamma_1 + \gamma_2 \sqrt{x_{t2}}$, para os dados do gás.



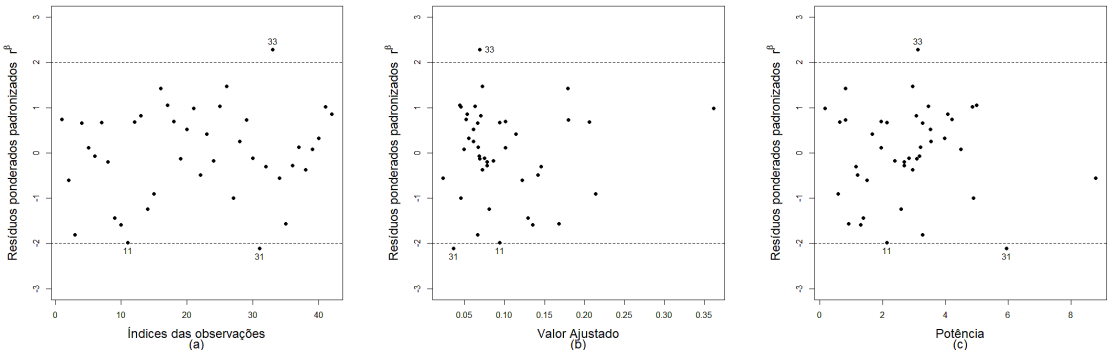
Fonte: Autoria própria.

Na Tabela 29 temos os valores das estatísticas de adequabilidade de ajuste, para cada um dos modelos aqui abordados. Observamos que o modelo M1, proposto inicialmente por Zerbinatti (2008) aparentemente apresenta ajuste razoável quando consideramos apenas os critérios do tipo R^2. Além disso, vemos que esse modelo apresenta o valor de $\hat{\alpha}$ mais próximo de 0.5. Entretanto, os valores dos critérios P^2 e P_c^2 para esse modelo, indicam baixo poder de predição para novas observações, e muito provavelmente, esse trata-se de um típico caso de *overfitting*. Esse problema é captado pela nossa estatística BV, que apresenta valor baixo para esse modelo, sendo igual a 0.5626. Dessa forma, com base na estatística BV, acreditamos que o modelo M1 não é adequado aos dados, e essa conclusão é confirmada quando observamos o gráfico de envelopes simulados com os resíduos ponderados padronizados, apresentado pela Figura 30a, nela vemos que alguns pontos encontram-se fora dos envelopes. Além disso, pela Figura 30b, com o gráfico dos resíduos *versus* os índices, vemos que três observações possuem resíduos mais elevados, fora do intervalo $(-2,2)$. Dessa forma, apesar de $\hat{\alpha} \approx 0.5$, o modelo M1 não está adequado. Isso confirma a ideia de que o diagnóstico de adequabilidade precisa considerar a concordância conjunta dos valores de BV e $\hat{\alpha}$.

Observando os valores das estatísticas na Tabela 29 para os modelos com dispersão variável, vemos que M4 é o modelo que apresenta o melhor desempenho para predizer novas observações. Além disso, esse modelo apresenta valor elevado para BV e $\hat{\alpha} \approx 0.5$, indicando ajuste razoável e aceitável desse modelo. Vale destacar que para esse conjuto de dados, o intervalo da média está localizado próximo de zero, cujos valores das medidas são tipicamente menores, mesmo quando o modelo está corretamente especificado. A partir dos gráficos de envelopes mostrados na Figura 31, para os modelos M2, M3 e M4, vemos pelas Figuras 31e e 31f, que o modelo M4 de fato apresenta melhor adequabilidade, tanto pelo gráfico com os resíduos r^{β}, como pelo gráfico com os resíduos do modelo da precisão, r^{γ}. Ademais, podemos observar que, apesar da Figura 31a não apresentar um ajuste tão bom do modelo M2, indo de acordo com o baixo valor da estatística BV, o gráfico da Figura 31b demonstra desempenho plausível dos resíduos no modelo da precisão, sendo melhor do que para o modelo M3. Isso justifica o fato do valor de $\hat{\alpha}$ para M2 ser mais próximo de 0.5 do que para M3.

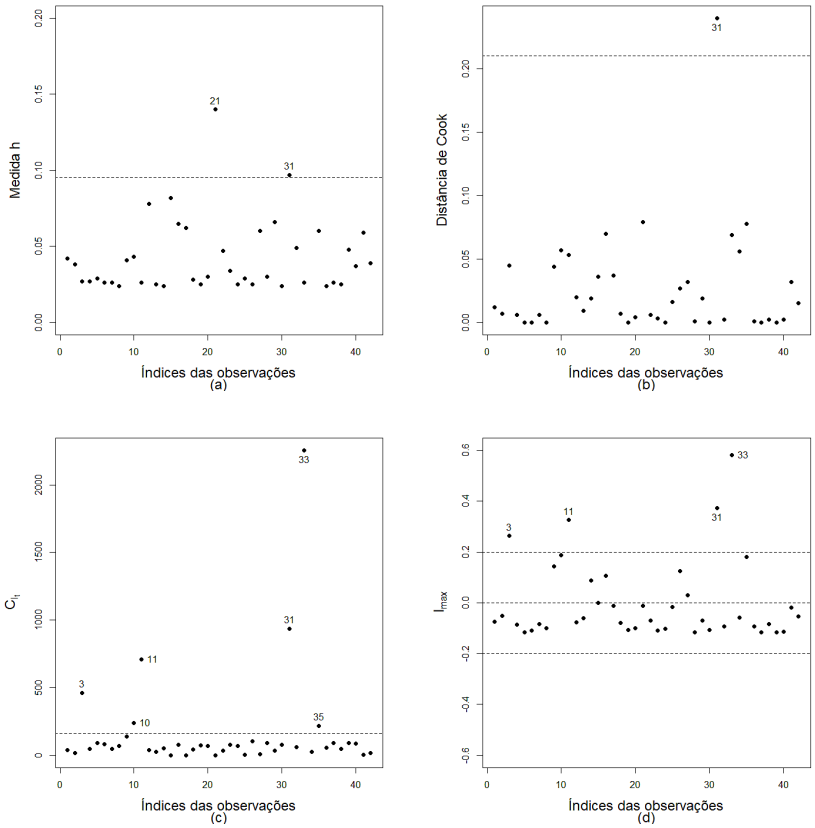
Seguiremos avaliando a adequabilidade do modelo M4 para o conjunto de dados do fator de simultaneidade. Na Figura 32 dispomos dos gráficos dos resíduos ponderados padronizados r^{β} *versus* os índices das observações, os valores ajustados e a covariável potência computada, em que observamos os resíduos distribuídos aleatoriamente dentro do intervalo $(-2,2)$, com exceção das observações 11, 31 e 33, que possuem resíduos mais elevados. Na Figura 33 temos mais alguns gráficos de diagnósticos, utilizados para detectar possíveis pontos influentes, os gráficos são referentes às medidas de alavancagem, distância de Cook e medidas de influência local. Então, de acordo com a Figura 33a vemos que as observações 21 e 31 apontaram altos graus de alavancagem, porém, pela Figura 33b, apenas o ponto 31 foi indicada como provavelmente influente. Entretanto, a partir dos gráficos de influência local, apresentados pelas Figura 33c e 33d, verificamos que pelas medidas de curvaturas normais, C_{I_t}, os pontos 3, 10, 11, 31, 33 e 35 são indicados como possíveis pontos influentes, e pelos vetores I_{max}, observamos que os pontos 3, 11, 31 e 33 são indicados como sendo conjuntamente influentes.

Figura 32 – Gráficos dos resíduos ponderados padronizados r^{β}, para o modelo M4: $\log\{-\log(\mu_t)\} = \beta_1 + \beta_2 \log(x_{t2})$ e $\log(\phi_t) = \gamma_1 + \gamma_2\sqrt{x_{t2}}$, para os dados do gás.



Fonte: Autoria própria.

Figura 33 – Gráficos dos possíveis pontos de alavanca e influentes, para o modelo M4: $\log\{-\log(\mu_t)\} = \beta_1 + \beta_2 \log(x_{t2})$ e $\log(\phi_t) = \gamma_1 + \gamma_2\sqrt{x_{t2}}$, para os dados do gás.



Fonte: Autoria própria.

Na Tabela 30 encontramos as mudanças relativas percentuais das estimativas dos parâmetros e os respectivos p-valores, verificados após a retirada das observações que apresentaram maiores resíduos e graus de alavancagem, e também daquelas indicadas como possívelmente influentes na estimativa dos parâmetros do modelo M4. Podemos então observar que, quando retiradas individualmente ou conjuntamente, não existem mudanças relativas significativas nas estimativas dos parâmetros. Ademais, os p-valores obtidos mantiveram a significância dos

parâmetros e consequentemente da covariável $\log(x_{t2})$ no submodelo da média e de $\sqrt{x_{t2}}$ no submodelo da precisão. Logo, as observações avaliadas individualmente ou conjuntamente não exercem influência demasiada nas estimativas dos parâmetros. Portanto, podemos concluir que o modelo M4 tem uma boa adequabilidade aos dados.

Tabela 30 – Mudanças relativas (M. R.) em porcentagens das estimativas dos parâmetros e respectivos p-valores, verificados com a retirada de observações, para os dados do gás, com o modelo M4: $\log\{-\log(\mu_t)\} = \beta_1 + \beta_2 \log(x_{t2})$ e $\log(\phi_t) = \gamma_1 + \gamma_2\sqrt{x_{t2}}$.

Pontos Retirados	Medidas	β_1	β_2	γ_1	γ_2
3	M. R. (%)	0.7904	0.1204	0.2887	2.8836
	p-valor	0.0000	0.0000	0.0004	0.0067
10	M. R. (%)	2.0583	2.3774	4.3498	4.3124
	p-valor	0.0000	0.0000	0.0003	0.0124
11	M. R. (%)	1.6302	1.3542	3.8687	2.2082
	p-valor	0.0000	0.0000	0.0002	0.0099
21	M. R. (%)	3.6061	5.4770	6.8443	8.9539
	p-valor	0.0000	0.0000	0.0005	0.0263
31	M. R. (%)	0.9386	3.6714	13.2651	23.4586
	p-valor	0.0000	0.0000	0.0025	0.0016
33	M. R. (%)	1.4984	0.6019	0.2156	9.0232
	p-valor	0.0000	0.0000	0.0004	0.0040
35	M. R. (%)	2.4535	3.1297	6.0128	6.6072
	p-valor	0.0000	0.0000	0.0003	0.0157
3, 11, 31 e 33	M. R. (%)	0.7410	0.7523	12.2373	42.3623
	p-valor	0.0000	0.0000	0.0023	0.0003

Fonte: Autoria própria.

7 CONSIDERAÇÕES FINAIS

7.1 CONCLUSÕES

Quando trabalhamos com o modelo linear beta com dispersão variável, vemos que para situações em que os modelos estão corretamente especificados, as estatísticas usuais detectam mais facilmente esses bons ajustes, quando os graus de heteroscedasticidade são baixos e os valores da média da resposta encontram-se em torno de 0.5. Desse modo, vemos que consequentemente, BV também apresenta bons desempenhos nesses cenários, apresentando valores mais elevados. Além disso, concluímos por meio dos estudos de simulação que para um modelo linear beta bem ajustado, com habilidades similares e elevadas em explicar bem a variabiliade da resposta e predizer bons valores dessa variável, espera-se que o valor de BV seja próximo de um e $\hat{\alpha} \approx 0.5$, porém, podendo haver desvios de aproximadamente 0.2 nos valores de $\hat{\alpha}$, quando o tamanho amostral é pequeno.

Ainda para os resultados sobre o modelo beta linear com dispersão variável, temos que quando o modelo ajustado omite uma determinada covariável no modelo da média, de modo que tal covariável segue uma distribuição muito diferente das demais, observamos drásticas mudanças nos valores médios das estatísticas usuais, da BV e também de $\hat{\alpha}$, principalmente quando o tamanho amostral cresce, indicando fortemente, que o modelo não está adequado. Entretanto, quando a distribuição da covariável omitida é similar às demais covariáveis presentes no modelo, as estatística P^2, R^2 e BV apresentam dificuldades em assimilar a falta dessa variável. Pudemos verificar que independentemente da distribuição da covariável omitida, as variâncias e precisões são mal estimadas, especialmente quando n cresce. Logo, em todas essas situações, o modelo ajustado apresenta problemas, devido à falta de uma covariável. Observamos então que a medida $\hat{\alpha}$ consegue captar a má especificação do modelo, apresentando em média valores consideravelmente mais distantes de 0.5. Prosseguindo, quando ajustado um modelo assumindo precisão constante, sob dados gerados a partir de um modelo com média e precisão variáveis, temos que as estatísticas P^2, R^2 e BV indicam a falta do modelo para a precisão quando o grau de heteroscedasticidade é bastante elevado, caso contrário, essas medidas apresentam dificuldades em assimilar o erro na especificação do modelo. Entretanto, o desempenho de $\hat{\alpha}$ novamente consegue se sobresair às demais estatísticas, apresentando sempre valores distantes de 0.5, indicando má adequabilidade dos modelos.

Para o modelo beta não linear com dispersão variável, quando ajustado corretamente tal modelo, vemos resultados similares ao caso linear, em que as estatísticas tradiocinais e também BV apresentam mais facilidade em assimilar a adequabilidade do modelo, quando μ é central. Entretanto, vale destacar que para os cenários em que μ é próximo de zero ou de um, e o graus de heteroscedasticidade são elevados, os coeficientes R^2_{FC}, $R^2_{FC_c}$, R^2_{LR} e $R^2_{LR_c}$ não conseguem assimilar bem a adequabilidade do ajuste, porém P^2 e P^2_c apresentam ainda bons valores. Logo, nesses cenários, os valores para BV são medianos, e aumentam de acordo com o tamanho amostral. Quanto à análise de $\hat{\alpha}$, vemos que seus valores médios encontram-se sempre próximos de 0.5, principalmente quando o tamanho da amostra aumenta, conforme o esperado. Além

disso, identificamos nesse estudo com o modelo beta não linear, que possivelmente, a nossa medida $\hat{\alpha}$ consegue fornecer informações a respeito dos vieses de $\hat{\phi}_t$, com $t = 1, \ldots, n$, de modo que, quanto melhor forem as estimativas das precisões ϕ_t, mais próximo de 0.5 será $\hat{\alpha}$.

Adicionalmente, foi feito um estudo sobre a eficácia de BV e $\hat{\alpha}$ e também das demais estatísticas, em detectar a omissão do fator de não linearidade existente no modelo da média para um tamanho amostral consideravelmente grande. Observamos então um mal desempenho das medidas P^2, R^2 e BV (quando avaliado sozinho), de modo que nenhuma delas pôde captar a omissão da não linearidade na média. Entretanto, $\hat{\alpha}$ nos apresentou um ótimo desempenho, cuja média dos valores das réplicas de Monte Carlo para essa medida, está bastante distante do esperado em um bom modelo.

Por fim, foram apresentados os estudos de simulação para o modelo linear beta com dispersão fixa, considerando modelos corretamente especificados. Pudemos ver que as estatísticas de qualidade de ajuste e de predição possuem melhores habilidades em indicar a adequabilidade do modelo, quando os valores verdadeiros da média estão centralizados e o parâmetro de precisão é elevado, consequentemente, BV também apresenta bons resultados sob esses cenários. Um ponto importante, observado nos modelos com precisão constante, é que para amostras pequenas, geralmente, $\hat{\alpha}$ apresenta valores consideravelmente menores que 0.5, mesmo com o modelo bem especificado e ϕ elevado, onde as demais medidas de adequabilidade são próximas de um. Esse fato ocorre devido o viés da estimativa para ϕ ser bastante alto, e como $\hat{\alpha}$ reflete a qualidade da estimativa da precisão, os seus valores tendem se aproximar de 0.5 quando o tamanho amostral cresce, pois nesse caso, os vieses de $\hat{\phi}$ tonam-se menores. Portanto, para o modelo linear beta com dispersão fixa é indicado considerar amostras maiores.

Durante as aplicações pudemos confirmar com dados reais nossas conclusões a respeito de BV e de $\hat{\alpha}$. Assim, podemos dizer que a medida BV junto com $\hat{\alpha}$, conseguem auxiliar bem nas escolhas de modelos de regressão beta, de modo que, embora BV possa ser considerada uma estatística rigorosa, cujo valor na maioria das vezes é obtido atribuíndo sempre pesos maiores ao mínimo, BV e $\hat{\alpha}$ conseguem conjuntamente detectar ambas as qualidades aqui investigadas em um modelo beta, que são: explicar bem a variabilidade da resposta e predizer bons valores de tal variável. Além disso, quando $BV \approx 1$, $\hat{\alpha}$ torna-se uma ótima ferramenta para investigar a qualidade das estimativas da precisão.

Finalmente, após os estudos de BV e $\hat{\alpha}$, propomos um método em que inicialmente o modelo seja selecionado com base em $\hat{\alpha}$. Modelos em que $\hat{\alpha} \approx 0.5$ devem ser considerados bem ajustados, e quando $BV \approx 1$ este modelo selecionado apresenta um equilíbrio entre predição e variabilidade.

7.2 TRABALHOS FUTUROS

Serão possíveis objetivos de nossas próximas pesquisas:

- Avaliar o desempenho de BV e $\hat{\alpha}$ quando omitimos o fator de não linearidade da média para diferentes tamanhos amostrais.

- Realizar avaliações numéricas no modelo com dispersão variável e amostras maiores.

- Avaliar o desempenho de $\hat{\alpha}$ quando realizamos correções de viés no estimador da precisão.

REFERÊNCIAS

AKAIKE, H. **Information theory and an extension of the maximum likelihood principle**. In PB N, C F (eds.), Proc. of the 2nd Int. Symp. on Information Theory, pp. 267–281, 1973.

ALLEN, D. M. **The prediction sum of squares as a criterion for selecting predictor variables**. University of Kentucky, Department of Statistics Technical Report, 23, 1971.

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

ATKINSON, A. C. **Two graphical display for outlying and influential observations in regression**. Biometrika, 68, 13-20, 1981.

BARNDORFF–NIELSEN, O. E.; JØRGENSEN, B. **Some parametric models on the simplex**. Journal of Multivariate Analysis, 39, 106–116, 1991.

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

BELSLEY, D. A.; EDWIN, K.; ROY, E. W. **Regression Diagnostics: identifying influential data and sources of collinearity**. New York, John Wiley and Sons, 1980.

BERAN, R. **Prepivoting test statistics: a bootstrap view of asymptotic refinements**. Journal of the American Statistical Association, 83(403), 687-697, 1988.

BÍBLIA. Isaías. Português. Bíblia Sagrada. Almeida Revista e Corrigida, 4ª edição. São Paulo: Sociedade Bíblica do Brasil, 2009. p. 717.

CARRASCO, J. M. F.; FERRARI, S. L. P.; ARELLANO-VALLE, R. B. **Errors-invariables beta regression models**. Journal of Applied Statistics, 41, 1530-1547, 2014.

COOK, R. D. **Detection of influential observations in linear regressions**. Technometrics, 19, 15-18, 1977.

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

COX, D.; SNELL, E. **A general definition of residuals**. Journal of the Royal Statistical Society B, 30, 248-275, 1968.

DOORNIK, J.A. **Ox: an Object-oriented matriz Programming Language**, 4th ed. London: Timberlake Consultants and Oxford, 2001. Disponível em <www.doornik.com>. Acesso em: 11 mar.2015.

DRAPER, N. R.; SMITH, H. **Applied Regression Analysis**. 2nd ed. New York. Wiley, 1981.

EFRON, B. **Bootstrap methods: another look at the jackknife**. Annals of Statistics, 7, 1-26, 1979.

EFRON, B. **Estimating the error rate of a prediction rule: improvement on cross-validation**. Journal of the American Statistical Association, 78(382), 316-331, 1983.

EFRON, B.; TIBSHIRANI, R.J. **An Introduction to the Bootstrap**. New York. Chapman and Hall, 1993.

ESPINHEIRA, P. L. **Regressão beta**. Tese de Doutorado em Estatística. Istituto de Matemática e Estatística. Universidade de São Paulo, 2007.

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

ESPINHEIRA, P. L.; FERRARI, S. L. P.; CRIBARI-NETO, F. **Bootstrap prediction intervals in beta regressions**. Computational Statistics, 29(5), 1263-1277, 2014.

ESPINHEIRA, P. L. et al. **Model Selection Criteria on Beta Regression for Machine Learning**. Machine Learning and Knowledge Extraction, 427–449, 2019.

FERRARRI, S. L. P.; CRIBARI-NETO, F. **Beta regression for modelling rates and proportions**. Journal of Applied Statistics, 31, 799–815, 2004.

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

GEISSER, S.; EDDY, W. F. **A predictive approach to model selection**. Journal of the American Statistical Association, 74, 153-60, 1979.

HALL, P.; MARTIN, M. A. **On bootstrap resampling and iteration**. Biometrika, 75(4), 661-671, 1988.

HOAGLIN, D. C.; WELSCH, R. E. **The hat matrix in regression and ANOVA**. The American Statistician, 32, 17-22, 1978.

HOCKING, R. R. **Criteria for selection of a subset regression: which one should be used**. Technometrics, 14, 967-970, 1972.

KNUTH, D. **The TEXbook**. New York, Adisson-Wesley, 1986.

LAMPORT, L. **A Document Preparation System**. 2nd ed. Massachusetts, Addison-Wesley, 1994.

LEMONTE, A. J.; BAZAN J. L. **A new class of Johnson S_B distribution and its associated regression model for rates and proportions**. Biometrical Journal, 58, 727-746, 2015.

LIMA, F. P. **Inferência Bootstrap em Modelos de Regressão Beta**. Tese de Doutorado, Universidade Federal de Pernambuco, 2017.

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

MEDIAVILLA, F.; SHAH, V. A. **A comparison of the coeffcient of predictive power, the coeffcient of determination and AIC for linear regression**. In: JE, K. (Ed.), Decision Sciences Institute, Atlanta, pp. 1261–1266, 2008.

MITNIK, P. A.; BAEK, S. **The Kumaraswamy distribution: Median-dispersion reparameterizations for regression modeling and simulation-based estimation**. Statistical Papers, 54, 177-192, 2013.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to Linear Regression Analysis** . 5nd ed, New York, John Wiley and Sons, 2012.

MOUSA, A. M.; EL-SHEIKN, A. A.; ABDEL-FATTAH, M. A. **A gamma regression for bounded continuous variables**. Advances and Applications in Statistics, 49, 305-326, 2016.

NAGELKERKE, N. J. D. **A note on a general definition of the coefficient of determination**. Biometrika, 78(3), 691–692, 1991.

NELDER, J. A.; WEDDERBURN, R. W. M. **Generalized linear models**. Journal of the Royal Statistical Society A, 135, 370-384, 1972.

OSPINA, R.; CRIBARI-NETO, F.; VASCONCELLOS, K. L. P. **Improved point and interval estimation for a beta regression model**. Computational Statistics and Data Analysis, 51(2), 960–981, 2006.

OSPINA, R.; FERRARI, S. L. P. **A general class of zero-or-one inflated beta regression models**. Computational Statistics and Data Analysis, 56, 1609-1623, 2012.

PAOLINO, P. **Maximum likelihood estimation of models with beta-distributed dependent variables**. Political Analysis, 9, 325-346, 2001.

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

PEREIRA, G. H. A.; BOTTER, D.A.; SANDOVAL, M.C. **The truncated inflated beta distribution**. Communications in Statistics - Theory and Methods, 41, 907-919, 2012.

PERTERNELLI, L. A.; MELLO, M. P. **Conhecendo o R :** uma visão estatística. Viçosa, Editora UFV, 2011.

PREGIBON, D. **Logistic regression diagnostics**. Annals of Statistics 9, 705-724, 1981.

QUAN T. N. **The Prediction Sum of Squares as a General Measure for Regression Diagnostics**. Journal of Business & Economic Statistics, 6, 501-04, 1988.

ROCHA, S. S.; ESPINHEIRA, P. L.; CRIBARI-NETO, F. **Residual and local influence analyses for unit gamma regressions**. Statistica Neerlandica, v.1, p. stan.12229, 2020.

SALAZAR, S. M. G. **Contribución as estudio de la reacción de descomposición de la zeolita "Y" en presencia de vapor de agua y vanadio**. Trabalho de Conclusão de Curso, Universidad Nacional de Colombia, Bogotá, 2005.

SCHWARZ, G. **Estimating the dimension of a model**. The Annals of Statistics, 6(2), 461–464, 1978.

SIMAS, A. B.; BARRETO-SOUZA, W.; ROCHA, A. V. **Improved estimators for a general class of beta regression models**. Computational Statistics and Data Analysis, 54(2), 348–366, 2010.

SMITHSON, M.; VERKUILEN, J. **A better lemon-squeezer? Maximum likelihood regression with beta-distribuited dependent variables**. Psychological Methods, 11, 54-71, 2006.

ZERBINATTI, L. F. M. **Predição de fator de simultaneidade através de modelos de regressão para proporções contínuas**. Dissertação de Mestrado, Universidade de São Paulo, 2008.