**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Modelo Rathie-Swamee: aplicações e extensão para
modelo de regressão**

Eduardo Monteiro de Castro Gomes

Tese apresentada para obtenção do título de Doutor em
Ciências. Área de concentração: Estatística e Experi-
mentação Agronômica

**Piracicaba
2013**

Eduardo Monteiro de Castro Gomes
Bacharel em Estatística

**Modelos Rathie-Swamee: Aplicações e extensão para
modelo de regressão**

Orientadora:

Prof^a Dr^a **CLARICE GARCIA BORGES DEMÉTRIO**

Tese apresentada para obtenção do título de Doutor em Ciências.
Área de concentração: Estatística e Experimentação Agronômica

**Piracicaba
2013**

**Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - ESALQ/USP**

Gomes, Eduardo Monteiro de Castro

Modelo Rathie-Swamee: aplicações e extensão para modelo de regressão /
Eduardo Monteiro de Castro Gomes.- - Piracicaba, 2013.
88 p: il.

Tese (Doutorado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2013.

1. Modelos logísticos 2. Modelos Rathie-Swamee 3. Modelos de regressão
4. Produto de variáveis aleatórias 5. Quociente de variáveis aleatórias 6. Simulação
Monte Carlo I. Título

CDD 519.536
G633m

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor"

AGRADECIMENTOS

Aos meus pais, Cíntia Monteiro de Castro e Vicente Gomes Neto por todo apoio e incentivo.

A minha avó Gina por todos seus ensinamentos e carinho.

A minha mulher e companheira Juliana por completar minha vida.

A Ana Maria e meus irmãos Mariana e Rodrigo por todo o carinho e apoio.

A professora Dra. Clarice Garcia Borges Demétrio por todo seu apoio e orientação.

Aos professores e funcionários do departamento de Ciências Exatas da ESALQ/USP.

Aos colegas do departamento de Estatística da Universidade de Brasília.

Aos amigos e colegas dos cursos de mestrado e doutorado do Departamento de Ciências Exatas da ESALQ/USP, em especial a Cássio Dessoti, Guilherme Biz, Braga Júnior, Epaminondas Couto, Fernanda Rizzato, Elizabeth Hashimoto, Marcelino Páscoa, Mariana Urbano, Pedro Henrique, Raphael Dias, Renato Pereira, Rodrigo Pescim, Bruno Henrique, Jalmar Carrasco, Simone Sartório, Ricardo Olinda e Tiago de Oliveira.

A todas as pessoas que contribuíram direta ou indiretamente para a realização deste trabalho.

SUMÁRIO

RESUMO	7
ABSTRACT	9
1 INTRODUÇÃO	11
2 REVISÃO BIBLIOGRÁFICA	15
3 METODOLOGIA	19
3.1 Modelo Rathie-Swamee	19
3.2 Modelo Rathie-Swamee Assimétrico	27
3.3 Produto e quociente de variáveis Rathie-Swamee	27
3.4 Modelo de regressão Rathie-Swamee	32
3.5 Modelos de misturas	33
3.6 Inferências	33
3.6.1 Método de Máxima verossimilhança	34
3.6.2 Método de barreiras	35
3.6.3 Intervalos de confiança	36
3.6.4 Testes de Hipóteses	40
3.6.5 Critério de informação de Akaike	40
3.7 Análise de resíduos	42
3.8 Influência Global	42
3.9 Influência Local	43
4 ESTUDOS DE SIMULAÇÃO E APLICAÇÕES	47
4.1 Estudo de simulação I	47
4.2 Estudo de simulação II	50
4.3 Análise dos dados de pesca de camarões na costa do Rio Grande do Norte	51
4.4 Análise dos dados de erupção de gêiser	62
5 CONSIDERAÇÕES FINAIS	71
REFERÊNCIAS	75
APÊNDICES	77

RESUMO

Modelo Rathie-Swamee: Aplicações e extensão para modelo de regressão

Neste trabalho são apresentadas aplicações estatísticas e extensões dos modelos Rathie-Swamee. Propostos em Rathie e Swamee (2006), os modelos Rathie-Swamee foram desenvolvidos a partir de uma generalização da distribuição logística. Esses modelos apresentam grande flexibilidade, assumindo formas unimodais e multimodais, e têm algumas aplicações exemplificadas neste trabalho com dados bimodais de pesca de camarões e de erupções de gêisers. Por meio de simulações desses modelos, são avaliados os desempenhos de diferentes métodos para obtenção de intervalos de confiança para os parâmetros dos modelos e dos estimadores de máxima verossimilhança. As extensões apresentadas para os modelos Rathie-Swamee são referentes à incorporação de covariáveis nos modelos, obtendo-se modelos de regressão. Esses novos modelos de regressão são utilizados para ajuste aos dados de pesca e de erupções, para exemplificar algumas aplicações dos modelos. Uma nova distribuição de probabilidades é apresentada como distribuição resultante de produtos e quocientes entre variáveis aleatórias independentes com distribuições Rathie-Swamee. Para essa nova distribuição é apresentada uma tabela com alguns quantis de interesse para diferentes valores do parâmetro, assim como os resultados de estimação por máxima verossimilhança obtidos para as simulações realizadas com diferentes valores para o parâmetro e tamanhos de amostra.

Palavras-chave: Modelos Rathie-Swamee; Modelos logísticos; Modelos de regressão; Simulação Monte Carlo; Produto de Variáveis aleatórias; Quociente de variáveis aleatórias

ABSTRACT

Rathie-Swamee Model: Applications and extension for regression models

Applications and extensions to the Rathie-Swamee models are presented in this work. Proposed by Rathie and Swamee (2006), the Rathie-Swamee models were developed as a generalization to the logistic distribution. These models have great flexibility, assuming unimodal and multimodal shapes, and have some of its applications exemplified with bimodal data of shrimp fishing and geyser eruptions. By the use of simulations, the performance of different methods to obtain confidence intervals are compared. The extensions presented for the Rathie-Swamee models refer to the inclusion of covariates, creating regression models. These new regression models are fitted to fishing and eruption data, to exemplify some applications of the models. A new probability distribution is presented as the resulting distribution of quotients and products between independent random variables with Rathie-Swamee distributions. For this new distribution are presented some simulation results along with a table of quantiles for some percentage points of interest.

Keywords: Rathie-Swamee models; Logistic models; Regression models; Monte Carlo Simulation; Products of random variables; Quotients of random variables

1 INTRODUÇÃO

Modelos estatísticos permitem descrever de forma simplificada o comportamento de variáveis aleatórias por meio de equações matemáticas. Pode-se representar a distribuição de probabilidades de uma variável utilizando modelos paramétricos de probabilidades, assim como descrever a relação entre duas ou mais variáveis por meio de modelo de regressão.

Por muito tempo, o modelo normal ou Gaussiano foi predominante nas aplicações de modelos estatísticos para análise de dados. Mesmo quando esse modelo não se mostrava adequado para representar o fenômeno em estudo, transformações nos dados eram consideradas para que o modelo normal pudesse ser utilizado (Atkinson, 1985).

Com a evolução do estudo de Estatística e suas aplicações nas mais diversas áreas do conhecimento, novos modelos são propostos com o objetivo de descrever de forma mais adequada o comportamento de diferentes tipos de variáveis aleatórias. Uma infinidade de diferentes tipos de modelos podem ser propostos, mas a qualidade ou utilidade de um modelo deve ser avaliada por sua capacidade de representar de forma adequada o comportamento de uma variável aleatória sob determinadas condições experimentais ou observacionais.

Os temas aqui abordados consideram algumas aplicações estatísticas do modelo Rathie-Swamee (Rathie e Swamee, 2006). Esse modelo foi proposto como uma generalização do modelo logístico e apresenta algumas características interessantes. Em sua versão mais simples, o modelo tem três parâmetros que proporcionam uma grande flexibilidade, possibilitando assumir formas simétricas com uma, duas ou três modas. Para determinados valores dos parâmetros desse modelo, algumas distribuições clássicas como a normal e a t-student podem ser aproximadas com grande precisão.

O modelo Rathie-Swamee, também, é apresentado como um modelo de locação e escala, de forma que a inclusão de parâmetros adicionais de locação e escala permitem um ganho na flexibilidade do modelo, podendo controlar a média e alterar a variabilidade descrita pelo modelo mais simples.

Neste trabalho, é apresentada uma proposta de extensão para o modelo Rathie-Swamee, com o intuito de incluir o efeito de variáveis explicativas ou regressoras sobre a média da variável em estudo. Nesta proposta, o parâmetro de locação é substituído por um vetor de parâmetros que descrevem o efeito das variáveis regressoras sobre a média

da variável resposta.

Além da proposta de extensão para modelo de regressão, aqui são apresentados resultados referentes às distribuições de quocientes e produtos entre variáveis aleatórias independentes com distribuição Rathie-Swamee. Uma nova distribuição é apresentada como caso geral das distribuições desses produtos e quocientes. Para essa nova distribuição, são apresentadas algumas propriedades matemáticas e resultados de estudo de simulação com aplicações estatísticas para o novo modelo.

A teoria de verossimilhança permite a obtenção de estimativas e a realização de inferências para os parâmetros do modelo a partir da observação de amostra aleatória da população considerada (Pawitan, 2001). Regiões e intervalos de confiança podem ser construídos a partir de aproximações da função de verossimilhança ou de estatísticas baseadas na verossimilhança.

A implementação desses métodos de estimação e inferência demandam métodos numéricos de otimização de funções e solução de sistemas não lineares de equações. O ambiente R de programação (R DEVELOPMENT CORE TEAM 2012) tem vasta biblioteca de algoritmos implementados para aplicações em estatística.

A validação das inferências obtidas para os modelos de regressão Rathie-Swamee pode ser realizada por uma análise de diagnóstico. Análise de resíduos e influência são algumas formas para avaliar a adequabilidade dos modelos utilizados em cada aplicação.

Os modelos e métodos aqui abordados têm suas aplicações exemplificadas com um conjunto de dados de pesca de camarões na costa do Rio Grande do Norte e um conjunto de dados de erupções do gêiser *Old faithful*, situado na Califórnia.

A pesca e o cultivo de peixes, crustáceos e moluscos são atividades de grande importância para a produção de alimentos. O elevado potencial de produção e a qualidade nutricional dos produtos gerados destacam-se como atrativos para o interesse comercial dessas atividades. A utilização de métodos e modelos estatísticos podem auxiliar nos processos de gerenciamento e regulação do cultivo e da pesca.

A espécie de camarões *Farfantepenaeus brasiliensis*, popularmente conhecida como camarão-rosa, é um dos recursos pesqueiros mais frequentes e explorados na costa brasileira. Essa espécie de camarões tem ampla distribuição geográfica, tendo sua presença sido registrada da costa da Carolina do Norte (EUA) até a costa do Rio Grande

do Sul (Flor, 2009). O aumento na produção de camarão, decorrente do maior número de embarcações envolvidas e ampliação do poder de pesca desses barcos, cria uma tensão nos estoques pesqueiros com sensíveis impactos ambientais.

A Instrução Normativa número 5, publicada no diário oficial da União do dia 21 de Maio de 2004, classificou a espécie de camarões *Farfantepenaeus brasiliensis* como sobre-explotadas. Essa instrução normativa determinou a elaboração de um plano nacional de gestão para o uso sustentável de camarões marinhos no Brasil. O plano proposto pelo Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis (Dias Neto, 2011) apresenta, de forma detalhada, tanto o marco legal e institucional referente à gestão e propriedade dos recursos pesqueiros e ambientais quanto os aspectos biológicos dos camarões considerados.

A preservação dos estoques selvagens desses camarões é fundamental ainda que se intensifique a criação em cativeiro. Flor (2009) aponta que o cultivo de várias espécies importantes ainda depende da coleta de reprodutores ou de sementes de populações naturais.

Modelos de regressão podem ser utilizados em estudos desenvolvidos para criação de camarões, estudos de caracterização das populações existentes, assim como para estudos ambientais ou climatológicos.

O parque nacional de Yellowstone, nos Estados Unidos da América, concentra em sua área a maior parte dos gêiseres existentes no mundo. Os gêiseres são fontes termais de origem vulcânica que lançam, intermitentemente, água e vapor em forma de esguicho que podem ultrapassar 60 metros de altura (MICHAELIS 2012). A observação das erupções dos gêiseres é um dos grandes atrativos turísticos do parque.

O gêiser *Old faithful* é o mais famoso e visitado dentre os demais. Sua atratividade é decorrente da grande frequência de erupções e regularidade nessa ocorrência. Apontado como o gêiser mais estudado do parque, suas erupções são continuamente registradas e, no site da "Associação de Estudo e Observação de Gêiser, estão disponíveis conjuntos de dados com registros das erupções desde 1970 até os dias atuais.

Um acompanhamento tão detalhado da atividade desse gêiser permite a criação de modelos para descrever e prever as ocorrências de erupções. A todo tempo, o serviço do parque de Yellowstone informa aos visitantes um horário predito para a próxima erupção do *Old faithful*.

Na Figura 1, pode-se ver um grande grupo de visitantes assistindo a uma erupção. Os visitantes se reúnem junto ao gêiser, conforme o horário predito para a erupção que se aproxima. A regularidade do gêiser e as previsões de horário de erupções apresentam-se como grandes atrativos turísticos, e essa demanda por modelos preditivos impulsiona o desenvolvimento dos modelos e técnicas estatísticas.

Figura 1 - Imagem de erupção do gêiser Old faithful 13/09/2012. Fonte: site do serviço nacional de parques dos EUA

Este trabalho está organizado de forma que na seção 2 é apresentada uma revisão da literatura sobre os modelos e métodos utilizados, e, na seção 3, esses modelos e métodos são detalhados. Exemplos de aplicações dos modelos em estudos de simulação e análises de dados reais são apresentados na seção 4, enquanto a seção 5 contém comentários sobre os resultados obtidos e propostas para pesquisas futuras.

2 REVISÃO BIBLIOGRÁFICA

Nesta seção, é apresentado um breve retrospecto da literatura sobre os modelos considerados e as metodologias utilizadas.

Alguns registros sobre a função logística a partir de 1838 indicam sua utilização inicial no estudo de curvas de crescimento. Uma posterior popularização do modelo ocorreu com a aplicação deles para diferentes problemas em estudos de agronomia, economia e diversas outras áreas.

Em Johnson et al. (1995), um dos capítulos é dedicado à distribuição de probabilidades logística, detalhando sua evolução histórica e apresentando os modelos e propriedades matemáticas. A distribuição é formulada considerando diferentes formas de parametrizações, e são apresentados resultados como momentos, estatísticas de ordem e métodos de inferência. Dentre as muitas generalizações da distribuição logística existentes, os autores apresentam funções de densidade e distribuição de probabilidades e alguns resultados para quatro tipos de generalizações da distribuição logística.

Balakrishnan (1991) apresenta e discute de forma bastante ampla os modelos logísticos, mostrando importantes resultados matemáticos para as distribuições logísticas e considerando diferentes aplicações e generalizações dessas distribuições. As relações entre as distribuições logísticas com outras distribuições são apresentadas e uma forma multivariada de distribuição logística é proposta.

Aplicações com produtos e quocientes entre variáveis aleatórias com distribuições logísticas têm despertado interesse para alguns estudos como pode ser visto em Nadarajah e Gupta (2006) e Nadarajah (2009). Os autores apresentam resultados para as distribuições das variáveis resultantes dos produtos ou quocientes entre variáveis independentes com distribuições logísticas. As funções resultantes, muitas vezes com fórmulas extensas e complicadas, podem ser representadas por meio de funções especiais, com forma mais compacta como pode ser visto em Rathie; Zornig e Gomes (2012).

Em Rathie e Swamee (2006), um outro tipo de generalização da distribuição logística foi proposta. A nova distribuição foi apresentada como uma alternativa inversível para a distribuição normal. O modelo proposto tem boa aproximação para a distribuição normal e tem sua aplicação proposta para simulação de variáveis aleatórias de forma simples, uma vez que sua função de distribuição de probabilidades é inversível.

Outra propriedade interessante dos modelos Rathie-Swamee é sua flexibili-

dade na função de densidade de probabilidade, que pode apresentar formas simétricas com uma, duas ou três modas. Aplicações dos modelos Rathie-Swamee para situações bimodais podem ser vistos em Rathie et. al(2011), no qual conjuntos de dados climatológicos e econômicos são analisados.

Considerando a proposta adotada por Azzalini (1985) para obtenção de distribuições assimétricas, Rathie e Coutinho (2011) propõe um modelo Rathie-Swamee assimétrico. Com o incremento de um parâmetro no modelo permite-se que as formas unimodais ou multimodais dos modelos possam ser assimétricas.

Uma forma alternativa para o ajuste de modelos para dados com comportamento multimodal e assimétrico é a utilização de misturas de distribuições. McLachlan e Peel (2004) apresentam uma extensa revisão de literatura considerando os aspectos teóricos e as aplicações dos modelos de misturas finitas. Dentre as aplicações são citados os conjuntos de dados com estrutura de grupos, que ocasionam ocorrência de duas ou mais modas.

Algumas aplicações computacionais dos modelos de mistura finitas são discutidas em Benaglia et al. (2009). Nesse artigo, os autores apresentam um pacote para o ambiente R de programação (R Development Core Team, 2011), que implementa diferentes métodos de estimação e ajuste de modelos de misturas finitas. Um dos exemplos exibe a aplicação de mistura de duas distribuições normais para ajuste de dados bimodais.

É comum, em estudos científicos, o interesse de relacionar o comportamento de duas ou mais variáveis. A análise de regressão é um método que procura estabelecer relações funcionais entre duas ou mais variáveis mediante a obtenção de um modelo que descreva da melhor forma o comportamento das variáveis em estudo. Os modelos de regressão normais são os modelos de regressão mais difundidos e utilizados, e uma vasta literatura detalha os principais aspectos teóricos e práticos da aplicação dos modelos de regressões normais.

A teoria da verossimilhança fornece diferentes ferramentas para a obtenção de estimativas e realização de inferências para os modelos probabilísticos utilizados. (AZZALINI, 1996; PAWITAN, 2001)

Para a utilização dos métodos baseados na função de verossimilhança, é importante considerar as restrições existentes no espaço paramétrico associado aos modelos Rathie-Swamee. Os métodos numéricos utilizados nas etapas de estimação e inferências

devem ser adaptados para considerar as restrições dos parâmetros do modelo. Zornig (2011) descreve alguns métodos de otimização para funções sujeitas a sistemas de restrições lineares. O método da barreira consiste em adaptar a função de verossimilhança incorporando um termo de barreira na função que impede que o algoritmo de otimização saia do espaço paramétrico associado ao modelo.

Lange (2011, 2010) descreve a utilização de função barreira logarítmica e propõe que o termo de barreira introduzido na função objetivo seja adaptativo e, dessa forma, seja alterado a cada passo do método iterativo de otimização utilizado. O chamado método da barreira adaptativa está implementado no ambiente R de programação estatística.

A formulação de intervalos de confiança para os parâmetros dos modelos Rathie-Swamee deve ser feita considerando as restrições existentes no espaço paramétrico desse modelo. Intervalos simétricos para parâmetros estritamente positivos podem criar intervalos inadequados e imprecisos. Meeker e Escobar (1998) propõem uma transformação nos parâmetros para contornar as restrições e utilizar aproximações quadráticas da superfície de verossimilhança. Regiões de confiança baseadas na função de verossimilhança permitem considerar possíveis relações existentes entre os parâmetros dos modelos (CASELLA; BERGER, 1990).

Quando existem no modelo alguns parâmetros de maior interesse, a função de verossimilhança perfilada permite isolar um parâmetro para realização de inferências isoladas para esse parâmetro, sem desconsiderar as possíveis relações existentes entre os parâmetros do modelo.

Metodologias de diagnóstico devem ser utilizadas para detectar problemas com o modelo ajustado assim como, analisar a sensibilidade das inferências a pequenas perturbações nos dados ou no modelo proposto. Se o modelo ajustado não apresentar uma boa descrição dos dados observados, podem ser obtidas conclusões errôneas.

Devido a esse fato, destaca-se a importância de se realizar um estudo sobre a robustez dos resultados obtidos, considerando-se vários aspectos que envolvem a formulação do modelo e as estimativas dos seus parâmetros.

A análise de sensibilidade é realizada por meio das metodologias de influência global, local e local total. A primeira é caracterizada pela deleção de casos, em que o i -ésimo indivíduo é retirado da análise. Cook (1977) propõe uma medida de in-

fluência conhecida como distância de Cook para avaliar a influência da deleção de um caso nas estimativas dos parâmetros em modelos de regressão linear. A distância de Cook mede a influência de cada observação na análise estatística. Outra medida utilizada em influência global é o afastamento da verossimilhança, que mede o quanto cada observação pode alterar a verossimilhança.

A metodologia de influência local proposta por Cook (1987) tem o objetivo de avaliar a influência conjunta das observações sob pequenas mudanças (perturbações) no modelo. A escolha da forma de perturbar o modelo proposto entre as inúmeras formas existentes, deve levar em consideração aspectos da análise que se deseja monitorar. Obviamente, deve-se pensar em esquemas de perturbação que sejam interpretáveis. Lesaffre e Verbeke (1998) desenvolvem a medida de influência local total.

Alguns autores têm estudado essas medidas. Por exemplo, Xie e Wei (2007) propõem medidas de diagnósticos baseadas na deleção de casos para o modelo de regressão log-Birnbaum-Saunders; Escobar e Meeker (1992) adaptaram métodos de influência local para os modelos de regressão paramétricos com dados censurados; Ortega; Cancho e Bolfarine (2006) desenvolvem uma análise de influência local no modelo de regressão Weibull - exponenciado; Fachini; Ortega e Louzada-Neto (2008) utilizam influência local nos modelos de riscos múltiplos.

No contexto de máxima verossimilhança com restrição nos parâmetros, Kwan e Fung (1998) utilizam influência local para análise fatorial sujeito a restrições; Gu e Fung (2001) consideram a abordagem de influência local na estrutura da verossimilhança restrita geral e afirmam que a curvatura encontrada para verossimilhança restrita pode ser usada seguindo a mesma abordagem indicada por Cook (1986). Paula e Cysneiros (2009) discutem, em trabalhos mais recentes, a avaliação de influência local e “leverage” sujeito a equações de restrições lineares nos parâmetros com extensão a inequações de restrições para modelos lineares generalizados.

3 METODOLOGIA

Neste capítulo estão apresentados os modelos e as principais metodologias utilizadas nos estudos de simulação e nas aplicações com dados de pesca de camarões e com dados de erupções do geiser Old Faithful.

3.1 Modelo Rathie-Swamee

O ponto de partida é a função de distribuição acumulada de uma variável aleatória Z com distribuição logística em sua forma padrão, aqui denominada $LG(1)$, dada por

$$F(z) = \frac{1}{1 + \exp(-z)}, \quad -\infty < z < \infty, \quad (1)$$

que tem função de densidade simétrica em torno de zero, na forma

$$f(z) = \frac{\exp(-z)}{[1 + \exp(-z)]^2}, \quad -\infty < z < \infty. \quad (2)$$

Pode-se verificar para essa variável aleatória que $E(Z) = 0$ e $V(Z) = \frac{\pi^2}{3}$.

Uma forma comum de generalização da distribuição logística é obtida mediante a inclusão de um parâmetro $b > 0$, na forma

$$F(z) = \frac{1}{1 + \exp(-bz)}, \quad -\infty < z < \infty. \quad (3)$$

O parâmetro b flexibiliza a variabilidade do modelo, sem influenciar na média e na simetria dele. Essa distribuição será denominada $LG(b)$, e é tal que $E(Z) = 0$ e $V(Z) = \frac{\pi^2}{3b^2}$, tendo função de densidade dada por

$$f(z) = \frac{b \exp(-bz)}{[1 + \exp(-bz)]^2}, \quad -\infty < z < \infty. \quad (4)$$

Com base nas equações (1) e (3), Rathie e Swamee (2006) propuseram uma generalização com três parâmetros para a função de distribuição acumulada do modelo logístico dada por:

$$F(z) = \frac{1}{1 + \exp[-z(a + b|z|^p)]}, \quad a, b, p \geq 0. \quad (5)$$

Esse modelo com três parâmetros não negativos, aqui denominado $RS(a, b, p)$, possui grande flexibilidade, podendo ter sua função de densidade assumindo formas unimodais e multimodais. A função de densidade correspondente tem a forma:

$$f(z) = \frac{[a + b(1 + p)|z|^p] \exp[-z(a + b|z|^p)]}{\{\exp[-z(a + b|z|^p)] + 1\}^2}. \quad (6)$$

Vale notar que as equações (1) e (2) representam casos particulares das equações (5) e (6) para algumas configurações paramétricas. Quando o parâmetro p assume valor igual a zero, existem infinitas combinações de valores que os demais parâmetros podem assumir para que as equações (5) e (6) sejam reduzidas aos modelos encaixados, com menor número de parâmetros. É simples perceber as igualdades entre $LG(1) = RS(a = 0, b = 1, p = 0) = RS(a = a^*, b = b^*, p = 0) = LG(1)$, $\forall (a^* + b^*) = 1$. Essas relações existentes entre os parâmetros indicam falta de identificabilidade do modelo (Casella e Berger, 1990).

Definição 1 *Um parâmetro θ para uma família de distribuições $\{f(x|\theta) : \theta \in \Theta\}$ é identificável se valores distintos de θ corresponderem a diferentes funções de densidade ou distribuição de probabilidades, ou seja, se $\theta \neq \theta^*$ então $f(x|\theta)$ não é a mesma função que $f(x|\theta^*)$.*

Nos casos multiparamétricos diferentes vetores de parâmetros implicam em diferentes funções de distribuição de probabilidades dos modelos identificáveis. Nos casos em que isso não ocorre, a falta de identificabilidade causa problemas para realização de inferências, pois diferentes vetores de estimativas para os parâmetros resultam nos mesmos valores da função de verossimilhança.

As Figuras 2, 3 e 4 apresentam algumas formas que a função de densidade do modelo RS pode assumir para diferentes configurações paramétricas. Os gráficos foram elaborados de forma que se possa explorar a contribuição de cada um dos parâmetros sobre a forma da função de densidade de probabilidade. Essas três figuras foram elaboradas com o objetivo de isolar o efeito de cada um dos três parâmetros, e nesse contexto as Figuras 2, 3 e 4 têm a , b e p , respectivamente como parâmetros de interesse. Os parâmetros de interesse são alterados de forma crescente nos gráficos dispostos no sentido da esquerda para a direita, enquanto que os demais parâmetros são mantidos constantes para os gráficos contidos em uma mesma linha de cada figura.

A Figura 2 ilustra, de forma principal, os efeitos que o parâmetro a tem sobre a função de densidade de probabilidade. Pode-se notar o efeito desse parâmetro sobre a unimodalidade ou multimodalidade da distribuição. Quando o parâmetro a se aproxima de zero, a função de densidade se aproxima de zero no ponto $z = \mu = 0$, causando efeito de bimodalidade. Conforme o valor de a aumenta, o valor da função de densidade no ponto $z = \mu = 0$ aumenta até atingir um ponto de máximo, implicando no efeito de unimodalidade da distribuição.

Figura 2 - Gráficos com diferentes formas da função densidade do modelo Rathie-Swamee, mostrando o efeito do parâmetro a

Na Figura 3, pode-se notar o efeito do parâmetro b sobre a contração e dilatação da base da forma da função de densidade. O aumento do valor desse parâmetro implica na contração da base da forma da função de densidade, causando a diminuição da variabilidade da distribuição obtida.

Figura 3 - Gráficos com diferentes formas da função densidade do modelo Rathie-Swamee, mostrando o efeito do parâmetro b

A Figura 4 ilustra os efeitos do parâmetro p sobre o afastamento ou aproximação das possíveis modas existentes nas distribuições obtidas. O incremento do parâmetro p afasta as diferentes modas, nos casos bimodais, podendo resultar em distribuições com até três modas.

Figura 4 - Gráficos com diferentes formas da função densidade do modelo Rathie-Swamee, mostrando o efeito do parâmetro p

Como a distribuição dada pela equação (5) é simétrica, os momentos da distribuição são dados por

$$E(Z^h) = \int_{-\infty}^{\infty} z^h f(z) dz = \begin{cases} 0, & \text{se } h \text{ é um inteiro ímpar} \\ 2 \int_0^{\infty} z^h f(z) dz, & \text{se } h \text{ é um inteiro par.} \end{cases} \quad (7)$$

Para o cálculo dos momentos inteiros pares é útil considerar uma apresentação diferente para a equação (6) na forma

$$f(z) = \sum_{r=0}^{\infty} s_r [a + b(1+p)|z|^p] \exp[-(1+r)z(a + b|z|^p)], \quad (8)$$

sendo $s_r = (-1)^r(1+r)$, e $(1+y)^{-2} = \sum_{r=0}^{\infty} s_r y^r$, e, então,

$$E(Z^h) = 2 \sum_{r=0}^{\infty} s_r [aI_{h,r} + b(1+p)I_{h+p,r}], \quad (9)$$

em que

$$I_{h,r} = \int_0^{\infty} z^h \exp[-(1+r)z(a+bz^p)] dz, \quad (10)$$

e, usando a equação (9) no caso $h = 2$, obtém-se

$$\begin{aligned} \text{Var}(Z) &= 2 \sum_{r=0}^{\infty} s_r \left\{ a[a(1+r)]^{-3} H_{11}^{11} \left[\frac{a^{p+1}(1+r)^p}{b} \middle|_{(3,p+1)}^{(1,1)} \right] \right. \\ &\quad \left. + b(1+p)[a(1+r)]^{-(p+3)} + H_{11}^{11} \left[\frac{a^{p+1}(1+r)^p}{b} \middle|_{(3,p+1)}^{(1,1)} \right] \right\}, \end{aligned} \quad (11)$$

sendo a função H de Fox definida pela seguinte integral de contorno

$$H_{p,q}^{m,n} \left[z \middle| \begin{matrix} (u_1, U_1) & \dots & (u_p, V_p) \\ (v_1, V_1) & \dots & (v_q, V_q) \end{matrix} \right] = \frac{1}{2\pi i} \int_L \frac{(\prod_{j=1}^m \Gamma(v_j + V_j s)) (\prod_{j=1}^n \Gamma(1 - u_j - U_j s))}{(\prod_{j=m+1}^q \Gamma(1 - v_j - V_j s)) (\prod_{j=n+1}^p \Gamma(u_j + U_j s))} z^{-s} ds$$

mais detalhes em Rathie et al. (2011) e Mathai et al. (2010).

Os cálculos dos resultados que envolvem a função H ainda não estão amplamente implementados nos softwares estatísticos e matemáticos e, dessa forma, a variância do modelo RS(a,b,p) pode ser apresentada na forma de somas infinitas duplas

$$\begin{aligned} \text{Var}(Z) &= 2 \sum_{r=0}^{\infty} \left\{ a[a(1+r)]^{-3} \sum_{k=0}^{\infty} \left[-\frac{b}{a^{p+1}(1+r)^p} \right]^k \frac{\Gamma[3 + (p+1)k]}{k!} \right. \\ &\quad \left. + b(1+p)[a(1+r)]^{-(p+3)} \sum_{k=0}^{\infty} \left[-\frac{b}{a^{p+1}(1+r)^p} \right]^k \frac{\Gamma[p + 3 + (p+1)k]}{k!} \right\}. \end{aligned} \quad (12)$$

Em muitas aplicações, obter o valor inverso da função de distribuição é de extrema importância. Em Swamee e Rathie (2007), o modelo Rathie-Swamee é apresentado como uma alternativa inversível para distribuições como normal e t-student. Em Rathie e Swamee (2006) e Rathie (2011) a inversa é apresentada na forma

$$z = \begin{cases} - \sum_{n=0}^{\infty} \frac{(-b/a)^n}{n!} \frac{\Gamma(np+n+1)}{\Gamma(np+2)} \left[\frac{1}{a} \ln \left(\frac{1-F}{F} \right) \right]^{np+1} & , F \leq 0,5 \\ \sum_{n=0}^{\infty} \frac{(-b/a)^n}{n!} \frac{\Gamma(np+n+1)}{\Gamma(np+2)} \left[\frac{1}{a} \ln \left(\frac{F}{1-F} \right) \right]^{np+1} & , F \geq 0,5. \end{cases} \quad (13)$$

Dependendo das configurações paramétricas, essas séries apresentam ins-

tabilidade e podem não convergir rapidamente. Na obtenção das fórmulas em (13), o teorema da inversão de Lagrange foi utilizado, sendo necessário fazer uma escolha do termo a ser isolado. Considerando uma escolha alternativa, pode-se escrever a função inversa de outras maneiras. A partir da equação (5), pode-se escrever

$$z^{p+1} = \frac{1}{b} \ln \left(\frac{F}{1-F} \right) - \frac{a}{b} z,$$

e isolando o termo z do lado esquerdo da equação, tem-se

$$z = \left[\frac{1}{b} \ln \left(\frac{F}{1-F} \right) - \frac{a}{b} z \right]^{\frac{1}{p+1}}.$$

Aplicando o teorema de inversão de Lagrange, tem-se

$$z = \sum_{n=1}^{\infty} \frac{1}{n!} \frac{d^{n-1}}{dy^{n-1}} \left[\frac{1}{b} \ln \left(\frac{F}{1-F} \right) - \frac{a}{b} y \right]^{\frac{n}{p+1}} \Big|_{y=0}.$$

Ao calcular as derivadas, o resultado obtido é

$$z = \sum_{n=1}^{\infty} \frac{1}{n!} \left(\frac{-a}{b} \right)^{n-1} \left(\frac{n}{p+1} \right) \left(\frac{n}{p+1} - 1 \right) \cdots \left(\frac{n}{p+1} - n + 2 \right) \left[\frac{1}{b} \ln \left(\frac{F}{1-F} \right) \right]^{\frac{n}{p+1} - n + 1},$$

e pode-se aplicar a propriedade de recursividade da função gama para escrever

$$z = \sum_{n=1}^{\infty} \frac{\left(\frac{-a}{b} \right)^{n-1}}{n!} \frac{\Gamma\left(\frac{n}{p+1} + 1\right)}{\Gamma\left(\frac{n}{p+1} - n + 2\right)} \left[\frac{1}{b} \ln \left(\frac{F}{1-F} \right) \right]^{\frac{n}{p+1} - n + 1}, F > 0, 5.$$

Para valores $F < 0, 5$, o desenvolvimento segue os mesmos passos e pode-se escrever a inversa da função de distribuição acumulada na forma

$$z = \begin{cases} - \sum_{n=0}^{\infty} \frac{(-a/b)^{n-1}}{n!} \frac{\Gamma\left(\frac{n}{p+1} + 1\right)}{\Gamma\left(\frac{n}{p+1} - n + 2\right)} \left[\frac{1}{b} \ln \left(\frac{1-F}{F} \right) \right]^{\frac{n}{p+1} - n + 1}, & F \leq 0, 5 \\ \sum_{n=0}^{\infty} \frac{(-a/b)^{n-1}}{n!} \frac{\Gamma\left(\frac{n}{p+1} + 1\right)}{\Gamma\left(\frac{n}{p+1} - n + 2\right)} \left[\frac{1}{b} \ln \left(\frac{F}{1-F} \right) \right]^{\frac{n}{p+1} - n + 1}, & F \geq 0, 5 \end{cases} \quad (14)$$

Mesmo com essa formulação alternativa ocorrem problemas de convergência da série. Em investigações preliminares, a forma dada em (13) tem melhor desempenho em situações em que $a > b$, enquanto a forma dada em (14) produz melhores resultados quando $a < b$. Em ambos os casos, à medida em que os valores de F se aproximam de

1/2 as séries não apresentam resultados satisfatórios. Como alternativa à utilização das séries apresentadas, nos estudos de simulação e demais casos em que a inversa da função de distribuição (5) é utilizada neste trabalho, métodos numéricos são utilizados.

Para possibilitar mais flexibilidade ao modelo RS, pode-se considerar a introdução de parâmetros de locação e escala no modelo. Considerando uma variável aleatória Z com distribuição RS(a,b,p), a variável transformada $Y = \mu + Z\sigma$ terá uma distribuição com cinco parâmetros cuja função de distribuição de probabilidades tem a forma

$$F(y) = \left\{ \exp \left[-\frac{y-\mu}{\sigma} (a + b \left| \frac{y-\mu}{\sigma} \right|^p) \right] + 1 \right\}^{-1}, \quad (15)$$

e função de densidade

$$f(y) = \frac{1}{\sigma} \frac{[a + b(1+p) \left| \frac{y-\mu}{\sigma} \right|^p] \exp \left[-\frac{y-\mu}{\sigma} (a + b \left| \frac{y-\mu}{\sigma} \right|^p) \right]}{\left\{ \exp \left[-\frac{y-\mu}{\sigma} (a + b \left| \frac{y-\mu}{\sigma} \right|^p) \right] + 1 \right\}^2}, \quad (16)$$

em que $-\infty < \mu < \infty$ representa o valor esperado da variável aleatória Y e permite deslocar a distribuição ao longo dos reais, enquanto $\sigma > 0$, chamado parâmetro de escala é um parâmetro que flexibiliza a variância de Y podendo alterar a escala de representação do modelo.

A inclusão desses parâmetros adicionais deve ser feita com cautela devido aos possíveis problemas de identificabilidade que podem ocorrer com a inclusão do parâmetro de escala σ . Inicialmente, pode-se fixar $\sigma = 1$, não incluir esse parâmetro adicional ou fazer uma reparametrização do modelo para contornar os possíveis problemas de identificabilidade.

A propriedade de simetria de um modelo pode ser limitante na utilização do mesmo para representar o comportamento de uma variável aleatória com comportamento assimétrico, e, assim, utilizando a proposta de Azzalini (1985) para criação do modelo Skew-Normal, Rathie e Coutinho (2011) propuseram o modelo Rathie-Swamee assimétrico (RSA).

3.2 Modelo Rathie-Swamee Assimétrico

A forma utilizada para se obterem distribuições assimétricas $h(z)$ a partir de distribuições simétricas é dada por

$$h(z) = 2f(z)G(\lambda z), \quad (17)$$

sendo f uma função de densidade simétrica em torno da origem e G uma função de distribuição cuja respectiva densidade seja simétrica em torno da origem. As funções f e G podem ser referentes a diferentes modelos de probabilidade simétricos ou também a um modelo comum às duas.

O modelo proposto por Rathie e Coutinho (2011) considera uma mesma distribuição $RS(a, b, p)$ para as duas componentes do lado direito da equação (17), considerando as equações (6) e (5) para obter

$$h(z) = 2f(z)F(cz) = \frac{2[a + b(1 + p)|z|^p] \exp[-z(a + b|z|^p)]}{\{\exp[-z(a + b|z|^p)] + 1\}^2 \{\exp[-cz(a + b|cz|^p)] + 1\}}. \quad (18)$$

Incluindo os parâmetros de locação e escala no modelo assimétrico (18), a função de densidade tem a forma

$$h(y) = 2f(y)F(cy) = \frac{2[a + b(1 + p)|\frac{y-\mu}{\sigma}|^p] \exp[-\frac{y-\mu}{\sigma}(a + b|\frac{y-\mu}{\sigma}|^p)]}{\{\sigma \exp[-\frac{y-\mu}{\sigma}(a + b|\frac{y-\mu}{\sigma}|^p)] + 1\}^2 \{\exp[-c\frac{y-\mu}{\sigma}(a + b|c\frac{y-\mu}{\sigma}|^p)] + 1\}}. \quad (19)$$

3.3 Produto e quociente de variáveis Rathie-Swamee

Em algumas aplicações, tem-se interesse em modelar o produto ou quociente entre duas variáveis aleatórias. Nadarajah (2009) e Nadarajah e Gupta (2006) citam exemplos de aplicações e apresentam resultados para produtos e quocientes entre variáveis logísticas. No caso de duas variáveis com distribuição Rathie-Swamee particularmente com $a = 0$, Rathie; Zornig e Gomes (2012) mostram que os quocientes e produtos resultam em uma mesma família de distribuições de probabilidade.

A função considerada é

$$F_\alpha(v) = 2 \int_0^1 \frac{1}{(1 + w)^2 [\exp(-v(-\ln v)^\alpha) + 1]} dw, \quad v \in R, \quad (20)$$

sendo α o parâmetro do modelo, tal que $\alpha \in R$. Pode-se mostrar que a função (20) é uma função de distribuição de probabilidades. Para valores $0 < w < 1$, a função

$$S(v) = \frac{1}{\exp(-v(-\ln w)^\alpha) + 1}, \quad (21)$$

tem a forma da distribuição logística dada em (3) e é simples notar que $\lim_{v \rightarrow \infty} S(v) = 1$ e $\lim_{v \rightarrow -\infty} S(v) = 0$. Dessa forma a função $F_\alpha(v)$ é monótona crescente com

$$\lim_{v \rightarrow \infty} F_\alpha(v) = 2 \int_0^1 \frac{1}{(1+w)^2} dw = 1 \quad \text{and} \quad \lim_{v \rightarrow -\infty} F(v) = 0, \quad (22)$$

e, consequentemente, $F_\alpha(v)$ é a função de distribuição de probabilidades da variável aleatória V . Derivando (20) em relação a v , obtem-se a função de densidade de probabilidade

$$f_\alpha(v) = 2 \int_0^1 \frac{(-\ln w)^\alpha \exp(-v(-\ln w)^\alpha)}{(1+w)^2 [\exp(-v(-\ln w)^\alpha) + 1]^2} dw, \quad v \in \mathbb{R}. \quad (23)$$

Essa função de densidade é simétrica, pois $S(v) + S(-v) = 1$ e, assim, obtem-se

$$F_\alpha(v) + F_\alpha(-v) = 2 \int_0^1 \frac{1}{(1+w)^2} (S(v) + S(-v)) dw = 1. \quad (24)$$

Considerando duas variáveis aleatórias simétricas e independentes X e Y com funções de distribuição F e G e funções de densidade f e g , respectivamente, o quociente $Z = \frac{Y}{X}$ tem distribuição de probabilidades dada por

$$H(z) = P(Y/X \leq z) = 2 \int_0^\infty f(x)G(zx)dx. \quad (25)$$

que é simétrica, pois

$$H(-z) = 2 \int_0^\infty f(x)(1 - G(zx))dx = 1 - H(z). \quad (26)$$

Se $X \sim RS(a = 0, b = b_1, p = p_1)$, $Y \sim RS(a = 0, b = b_2, p = p_2)$ e $z \geq 0$ então

$$H(z) = 2 \int_0^\infty \frac{b_1(p_1 + 1)x^{p_1} \exp(-b_1x^{p_1+1})}{[\exp(-b_1x^{p_1+1}) + 1]^2 [\exp(-b_2(xz)^{p_2+1}) + 1]} dx, \quad (27)$$

e fazendo a substituição $w = \exp(-b_1 x^{p_1+1})$, obtém-se

$$\begin{aligned} H(z) &= \int_0^1 \frac{2}{(w+1)^2 [\exp(-b_2 z^{p_2+1} x^{p_2+1}) + 1]} dw \\ &= \int_0^1 \frac{2}{(w+1)^2 \left\{ \exp \left[-b_2 z^{p_2+1} \left(\frac{-\ln w}{b_1} \right)^{\frac{p_2+1}{p_1+1}} \right] + 1 \right\}} dw. \end{aligned} \quad (28)$$

Pode-se notar que a forma em (28) é a forma da distribuição (20) com $\alpha = \frac{p_2+1}{p_1+1}$ e $v = \frac{b_2 z^{p_2+1}}{b_1^\alpha}$.

Assim, a distribuição de um quociente de variáveis com distribuição $RS(a = 0, b, p)$ é

$$H(z) = F_\alpha(b_2 b_1^{-\alpha} z^{p_2+1}). \quad (29)$$

No caso particular em que $p_1 = p_2 = 0$, isto é, $X \sim LG(b_1)$ e $Y \sim LG(b_2)$, o resultado (29) corresponde a

$$H(z) = F_1\left(\frac{b_2}{b_1} z\right). \quad (30)$$

Nos casos em que $\alpha = 1$, como no caso do quociente de variáveis com distribuição logística (3), a função de distribuição de probabilidades (20) pode ser representada de diferentes formas com o uso de funções especiais. Para valores $v > 0$, obtêm-se

$$F_1(v) = 2 \sum_{r=0}^{\infty} (-1)^r {}_2F_1\left(1, \frac{r+1}{v}; \frac{r+1}{v} + 1; -1\right) \quad (31)$$

$$F_1(v) = 2 \sum_{s=0}^{\infty} \sum_{r=0}^{\infty} (-1)^{s+r} \frac{r+1}{sv+r+1} \quad (32)$$

$$F_1(v) = 2 \sum_{r=0}^{\infty} \frac{(-1)^r}{1+rv} {}_2F_1(2, 1+rv; 2+rv; -1) \quad (33)$$

$$F_1(v) = \frac{1}{v} \sum_{r=1}^{\infty} (-1)^{r-1} r \left[\Psi\left(\frac{r}{2v} + \frac{1}{2}\right) - \Psi\left(\frac{r}{2v}\right) \right] \quad (34)$$

$$F_1(v) = \frac{1}{2} + v \sum_{r=1}^{\infty} (-1)^r r \left[\Psi\left(\frac{rv+1}{2}\right) - \Psi\left(\frac{rv}{2} + 1\right) \right], \quad (35)$$

sendo ${}_2F_1(a, b; c; x)$ a função Gauss hypergeométrica e $\Psi(x)$ a função digamma (Luke, 1969). A obtenção das formulações alternativas está detalhada em Rathie; Zornig e Gomes (2012).

A apresentação nessas formas compactas pode simplificar a obtenção de resultados analíticos e de computações para resultados numéricos nas aplicações do modelo.

A distribuição de probabilidades referente à equação (20), também pode ser utilizada para descrever a distribuição resultante do produto entre variáveis aleatórias com distribuição Rathie-Swamee no caso particular de $a = 0$. Sejam X e Y variáveis aleatórias independentes com funções de distribuição F e G e funções densidade f e g , pode-se escrever a distribuição de $Z = XY$ como

$$H(z) = P(XY \leq z) = 2 \int_0^\infty f(x)G\left(\frac{z}{x}\right) dx. \quad (36)$$

No caso específico de variáveis com distribuição Rathie-Swamee, tem-se $X \sim RS(a = 0, b_1, p_1)$ e $Y \sim RS(a = 0, b_2, p_2)$, e a função obtida para valores $z \geq 0$ é

$$H(z) = 2 \int_0^1 \frac{1}{(w+1)^2 \left\{ \exp \left[-b_2 z^{p_2+1} \left(\frac{b_1}{-\ln w} \right)^{\frac{p_2+1}{p_1+1}} \right] + 1 \right\}} dw. \quad (37)$$

Considerando $\alpha = -\frac{p_2+1}{p_1+1}$, pode-se definir $H(z)$ como

$$H(z) = F_\alpha(b_2 b_1^{-\alpha} z^{p_2+1}), \quad (38)$$

e, assim, o produto entre variáveis com distribuição Rathie-Swamee pode ser escrito na forma da distribuição dada pela equação (20).

No caso de distribuições LG(b), tomando $p_1 = p_2 = 0$, os produtos entre variáveis podem ser representados por

$$H(z) = F_{-1}(b_1 b_2 z) = 1 + 4\sqrt{b_1 b_2 z} \sum_{r=0}^{\infty} \sum_{s=1}^{\infty} (-1)^{r+s} \sqrt{s(r+1)} K_1 \left(2\sqrt{(r+1)s b_1 b_2 z} \right). \quad (39)$$

Vale notar que para os casos de quocientes entre variáveis, a função (20) tem valores positivos para o parâmetro α , e, nos casos de produtos entre variáveis, os valores para o parâmetro α são negativos. As Figuras 5 e 6 apresentam formas da função de densidade (23) para alguns valores de α selecionados, junto à forma da função de

densidade de uma distribuição Normal padrão para efeito de comparações.

Figura 5 - Função de densidade de probabilidade de quocientes entre variáveis Rathie-Swamee

Figura 6 - Função de densidade de probabilidade de produtos entre variáveis Rathie-Swamee

O parâmetro α é um parâmetro de forma da distribuição. No caso particular em que $\alpha = 0$, o modelo fica reduzido ao modelo logístico em sua forma padrão descrito em (1). Quanto maior o valor absoluto de α , mais pesadas se tornam as caudas da distribuição.

Na Tabela 1, são apresentados alguns quantis da distribuição dada pela equação (20), para alguns valores selecionados do parâmetro α . Note que o valor dos quantis, a partir de $p = 0,95$ no topo e na base da tabela, onde estão os maiores valores absolutos de α , são muito grandes.

Tabela 1 - Quantis v para o modelo (20) tais que $F_\alpha(v) = p$

α	$p = 0,6$	$p = 0,7$	$p = 0,8$	$p = 0,85$	$p = 0,9$	$p = 0,95$	$p = 0,975$	$p = 0,99$
-3	0,0477	0,4838	2,4112	5,0972	11,3389	30,8251	65,3220	142,6832
-2,5	0,0723	0,5215	2,0764	3,9628	7,9118	18,7996	36,0093	70,7952
-2	0,1084	0,5630	1,8015	3,1110	5,5861	11,6290	20,1669	35,7451
-1,5	0,1605	0,6104	1,5830	2,4811	4,0184	7,3532	11,5731	18,5401
-1	0,2335	0,6690	1,4237	2,0346	2,9849	4,8261	6,9210	10,0643
-0,5	0,3293	0,7503	1,3422	1,7638	2,3635	3,4102	4,4908	5,9787
0	0,4055	0,8473	1,3863	1,7346	2,1972	2,9444	3,6636	4,5951
0,5	0,3815	0,8200	1,4181	1,8574	2,5257	3,8819	5,6576	9,0461
1	0,3041	0,7092	1,4100	2,0565	3,2882	6,8392	13,8163	34,6386
1,5	0,2226	0,5948	1,4328	2,3960	4,6571	13,6141	38,8223	153,8091
2	0,1568	0,4973	1,4873	2,8812	6,8702	28,4222	114,6355	718,1314
2,5	0,1085	0,4168	1,5653	3,5269	10,3492	60,7128	346,5301	3433,0067
3	0,0744	0,3503	1,6616	4,3631	15,7776	131,3622	1061,2661	16627,7410

3.4 Modelo de regressão Rathie-Swamee

Modelos de regressão permitem analisar a relação entre uma variável dependente e um conjunto de variáveis independentes. Existem diferentes técnicas e modelos propostos para análise de regressão. De uma forma geral, pode-se propor um modelo de regressão como uma relação funcional entre a variável dependente Y , as variáveis independentes $\mathbf{x}$ e um conjunto de parâmetros desconhecidos $\boldsymbol{\beta}$. Usualmente, essa relação é proposta na forma:

$$E(Y|\mathbf{x}) = g(\mathbf{x}, \boldsymbol{\beta}) \quad (40)$$

Considerando os modelos Rathie-Swamee simétricos com parâmetros de locação e escala, de forma que $Y = \mu + \sigma Z$, pode-se incorporar na média o efeito das covariáveis e reescrever o modelo como $Y = \mathbf{x}^T \boldsymbol{\beta} + \sigma Z$ e, assim, as funções de distribuição e densidade são reescritas, respectivamente, como

$$F(y) = \left\{ \exp \left[- \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma} (a + b \left| \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma} \right|^p) \right] + 1 \right\}^{-1}, \quad (41)$$

e

$$f(y) = \frac{1}{\sigma} \frac{[a + b(1 + p) \left| \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma} \right|^p] \exp \left[- \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma} (a + b \left| \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma} \right|^p) \right]}{\{ \exp \left[- \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma} (a + b \left| \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma} \right|^p) \right] + 1 \}^2}, \quad (42)$$

sendo $\mathbf{x}$ um vetor de covariáveis e $\boldsymbol{\beta}$ um vetor de parâmetros referente aos efeitos das covariáveis sobre a média.

No caso do modelo assimétrico, a função de densidade que incorpora o efeito

de covariáveis na média da variável aleatória tem a forma

$$h(y) = \frac{2[a + b(1 + p)|\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}|^p] \exp[-\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}(a + b|\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}|^p) \exp[-\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}(a + b|\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}|^p)]}{\{\exp[-\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}(a + b|\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}|^p)] + 1\}^2 \{exp[-c\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}(a + b|c\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}|^p)] + 1\}}. \quad (43)$$

Nas aplicações desenvolvidas neste trabalho, os modelos de regressão considerados são referentes aos casos particulares dos modelos (42) e (43) em que $\sigma = 1$, para evitar problemas de falta de identificabilidade nos modelos.

3.5 Modelos de misturas

Como indicado na seção de revisão, os modelos de mistura permitem o ajuste de modelos para dados com comportamento multimodal. Para definir uma mistura finita com g componentes considera-se uma variável aleatória Y , com a seguinte estrutura para sua função de densidade:

$$f(y; \Psi) = \sum_{j=1}^g \pi_j f_j(y; \boldsymbol{\theta}_j). \quad (44)$$

Nesse contexto, a expressão (44) é uma mistura finita de g funções de densidade de probabilidade, os termos π_j são os pesos de mistura, tais que $0 \leq \pi_j \leq 1$ e $\sum_{j=1}^g \pi_j = 1$, e as funções $f_j(y; \boldsymbol{\theta}_j)$ são as funções de densidade que compõem a mistura tendo $\boldsymbol{\theta}_j$ como parâmetros.

O número g de componentes pode ser definido, a priori, de forma intrínseca conforme o conhecimento de características dos dados, assim como pode ser uma quantidade desconhecida a ser estimada. Como o objetivo da utilização de misturas neste trabalho é comparar o desempenho obtido pelos modelos de mistura com o desempenho dos modelos Rathie-Swamee para dados com bimodalidade, o número de componentes g é considerado fixo e igual a 2.

As diferentes funções de densidade de probabilidade que compõem a mistura podem ser de uma mesma família paramétrica ou não. Aqui serão consideradas misturas de duas funções de densidade normais, com parâmetros $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2)$.

3.6 Inferências

A teoria de verossimilhança é a base para as etapas de estimação e inferências sobre os parâmetros dos modelos. A função de verossimilhança é uma ferramenta

para analisar de forma objetiva um conjunto de dados, pois a mesma captura todas as informações da amostra sobre o vetor de parâmetros do modelo. Diversos métodos e conceitos são desenvolvidos a partir da função de verossimilhança. Para realizar estimativas pontuais para os parâmetros, o método da máxima verossimilhança torna-se uma escolha natural.

3.6.1 Método de Máxima verossimilhança

Definição 3 *Seja $L(\boldsymbol{\theta}; y)$ a função de verossimilhança. O valor $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(y)$ é a estimativa de máxima verossimilhança para $\boldsymbol{\theta}$, se $L(\hat{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta})$, $\forall \boldsymbol{\theta}$.*

Para efeitos práticos, é conveniente a utilização da função referente ao logaritmo natural da função de verossimilhança

$$l(\boldsymbol{\theta}; y) = \ln L(\boldsymbol{\theta}; y) = \sum_{i=1}^n \ln f(y_i | \boldsymbol{\theta}), \quad (45)$$

por ser uma transformação monótona, o mesmo $\hat{\boldsymbol{\theta}}$ que maximiza a função de verossimilhança, maximiza, também, o seu logaritmo. É necessário resolver o sistema de equações para encontrar o estimador de máxima verossimilhança de $\boldsymbol{\theta}$:

$$U(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0. \quad (46)$$

Para grandes amostras, a distribuição amostral dos estimadores de máxima verossimilhança $\hat{\boldsymbol{\theta}}$ é a distribuição normal com média $\boldsymbol{\theta}$ e matriz de covariâncias dada pelo inverso da matriz de informação de Fisher, ou seja, $\text{cov}(\boldsymbol{\theta}) = I^{-1}(\boldsymbol{\theta})$, sendo $I(\boldsymbol{\theta}) = E\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right]$. A matriz de covariâncias pode ser estimada de forma consistente utilizando a matriz de informação observada, $\ddot{L}(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Big|_{\hat{\boldsymbol{\theta}}}$, de forma que $\widehat{\text{cov}(\hat{\boldsymbol{\theta}})} = \ddot{L}^{-1}(\hat{\boldsymbol{\theta}})$.

Como o sistema de equações (46) é, usualmente, não linear, é necessária a utilização de método numérico para obter as estimativas de máxima verossimilhança. Dentre os métodos disponíveis devemos escolher aqueles que considerem as especificidades das funções envolvidas. Os modelos Rathie-Swamee simétrico e assimétrico possuem parâmetros cujo espaço paramétrico é restrito e, por isso, os métodos de maximização devem considerar essas restrições. Os métodos de barreira adaptam o problema de otimização para incorporar as restrições na função objetivo.

3.6.2 Método de barreiras

O método da barreira visa a constituir uma sequência de problemas irrestritos, cujos valores ótimos que convergem para o ponto de ótimo da função objetivo. Nesse caso, a função de verossimilhança com restrições nos parâmetros. Adicionando uma função barreira à função objetivo impede-se que os algoritmos de busca saiam do espaço paramétrico do modelo (Zörnig, 2011). O interesse inicial é maximizar o logaritmo da função de verossimilhança (45) em relação a θ , mas a função objetivo do problema de maximização é alterada na seguinte forma

$$l_R(\theta; x) = \sum_{i=1}^n \ln(f(x_i|\theta)) - \phi B(\theta).$$

em que $\phi > 0$ é uma constante positiva e a função barreira $B(\theta)$ pode assumir diferentes formas.

Supondo que um modelo tenha m restrições que podem ser escritas na forma $g_i(\theta) = u_i\theta - c_i \geq 0$, $i = 1, \dots, m$ e considerando a função barreira logarítmica, a função objetivo modificada é dada por

$$l_R(\theta; x) = \sum_{i=1}^n \ln(f(x_i|\theta)) - \phi \sum_{i=1}^m \ln g_i(\theta),$$

em que u é a matriz $(m \times k)$ de incidência de restrições sobre os parâmetros e c o vetor $(1 \times m)$ de restrições. Essa função é otimizada para uma sequência de constantes ϕ tendendo para zero, fazendo com que a barreira tenha menor contribuição na função objetivo.

Questões como a arbitrariedade na escolha da velocidade com que a constante ϕ deve ir para zero, assim como a necessidade de repetir todo o processo de otimização da função para cada valor de ϕ são apontados como desvantagens do método barreira em Lange (2001), que apresenta o método da barreira adaptativa em que o valor de ϕ é mantido constante enquanto os coeficientes da função barreira são modificados a cada iteração. Assim, o problema de otimização é transferido para a função

$$R(\theta|\theta_j) = l(\theta) - \phi \sum_{i=1}^m [g_i(\theta_k) \ln g_i(\theta) - u_i^t \theta], \quad (47)$$

e o termo adaptivo $g_i(\boldsymbol{\theta}_k) \ln g_i(\boldsymbol{\theta})$ permite que os valores se aproximem da fronteira, sem a necessidade de criar a sequência de estimativas para diferentes valores de ϕ .

Considerando o modelo RS(a=0,b,p), as restrições do modelo são $b \geq 0$ e $p \geq 0$ e podem ser definidas

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} b \\ p \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Após essa modificação na função objetivo, é necessária a utilização de algum método numérico para obtenção das estimativas para o modelo. No ambiente R de programação estatística (R Development Core Team, 2011) a função *optim* tem diferentes métodos de otimização disponíveis, dentre eles o método BFGS, que é utilizado nas aplicações aqui apresentadas.

A função *constrOptim* pode ser utilizada no contexto do método da barreira adaptiva, reescrevendo o problema conforme a equação (47) para utilização da função *optim*.

Para construção de testes de hipóteses e intervalos de confiança, a estatística da razão de verossimilhança é de grande utilidade e pode ser definida mesmo para os casos em que as estimativas de máxima verossimilhança sejam obtidas numericamente.

$$\lambda(\mathbf{y}) = \frac{\sup_{\Theta_0} L(\theta|\mathbf{y})}{\sup_{\Theta} L(\theta|\mathbf{y})}. \quad (48)$$

Sob as condições de regularidade do modelo $f(y|\boldsymbol{\theta})$, a estatística $-2 \ln \lambda(\mathbf{Y})$ converge, assintoticamente, para uma distribuição qui-quadrado. O número de graus de liberdade da distribuição é dado pela diferença do número de parâmetros livres existentes em $\boldsymbol{\theta} \in \Theta_0$ e o número de parâmetros livres existentes em $\boldsymbol{\theta} \in \Theta$, ou seja, a dimensão do vetor de parâmetros sobre o qual se fazem as inferências.

3.6.3 Intervalos de confiança

Com a construção intervalos de confiança, pode-se incorporar na estimativa dos parâmetros a incerteza associada à aleatoriedade dos resultados amostrais. No caso dos modelos multiparamétricos, as regiões de confiança representam a variabilidade conjunta dos parâmetros.

Definição 4 Um intervalo de confiança $(1 - \alpha)$ para θ é um intervalo $C_n(a, b)$, sendo $a = a(X_1, \dots, X_n)$ e $b = b(X_1, \dots, X_n)$ funções das observações amostrais, tais que $P_{\theta}(\theta \in C_n) \geq (1 - \alpha)$, $\forall \theta \in \Theta$. Sendo o parâmetro θ fixo e o intervalo aleatório.

Com base na propriedade de normalidade assintótica dos estimadores de máxima verossimilhança, podem-se construir intervalos de confiança aproximados para os parâmetros. Como $\hat{\theta} \sim N(\theta, V(\hat{\theta}))$, tem-se que $\frac{\hat{\theta} - \theta}{\sqrt{V(\hat{\theta})}} = Z \sim N(0, 1)$, e para um coeficiente de confiança $\gamma = (1 - \alpha)$ qualquer pode-se escrever

$$\begin{aligned} \gamma &= P(-z_\gamma < Z < z_\gamma) = P(-z_\gamma < \frac{\hat{\theta} - \theta}{\sqrt{V(\hat{\theta})}} < z_\gamma) \\ &= P(\hat{\theta} - z_\gamma \sqrt{V(\hat{\theta})} < \theta < \hat{\theta} + z_\gamma \sqrt{V(\hat{\theta})}). \end{aligned}$$

Assim, para criar um intervalo aproximado com coeficiente de confiança, γ , tem-se:

$$IC[\theta, \gamma] = [\hat{\theta} - z_\gamma \sqrt{V(\hat{\theta})}; \hat{\theta} + z_\gamma \sqrt{V(\hat{\theta})}]. \quad (49)$$

É importante observar que os intervalos simétricos obtidos por (49) podem possuir limites fora do espaço paramétrico devido às restrições nos parâmetros, criando, assim, intervalos inadequados. Uma ideia para contornar essa situação é usar a propriedade de invariância dos estimadores de máxima verossimilhança e buscar uma transformação no parâmetro para eliminar as restrições e modificar a superfície de verossimilhança, tornando-a mais quadrática e, possivelmente, melhorar a aproximação da curva por uma parábola.

No caso de parâmetros que assumem apenas valores positivos, uma transformação intuitiva sugerida por Meeker e Escobar (1998) é $\tau = \ln \theta$. Sabendo-se que o estimador de máxima verossimilhança para ϕ pode ser obtido por $\hat{\phi} = \ln \hat{\theta}$ e que pelo método delta pode-se obter $V(\hat{\tau}) = \frac{V(\hat{\theta})}{\hat{\theta}}$, pode-se obter um intervalo de confiança aproximado para τ como na forma (49) e aplicar a transformação inversa $\theta = \exp(\tau)$ para obter um intervalo para θ que não possua limites inadequados para os parâmetros restritos. O intervalo terá a forma:

$$IC[\theta, \gamma] = \left[\frac{\hat{\theta}}{\exp\left(z_\gamma \frac{V(\hat{\theta})}{\hat{\theta}}\right)}; \hat{\theta} \exp\left(z_\gamma \frac{V(\hat{\theta})}{\hat{\theta}}\right) \right]. \quad (50)$$

Vale notar que o intervalo é aproximado e a cobertura nominal do mesmo é objeto de estudo neste trabalho.

Como visto, o método de aproximação quadrática da função de verossimilhança permite construir intervalos de confiança para cada parâmetro de forma marginal, criando sempre intervalos simétricos sem considerar restrições no espaço paramétrico, dependendo de transformações para melhorar as aproximações ou adaptar as restrições contidas no espaço paramétrico. A estatística da razão de verossimilhança e o conhecimento de sua distribuição assintótica possibilita a criação de regiões de confiança baseadas na função de verossimilhança.

A ideia é determinar uma região de valores para os parâmetros que apresentem determinada compatibilidade com a amostra. Para que os intervalos tenham uma interpretação probabilística utiliza-se a distribuição assintótica associada à estatística de razão de verossimilhança para estabelecer os limites da região de confiança.

Considera-se a função $D(\Theta) = -2 \ln \lambda(\mathbf{y})$, e a região definida por $[\theta \in \Theta : D(\Theta) \leq m]$, tal que c seja definido a partir de um coeficiente confiança $\gamma = (1 - \alpha)$ de forma que $P(\chi_d^2 \geq m) = \alpha$. A obtenção dos valores de θ que satisfaçam a condição citada dependem da solução de equações não lineares implicando certo custo computacional.

Nesta abordagem, as regiões de confiança consideram simultaneamente todos os parâmetros e as relações existentes entre eles, enquanto os intervalos obtidos marginalmente desconsideram as relações existentes entre os parâmetros, tratando-os como se fossem ortogonais. No caso de modelos com muitos parâmetros, as regiões de confiança têm difícil interpretação e visualização. O perfilamento da função verossimilhança em relação aos parâmetros permite a criação de intervalos de confiança individuais para cada parâmetro sem desconsiderar as relações existentes entre os parâmetros do modelo.

Definição 5 *A verossimilhança perfilada de θ_1 é definida por*

$$L_p(\theta_1) = L(\theta_1, \hat{\theta}_{2_{\theta_1}}). \quad (51)$$

A maximização da função para obtenção dos estimadores é realizada em duas etapas. Primeiro, deve-se obter o valor $\hat{\theta}_{\theta_1}$ que maximiza $l(\theta_1, \theta_2)$ com respeito a θ_2 para um conjunto de valores de θ_1 , considerados fixos. Então, maximiza-se $l_p(\theta_1)$.

Além dos intervalos obtidos de forma analítica a partir da função de verossimilhança ou de aproximações quadráticas da mesma, métodos computacionalmente

intensivos podem ser utilizados para avaliar a incerteza associada aos estimadores.

O método de bootstrap consiste em criar replicações dos dados observados e com essas replicações avaliar a variabilidade dos estimadores, sem a necessidade de suposições ou resultados analíticos sobre os modelos considerados Hinkley (1997). As replicações dos dados podem ser feitas sob uma abordagem paramétrica ou não paramétrica.

Supondo que se considera um determinado modelo paramétrico para descrever a distribuição de um conjunto de dados $y_1, \dots, y_n$, pode-se representar esse modelo pelas funções de densidade de probabilidade $f(y; \theta)$ e função de distribuição de probabilidade $F(y; \theta)$. A partir da obtenção de estimativas $\hat{\theta}$ para os parâmetros, o modelo representado por $f(y; \hat{\theta})$ e $F(y; \hat{\theta})$ é chamado de modelo ajustado. B simulações de diferentes amostras de variáveis com distribuição dada pelo modelo ajustado, permitem que para cada uma das B amostras possam ser reestimados os parâmetros θ e, assim, avaliar de forma empírica a distribuição e características como viés e variabilidade associada ao estimador $\hat{\theta}$ utilizado.

Sob a abordagem não paramétrica, a forma de replicar os dados é por meio de B amostragens com reposição dos dados observados. De forma semelhante ao caso paramétrico, para cada uma das replicações, podem-se reestimar os parâmetros de interesse para assim obter uma distribuição empírica do estimador utilizado.

A obtenção de intervalos de confiança a partir das distribuições empíricas resultantes da utilização do método de bootstrap, é bastante simples. Será considerado o método percentil de obtenção de intervalo de confiança, em que as estimativas obtidas nas B replicações são ordenadas e tomam-se como limite dos intervalos de confiança os quantis referentes ao coeficiente confiança desejado para os intervalos.

Definição 5 *Sejam $\lambda_{(1)}^*, \dots, \lambda_{(B)}^*$ os valores ordenados das estimativas obtidas para um parâmetro λ em B replicações bootstrap. Um intervalo de confiança percentílico para o parâmetro λ é obtido por:*

$$IC(\lambda, 1 - 2\alpha) = [\lambda_{((B+1)(1-\alpha))}^*, \lambda_{((1+B)\alpha)}^*], \quad (52)$$

sendo B escolhido para que $((R+1)(1-\alpha))$ e $((1+B)\alpha)$ sejam números inteiros.

3.6.4 Testes de Hipóteses

Nas aplicações de estatística, as hipóteses são afirmações sobre a população em estudo. Dada a utilização de modelos para descrever o comportamento da população, as hipóteses são afirmações sobre as características do modelo. A utilização dos testes de hipótese tem como objetivo avaliar a compatibilidade de uma amostra observada em relação a uma hipótese formulada.

O teste da razão de verossimilhança compara valores da função de verossimilhança em seu máximo, considerando os diferentes subconjuntos do espaço paramétrico definidos pelas hipóteses nula e alternativa.

Definição 5 *A estatística do teste de razão de verossimilhança para testar $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ é*

$$\lambda(y) = \frac{\sup_{\Theta_0} L(\theta|y)}{\sup_{\Theta} L(\theta|y)} \quad (53)$$

e o teste da razão de verossimilhança é qualquer teste que tenha região de rejeição na forma $\{y : \lambda(y) \geq m\}$, para $0 \leq m \leq 1$.

Para que se possam avaliar os possíveis erros que podem ser cometidos ao aplicar um teste de razão de verossimilhança é necessário que a quantidade m seja escolhida levando em consideração as propriedades probabilísticas da estatística de teste. Para grandes amostras, pode-se supor que $-2 \ln \lambda(y) \sim \chi_d^2$, e, dessa forma, para uma determinada amostra aleatória $\mathbf{Y}$ rejeita-se a hipótese nula se e somente se $\lambda(\mathbf{Y}) \geq \exp(-\chi_d^2/2)$, sendo d o número de graus de liberdade, resultante da diferença entre a dimensão do espaço paramétrico e o número de parâmetros livres sob a hipótese nula.

Uma outra metodologia utilizada para definir testes de hipóteses é utilizar as propriedades assintóticas dos estimadores de máxima verossimilhança, que permitam a criação de testes de hipóteses aproximados, semelhantes aos resultados obtidos para os intervalos de confiança. Considerando que os estimadores de máxima verossimilhança têm distribuição normal assintótica, $\hat{\theta} \sim N(\theta, V(\hat{\theta}))$ pode-se utilizar a estatística de teste $W = \frac{\hat{\theta} - \theta_0}{V(\hat{\theta})}$ com distribuição Normal padrão.

3.6.5 Critério de informação de Akaike

Além dos testes de hipóteses construídos com o intuito de selecionar modelos, relacionando os modelos encaixados, a partir de hipóteses formuladas sobre os

parâmetros dos modelos, outras medidas permitem classificar e selecionar modelos de diferentes formas e famílias. Akaike (1973) propõe a utilização de uma medida relacionada à estimativa de uma aproximação da distância de Kullback-Leiber entre os modelos considerados e um verdadeiro modelo gerador do fenômeno observado.

A medida denominada de informação de Akaike foi definida como:

$$AIC = -2\log(L(\hat{\theta}|\mathbf{y})) + 2k, \quad (54)$$

e considera o valor do logaritmo da função de verossimilhança no ponto de máximo, referente aos estimadores de máxima verossimilhança para um determinado modelo, e o número k de parâmetros estimados no modelo.

Dentre um conjunto de diferentes modelos considerados para um determinado problema, deve-se escolher aquele que tiver o menor AIC . Note que o termo $2k$ na equação (54), atua como uma penalização para a inclusão de parâmetros a serem estimados nos modelos, uma vez que existe a incerteza associada à estimação deles. O princípio de parcimoniosidade na formulação dos modelos estatísticos deve ser observado, de forma que os modelos sejam os mais simples possíveis em relação à inclusão de variáveis, estrutura do modelo e número de parâmetros a serem estimados (Burnham e Anderson, 2002).

O valor absoluto do AIC não é relevante isoladamente. Deven-se avaliar as diferenças entre os valores de AIC calculados para os diferentes modelos considerados na etapa de seleção de modelos.

A determinação de que um modelo seja melhor dentre um conjunto de diferentes modelos considerados para representar um conjunto de dados não significa que aquele modelo seja adequado para esse conjunto de dados. Para validação dos resultados obtidos na realização de inferências para um determinado modelo é necessário verificar se ele é adequado. Análises de resíduos e sensibilidade são algumas das formas para avaliar a adequabilidade dos modelos considerados. As técnicas de diagnóstico são utilizadas com o objetivo de detectar desvios nas suposições adotadas, existência de valores extremos ou influentes e possíveis problemas no ajuste do modelo aos dados.

3.7 Análise de resíduos

Análise de resíduo tem papel central na validação de um modelo. Diversas formas de resíduo foram propostas para os diferentes tipos de modelos de regressão. O resíduos ordinários $e_i = y_i - x_i^t \hat{\beta}$ e suas padronizações são os mais utilizados para modelos lineares normais. Resíduos de Pearson e do tipo componentes de *deviance* são algumas das propostas utilizadas para os modelos lineares generalizados. A utilização dos resíduos ordinários em modelos não normais depende da obtenção da distribuição de probabilidades dos resíduos para analisar o ajuste do modelo.

Dunn e Smith (1996) propõem a utilização de resíduos quantílicos como um caso especial do resíduo de Cox-Snell. Considerando que a distribuição de probabilidade de uma variável pode ser descrita por uma função $F(y; \theta)$, se essa função F for contínua, então, $F(y; \theta)$ tem distribuição uniforme(0, 1). O resíduo quantílico é definido como: $rq_i = \Phi^{-1}\{F(y_i; \theta_i)\}$ sendo Φ a função de distribuição normal padrão acumulada. Fora a variabilidade amostral de $\hat{\theta}$ os resíduos rq_i têm distribuição normal padrão exata, indicando que a distribuição de rq_i converge para normal padrão se θ for estimado de forma consistente.

3.8 Influência Global

A metodologia de influência global descrita por Cook (1977), também denominada deleção de casos, é o procedimento que avalia o efeito da i -ésima observação nas estimativas dos parâmetros. Esse procedimento baseia-se na diferença entre as estimativas de máxima verossimilhança relativas à amostra completa e às estimativas obtidas da amostra reduzida por exclusão de um indivíduo.

Ao considerar que a i -ésima observação foi retirada da amostra, o logaritmo da função de verossimilhança com restrição nos parâmetros é denotado por $R_{(i)}(\theta|\theta_j)$. Seja $\hat{\theta}_{(i)}$ o estimador de máxima verossimilhança, sujeito às restrições, obtido a partir de $R_{(i)}(\theta|\theta_j)$, a influência da i -ésima observação no estimador de máxima verossimilhança é avaliada por meio da diferença $\hat{\theta}_{(i)} - \hat{\theta}$. Quando essa diferença $\hat{\theta}_{(i)} - \hat{\theta}$ é relativamente grande, pode-se considerar a observação como influente e maior atenção deve ser dada a esse indivíduo.

As medidas de influência global mais utilizadas são denominadas distância de Cook Generalizada e afastamento da verossimilhança. A distância de Cook é definida

como a norma padronizada de $\hat{\theta}_{(i)} - \hat{\theta}$ e é descrita por:

$$CD_i(\theta) = (\hat{\theta}_{(i)} - \hat{\theta})^T \mathbf{M}(\hat{\theta}_{(i)} - \hat{\theta}), \quad (55)$$

em que podem ser consideradas várias escolhas de $\mathbf{M}$, segundo Cook e Weisberg (1982). Entretanto, as escolhas mais utilizadas entre os pesquisadores é expressa por $\mathbf{M} = -\ddot{L}(\hat{\theta})$ ou $\mathbf{M} = [-\ddot{L}(\hat{\theta})]^{-1}$.

Uma partição do vetor de parâmetros $\theta = (\theta_1^T, \theta_2^T)^T$ pode ser considerada e uma alternativa é utilizar os valores de $CD_i(\theta_1)$ e $CD_i(\theta_2)$ que revelam o impacto do i -ésimo caso nas estimativas de θ_1 e θ_2 , respectivamente. Ver, por exemplo, Zhu e Zhang (2004).

A seguir, a medida afastamento da verossimilhança é expressa em função de $R(\hat{\theta}|\theta_j)$ e $R_{(i)}(\hat{\theta}|\theta_j)$ e é definida por:

$$LD_i(\theta) = 2[R(\hat{\theta}|\theta_j) - R_{(i)}(\hat{\theta}|\theta_j)]. \quad (56)$$

Uma das dificuldades na implementação do método de influência global é a necessidade de recalcular estimativas após a deleção dos casos. As medidas de influência local, definidas a seguir, podem auxiliar na escolha de possíveis subconjuntos de observações a serem submetidos à deleção de casos, podendo-se optar por fazer a deleção de casos individualmente ou em grupos, sem a necessidade de implementar o método de deleção de casos para todas as observações contidas em um banco de dados.

3.9 Influência Local

Proposta por Cook (1986), a metodologia de influência local consiste em avaliar o efeito de pequenas perturbações, no modelo ou nos dados, sobre os resultados de estimação e inferência. Usando métodos de geometria diferencial, busca-se avaliar a superfície de verossimilhança de forma multivariada. Nesse contexto, incorpora-se, na função de verossimilhança, uma ou mais dimensões relacionadas às suposições ou aos dados.

Considere $L(\theta)$ o logaritmo da verossimilhança do modelo postulado, sendo θ um vetor p dimensional de parâmetros desconhecidos. Um vetor ω , contido em um subespaço aberto $\Omega \subset R^q$, de perturbações é incorporado à função de verossimilhança.

Nas aplicações aqui utilizadas, a dimensão de $q = n$, sendo uma dimensão para cada observação, considerando uma função de verossimilhança em que as observações y não são consideradas fixas.

Os aspectos que podem ser avaliados em relação ao ajuste do modelo aos dados com os esquemas de perturbação adotados são ocorrências de observações atípicas ou influentes e observações mal-ajustadas pelo modelo considerado.

Considera-se que $L(\theta|\omega)$ corresponde ao logaritmo da função verossimilhança do modelo perturbado para um dado ω , de forma que exista um vetor ω_0 em Ω tal que $L(\theta|\omega_0) = L(\theta)$. Os estimadores de máxima verossimilhança sob $L(\theta)$ e $L(\theta|\omega)$ são denotados por $\hat{\theta}$ e $\hat{\theta}_\omega$. Quando $\omega = \omega_0$ as superfícies de verossimilhança e verossimilhança perturbada se tocam e os métodos de influência local consistem em medir a curvatura do afastamento das verossimilhanças a partir desse ponto onde as superfícies se encontram.

Para analisar o efeito das perturbações considera-se, inicialmente, a função de afastamento de verossimilhança

$$LD(\omega) = 2[L(\hat{\theta}) - L(\hat{\theta}_\omega)], \quad (57)$$

e a partir dessa função analisa-se a superfície de influência obtida por $LD(\omega)$ conforme ω varia por Ω .

A análise dessa superfície é realizada em torno do ponto ω_0 , considerando uma linha em Ω que passa por ω_0 dada por $\omega(a) = \omega_0 + al$, em que $a \in R^1$ e l é um vetor de norma unitária em R^q . As curvaturas normais C_l da linha $\omega(a)$ na direção l podem ser obtidas por:

$$C_l = 2|l^T \Delta^T (\ddot{L})^{-1} \Delta l|, \quad (58)$$

sendo $-\ddot{L}$ a matriz de informação observada do modelo postulado e $\Delta_{ij} = \frac{\delta^2 L(\theta|\omega)}{\delta \theta_i \delta \omega_j}$.

Cook (1986) propõe que o autovetor l_{max} associado ao autovalor C_{max} da matriz $\ddot{F} = \Delta^T (\ddot{L})^{-1} \Delta$ indica como as perturbações no modelo postulado atingem o afastamento máximo da verossimilhança. É necessário definir a estrutura de perturbação e a forma da função de verossimilhança perturbada. Neste trabalho, três formas de perturbação são consideradas:

Ponderação de casos: Esse esquema de perturbação tem certa relação com o método de deleção de casos. O logaritmo da função de verossimilhança é escrito na forma: $l(\theta|\omega) = \sum_{i=1}^n \omega_i \log(f(y; x, \theta))$ e $\omega = (1, 1, \dots, 1) = \omega_0$ representa o vetor de não perturbação.

Perturbação da resposta: Nesse caso, a perturbação é obtida ao substituir y_i por $y_i^* = y_i + \sigma\omega_i$ na função de verossimilhança. σ é o desvio padrão da resposta e o vetor de não perturbação tem a forma $\omega_0 = (0, 0, \dots, 0)$.

Perturbação de covariável contínua: Semelhante à perturbação da resposta. Nesse caso, a substituição é de uma covariável contínua x_{ji} por $x_{ji}^* = x_{ji} + \sigma\omega_i$ na função de verossimilhança.

Para todos esses esquemas de perturbação, pode-se avaliar a influência de cada observação de um conjunto de dados sobre as inferências. O cálculo das curvaturas normais (58) pode ser realizado numericamente e um gráfico dos valores obtidos pelo vetor l_{max} permite identificar pontos possivelmente influentes.

Com base nos resultados obtidos, utilizando os métodos de influência local, uma estratégia que pode ser adotada é comparar as estimativas de $\hat{\theta}$ com $\hat{\theta}_{(i)}$, sendo $\hat{\theta}_{(i)}$ o estimador de máxima verossimilhança de θ obtido após a deleção dos casos indexados pelo conjunto $I = \{i_1, i_2, \dots, i_m\}$. Essas comparações evidenciam a sensibilidade das estimativas obtidas para os parâmetros do modelo. A existência de subconjuntos dos dados que resultem em estimativas muito diferentes das obtidas para o conjunto completo de dados indica uma sensibilidade no ajuste do modelo, e a possível presença de valores extremos ou inconsistentes.

4 ESTUDOS DE SIMULAÇÃO E APLICAÇÕES

Nesta seção são apresentados exemplos de aplicações dos modelos e métodos considerados. Estudos de simulação com os modelos Rathie-Swamee simétrico e com a nova distribuição definida pela equação (20) foram realizados para avaliar o desempenho dos métodos de estimação e inferência utilizados. Para exemplificar aplicações do modelo de regressão Rathie-Swamee, conjuntos de dados de pesca de camarões e de erupção de geiser são analisados, utilizando esses modelos.

4.1 Estudo de simulação I

O primeiro estudo de simulação foi realizado para avaliar o processo de simulação do modelo Rathie-Swamee e estimação de seus parâmetros pelo método de verossimilhança e comparar o desempenho dos diferentes métodos para obtenção de intervalos de confiança descritos na seção 3.6.3. O caso particular dos modelos $RS(a=0, b, p)$, com o parâmetro a fixado em zero foi considerado.

Três cenários de diferentes valores para os parâmetros foram considerados e os casos escolhidos foram $RS(a = 0, b = 1, p = 1)$, $RS(a = 0, b = 0, 1, p = 0, 5)$ e $RS(a = 0, b = 0, 5, p = 0, 1)$. Em cada um desses casos foram geradas 500 repetições de amostras com tamanhos 30, 50 e 100.

A partir das 500 amostras simuladas para cada caso, obtiveram-se as estimativas dos parâmetros e distribuições empíricas para os estimadores de máxima verossimilhança, calculando valores médio e erro padrão, e calcularam-se as cobertura nominais e as amplitudes médias dos diferentes tipos de intervalos de confiança. Os parâmetros dos modelos foram estimados pelo método da máxima verossimilhança considerando uma função barreira adaptativa, conforme descrito na seção de metodologia. A função *constrOptim* foi utilizada, vide rotinas contidas no Apêndice A.

A Tabela 2 apresenta as médias das estimativas obtidas e os erros-padrão para os parâmetros b e p considerados. Os valores médios das estimativas estão em concordância com os valores definidos para os parâmetros nas simulações realizadas. O método de otimização utilizado foi robusto em relação à escolha dos valores iniciais, no sentido em que os diferentes valores iniciais aleatórios adotados resultam na convergência para o mesmo ponto de máximo global.

Tabela 2 - Médias e (erros-padrão) das estimativas de máxima verossimilhança $\hat{b}$ e $\hat{p}$ para o modelo (6) com os diferentes tamanhos de amostra e valores verdadeiros selecionados para os parâmetros

n	Caso 1 ($b = 1$ $p = 1$)		Caso 2 ($b = 0,1$ $p = 0,5$)		Caso 3 ($b = 0,5$ $p = 0,1$)	
	$\hat{b}$	$\hat{p}$	$\hat{b}$	$\hat{p}$	$\hat{b}$	$\hat{p}$
30	0,9977 (0,2018)	1,0933 (0,3218)	0,0999 (0,0480)	0,5786 (0,2581)	0,4802 (0,1205)	0,1753 (0,1652)
50	0,9932 (0,1510)	1,0578 (0,2432)	0,0983 (0,0375)	0,5529 (0,1910)	0,4939 (0,0911)	0,1321 (0,1155)
100	1,0016 (0,1070)	1,0184 (0,1642)	0,0993 (0,0268)	0,5224 (0,1260)	0,4947 (0,0701)	0,1185 (0,0894)

Os aspectos avaliados nos intervalos de confiança calculados foram as coberturas nominais e as amplitudes médias dos intervalos de confiança. Um coeficiente de confiança de 95% foi definido nos cálculos para obtenção dos intervalos. As Tabelas 3, 4 e 5 exibem os coeficientes de coberturas nominais e os valores das médias das amplitudes dos diferentes intervalos de confiança obtidos para cada caso simulado.

Tabela 3 - Coberturas nominais e (médias de amplitudes) dos intervalos de confiança com os métodos: bootstrap paramétrico (BP), bootstrap não-paramétrico (BNP), Transformação dos Parâmetros (TP), normal e perfilado

n		BP	BNP	TP	Normal	Perfilado
30	$b = 1$	0,9280 (0,7968)	0,9380 (0,7855)	0,9420 (0,7714)	0,9360 (0,7529)	0,9380 (0,7530)
		0,9040 (1,3493)	0,9260 (1,3748)	0,9300 (1,3174)	0,9580 (1,2442)	0,9500 (1,2450)
	$p = 1$	0,9560 (0,5993)	0,9460 (0,5961)	0,9500 (0,5897)	0,9460 (0,5812)	0,9540 (0,5813)
		0,9240 (0,9795)	0,9400 (1,0039)	0,9500 (0,9775)	0,9580 (0,9444)	0,9560 (0,9445)
50	$b = 1$	0,9460 (0,4181)	0,9440 (0,4181)	0,9480 (0,4156)	0,9520 (0,4127)	0,9520 (0,4127)
		0,9280 (0,6610)	0,9400 (0,6749)	0,9380 (0,6640)	0,9580 (0,6526)	0,9520 (0,6526)
	$p = 1$	0,9280 (0,6610)	0,9400 (0,6749)	0,9380 (0,6640)	0,9580 (0,6526)	0,9520 (0,6526)
		0,9280 (0,6610)	0,9400 (0,6749)	0,9380 (0,6640)	0,9580 (0,6526)	0,9520 (0,6526)
100	$b = 1$	0,9460 (0,4181)	0,9440 (0,4181)	0,9480 (0,4156)	0,9520 (0,4127)	0,9520 (0,4127)
		0,9280 (0,6610)	0,9400 (0,6749)	0,9380 (0,6640)	0,9580 (0,6526)	0,9520 (0,6526)
	$p = 1$	0,9280 (0,6610)	0,9400 (0,6749)	0,9380 (0,6640)	0,9580 (0,6526)	0,9520 (0,6526)
		0,9280 (0,6610)	0,9400 (0,6749)	0,9380 (0,6640)	0,9580 (0,6526)	0,9520 (0,6526)

Os resultados das Tabelas 3, 4 e 5 indicam que nenhum método de obtenção de intervalos se sobressai aos demais. De uma forma geral, os diferentes métodos apresentaram resultados similares, com coberturas nominais abaixo dos 95% fixados. O método de transformação dos parâmetros pode resultar em intervalos com amplitude infinita, devido à escolha da transformação, nos casos em que as estimativas para o parâmetro considerado se aproximam de zero. Esses casos foram desconsiderados nos cálculos de coberturas nominais e médias de amplitudes e ocorreram apenas nas situações em que se fixou o parâmetro $p = 0, 1$.

Tabela 4 - Coberturas nominais e (médias de amplitudes) dos intervalos de confiança com os métodos: bootstrap paramétrico (BP), bootstrap não-paramétrico (BNP), Transformação dos Parâmetros (TP), normal e perfilado

n		BP	BNP	TP	Normal	Perfilado
30	$b = 0,1$	0,9040	0,9260	0,9480	0,8840	0,9360
		(0,1771)	(0,1782)	(0,2154)	(0,1857)	(0,1834)
	$p = 0,5$	0,8880	0,8940	0,8960	0,9480	0,9400
		(1,0231)	(1,0278)	(1,0957)	(0,9346)	(0,9084)
50	$b = 0,1$	0,9060	0,9280	0,9380	0,8880	0,9380
		(0,1398)	(0,1414)	(0,1574)	(0,1437)	(0,1469)
	$p = 0,5$	0,8920	0,9100	0,9140	0,9400	0,9280
		(0,7408)	(0,7536)	(0,7766)	(0,7121)	(0,7108)
100	$b = 0,1$	0,9420	0,9460	0,9460	0,9360	0,9460
		(0,1017)	(0,1025)	(0,1079)	(0,1031)	(0,1052)
	$p = 0,5$	0,9400	0,9500	0,9480	0,9480	0,9460
		(0,5018)	(0,5078)	(0,5119)	(0,4918)	(0,4921)

Tabela 5 - Coberturas nominais e (médias de amplitudes) dos intervalos de confiança com os métodos: bootstrap paramétrico (BP), bootstrap não-paramétrico (BNP), Transformação dos Parâmetros (TP), normal e perfilado

n		BP	BNP	TP	Normal	Perfilado
30	$b = 0,5$	0,9160	0,9420	0,9720	0,9300	0,9080
		(0,4535)	(0,4574)	(0,5388)	(0,5130)	(0,4121)
	$p = 0,1$	0,8860	0,9100	0,8200	0,9740	0,9600
		(0,6291)	(0,6576)	(1,4772)	(0,6987)	(0,5288)
50	$b = 0,5$	0,9300	0,9480	0,9720	0,9520	0,9300
		(0,3613)	(0,3653)	(0,4164)	(0,4047)	(0,3326)
	$p = 0,1$	0,9260	0,9500	0,8960	0,9780	0,9700
		(0,4360)	(0,4622)	(1,4565)	(0,5200)	(0,3911)
100	$b = 0,5$	0,9340	0,9500	0,9600	0,9480	0,9060
		(0,2589)	(0,2613)	(0,2900)	(0,2860)	(0,2489)
	$p = 0,1$	0,9260	0,9440	0,8820	0,9760	0,9560
		(0,3418)	(0,3454)	(0,8783)	(0,3700)	(0,3208)

Pelo método de aproximação quadrática da superfície de verossimilhança, em que a distribuição normal é utilizada, alguns casos resultaram em intervalos de confiança com limites inferiores negativos, apesar do espaço paramétrico ser restrito aos números reais não negativos. Nesses casos, a amplitude dos intervalos foi considerada com limites inferiores iguais a zero. Apesar desse truncamento na amplitude dos intervalos normais, esses não tiveram amplitudes médias menores que os demais métodos.

Nos casos em que o parâmetro b foi definido com valor $b = 0,1$, o desempenho dos intervalos normais foi pior que dos demais métodos. Nos demais casos e valores definidos para os parâmetros, os intervalos normais tiveram desempenho igual ou superior aos outros métodos. Considerando o custo computacional elevado associado aos intervalos do tipo Bootstrap e do tipo perfilado, e o desempenho semelhante desses métodos, os intervalos normais se mostram muito vantajosos, devido a sua simplicidade de cálculo, de-

pendendo apenas dos resultados de estimação pontual, com erros padrão obtidos a partir da matriz hessiana da superfície de verossimilhança no ponto de ótimo.

4.2 Estudo de simulação II

Este estudo de simulação foi realizado com o objetivo de avaliar a performance do método de estimação por máxima verossimilhança para o modelo (20), que representa uma classe de modelos que inclui as distribuições resultantes de produtos e quocientes entre variáveis independentes com distribuições RS(a=0,b,p). Ao todo foram geradas 19500 amostras de valores pseudo-aleatórios, seguindo o modelo probabilístico descrito pelas equações (20) e (23). Treze valores diferentes para o parâmetro α foram escolhidos, $\alpha = -3, -2, 5, 2, \dots, 3$. Para cada valor do parâmetro, foram simuladas amostras de tamanhos 30, 50 e 100, e em cada amostra obteve-se uma estimativa para o parâmetro, obtendo-se uma distribuição amostral empírica. O ambiente R de programação (R Development Core Team, 2011) foi utilizado para as simulações e estimações.

As Tabelas 6 e 7 apresentam as médias e erros-padrão das estimativas obtidas para cada tamanho de amostra e valor fixado para o parâmetro. Na Tabela 6, estão exibidos os resultados para os valores não negativos de α , enquanto a Tabela 7 contém os resultados para os valores negativos de α .

Tabela 6 - Médias e (erros-padrão) dos estimadores de máxima verossimilhança $\hat{\alpha}$ para o modelo de quocientes entre variáveis Rathie-Swamee

	$\alpha = 0$	$\alpha = .5$	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 2.5$	$\alpha = 3$
$n = 30$	0,0356 (0,2287)	0,4053 (0,3136)	0,9884 (0,4126)	1,4881 (0,3525)	1,9960 (0,4130)	2,4462 (0,4160)	2,8530 (0,4086)
$n = 50$	0,0209 (0,2111)	0,4666 (0,3379)	0,9983 (0,2468)	1,5116 (0,2833)	1,9842 (0,3094)	2,4280 (0,3118)	2,8438 (0,3340)
$n = 100$	0,0447 (0,3119)	0,4923 (0,2657)	1,0251 (0,2390)	1,5057 (0,2279)	1,9818 (0,2301)	2,4525 (0,2402)	2,8656 (0,2693)

O parâmetro foi estimado satisfatoriamente, e em todos os casos, as estimativas estão em concordância com os valores verdadeiros dos parâmetros. Como previsto, os erros-padrão são inversamente relacionados com os tamanhos amostrais, de forma que quanto maior o tamanho da amostra, menor o valor do erro padrão, indicando maior precisão das estimativas.

As equações (31) a (35) apresentam formulações distintas para o cálculo da

Tabela 7 - Médias e (erros-padrão) dos estimadores de máxima verossimilhança $\hat{\alpha}$ para o modelo de produtos entre variáveis Rathie-Swamee

Valores Negativos	$\alpha = -.5$	$\alpha = -1$	$\alpha = -1.5$	$\alpha = -2$	$\alpha = -2.5$	$\alpha = -3$
$n = 30$	-0,4323 (0,2928)	-1,0011 (0,3569)	-1,4815 (0,3818)	-2,0437 (0,5381)	-2,5823 (0,5182)	-3,0282 (0,4752)
$n = 50$	-0,4541 (0,2547)	-0,9611 (0,2584)	-1,4810 (0,3306)	-2,1098 (0,5376)	-2,6713 (0,5717)	-3,1313 (0,4781)
$n = 100$	-0,4417 (0,2267)	-0,9648 (0,2339)	-1,4824 (0,2703)	-2,0592 (0,4253)	-2,7195 (0,5676)	-3,2733 (0,4144)

função de distribuição acumulada descrita em (20) no caso particular em que $\alpha = 1$. As diferenças entre essas formas são referentes ao tipo de função especial utilizada em cada uma. Para avaliar diferenças entre essas formas, elas foram comparadas em relação ao tempo gasto para computação das mesmas. Dessa forma, as equações (31), (33), (34) e (35) foram usadas no cálculo das probabilidades acumuladas para os quantis apresentados na Tabela 1. Usando o software Maple, as diferentes formulações tiveram seus tempos de computação calculados para os diferentes quantis.

A equação (32) não foi utilizada devido à lentidão observada para a convergência das somas infinitas duplas envolvidas, demorando 100 vezes mais do que com a utilização das outras formas. Apesar de sua lenta computação, essa equação (32), é de interesse teórico, pois expressa a função $F_1(v)$ de forma simples, sem envolver o uso de funções especiais.

A precisão obtida pelas diferentes formas de expressar as funções foi bastante semelhante e os valores calculados são essencialmente iguais. As diferenças observadas são da ordem de 10^{-5} .

A Tabela 8 apresenta os tempos gastos por método para cada quantil selecionado, e as médias de tempos gastos por método. As equações usando as funções hipergeométricas foram calculadas mais rapidamente do que aquelas envolvendo as funções Ψ ou integração numérica.

4.3 Análise dos dados de pesca de camarões na costa do Rio Grande do Norte

Os dados utilizados nesta seção são referentes a medidas biométricas realizadas em camarões da espécie *farfantepenaeus brasiliensis* coletados em três regiões do estado do Rio Grande do Norte. Os dados foram coletados por Pinheiro (2008), e apresentados junto à análise morfométrica e molecular dos camarões. O autor também utilizou

Tabela 8 - Tempo de CPU, em segundos, gasto para calcular a distribuição de probabilidade $F_1(v)$ para quocientes entre variáveis logísticas, usando as diferentes equações propostas na seção 3.3

v	(IntNum)	(31)	(33)	(34)	(35)
0.304	0.327	0.078	0.078	0.655	0.717
0.709	0.343	0.062	0.078	0.655	0.639
1.41	0.358	0.078	0.078	0.702	0.624
2.057	0.358	0.078	0.062	0.639	0.639
3.288	0.343	0.078	0.078	0.670	0.577
6.839	0.343	0.109	0.078	0.717	0.546
13.816	0.343	0.109	0.062	0.639	0.436
34.639	0.358	0.124	0.062	0.702	0.733
Tempo médio	0.346	0.089	0.072	0.672	0.613

os dados de biometria para fazer algumas análises de variância após transformações nos dados.

O banco de dados, disponível no Apêndice B, é constituído por observações de 4 variáveis referentes ao peso, comprimento, sexo e localidade de pesca de 120 camarões *farfantepenaeus brasiliensis*. Os camarões observados tinham pesos entre 1,98 g e 35,78 g, com média de 15,75 g e desvio padrão de 8,38 g. Os comprimentos apresentaram variações entre 39,6 mm e 99,2 mm, com média de 73,16 mm e desvio padrão de 13,42 mm. A Tabela 9 apresenta a distribuição dos camarões de cada sexo por localidade de pesca.

Tabela 9 - Distribuição dos camarões de cada sexo por localidade de pesca

Localidade	Sexo		Total
	Feminino	Masculino	
Baia Formosa	22	15	37
Diogo Lopes	21	11	32
Touros	42	9	51
Total	85	35	120

Como a medida para comercialização do camarão é o peso, essa é a medida adotada como variável resposta, e a proposta é ajustar um modelo de regressão considerando as demais variáveis como explicativas para o modelo. A correlação linear entre peso e comprimento nos dados observados é de 0,9637. A Figura 7 exhibe o gráfico de dispersão do peso pelo comprimento dos 120 camarões observados, confirmando a grande associação existente entre as variáveis.

Na Figura 8 podem-se observar os diagramas de Box-plot do peso dos camarões separados por localidade de pesca e por sexo. Podem-se notar grandes diferenças nas médias e variabilidades dos diferentes grupos. Na localidade de Touros, foram co-

letados camarões em média mais pesados que nas outras localidades, fato possivelmente decorrente da predominância de camarões do sexo feminino coletados naquela localidade. Os camarões do sexo feminino apresentaram maior média e variabilidade dos pesos observados.

Figura 7 - Gráfico de dispersão entre as variáveis comprimento em milímetros e peso em gramas dos 120 camarões *farfantepenaeus brasiliensis* observados

Figura 8 - Boxplot dos pesos dos camarões separados por localidades e por sexo

O histograma dos pesos dos camarões, apresentado na Figura 9, evidencia a bimodalidade na distribuição desses pesos. Os modelos Rathie-Swamee simétrico e assimétrico permitem considerar essa bimodalidade e se ajustam aos dados, como pode-se ver pelas linhas referente às funções de densidade ajustadas que acompanham o histograma na Figura 9. Existe uma pequena assimetria nos dados observados, mas as funções de densidade dos modelos simétrico e assimétrico ajustados não são muito distintas. Na Figura 10, está apresentada a função de distribuição empírica das funções de distribuição ajustadas com os modelos simétrico e assimétrico para os dados de peso. Pode-se notar nessa figura que as distribuições ajustadas estão bem próximas da distribuição empírica dos dados.

Na Tabela 10, são apresentadas as estimativas dos parâmetros e seus erros-padrão, obtidos no ajuste aos dados de peso dos camarões. As medidas de AIC calculadas para os modelos, também, são apresentadas na mesma tabela e indicam melhor desempenho do modelo simétrico, validando a impressão causada pela inspeção dos gráficos das Figuras 9 e 10. Apesar de haver uma pequena assimetria nos dados observados, o parâmetro de assimetria não melhora o ajuste do modelo significativamente, considerando os valores de AIC observados e TRVs construídos.

Figura 9 - Histograma dos pesos dos camarões observados e funções de densidade dos modelos Rathie-Swamee ajustados

Tabela 10 - Estimativas para os parâmetros dos modelos Rathie-Swamee simétrico e assimétrico ajustados aos pesos dos camarões

	Estimativa	Erro padrão
Modelo RS		
b	0,0477	0,0141
p	0,6209	0,1212
μ	15,72	0,2742
$l(\theta)$	-413,638	
AIC	833,276	
Modelo RSA		
b	0,0473	0,0140
p	0,6248	0,1217
c	0,1730	0,2092
μ	15,6533	0,3184
$l(\theta)$	-412,5929	
AIC	833,1858	

Na Tabela 11, estão apresentadas as estimativas dos parâmetros do modelo de mistura de distribuições normais ajustado aos dados de peso dos camarões. Esse modelo com a mistura de dois componentes normais é composto por 5 parâmetros, sendo dois parâmetros de média, dois parâmetros de variância e um parâmetro que define o peso dos componentes na composição da mistura das distribuições.

Pode-se comparar o desempenho do modelo Rathie-Swamee, no ajuste de dados com comportamento bimodal, com o modelo de mistura. Apesar dos diferentes modelos apresentarem valores próximos para as funções de verossimilhança, o aumento

Figura 10 - Funções de distribuições acumuladas ajustadas com modelos Rathie-Swamee e empírica, para pesos dos camarões observados

Tabela 11 - Estimativas dos parâmetros do modelo de mistura de distribuições normais ajustado aos pesos dos camarões

Modelo de mistura de Normais	$\hat{\mu}$	$\hat{\sigma}^2$	λ
Componente 1	8.6249	3.5275	0.497
Componente 2	22.7944	5.1465	0.503
$l(\theta) = 412.5957$			
$AIC = 835.1915$			

no número de parâmetros ocasionado pela mistura de distribuições normais mostra-se desvantajoso pelo cálculo do critério de AIC .

Com o objetivo de verificar o efeito das variáveis de localidade, sexo e comprimento sobre as médias de pesos dos camarões coletados, consideraram-se os modelos de regressão baseados nas equações (42) e (43). No processo de ajuste dos modelos, notou-se grande instabilidade no método de otimização que obtém convergência para pontos de máximo locais. A superfície de verossimilhança aparenta ser achatada, tornando a otimização altamente dependente dos valores iniciais escolhidos, especialmente para o modelo completo, dada a dimensão da superfície e não ortogonalidade com possíveis problemas de identificabilidade no modelo.

Nas diferentes aplicações dos modelos de regressão Rathie-Samee,

sequências aleatórias de valores iniciais foram utilizadas para busca dos pontos de máximo global das funções de verossimilhança.

Considerando a inclusão da covariável comprimento, fortemente relacionada aos pesos dos camarões, buscou-se, dentre a classe de modelos RS, o modelo de regressão que mais se adequasse aos dados observados. O modelo escolhido foi o caso mais simples do modelo Rathie-Swamee, em que os parâmetros a e p assumem valor igual a zero, resultando no modelo logístico como caso particular. A escolha desse modelo foi realizada considerando os valores de AIC dos modelos, e testes de razão de verossimilhança para esses modelos encaixados, para os quais não se rejeitou a hipótese de que os parâmetros a , p e c fossem iguais a zero nessa aplicação.

É interessante notar que, ao se introduzirem as covariáveis no modelo, incorporando informações adicionais dessas variáveis, o modelo ajustado é um caso unimodal, apesar do comportamento bimodal dos dados e dos modelos ajustados sem covariáveis. Nesse sentido, pode-se interpretar que a incorporação das covariáveis agregou informações que permitiram a utilização de um modelo mais simples e melhorar a qualidade do ajuste, reduzindo os erros e melhorando o poder preditivo do modelo. Nesse caso específico, a bimodalidade dos dados pode ter sido causada pela grande diferença entre os pesos dos diferentes sexos e essa covariável foi observada, mas nos casos em que a variável que distingue os grupos for latente ou não observada, os modelos com comportamento multimodal podem descrever o comportamento dos diferentes grupos.

A Tabela 12 apresenta os resultados dos modelos de regressão ajustados, considerando a inclusão de diferentes conjuntos de covariáveis. Os parâmetros foram atribuídos de forma que β_0 é o intercepto do modelo linear e β_1 e β_2 estão associados às covariáveis *comprimento* e *comprimento*², respectivamente. Os parâmetros β_3 , β_4 e β_5 são referentes, respectivamente, às variáveis qualitativas *sexo = masculino*, *local = DiogoLopes* e *local = Touros*, de forma que as covariáveis foram codificadas e assumem valor 1 quando a observação pertence àquela categoria e 0, caso contrário.

Tabela 12 - Estimativas e (erros padrão) para os parâmetros dos modelos de regressão ajustados para o conjunto de dados de camarões

Modelo		b	β_0	β_1	β_2	β_3	β_4	β_5
$RS(b, \beta_0, \beta_1)$	$l(\hat{\theta}) = -258,1953$	0,8620	-28,0524	0,4242	—	—	—	—
	$AIC = 522,3906$	(0,0661)	(1,1835)	(0,0112)	—	—	—	—
$RS(b, \beta_0, \beta_1, \beta_2)$	$l(\hat{\theta}) = -163,5567$	1,9144	14,9679	-0,4690	0,0045	—	—	—
	$AIC = 355,1134$	(0,1469)	(1,8013)	(0,0377)	(0,0002)	—	—	—
$RS(b, \beta_0, \beta_1, \beta_2, \beta_3)$	$l(\hat{\theta}) = -159,5285$	1,9833	13,8781	-0,4365	0,0043	-0,5683	—	—
	$AIC = 329,057$	(0,1524)	(1,7338)	(0,0372)	(0,0002)	(0,1964)	—	—
$RS(b, \beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$	$l(\hat{\theta}) = -157,7024$	2,0108	13,3202	-0,5485	-0,4264	0,0042	0,1800	0,3930
	$AIC = 329,4048$	(0,1543)	(1,9904)	(0,2010)	(0,0395)	(0,0002)	(0,2626)	(0,2044)

Na Tabela 12, nota-se o efeito da inclusão das covariáveis sobre os valores dos logaritmos das verossimilhanças dos modelos, indicando a significância delas em relação à média de peso dos camarões observados. O modelo que inclui as covariáveis *comprimento*, *comprimento*² e *sexo* foi significativamente superior aos demais e tem seus resultados apresentados na Tabela 13, que inclui estimativas pontuais e intervalares para os parâmetros do modelo selecionado.

Tabela 13 - Estimativas e intervalos de confiança ($\gamma = 95\%$) para os parâmetros do modelo de regressão selecionado para o conjunto de dados de camarões

Parâmetro	Estimativa	Limite inferior	Limite superior
b	1.9833	1.6846	2.2820
<i>Intercepto</i>	13.8781	10.4799	17.2763
<i>Comprimento</i>	-0.4365	-0.5094	-0.3636
<i>Comprimento</i> ²	0.0043	0.0039	0.0047
<i>Sexo = Masculino</i>	-0.5683	-0.9532	-0.1834

Para avaliar a qualidade do ajuste do modelo utilizado, as medidas de resíduos ordinários padronizados e resíduos quantílicos foram calculadas e apresentadas de forma gráfica nas Figuras 11 e 12.

Figura 11 - Gráfico dos comprimentos dos camarões observados versus resíduos padronizados obtidos pelo modelo selecionado

Figura 12 - Gráfico de probabilidade normal dos resíduos quantílicos

Os resíduos padronizados estão distribuídos em torno de zero de forma aleatória como esperado, destacando as observações 92 e 100 das demais. O gráfico de probabilidade normal com os resíduos quantílicos apresenta uma concordância entre os quantis teóricos e os observados, indicando a adequação do modelo probabilístico ajustado aos dados. As observações 92 e 100 também são destacadas dentre os resíduos quantílicos e demandam atenção especial.

As observações 92 e 100 são referentes a camarões do sexo feminino mais pesadas que os demais animais com comprimentos semelhantes. Na Figura 7, essas observações são as que se encontram acima da reta que relaciona as variáveis, estando um pouco fora do padrão apresentado pelas demais

As medidas de influência local permitem avaliar o efeito de cada uma das observações sobre as estimativas obtidas para o modelo. A Figura 13 apresenta os resultados da análise de influência local para os esquemas de ponderação de casos, perturbação da covariável comprimento e perturbação da variável resposta.

Figura 13 - Gráficos das medidas de influência local para o modelo ajustado para os pesos de camarões

Sob o esquema de perturbação com ponderação de casos, as observações 92 e 100 mais uma vez se destacaram das demais, indicando que essas observações podem estar influenciando de forma desproporcional as estimativas obtidas para o modelo.

Dessa forma é interessante considerar a metodologia de deleção de casos para comparar os resultados obtidos para o modelo de regressão quando essas observações são excluídas individualmente ou de forma conjunta do banco de dados em análise.

Uma forma tradicional de realizar a análise de deleção de casos é excluir uma a uma, todas as observações do conjunto de dados e graficamente comparar as medidas (55) e (56) para todas as observações para visualizar se uma ou mais observações se destacam das demais, indicando uma possível influência desproporcional sobre as estimativas obtidas para os parâmetros do modelo.

Devido à instabilidade na obtenção dos pontos de máximo global para os modelos Rathie-Swamee de regressão, esse método extensivo de deleção de casos se torna desaconselhável ou impraticável, pois para cada uma das $n + 1$ vezes em que o modelo é estimado seria necessário utilizar sequências aleatórias de valores iniciais para busca do ponto de máximo global na superfície de verossimilhança. Dessa forma serão considerados para deleção de casos apenas as observações identificadas como possivelmente influente nas análises de resíduos e influência local.

A Tabela 14 apresenta as estimativas para os parâmetros e os respectivos erros-padrão considerando três diferentes sub-amostras dos dados de camarões. A sub-amostra indexada por $I(92, 100)$ tem as observações 92 e 100 excluídas conjuntamente dos dados, enquanto as sub-amostras indexadas por $I(92)$ e $I(100)$ têm as observações 92 e 100 excluídas de forma individual.

Tabela 14 - Estimativas e (erros-padrão) para os parâmetros dos modelos de regressão ajustados para diferentes sub-amostras do conjunto de dados de camarões

Sub-amostra	p	<i>Intercepto</i>	<i>Sexo</i>	<i>Comprimento</i>	<i>Comprimento²</i>
I=(92,100)	2.1613 (0.1660)	13.6789 (1.6348)	-0.5603 (0.1839)	-0.4316 (0.0351)	0.0043 (0.0002)
I=(92)	2.0548 (0.1579)	13.3894 (1.6968)	-0.5898 (0.1913)	-0.4247 (0.0364)	0.0042 (0.0002)
I=(100)	2.0818 (0.1598)	14.1599 (1.6736)	-0.5388 (0.1891)	-0.4432 (0.0359)	0.0043 (0.0002)

Comparando os resultados na Tabela 14, obtidos após a exclusão das observações possivelmente influentes com os resultados obtidos para o banco de dados completo na Tabela 12, pode-se perceber que existem pequenas variações nas estimativas dos parâmetros, mas sem alterações na interpretação dos resultados e inferências. Dessa

forma, as observações 92 e 100 não estão influenciando os resultados de forma desproporcional e devem ser mantidas no ajuste do modelo, que se mostrou adequado para descrever o peso dos camarões observados.

4.4 Análise dos dados de erupção de gêiser

Os dados utilizados nesta seção são referentes à duração e ao intervalo entre ocorrências de erupções em um gêiser. O serviço do parque nacional de Yellowstone indica para os visitantes um valor predito para o tempo até a próxima erupção, e, considerando esse problema, escolhe-se esta variável como variável resposta nesta aplicação. O banco de dados está disponível no pacote *datasets* do ambiente R (R Development Core Team, 2011) com o nome *faithful*, e possui 272 observações de 2 variáveis referentes à duração das erupções em minutos e tempo de espera até as erupções em minutos, coletadas no mês de Outubro de 1980.

Nos registros, foram observados tempos de espera entre 43 e 96 minutos, com média de 70,9 minutos e desvio padrão de 13,5949 minutos. Os tempos de duração das erupções variaram entre 1,6 e 5,1 minutos, com média de 3,48 minutos e desvio padrão de 1,1413. A correlação linear entre as variáveis observadas tempo de duração e tempo de espera foi de 0,9, e a Figura 14 apresenta gráfico de dispersão das observações, indicando uma relação linear existente entre as variáveis observadas.

Figura 14 - Gráfico de dispersão entre as variáveis tempo de duração e tempo de espera referentes a erupções do gêiser Old Faithful

Para estabelecer qual tipo de modelo de regressão utilizar para descrever a relação linear entre as variáveis, é necessário pensar em possíveis distribuições de probabilidades que possam ser utilizadas para modelar a variável resposta. Um comportamento bimodal das observações pode ser visualizado na Figura 15, que apresenta um histograma das observações. As distribuições RS com sua flexibilidade e possíveis formas com uma ou mais modas podem ser adequadas para modelar o tempo de espera entre erupções, enquanto a distribuição normal e outras clássicas da família exponencial de distribuições demandariam transformações nos dados para que os modelos tivessem as suposições satisfeitas.

Figura 15 - Histograma dos tempos de espera até a ocorrência das erupções e funções de densidade dos modelos Rathie-Swamee ajustados

O ajuste dos modelos Rathie-Swamee simétrico e assimétrico aos tempos de espera pode ser visualizado nas Figuras 15 e 16. Os parâmetros dos modelos (16) e (19) foram estimados pelo método da máxima verossimilhança, considerando uma função barreira adaptativa, conforme descrito na seção de metodologia. A função *constrOptim* com o método *BFGS* foi utilizada, vide programa contido no apêndice A. A Tabela 15 apresenta as estimativas obtidas e erros-padrão associados, assim como o valor do logaritmo da função de verossimilhança no ponto de máximo para comparação dos modelos.

O melhor ajuste do modelo assimétrico, visível nas Figuras 15 e 16, é confirmado. Fazendo um teste de razão de verossimilhança para testar a significância do

Figura 16 - Funções de distribuições acumuladas ajustadas com modelos Rathie-Swamee e empírica, para tempos de espera até as erupções

Tabela 15 - Estimativas e (erros padrão) para parâmetros dos modelos Rathie-Swamee simétrico (RS) e assimétrico (RSA) ajustados aos tempos de espera até as erupções do geiser Old Faithful

	Estimativa	Erro padrão
Modelo RS		
a	0.0226	0.0117
b	0.0015	0.0014
p	1.4719	0.2811
μ	67.5680	0.4020
$l(\theta)=-1043.59$		
$AIC = 2095,18$		
Modelo RSA		
a	0.0277	0.0091
b	0.0007	0.0006
p	1.6604	0.2586
c	0.6907	0.0946
μ	65.9342	0.6183
$l(\theta)=-1033.643$		
$AIC = 2077,286$		

Tabela 16 - Estimativas para modelo de mistura de distribuições normais com dois componentes

modelo de misturas normais	$\hat{\mu}$	$\hat{\sigma}^2$	λ
Componente 1	54.6149	5.8712	0.3608
Componente 2	80.0910	5.8677	0.6391
$l(\theta) - 1034.0017$			
$AIC = 2078.0034$			

parâmetro de assimetria obtém-se o valor de p igual a 8×10^{-6} , rejeitando a hipótese de que o parâmetro de assimetria seja igual a zero.

Na Tabela 16, estão apresentadas as estimativas dos parâmetros do modelo de mistura de distribuições normais ajustado aos dados de erupção do geiser. Esse modelo com a mistura de dois componentes normais é composto por 5 parâmetros, sendo dois parâmetros de média, dois parâmetros de variância e um parâmetro que define o peso dos componentes na composição da mistura das distribuições.

O desempenho do modelo Rathie-Swamee pode ser comparado no ajuste de dados com comportamento bimodal, com o modelo de mistura. Pode-se notar pelo cálculo do critério AIC que o modelo Rathie-Swamee assimétrico tem ajuste superior ao do modelo de mistura de duas distribuições normais.

Adicionando as informações sobre a duração das erupções podem-se criar modelos de regressão e realizar predições sobre os tempos de espera até erupções. Os modelos Rathie-Swamee de regressão nas formas (42) e (43) são utilizados considerando um parâmetro fixo $\sigma = 1$, para evitar problemas de falta de identificabilidade. A matriz de covariáveis X é formada pela primeira coluna com todos os elementos iguais a um e a segunda coluna contendo os tempos de duração das erupções observadas. Desta forma o vetor de parâmetros associado às covariáveis é $\beta = (\beta_0, \beta_1)^T$ sendo β_0 o parâmetro referente ao intercepto da relação linear entre o tempo de espera e a duração das erupções e β_1 o parâmetro referente ao efeito de cada minuto de duração de uma erupção sobre o tempo de espera até a próxima erupção.

Na Tabela 17, estão apresentadas as estimativas e erros padrão dos parâmetros dos modelos de regressão Rathie-Swamee. Foram considerados diversos casos particulares dos modelos simétricos e assimétricos. Para cada um dos casos, além das estimativas, está apresentado o valor do logaritmo da função de verossimilhança nos pontos de máximo de cada modelo ajustado. Pode-se notar um aumento significativo do

logaritmo da função de verossimilhança mediante a inclusão de cada um dos parâmetros considerados entre os diferentes modelos. Dessa forma, o modelo selecionado, utilizando testes de razão de verossimilhança, é o modelo de regressão Rathie-Swamee assimétrico completo, conforme descrito na equação (43).

Tabela 17 - Estimativas e (erros padrão) para os parâmetros dos modelos de regressão Rathie-Swamee simétrico (RS) e assimétrico (RSA) ajustados aos tempos de espera até as erupções do geiser Old Faithful

Modelo		a	b	p	c	β_0	β_1
$RSA(a, b, p, c, \beta_0, \beta_1)$	$l(\theta) = -862,325$	0,2358	5×10^{-7}	4,7347	-0,5862	35,9039	10,7521
	$AIC = 1664,651$	(0,0229)	(2×10^{-6})	(1,5935)	(0,2777)	(1,3941)	(0,2715)
$RSA(b, p, c, \beta_0, \beta_1)$	$l(\theta) = -865,317$	—	0,0196	0,6964	11,3839	22,6010	10,6186
	$AIC = 1740,635$	—	(0,0053)	(0,0922)	(4,2085)	(0,5732)	(0,1433)
$RSA(b, c, \beta_0, \beta_1)$	$l(\theta) = -870,890$	—	0,2540	—	0,8198	29,9598	10,7157
	$AIC = 1749,79$	—	(0,0268)	—	(0,3930)	(1,6320)	(0,3064)
$RS(a, b, p, \beta_0, \beta_1)$	$l(\theta) = -866,128$	0,2575	2×10^{-5}	3,2326	—	33,7579	10,7371
	$AIC = 1742,254$	(0,0189)	(7×10^{-5})	(1,0798)	—	(1,1271)	(0,3043)
$RS(b, p, \beta_0, \beta_1)$	$l(\theta) = -869,922$	—	0,2071	0,1730	—	32,2563	10,9431
	$AIC = 1747,845$	—	(0,0274)	(0,0609)	—	(0,5265)	(0,1281)
$RS(b, \beta_0, \beta_1)$	$l(\theta) = -873,772$	—	0,2921	—	—	33,4461	10,6989
	$AIC = 1753,546$	—	(0,0146)	—	—	(1,1646)	(0,3168)

A Tabela 18 apresenta os resultados do modelo selecionado. São apresentadas as estimativas pontuais e intervalares para os parâmetros do modelo. Os intervalos de confiança com $\gamma = 95\%$ foram construídos considerando a normalidade assintótica dos estimadores de máxima verossimilhança, uma vez que esse método se mostrou adequado no estudo de simulação. Para o parâmetro b , que foi estimado com valor muito próximo de zero, o intervalo foi truncado em seu limite inferior para zero, uma vez que esse parâmetro só pode assumir valores maiores do que ou iguais a zero.

Tabela 18 - Estimativas e intervalos de confiança ($\gamma = 95\%$) para os parâmetros do modelo de regressão selecionado para o conjunto de dados de erupção do geiser Old Faithful

Parâmetro	Estimativa	Limite inferior	Limite superior
a	0.2358	0.1909	0.2807
b	5×10^{-7}	0.0000	4×10^{-6}
p	4.7347	1.6115	7.8579
c	-0.5862	-1.1305	-0.0419
Intercepto	35.9039	33.1715	38.6363
Duração	10.7521	10.2200	11.2842

A validade do modelo ajustado e de resultados de predições e inferências

depende da qualidade do ajuste do modelo e das suposições adotadas em relação ao modelo e aos dados observados. A seguir, alguns gráficos com as medidas de resíduo e influência descritas na seção 3 são apresentados para análise de diagnóstico do modelo ajustado.

As Figuras 17 e 18 apresentam gráficos com os resíduos padronizados e resíduos quantílicos. Nessas figuras, não aparecem indícios de falta de ajuste ou violação das suposições para o modelo probabilístico adotado. Algumas observações são destacadas das demais por apresentarem comportamento diferente do esperado.

Figura 17 - Gráfico dos tempos de duração versus os resíduos padronizados obtidos pelo modelo selecionado

Dentre as observações destacadas, as observações 158 e 197 apresentaram tempos grandes de espera até a erupção, substancialmente maiores do que os tempos esperados dados os tempos de duração registrados para as mesmas observações. As observações 161, 242 e 265 apresentaram resultados opostos, no sentido em que tiveram tempos de espera abaixo dos esperados, dados os tempos de duração observados. De uma forma geral, todas essas observações identificadas apresentaram uma relação entre o tempo de espera e a duração da erupção diferente da relação apresentada pela maioria das observações e captada pelo modelo de regressão ajustado.

Figura 18 - Gráfico de probabilidade normal dos resíduos quânticos

Figura 19 - Gráficos das medidas de influência local para o modelo ajustado para tempos de espera de erupção

Os métodos de análise de sensibilidade, como influência local e global, permitem avaliar a sensibilidade das estimativas para o modelo ajustado em relação à presença de observações atípicas ou violações das suposições adotadas. A Figura 19 apresenta algumas medidas da análise de influência local para os esquemas de ponderação de casos,

perturbação da covariável, comprimento e perturbação da variável resposta.

As observações que se destacaram nos gráficos com as medidas de influência local foram as mesmas apontadas nos gráficos de resíduos, indicando certa concordância entre os diferentes métodos de diagnóstico. A observação 76 aparece de forma destacada das demais em todos os esquemas de perturbação considerados. Essa observação tem o segundo maior tempo de duração de erupção e tempo de espera mediano, fugindo do padrão linear de relação entre essas variáveis, em que o previsto seria que essa observação tivesse também um grande tempo de espera.

Como no exemplo de aplicação anterior, a instabilidade na estimação inviabiliza a deleção de todos os casos. Excluíram-se apenas os valores apontados como possivelmente influentes pelas outras técnicas, e, em cada caso, a estimação é realizada, utilizando sequências aleatórias de valores iniciais e buscando em cada caso a obtenção de máximos globais para as superfícies de verossimilhança.

Na Tabela 19, estão apresentados alguns resultados obtidos pelo método de deleção de casos. Para avaliar o impacto das observações destacadas nas análises de resíduo e influência, o modelo de regressão é reajustado para diferentes sub-amostras dos dados de erupção. Cada sub-amostra é obtida por meio da deleção de uma das observações destacadas nas análises prévias. Para cada sub-amostra considerada as estimativas para os parâmetros do modelo e seus erros-padrão são apresentados. O objetivo é avaliar se a deleção dessas observações causa mudanças substanciais nas estimativas ou inferências sobre o modelo.

Tabela 19 - Estimativas e (erros padrão) para os parâmetros dos modelos de regressão ajustados para diferentes sub-amostras do conjunto de dados de erupções

Sub-amostra	a	b	p	c	β_0	β_1
$I(76)$	0.2349 (0.0232)	3.0×10^{-7} (1.3×10^{-6})	4.9904 (1.7228)	-0.6175 (0.2774)	35.6129 (1.3917)	10.8792 (0.2760)
$I(158)$	0.2436 (0.0204)	1.5×10^{-6} (5.6×10^{-6})	4.4172 (1.3807)	-0.4721 (0.2304)	35.5131 (1.2228)	10.7041 (0.2699)
$I(161)$	0.2342 (0.0230)	4.7×10^{-7} (1.6×10^{-6})	4.8768 (1.5430)	-0.6219 (0.2825)	36.4060 (1.3893)	10.6559 (0.2639)
$I(197)$	0.2450 (0.0202)	1.2×10^{-6} (4.7×10^{-6})	4.5132 (1.4449)	-0.4537 (0.2188)	35.2827 (1.2290)	10.7485 (0.2680)
$I(242)$	0.2342 (0.0233)	6.4×10^{-7} (2.5×10^{-6})	4.6938 (1.4811)	-0.6136 (0.2862)	36.2936 (1.4045)	10.6783 (0.2662)
$I(265)$	0.2342 (0.0232)	6.0×10^{-7} (2.3×10^{-6})	4.7183 (1.4699)	-0.6154 (0.2857)	36.3988 (1.4107)	10.6503 (0.2685)

Em nenhuma das sub-amostras consideradas, houve mudanças substanciais nas estimativas obtidas para o modelo de regressão. Dessa forma, as observações destacadas não estão influenciando de forma desproporcional as estimativas dos parâmetros do modelo ajustado. Apesar de terem comportamentos um pouco diferentes do padrão geral dos dados, essas observações também não devem ser consideradas *outliers* ou serem removidas da análise.

5 CONSIDERAÇÕES FINAIS

Neste trabalho, foram considerados os modelos propostos por Rathie e Swamee, com aplicações estatísticas e propostas de extensões para os modelos. Nos estudos de simulação realizados, a metodologia de estimação por máxima verossimilhança se mostrou satisfatória, de forma que os valores estimados estiveram em concordância com os valores definidos para os parâmetros dos modelos considerados.

As comparações feitas para os diferentes métodos de obtenção de intervalos de confiança para os parâmetros dos modelos Rathie-Swamee não apontaram diferenças consideráveis no desempenho dos diferentes métodos. Ao avaliar as coberturas nominais e amplitudes médias dos intervalos obtidos, os diferentes métodos considerados apresentaram resultados similares. Uma vez que os métodos têm resultados semelhantes, os intervalos calculados considerando a distribuição normal assintótica dos estimadores de máxima verossimilhança mostram-se vantajosos. Apesar de ocorrerem situações em que esses intervalos tenham limites inferiores negativos para parâmetros que podem assumir apenas valores não negativos, esse método se mostra vantajoso, pela simplicidade na forma de obtenção desses intervalos.

As técnicas bootstrap e perfilhamento da verossimilhança demandam computação intensiva, mas não resultam em coberturas nominais mais próximas das fixadas no experimento realizado ou menores amplitudes dos intervalos obtidos. Os intervalos descritos pela forma 50, obtidos após transformações nos parâmetros podem apresentar amplitudes infinitas dada a transformação escolhida e a configuração do espaço paramétrico com borda de fronteira em zero para o parâmetro p . Nos casos em que a estimativa do parâmetro assume valor igual ou próximo de zero, a transformação escolhida tende para infinito.

Nas aplicações com os dados de pesca de camarões e erupção de gêiser, os modelos Rathie-Swamee simétricos e assimétricos tiveram desempenho melhor que os demais modelos considerados, incluindo modelo de misturas de distribuições normais. Nas aplicações dos modelos de regressão, o processo de estimação por verossimilhança mostrou-se bastante sensível à escolha dos valores iniciais para os algoritmos de maximização. Diferentes escolhas para os valores iniciais, resultaram na convergência para diferentes pontos de máximo locais. A superfície de verossimilhança aparenta ser achatada, dificultando o processo de estimação e realização de inferências.

O modelo de regressão selecionado para descrever o peso dos camarões em função de seus comprimentos e sexos, foi o caso mais simples do modelo Rathie-Swamee, representando o caso particular da distribuição logística com um parâmetro. O modelo com forma bimodal, contendo dois parâmetros que fora selecionado para descrever os pesos isoladamente, pode ser simplificado para um caso particular, mediante a inclusão das covariáveis. Os ajustes e testes entre modelos Rathie-Swamee encaixados, considerando seus casos particulares, permitiu a obtenção de um modelo simples que se adequasse aos dados.

No caso da aplicação com os dados de erupção de gêiser, o modelo Rathie-Swamee assimétrico completo foi selecionado tanto para modelar os tempos de espera até erupção isoladamente, quanto para o modelo de regressão que incorpora o tempo de duração das erupções. Essa aplicação exemplifica de forma clara a utilidade dos modelos Rathie-Swamee de regressão, mostrando o desempenho superior do mesmo em relação aos outros modelos considerados.

As medidas de resíduo utilizadas não indicaram problemas nas aplicações dos modelos RS para os dados considerados, e as medidas de influência local possibilitaram a identificação de observações com comportamentos diferentes da tendência geral das observações, apontando essas observações como possíveis pontos com influência desproporcional sobre as estimativas.

A metodologia de deleção de casos apresentou problemas em sua execução devido à grande instabilidade na etapa de estimação dos parâmetros. Devido à sensibilidade do método de otimização em relação à escolha dos valores iniciais, o método de deleção de casos, que já depende de computação intensiva para estimar os parâmetros dos modelos considerando $n + 1$ conjuntos de dados, torna-se proibitivo. Para cada uma das $n + 1$ otimizações realizadas seria necessário repetir o algoritmo diversas vezes, considerando diferentes valores iniciais para buscar os pontos de máximo globais das funções de verossimilhança.

Considerando os resultados obtidos e as dificuldades encontradas são propostas algumas ideias para pesquisa futura, como se segue.

1. Utilização de métodos de inferência Bayesiana para os modelos de regressão Rathie-Swamee.
2. Elaboração de método de aproximação para estimativas dos parâmetros dos modelos

RS para utilização da metodologia de deleção de casos.

3. Obtenção das distribuições resultantes de combinações lineares entre variáveis com distribuição Rathie-Swamee independentes
4. Documentação e organização das rotinas computacionais utilizadas para criação de pacote para ambiente R.

REFERÊNCIAS

- ATKINSON, A. **Plots, transformations, and regression**. Oxford: Clarendon Press, 1985. 282 p.
- AZZALINI, A. A Class of Distributions Which Includes the Normal Ones. **Scandinavian Journal of Statistics**, Hoboken, v.12, n.2, p. 171–178, 1985.
- BALAKRISHNAN, N. **Handbook of the Logistic Distribution**. London: Taylor & Francis, 1991. 624 p.
- BENAGLIA, T.; CHAUVEAU, D.; HUNTER, D. R.; YOUNG, D. S. mixtools: An R Package for Analyzing Mixture Models. **Journal of Statistical Software**, Los Angeles, v.32, n.6, p. 1–28, 2009.
- BURNHAM, K.; ANDERSON, D. **Model Selection and Multi-Model Inference: a Practical Information-Theoretic Approach**. Vienna: Springer, 2002.
- CASELLA, G.; BERGER, R. **Statistical inference**. Belmont: Duxbury Press, 1990. 650 p.
- COOK, R. Assessment of local influence (with discussion). **Journal of the Royal Statistical Society**, Oxford, v.48, p. 133–169, 1986.
- COOK, R. D. Detection of Influential Observation in Linear Regression. **Technometrics**, Alexandria, v.19, n.1, p. 15–18, 1977.
- COOK, R. D. Influence assessment. **Journal of Applied Statistics**, Oxford, v.14, n.2, p. 117–131, 1987.
- DIAS NETO, J.(Org.). **Proposta de Plano Nacional de Gestão para o uso sustentável de camarões marinhos do Brasil**. Brasília: IBAMA, 2011. 242 p. (Série Plano de Gestão Recursos Pesqueiros, 3).
- FACHINI, J.; ORTEGA, E.; LOUZADA-NETO, F. Influence diagnostics for polyhazard models in the presence of covariates. **Statistical Methods and Applications**, New York, v.17, p. 413–433, 2008.
- FLOR, H. **Desempenho reprodutivo do camarão-rosa Farfantepenaeus brasiliensis (Latreille, 1817) (Crustacea, Decapoda, Penaeidae) em cativeiro: Efeito da alimentação e proporção sexual**. 2009. 57 p. Dissertação (Mestrado em Ciências) - Universidade Federal Rural do Rio de Janeiro, Seropédica, 2009.
- JOHNSON, N.; KOTZ, S.; BALAKRISHNAN, N. **Continuous univariate distributions**. Michigan: John Wiley & Sons, 1995. v.2. 752 p.
- LANGE, K. **Numerical Analysis for Statisticians**. New York: Springer, 2001. 356 p.
- LESAFFRE, E.; VERBEKE, G. Local Influence in Linear Mixed Models. **Biometrics**, Washington, v.54, n.2, p. 570–582, 1998.
- LUKE, Y. **The Special Functions and Their Approximations**. New York: Academic Press, 1969. 349 p.
- MATHAI, A.; SAXENA, R.; HAUBOLD, H. **The H-Function: Theory and Applications**. Trivandrum: Springer, 2010. 284 p.
- MCLACHLAN, G.; PEEL, D. **Finite Mixture Models**. Michigan: John Wiley & Sons, 2004. 456 p.
- MEEKER, W.; ESCOBAR, L. **Statistical Methods for Reliability Data**. Michigan: John Wiley & Sons, 1998. 712 p.

NADARAJAH, S. Exact Distribution of the Product of Two or More Logistic Random Variables. **Methodology and Computing in Applied Probability**, Vienna, v.11, p. 651–660, 2009.

NADARAJAH, S.; GUPTA, A. K. On the ratio of logistic random variables. **Computational Statistics & Data Analysis**., Amsterdam, v.50, n.5, p. 1206–1219, 2006.

ORTEGA, E. M. M.;CANCHO, V. G.; BOLFARINE, H. Influence diagnostics in exponentiated-Weibull regression models with censored data. **Statistics and Operation Research Transactions**, Catalunya, v.30, n.2, p. 171–192, 2006.

PAWITAN, Y. **In All Likelihood: Statistical Modelling and Inference Using Likelihood**. Oxford: Oxford Science Publications, 2001. 528 p.

PINHEIRO, A. **Caracterização genética e biométrica das populações de camarão rosa *Farfantepenaeus brasiliensis* de três localidades da costa do Rio Grande do Norte**. 2008. 85 p. Tese (Doutorado em Ecologia e Recursos Naturais) - Universidade Federal De São Carlos, São Carlos, 2008.

R DEVELOPMENT CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna. 2011. Disponível em: <<http://www.R-project.org>>. Acesso em: 15 jan. 2013.

RATHIE, P.; COUTINHO, M. A new Skew Generalized Logistic distribution and Approximations to Skew Normal Distribution. **Aligarh Journal of Statistics**, Aligarh, v.31, p. 1–12, 2011.

RATHIE, P.;SWAMEE, P. K.; SILVA, M. Generalized Logistic Distribution and its Applications. In: INTERNATIONAL CONFERENCE ON SPECIAL FUNCTIONS AND THEIR APPLICATIONS. 2011, Jodhpur. **Proceedings...** Jodhpur: ICSFA. 2011. p. 1-11.

RATHIE, P.;ZORNIG, P.; GOMES, E. On the distributions of quotient and product of generalized logistic random variables. **Computational Statistics & Data Analysis**, 2012, in press.

RATHIE, P. N.; SWAMEE, P. K. **On new invertible generalized logistic distribution approximation to normal distribution**. Brasília; UNB; Departamento de estatística, 2006. 7 p. Relatório Técnico de Pesquisa em Estatística No. 07/2006.

SWAMEE, P. K.; RATHIE, P. Invertible alternatives to normal and lognormal distribution. **Journal of Hydrologic Engineering**, v.12, p. 218–221, 2007.

XIE, F.; WEI, B. Diagnostics analysis for log-Birnbaum-Saunders regression models. **Computational Statistics & Data Analysis**, Amsterdam, v.51, n.9, p. 4692–4706, 2007.

ZÖRNIG, P. **Introdução à programação não linear**. Brasília: UnB, 2011. 395 p.

APÊNDICES

APÊNDICE A - Exemplos de rotinas em R.

```
####Principais rotinas em R
```

```
####Funções de probabilidade
```

```
qswamee<-function(novo,a,b,p,mu,sigma){
  desvio<-function(z){
    novo-(exp(-((z-mu)/sigma)*(a+b*abs(((z-mu)/sigma))^p))+1)^(-1)
  }
  uniroot(desvio,c(-1000+mu,1000+mu))$root
}
```

```
hswamee<-function(z,a,b,p,mu,sigma){
  (a+b*(1+p)*abs((z-mu)/sigma)^p)/(exp(-(z-mu)/sigma*
  (a+b*abs((z-mu)/sigma)^p))+1)/sigma
}
```

```
dswamee<-function(z,a,b,p,mu,sigma){
  (a+b*(1+p)*abs((z-mu)/sigma)^p)*exp(-abs(z-mu)*
  (a+b*abs((z-mu)/sigma)^p)/sigma)/
  (sigma*(exp(-abs(z-mu)*(a+b*abs((z-mu)/sigma)^p)/sigma)+1)^2)
}
```

```
rswamee<-function(n,a,b,p,mu,sigma){
  uu<-as.matrix(runif(n))
  apply(uu,1,qswamee,a=a,b=b,p=p,mu=mu,sigma=sigma)
}
```

```
pswamee<-function(z,a,b,p,mu,sigma){
  (exp(-((z-mu)/sigma)*(a+b*abs(((z-mu)/sigma))^p))+1)^(-1)
}
```

```
dSkew<-function(z,a,b,p,mu,sigma,cc){
  densi<-((2*(a+b*(1+p)*abs(((z-mu)/sigma))^p)*exp(-((z-mu)/sigma)*
  (a+b*abs(((z-mu)/sigma))^p)))/(sigma*(exp(-((z-mu)/sigma)*
  (a+b*abs(((z-mu)/sigma))^p))+1)^2*
  (exp(-cc*((z-mu)/sigma)*(a+b*abs(cc*((z-mu)/sigma))^p))+1))
  ifelse(is.na(densi),0,densi)
}
```

```
pSkew<-function(z,a,b,p,mu,sigma,cc){
  integra<-function(z){
    integrate(dSkew,-100,z,a,b,p,mu,sigma,cc)$value
  }
  apply(as.matrix(z),1,integra)
}
```

```
qSkew<-function(novo,a,b,p,mu,sigma,cc){
  desvio<-function(z,a,b,p,mu,sigma,cc){
    novo-pSkew(z,a,b,p,mu,sigma,cc)
  }
}
```



```

    }
    uniroot(desvio,c(-100+mu,100+mu),a,b,p,mu,sigma,cc)$root
  }

rSkew<-function(n,a,b,p,mu,sigma,cc){
  uu<-as.matrix(round(runif(n),4))
  apply(uu,1,qSkew,a=a,b=b,p=p,mu=mu,sigma=sigma,cc=cc)
}

###Modelos de regressão
vero<-function(param,z,X){
  b<-param[1]
  p<-param[2]
  Betas<-param[-c(1,2)]
  if(dim(X)[2]==1){mu<-Betas}
  else{mu<-X%*%Betas}
  -sum(log((exp(-(z-mu)*b*abs(z-mu)^p)*(p+1)*b*abs(z-mu)^p)/
    (exp(-b*(z-mu)*abs(z-mu)^p)+1)^2))
}

regmodel<-function(formula1,vb,vp){
  z<-model.frame(formula1)[,1]
  X<-model.matrix(formula1)
  p<-dim(X)[2]
  ccc<-constrOptim(c(vb,vp,lm(z~X-1)$coefficients),vero,grad=NULL,
    control=list(maxit=20000,reltol=.Machine$double.eps^4),
    ui=rbind(c(1,0,rep(0,p)),c(0,1,rep(0,p))),ci=c(0,0),X=X,z=z)
  if(p==1){preditos<-X*ccc$par[3]}
  else{preditos<-X%*%ccc$par[-c(1,2)]}
  resultado<-list()
  resultado$par<-ccc$par
  resultado$convergence<-ccc$convergence
  resultado$value<-ccc$value*(-1)
  resultado$residuo<-preditos-z
  resultado$predito<-preditos
  return(resultado)
}

veroSkew<-function(param,X,z){
  b<-param[1]
  p<-param[2]
  cc<-param[3]
  Betas<-param[-c(1,2,3)]
  if(dim(X)[2]==1){mu<-Betas}
  else{mu<-X%*%Betas}
  -sum(log((2*b*(1+p)*abs(z-mu)^p*exp(-(z-mu)*b*abs(z-mu)^p)/
    ((exp(-(z-mu)*b*abs(z-mu)^p)+1)^2*(exp(-cc*(z-mu)*b*abs(cc*(z-mu))^p)+1))))
}

regmodelSkew<-function(formula1,vb,vp,vc){
  z<-model.frame(formula1)[,1]

```

```

X<-model.matrix(formula1)
p<-dim(X)[2]
ccc<-constrOptim(c(vb, vp, vc, lm(z~X-1)$coefficients), veroSkew, grad=NULL,
  control=list(maxit=20000, reltol=.Machine$double.eps^4),
  ui=rbind(c(1,0,0,rep(0,p)), c(0,1,0,rep(0,p))), ci=c(0,0), X=X, z=z)
if(p==1){preditos<-X*ccc$par[4]}
else{preditos<-X*%%ccc$par[-c(1,2,3)]}
resultado<-list()
resultado$par<-ccc$par
resultado$convergence<-ccc$convergence
resultado$value<-ccc$value*(-1)
resultado$residuo<-preditos-z
resultado$predito<-preditos
return(resultado)
}

veroP0<-function(param, z, X){
  b<-param[1]
  p<-0
  Betas<-param[-1]
  if(dim(X)[2]==1){mu<-Betas}
  else{mu<-X*%%Betas}
  -sum(log((exp(-(z-mu)*b*abs(z-mu)^p)*(p+1)*b*abs(z-mu)^p)/
    (exp(-b*(z-mu)*abs(z-mu)^p)+1)^2))
}

regmodelP0<-function(formula1, vb){
  z<-model.frame(formula1)[,1]
  X<-model.matrix(formula1)
  p<-dim(X)[2]
  ccc<-constrOptim(c(vb, lm(z~X-1)$coefficients), veroP0, grad=NULL,
    control=list(maxit=20000, reltol=.Machine$double.eps^4),
    ui=c(1,rep(0,p)), ci=0, X=X, z=z)
  if(p==1){preditos<-X*ccc$par[2]}
  else{preditos<-X*%%ccc$par[-1]}
  resultado<-list()
  resultado$par<-ccc$par
  resultado$convergence<-ccc$convergence
  resultado$value<-ccc$value*(-1)
  resultado$residuo<-preditos-z
  return(resultado)
}

veroSkewP0<-function(param, X, z){
  b<-param[1]
  p<-0
  cc<-param[2]
  Betas<-param[-c(1,2)]
  if(dim(X)[2]==1){mu<-Betas}
  else{mu<-X*%%Betas}
  -sum(log((2*b*(1+p)*abs(z-mu)^p*exp(-(z-mu)*b*abs(z-mu)^p))/

```

```

((exp(-(z-mu)*b*abs(z-mu)^p)+1)^2*(exp(-cc*(z-mu)*b*abs(cc*(z-mu))^p)+1)))
}

regmodelSkewP0<-function(formula1,vb,vc){
  z<-model.frame(formula1)[,1]
  X<-model.matrix(formula1)
  p<-dim(X)[2]
  ccc<-constrOptim(c(vb,vc,lm(z~X-1)$coefficients),veroSkewP0,grad=NULL,
    control=list(maxit=20000,reltol=.Machine$double.eps^4),
    ui=rbind(c(1,0,rep(0,p))),ci=0,X=X,z=z)
  if(p==1){preditos<-X*ccc$par[3]}
  else{preditos<-X%*%ccc$par[-c(1,2)]}
  resultado<-list()
  resultado$par<-ccc$par
  resultado$convergence<-ccc$convergence
  resultado$value<-ccc$value*(-1)
  resultado$residuo<-preditos-z
  resultado$predito<-preditos
  return(resultado)
}

veroCompleta<-function(param,z,X){
  a<-param[1]
  b<-param[2]
  p<-param[3]
  Betas<-param[-c(1,2,3)]
  if(dim(X)[2]==1){mu<-Betas}
  else{mu<-X%*%Betas}
  -sum(log(dswamee(z,a,b,p,mu,1)))
}

regmodelCompleto<-function(formula1,va,vb,vp){
  z<-model.frame(formula1)[,1]
  X<-model.matrix(formula1)
  p<-dim(X)[2]
  ccc<-constrOptim(c(va,vb,vp,lm(z~X-1)$coefficients),veroCompleta,grad=NULL,
    control=list(maxit=20000,reltol=.Machine$double.eps^4),
    ui=rbind(c(1,0,0,rep(0,p)),c(0,1,0,rep(0,p)),
      c(0,0,1,rep(0,p))),ci=c(0,0,0),X=X,z=z)
  if(p==1){preditos<-X*ccc$par[4]}
  else{preditos<-X%*%ccc$par[-c(1,2,3)]}
  resultado<-list()
  resultado$par<-ccc$par
  resultado$convergence<-ccc$convergence
  resultado$value<-ccc$value*(-1)
  resultado$residuo<-preditos-z
  resultado$predito<-preditos
  return(resultado)
}

```

```

veroSkewCompleta<-function(param,X,z){
  a<-param[1]
  b<-param[2]
  p<-param[3]
  cc<-param[4]
  Betas<-param[-c(1,2,3,4)]
  if(dim(X)[2]==1){mu<-Betas}
    else{mu<-X%%Betas}
  -sum(log(dSkew(z,a,b,p,mu,1,cc)))
}

regmodelSkewCompleta<-function(formula1,va,vb,vp,vc){
  z<-model.frame(formula1)[,1]
  X<-model.matrix(formula1)
  p<-dim(X)[2]
  ccc<-constrOptim(c(va,vb,vp,vc,lm(z~X-1)$coefficients),
    veroSkewCompleta,grad=NULL,
    control=list(maxit=20000,reltol=.Machine$double.eps^4),
    ui=rbind(c(1,0,0,0,rep(0,p)),c(0,1,0,0,rep(0,p)),
      c(0,0,1,0,rep(0,p))),ci=c(0,0,0),X=X,z=z)
  if(p==1){preditos<-X*ccc$par[5]}
  else{preditos<-X%%ccc$par[-c(1,2,3,4)]}
  resultado<-list()
  resultado$par<-ccc$par
  resultado$convergence<-ccc$convergence
  resultado$value<-ccc$value*(-1)
  resultado$residuo<-preditos-z
  resultado$predito<-preditos
  return(resultado)
}

####Estudo com intervalos de confiança
require(boot)
veroBp<-function(param,z){
  b<-param[1]
  p<-param[2]
  -sum(log(b*(1+p)*abs(z)^p*exp(-z*b*abs(z)^p)/(exp(-z*b*abs(z)^p)+1)^2))}

grrBp<-function(param,z){
  b<-param[1]
  p<-param[2]
  gradb<--sum(((exp(-b*z*abs(z)^p)+1)^2*exp(b*z*abs(z)^p)*(-(b*(p+1)*z*
    exp(-b*z*abs(z)^p)*abs(z)^(2*p)))/
    (exp(-b*z*abs(z)^p)+1)^2+(2*b*(p+1)*z*exp(-2*b*z*abs(z)^p)*abs(z)^(2*p))/
    (exp(-b*z*abs(z)^p)+1)^3+((p+1)*exp(-b*z*abs(z)^p)*abs(z)^p)/
    (exp(-b*z*abs(z)^p)+1)^2))/(b*(p+1)*abs(z)^p))
  gradp<--sum(((exp(-b*z*abs(z)^p)+1)^2*exp(b*z*abs(z)^p)*(-(b^2*(p+1)*z*
    exp(-b*z*abs(z)^p)*abs(z)^(2*p)*log(abs(z))))/
    (exp(-b*z*abs(z)^p)+1)^2+(2*b^2*(p+1)*z*exp(-2*b*z*abs(z)^p)*
    abs(z)^(2*p)*log(abs(z)))/(exp(-b*z*abs(z)^p)+1)^3
    +(b*(p+1)*exp(-b*z*abs(z)^p)*abs(z)^p*log(abs(z)))/

```

```

      (exp(-b*z*abs(z)^p)+1)^2+(b*exp(-b*z*abs(z)^p)*abs(z)^p)/
      (exp(-b*z*abs(z)^p)+1)^2)/(b*(p+1)*abs(z)^p))
    c(gradb,gradp)
  }

estimativas<-function(x,indice,param){
  bbb<-constrOptim(param,veroBp,grrBp,control=list(maxit=20000),
    ui=rbind(c(1,0),c(0,1)), ci=c(0,0),z=x[indice])
  round((bbb$par),4)
}

###Pra definir o bootParametrico

sim.boot<-function(dados,mle){
  n<-length(dados)
  rswamee(n=n,0,mle[1],mle[2],0,1)
}

estimativas1<-function(x,param){
  dados<-x
  bbb<-constrOptim(param,veroBp,grrBp,control=list(maxit=20000),
    ui=rbind(c(1,0),c(0,1)), ci=c(0,0),z=dados)
  round((bbb$par),4)
}

perf.b<-function(p,b,dados){
  ll<-sum(log(dswamee(dados,0,b,p,0,1)))
  return(ll)
}

perf.p<-function(b,p,dados){
  ll<-sum(log(dswamee(dados,0,b,p,0,1)))
}

dev.perf.b<-function(b,theta,dados){
  2*(-veroBp(theta,dados)-optimize(perf.b,
    interval=c(theta[2]/3,theta[2]*3),b,dados,maximum=TRUE)$objective)
}

dev.perf.p<-function(p,theta,dados){
  2*(-veroBp(theta,dados)-optimize(perf.p,
    interval=c(theta[1]/3,theta[1]*3),p,dados,maximum=TRUE)$objective)
}

ic.dev.perf.b<-function(b,theta,dados){
  dev.perf.b(b,theta,dados)-qchisq(.95,1)
}

ic.dev.perf.p<-function(p,theta,dados){
  dev.perf.p(p,theta,dados)-qchisq(.95,1)
}

```

```

contemB<-function(linha,vb){
  intervalos<-length(linha)
  acerto<-logical()
  j<-1
  for(i in seq(1,intervalos,by=2)){
    acerto[j]<-(linha[i]<vb && vb<linha[i+1])
    j<-j+1
  }
  return(acerto)
}

interval<-function(vb,vp,amostras,n){
  flagB<-numeric()
  flagP<-numeric()
  estfim<-matrix(0,ncol=2,nrow=amostras)
  final.b<-matrix(0,ncol=10,nrow=amostras)
  final.p<-matrix(0,ncol=10,nrow=amostras)
  matrizdados<-matrix(rswamee(n*amostras,0,vb,vp,0,1),nrow=n,ncol=amostras)
  for(i in 1:amostras){
    dados<-matrizdados[,i]
    aaa<-constrOptim(c(vb,vp),veroBp,grrBp,
      control=list(maxit=20000), ui=rbind(c(1,0),c(0,1)),
      ci=c(0,0),z=dados,hessian=TRUE)
    erropad<-sqrt(diag(solve(aaa$hessian)))
    resultado<-boot(data=dados,statistic=estimativas,R=1000,param=aaa$par)
    intervalos.b<-boot.ci(resultado,var.t0=erropad[1]^2,
      type = "perc",index=1)
    intervalos.p<-boot.ci(resultado,var.t0=erropad[2]^2,
      type = "perc",index=2)
    resultadoPar<-boot(data=dados,
      statistic=estimativas1,R=1000,
      sim = "parametric",ran.gen = sim.boot,mle=aaa$par, param=aaa$par)
    intervalosPar.b<-boot.ci(resultadoPar,
      var.t0=erropad[1]^2, type = "perc",index=1)
    intervalosPar.p<-boot.ci(resultadoPar,
      var.t0=erropad[2]^2, type = "perc",index=2)
    intervalo.transf.b<-c(aaa$par[1]/
      exp(1.96*erropad[1]/aaa$par[1]),
      aaa$par[1]*exp(1.96*erropad[1]/aaa$par[1]))
    intervalo.transf.p<-c(aaa$par[2]/
      exp(1.96*erropad[2]/aaa$par[2]),
      aaa$par[2]*exp(1.96*erropad[2]/aaa$par[2]))
    intervalo.norm.b<-c(aaa$par[1]-1.96*erropad[1],aaa$par[1]+1.96*erropad[1])
    intervalo.norm.p<-c(aaa$par[2]-1.96*erropad[2],aaa$par[2]+1.96*erropad[2])
    ##Criei uma condição que o limite inferior fica truncada no Zero
    if((dev.perf.b(.00001,aaa$par,dados)-qchisq(.95,1))>0)
      {intervalo.perf.b<-c(uniroot(ic.dev.perf.b,
        interval=c(.00001,aaa$par[1]),theta=aaa$par,
        dados=dados)$root,uniroot(ic.dev.perf.b,
        interval=c(aaa$par[1],10),theta=aaa$par,dados=dados)$root)
        flagB[i]<-0
  }
}

```

```

    }
    else{intervalo.perf.b<-c(0,uniroot(ic.dev.perf.b,
    interval=c(aaa$par[1],10),theta=aaa$par,dados=dados)$root)
      flagB[i]<-1
    }
    if((dev.perf.p(.00001,aaa$par,dados)-qchisq(.95,1))>0)
    {intervalo.perf.p<-c(uniroot(ic.dev.perf.p,
    interval=c(.00001,aaa$par[2]),theta=aaa$par,dados=dados)$root,
    uniroot(ic.dev.perf.p,interval=c(aaa$par[2],10),
    theta=aaa$par,dados=dados)$root)
      flagP[i]<-0
    }
    else{intervalo.perf.p<-c(0,uniroot(ic.dev.perf.p,
    interval=c(aaa$par[2],10),theta=aaa$par,dados=dados)$root)
      flagP[i]<-1
    }
    estfim[i,]<-aaa$par
    final.b[i,]<-c(intervalos.b$percent[4:5],
    intervalosPar.b$percent[4:5],intervalo.transf.b,
    intervalo.norm.b,intervalo.perf.b)
    final.p[i,]<-c(intervalos.p$percent[4:5],
    intervalosPar.p$percent[4:5],intervalo.transf.p,
    intervalo.norm.p,intervalo.perf.p)
    print(i)
  }
  return(list(matrizdados,final.b,final.p,estfim,flagB,flagP))
}

contemP<-function(linha,vp){
  intervalos<-length(linha)
  acerto<-logical()
  j<-1
  for(i in seq(1,intervalos,by=2)){
    acerto[j]<-(linha[i]<vp && vp<linha[i+1])
    j<-j+1
  }
  return(acerto)
}

amplitude<-function(linha){
  intervalos<-length(linha)
  ampli<-numeric()
  j<-1
  for(i in seq(1,intervalos,by=2)){
    ampli[j]<-(linha[i+1]-linha[i])
    j<-j+1
  }
  return(ampli)
}

```

###Funções de probabilidade para modelo de produtos e quocientes

```

dzoring<-function(x,alpha){
  x<-as.matrix(x)
  laco<-function(z,...){
    z<-abs(z)
    nucleo<-function(w){
      ((-log(w))^alpha*exp(-z*(-log(w))^alpha))/
      ((1+w)^2*(exp(-z*(-log(w))^alpha)+1)^2)
    }
    2*integrate(nucleo,0,.9999)$value
  }
  apply(x,1,laco)
}

pzoring<-function(x,alpha){
  x<-as.matrix(x)
  laco<-function(z,...){
    sinal<-ifelse(z<0,-1,1)
    z<-abs(z)
    nucleoAcum<-function(w){
      1/((1+w)^2*(exp(-z*(-log(w))^alpha)+1))
    }
    if(sinal>0){2*integrate(nucleoAcum,0,1)$value}
    else {1-(2*integrate(nucleoAcum,0,1)$value)}
  }
  apply(x,1,laco)
}

qzoring<-function(p,alpha){
  desvio<-function(z){
    p-pzoring(z,alpha)
  }
  sinal<-ifelse(p<.5,-1,1)
  if(p<.5){ p<-(1-p)}
  uniroot(desvio,c(0,900000000))$root*sinal
}

rzoring<-function(n,alpha){
  uu<-as.matrix(round(runif(n),4))
  apply(uu,1,qzoring,alpha)
}

```

APÊNDICE B - Dados de pesca de camarões.

Tabela 20 - Dados de pesca de camarões coletados disponibilizados em Pinheiro (2008)

Obs	Local	Sexo	Comprimento	Peso	Obs	Local	Sexo	Comprimento	Peso
1	BaiaFormosa	M	94.6	9.67	61	DiogoLopes	M	55.8	1.98
2	BaiaFormosa	F	133.0	29.32	62	DiogoLopes	F	134.3	32.08
3	BaiaFormosa	F	113.5	18.40	63	DiogoLopes	M	74.3	5.14
4	BaiaFormosa	F	120.8	21.70	64	DiogoLopes	F	69.6	4.44
5	BaiaFormosa	F	116.4	21.60	65	DiogoLopes	F	73.2	4.85
6	BaiaFormosa	M	91.0	8.96	66	DiogoLopes	M	98.9	11.65
7	BaiaFormosa	F	102.0	13.86	67	DiogoLopes	M	62.8	3.32
8	BaiaFormosa	F	100.0	12.93	68	DiogoLopes	F	80.3	7.10
9	BaiaFormosa	F	112.7	19.02	69	DiogoLopes	F	80.0	6.77
10	BaiaFormosa	M	88.0	7.20	70	Touros	F	117.7	21.34
11	BaiaFormosa	F	113.0	18.20	71	Touros	F	115.9	21.18
12	BaiaFormosa	F	87.2	7.92	72	Touros	F	123.2	24.79
13	BaiaFormosa	M	91.9	10.07	73	Touros	F	130.0	27.80
14	BaiaFormosa	F	114.9	19.86	74	Touros	F	132.5	31.22
15	BaiaFormosa	M	91.4	10.46	75	Touros	M	100.0	12.03
16	BaiaFormosa	F	115.4	22.95	76	Touros	F	122.0	26.07
17	BaiaFormosa	M	99.7	11.47	77	Touros	F	138.2	35.78
18	BaiaFormosa	F	114.2	21.71	78	Touros	F	122.9	24.33
19	BaiaFormosa	M	93.8	10.71	79	Touros	F	119.7	23.13
20	BaiaFormosa	M	95.4	9.86	80	Touros	F	128.4	29.73
21	BaiaFormosa	F	121.9	22.71	81	Touros	F	130.7	29.04
22	BaiaFormosa	M	97.4	10.95	82	Touros	F	119.6	24.23
23	BaiaFormosa	F	122.6	23.46	83	Touros	F	115.9	20.36
24	BaiaFormosa	M	92.0	10.61	84	Touros	F	130.0	30.19
25	BaiaFormosa	F	114.4	20.82	85	Touros	F	123.2	24.87
26	BaiaFormosa	M	92.5	9.27	86	Touros	F	112.1	18.82
27	BaiaFormosa	F	123.2	23.93	87	Touros	F	110.2	18.27
28	BaiaFormosa	F	116.1	21.06	88	Touros	F	128.8	29.45
29	BaiaFormosa	M	94.5	9.99	89	Touros	F	124.3	27.17
30	BaiaFormosa	F	116.9	22.47	90	Touros	F	123.0	26.58
31	BaiaFormosa	F	102.3	13.05	91	Touros	F	122.0	24.21
32	BaiaFormosa	M	97.8	12.43	92	Touros	F	133.3	35.30
33	BaiaFormosa	F	107.3	16.19	93	Touros	F	117.4	22.63
34	BaiaFormosa	F	100.0	12.92	94	Touros	F	105.9	16.21
35	BaiaFormosa	M	82.4	6.74	95	Touros	F	125.9	27.30
36	BaiaFormosa	M	93.2	10.20	96	Touros	F	117.9	22.11
37	BaiaFormosa	F	124.1	25.37	97	Touros	F	114.8	21.14
38	DiogoLopes	F	102.9	14.57	98	Touros	F	114.5	20.17
39	DiogoLopes	F	68.7	3.69	99	Touros	F	114.9	20.87
40	DiogoLopes	F	85.3	8.07	100	Touros	F	103.9	19.27
41	DiogoLopes	F	61.9	2.74	101	Touros	F	124.5	26.05
42	DiogoLopes	M	90.6	9.60	102	Touros	F	124.3	27.02
43	DiogoLopes	M	92.2	9.47	103	Touros	F	114.1	19.73
44	DiogoLopes	M	84.8	7.95	104	Touros	F	107.1	17.00
45	DiogoLopes	F	121.1	23.23	105	Touros	F	110.5	18.02
46	DiogoLopes	M	101.5	14.14	106	Touros	F	104.5	14.32
47	DiogoLopes	F	86.3	7.60	107	Touros	F	112.5	17.97
48	DiogoLopes	F	71.6	5.54	108	Touros	M	94.5	10.72
49	DiogoLopes	F	92.6	10.01	109	Touros	F	100.2	12.98
50	DiogoLopes	F	74.4	4.80	110	Touros	M	98.3	11.15
51	DiogoLopes	M	89.6	7.68	111	Touros	F	118.3	21.72
52	DiogoLopes	F	89.3	8.65	112	Touros	F	103.3	14.42
53	DiogoLopes	F	81.3	6.58	113	Touros	M	89.3	8.10
54	DiogoLopes	F	80.7	7.02	114	Touros	F	121.5	22.40
55	DiogoLopes	M	71.2	3.96	115	Touros	M	93.7	10.30
56	DiogoLopes	F	68.6	3.66	116	Touros	F	101.8	12.99
57	DiogoLopes	F	68.3	4.21	117	Touros	M	89.0	9.98
58	DiogoLopes	M	71.7	2.31	118	Touros	M	101.4	12.59
59	DiogoLopes	F	56.7	2.56	119	Touros	M	99.1	11.54
60	DiogoLopes	F	59.4	3.76	120	Touros	M	89.4	8.50