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

MODELOS INGARCH LOG-LINEARES COM INOVAÇÕES POISSON MISTAS

Valdemi Nunes Costa

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

Orientadores: Rodrigo Bernardo da Silva
Pedro Rafael Diniz Marinho

João Pessoa
Julho de 2024

Ata da Sessão Pública de Defesa de Dissertação de Mestrado de **VALDEMI NUNES COSTA**, candidato ao título de Mestre em Modelagem Matemática e Computacional, realizada no dia 12 de julho de 2024.

Aos doze dias do mês de julho do ano de dois mil e vinte e quatro, às 14h, via videoconferência, reuniram-se os membros da Banca Examinadora constituída para julgar o Trabalho Final do discente VALDEMI NUNES COSTA, vinculado a Universidade Federal da Paraíba sob a matrícula nº 20221017938, candidato ao grau de Mestre em "*Modelagem Matemática e Computacional*", na linha de pesquisa *"Modelagem Probabilística"* do Programa de Pós-Graduação em Modelagem Matemática e Computacional. A comissão examinadora foi composta pelos professores Rodrigo Bernardo da Silva, Orientador e Presidente da Banca; Pedro Rafael Diniz Marinho, Coorientador; Tatiene Correia de Souza, Examinadora Interna ao Programa; e Marcelo Bourguignon Pereira, Examinador Externo à Instituição. Dando início aos trabalhos, o Presidente da Banca cumprimentou os presentes, comunicou a finalidade da reunião e passou a palavra ao candidato para que fizesse, oralmente, a exposição do trabalho de dissertação intitulado *"Modelos INGARCH log-lineares com inovações Poisson mistas"*. Concluída a exposição, o candidato foi arguido pela Banca Examinadora, que emitiu o seguinte parecer: "***aprovado***". Do ocorrido, eu, Gean Paulo P. M. de Barros, secretário do Programa de Pós-Graduação em Modelagem Matemática e Computacional (PPGMMC), lavrei a presente ata, que vai assinada por mim e pelos membros da Banca Examinadora. João Pessoa, 12 de julho de 2024.

Gean Paulo Pereira Maurício de Barros
Secretário do PPGMMC
SIAPE 2326476

Prof. Dr. Rodrigo Bernardo da Silva
Orientador (PPGMMC)

gov.br Documento assinado digitalmente
RODRIGO BERNARDO DA SILVA
Data: 22/07/2024 11:59:44-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Pedro Rafael Diniz Marinho
Coorientador (PPGMMC)

gov.br Documento assinado digitalmente
PEDRO RAFAEL DINIZ MARINHO
Data: 20/07/2024 17:33:14-0300
Verifique em https://validar.iti.gov.br

Profª. Drª. Tatiene Correia de Souza
Examinadora Interna ao Programa (PPGMMC)

gov.br Documento assinado digitalmente
TATIENE CORREIA DE SOUZA
Data: 17/07/2024 10:09:07-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Marcelo Bourguignon Pereira
Examinador Externo à Instituição (UFRN)

gov.br Documento assinado digitalmente
MARCELO BOURGUIGNON PEREIRA
Data: 15/07/2024 19:55:36-0300
Verifique em https://validar.iti.gov.br

Catalogação na publicação
Seção de Catalogação e Classificação

C838m Costa, Valdemi Nunes.
Modelos INGARCH log-lineares com inovações Poisson mistas / Valdemi Nunes Costa. - João Pessoa, 2024.
76 f. : il.

Orientação: Rodrigo Bernardo da Silva.
Coorientação: Pedro Rafael Diniz Marinho.
Dissertação (Mestrado) - UFPB/CI.

1. Modelagem matemática. 2. Séries temporais de contagem. 3. Modelos INGARCH. 4. Algoritmo EM. I. Silva, Rodrigo Bernardo da. II. Marinho, Pedro Rafael Diniz. III. Título.

UFPB/BC CDU 519.673(043)

Elaborado por Larissa Silva Oliveira de Mesquita - CRB-15/746

A meus pais Francisco e Helena,
a meus irmãos, a Walkiria
Maria Rodrigues (in
memoriam) e minhas filhas
Clara Raquel e Maria Eduarda.

Agradecimentos

Gostaria de agradecer primeiramente a meus pais Francisco Araújo Costa e Helena Nunes Costa, por todo cuidado, ensinamentos de vida e construção de valores. A meus irmãos Clarete, Valdeci, Valderi, Valmir e Dilza que sempre se fazem presentes em todos os momentos e estão sempre prontos para ajudar nas horas difíceis. A todos os professores do programa de mestrado e em especial a meu orientador Rodrigo Bernardo da Silva que sempre se mostrou presente, guiando nas orientações, ajudando a superar as dificuldades em todas as etapas de construção deste trabalho. Agradeço ainda a meus colegas de IFPI (Instituto Federal de Educação, Ciência e Tecnologia do Piauí) que se prontificaram a abraçar minha carga horária por não ter sido contemplado com professor substituto, além de sempre ajudarem no que fosse preciso. A meu coorientador Pedro Rafael Marinho Diniz por estar sempre pronto a ajudar. Agradeço a todos que contribuíram direto e indiretamente ao longo de minha vida acadêmica.

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

MODELOS INGARCH LOG-LINEARES COM INOVAÇÕES POISSON MISTAS

Valdemi Nunes Costa

Julho/2024

Orientadores: Rodrigo Bernardo da Silva
Pedro Rafael Diniz Marinho

Programa: Modelagem Matemática e Computacional

Neste trabalho, será proposta uma estrutura geral para inferência em modelagem de conjuntos de dados discretos, através de um modelo INGARCH de séries temporais de contagem com estrutura log linear com inovações Poisson mistas. Para tanto, será utilizada classe de distribuições de probabilidade, cujo objetivo principal é de modelar dados de contagem no tempo que apresentem condição de sobredispersão. De forma mais específica, o trabalho apresenta dois casos particulares que são as distribuições log linear Poisson Inversa Gaussiana e a distribuição log linear Binomial Negativa, que são obtidas ao considerar os casos de dados latentes que seguem as distribuições Inversa Gaussiana e Gama, respectivamente. As distribuições, inseridas através da média, têm como ponto comum o fato de serem membros da família exponencial de distribuições. Para estimar os parâmetros do modelo será utilizado o método iterativo de máxima verossimilhança por meio do algoritmo EM. A performance dos estimadores é avaliada através de estudos de simulação pelo método de Monte Carlo, considerando diferentes tamanhos amostrais para avaliar o comportamento assintótico destes estimadores. Na parte de aplicações do modelo proposto a conjuntos de dados reais, foram considerados para análise três bancos de dados: o primeiro deles lista o número de internações por abuso de álcool no estado da Paraíba, o segundo avalia a mesma problemática, porém com os dados apresentados para o estado do Piauí e, por fim, foi avaliado o banco de dados composto pelo número de casos de infecções por Campilobactéria na província de Quebec no Canadá, fechando assim a parte de aplicações a dados reais. Os dados de simulação foram testados nas duas extensões propostas e no modelo de comparação denominado log linear Poisson proposto por [9], levando em conta de início, uma análise gráfica para o comportamento da amostra, a autocorrelação e autocorrelação parcial,

o estudo de simulação por meio de convergência levando em conta os valores obtidos e a observação de gráficos que representam uma visão generalizada da disposição dos dados de simulação. Posteriormente foi feita uma reflexão sobre sua eficácia através de critérios de informação e do erro quadrático médio usados no processo de avaliação e escolha do melhor modelo de regressão para o ajuste dos dados.

Palavras-chave: Séries temporais de contagem, modelos INGARCH, Algoritmo EM.

Abstract of Qualifying presented to PPGMMC/CI/UFPB as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

LOG-LINEAR INGARCH MODELS WITH MIXED POISSON INNOVATIONS

Valdemi Nunes Costa

July/2024

Advisors: Rodrigo Bernardo da Silva
Pedro Rafael Diniz Marinho

Program: Computational Mathematical Modelling

This work proposes a general structure for inference in modelling discrete data sets using a count time series model with INGARCH models with a log-linear structure and mixed Poisson innovations. To this end, a class of probability distributions will be used, the main objective of which is to model count data over time that present a condition of overdispersion. More specifically, the work presents two particular cases: the Inverse Gaussian Poisson log-linear distribution and the Negative Binomial log-linear distribution, which are obtained by considering cases of unobservable data that follow the Inverse Gaussian and Gamma distributions, respectively. The distributions inserted through the mean have as a common point the fact that they are members of the exponential family of distributions. The iterative maximum likelihood method will be used to estimate the model parameters using the EM algorithm. The performance of the estimators will be evaluated through simulation studies using the Monte Carlo method, considering different sample sizes to evaluate the asymptotic behaviour of these estimators.In the section on applying the proposed model to real data sets, three databases were considered for analysis: the first lists the number of hospitalisations due to alcohol abuse in the state of Paraíba, the second evaluates the same problem, but with the data presented for the state of Piauí and, finally, the database consisting of the number of cases of Campylobacter infections in the province of Quebec in Canada was evaluated, thus closing the section on applications to real data. The simulation data was tested using the two proposed extensions and the comparison model called log-linear Poisson proposed by [9], initially taking into account a graphical analysis of the behaviour of the sample, autocorrelation and partial autocorrelation, the study of simulation by means of convergence taking into account the values obtained and the observation of graphs

representing a generalised view of the layout of the simulation data. Subsequently, a reflection was made on its effectiveness through information criteria and the mean square error used in the process of evaluating and choosing the best regression model to adjust the data.

Keywords: Count time series, INGARCH models, EM algorithm.

Índice

Lista de Figuras xii

Lista de Tabelas xiv

1 Introdução 1

2 Revisão Bibliográfica 6

2.1 Breve abordagem sobre as séries temporais . . . 6
2.2 Estacionaridade de séries temporais . . . 7
2.3 Modelos lineares e log-lineares . . . 8
2.4 Distribuições de probabilidade mistas . . . 9
2.5 Distribuições Poisson mistas . . . 10
2.6 Poisson mistas com família exponencial . . . 12
2.7 Casos especiais: Poisson Inversa Gaussiana e Binomial Negativa . . . 13
2.8 Propriedades das Poisson mista . . . 15
2.9 Avaliação de viés de um estimador . . . 16
2.10 Critérios de comparação de modelos . . . 16

3 Método Proposto 18

3.1 Modelo INGARCH log-linear Poisson misto . . . 18
3.2 Aspectos inferenciais . . . 19
3.2.1 Algoritmo EM . . . 19
3.3 Estudos de simulação . . . 26
3.3.1 Cenários de simulação para os parâmetros . . . 26
3.3.2 Simulações para o modelo INGARCH log-linear PIG . . . 28
3.3.3 Simulações para o modelo log-linear BN . . . 32
3.4 Custo computacional de simulação . . . 36

4 Aplicação do modelo a dados reais 37

4.1 Internações por abuso de álcool . . . 37
4.2 Dados de internação por abuso de álcool no estado da Paraíba . . . 38
4.3 Dados de internação por abuso de álcool no estado do Piauí . . . 41

4.3.1 Casos de contaminação por campilobactérias 44

5 Conclusões **48**

Referências Bibliográficas **50**

Lista de Figuras

3.1 Série temporal de contagem, gráfico de autocorrelação e autocorrelação parcial para a PIG, cenário A, tamanho amostral 100. 28
3.2 Boxplots das estimativas dos parâmetros α, ρ, η e ϕ para os cenários [A], [B], [C] e [D] - Modelo INGARCH log-linear PIG. 31
3.3 Série temporal de contagem, gráfico de autocorrelação e autocorrelação parcial para a BN, cenário A, tamanho amostral 100. 32
3.4 Boxplots das estimativas dos parâmetros α, ρ, η e ϕ para os cenários [A], [B], [C] e [D] - Modelo INGARCH log-linear BN. 35

4.1 Série temporal de contagem do número de internações por abuso de álcool na Paraíba, autocorrelação e autocorrelação parcial 38
4.2 Ajuste dos dados pelo modelo log-linear Poisson para dados de internação na Paraíba. 39
4.3 Ajuste dos dados pelo modelo log-linear PIG para dados de internação na Paraíba. 39
4.4 Resíduos para os dados de internação por abuso de álcool na Paraíba. ACF, PACF, densidade estimada e Q-Q plot. 41
4.5 Série temporal de contagem do número de internações por abuso de álcool no Piauí, autocorrelação e autocorrelação parcial 42
4.6 Ajuste dos dados pelo modelo log-linear Poisson para dados de internação no Piauí. 42
4.7 Ajuste dos dados pelo modelo log-linear PIG para dados de internação no Piauí. 43
4.8 Resíduos para os dados de internação por abuso de álcool no Piauí. ACF, PACF, densidade estimada e Q-Q plot. 44
4.9 Série temporal de contagem do número de casos de infecções por campilobactérias, autocorrelação e autocorrelação parcial 45
4.10 Ajuste dos dados pelo modelo log-linear Poisson para dados infecção por campilobactéria. 45
4.11 Ajuste dos dados pelo modelo INGARCH log-linear PIG para dados infecção por campilobactéria. 46

4.12 Ajuste dos dados pelo modelo INGARCH log-linear BN para dados infecção por campilobactéria. 46
4.13 Resíduos para os dados de infecção por campilobactéria. ACF, PACF, densidade estimada e Q-Q plot. 47

Lista de Tabelas

2.1 Distribuições de probabilidade mistas de Poisson. 11

3.1 Valores e funções auxiliares para as distribuições PIG e BN tomando $\xi_0 = -\frac{1}{2}$ e $\xi_0 = -1$ respectivamente 26
3.2 Valores teóricos dos parâmetros para os cenários de simulação da PIG pelo algoritmo EM. 27
3.3 Valores teóricos dos parâmetros para os cenários de simulação da BN pelo algoritmo EM. 27
3.4 Resultados das simulações para os cenários [A], [B], [C] e [D] para a distribuição PIG. 29
3.5 Viés dos estimadores para os cenários [A], [B], [C] e [D] para a PIG. . 30
3.6 Resultados das simulações para os cenários [A], [B], [C] e [D] para a distribuição BN. 33
3.7 Viés dos estimadores para os cenários [A], [B], [C] e [D] para a BN. . 34
3.8 Tempo total de simulação dos cenários [A], [B], [C] e [D] para as distribuições PIG e BN 36

4.1 Valores estimados para os parâmetros e critérios de comparação AIC, BIC e EQM para os dados de estado da Paraíba. 40
4.2 Valores estimados para os parâmetros e critérios de comparação AIC, BIC e EQM para os dados de estado do Piauí. 43
4.3 Valores estimados para os parâmetros e critérios de comparação AIC, BIC e EQM para os dados de infecção por campilobactéria. 47

Capítulo 1

Introdução

Variáveis aleatórias compostas por dados de contagem, ordenadas em escala temporal são frequentemente encontradas nos mais diversos campos de pesquisa, como é o caso das ciências sociais, ciências humanas e da saúde, na astronomia, na física, ou seja, basicamente em todos os ramos do conhecimento. A importância do estudo das séries temporais é indiscutível, tendo em vista sua aplicabilidade e a quantidade de pesquisas e estudos produzidos ao longo da história. Segundo [21], o primeiro registro de análise de dados de séries temporais ocorreu no ano de 1906, quando foi realizado um estudo avaliando o número mensal de manchas solares. O estudo de problema que podem ser expressos como uma série temporal, como sugere [17], está diretamente ligado à necessidade de predições de eventos futuros, tomando como base as informações do passado.

Aplicações de séries temporais de contagem podem ser observadas, por exemplo, ao se estudar o comportamento do número de casos diários de determinada doença e, com isso, estruturar conjecturas que possibilitem elaborar estratégias de combate à enfermidade de forma mais eficaz, como por exemplo é discutido por [10]. Para avaliar a relevância de determinado conteúdo vinculado em um site, é importante verificar a quantidade de acessos por período de tempo a url deste site e, com isso, traçar perfil das preferências dos tipos de publicações a serem veiculadas. Em uma empresa de vendas de pacotes turísticos, se faz necessário, para o planejamento de vendas, observar a série temporal de contagem de modo a ter um panorama que identifique os períodos anuais de maior concentração de vendas e, deste modo, programar suas campanhas promocionais para os períodos cuja procura seja em menor escala. Outras aplicações são discutidas em [25] e [21].

Trabalhos voltados especificamente para séries temporais de contagem são recorrentes na literatura, principalmente nas últimas décadas, como é o caso do trabalho de [25], que discute os principais modelos voltados para o tema. Aplicações das séries temporais de contagem são diversas e abrem uma gama de possibilidades para o estudo de modelos capazes de descrever comportamento de conjuntos de dados

discretos no tempo, afim de melhor compreendê-los, além de avaliar a possibilidade de fazer predições futuras.

São muitos os trabalhos publicados que abordam conjuntos de dados de contagem dispostos no tempo que tem por objetivo descrever, de forma cada vez mais próxima do real, o comportamento deste tipo de variáveis aleatórias. Muitas das publicações trazem propostas de estudo das séries temporais de contagem, e alguns modelos matemáticos que se propõem a modelar estas séries temporais são apresentados com o objetivo de preencher lacunas deixadas por outros publicados em trabalhos anteriores.

Outro estudo que também se propõe a analisar a correlação de dados que apresentam variância maior que a média, denominado ACP (*Poisson Condicional Autoregressivo*), foi apresentado por [7] e teve como aplicação, uma análise do comportamento temporal para o número diário de mudança no valor da ação de uma empresa. Outras abordagens para conjuntos de dados de séries temporais de contagem surgiram no decorrer do tempo, de modo que o modelo INGARCH (*integer-valued generalized autoregressive conditional heteroscedasticity*), proposto por [8], chama a atenção, pois estabelece condições aos parâmetros para um processo estacionário de primeira e segunda ordem. Neste trabalho, [8] apresenta propriedades relacionadas à estimação por máxima verossimilhança e utiliza os resultados para estudar como exemplo de aplicação, o número de casos de infecção por campilobactoriose. Uma restrição apresentada pelo modelo linear de [8] é discutida por [9] e está relacionada ao fato de que o modelo não admite correlações negativas. Outras publicações, como a de [22], que propôs uma classe de modelos INGARCH ampliando a classe MPINGARCH (*mixer Poisson integer-valued generalized autoregressive conditional heteroscedasticity*) em um trabalho de [27], em que foi utilizado o algoritmo EM (*expectation–maximization*), de modo a utilizar este algoritmo para fazer inferência de verossimilhança. O modelo INGARCH log-linear Poisson de [9], contorna a restrição daquele que foi proposto por [8], no entanto, é importante observar que a proposta de [9] não é indicada para conjuntos de dados que apresentam sobredispersão, já que tem como distribuição de probabilidade Poisson como base de formulação. Como já difundido na literatura, este tipo de distribuição é a que melhor se ajusta a dados equidispersos por ter média e variância iguais. Outras publicações discutem subdispersão de dados, como é o caso de [28] denominada COM -Poisson (Conway-Maxwell), que é composta por dois parâmetros e que segundo o autor, "é versátil para dados sub e superdispersos".

A literatura recente também contempla o estudo de séries temporais de contagem, mostrando a importância desses modelos e discutindo sua aplicabilidade nos mais diversos tipos de problemas relacionados a contagem no tempo. Dentre os mais recentes publicados, vale a pena citar o trabalho de [15], que discute a eficiência do

modelo de predição INGARCH para investigar o tráfego de dados de rede; [14], que avalia séries temporais de contagens de crimes utilizando um modelo bivariádo de Poisson; e [13], que utiliza o modelo INGARCH para avaliar riscos climáticos e sua relação com preço e aumento no volume de negociação de ouro. Os trabalhos citados acima são apenas alguns exemplos de abordagem de modelos INGARCH que justificam a importância de propostas de ramificações e adaptações deste tipo de abordagem para séries temporais de dados de contagem.

O modelo INGARCH log-linear Poisson proposto por [9] supera a limitação apresentada pelo modelo linear proposto por [8] de modo a aceitar correlações positivas e negativas, além disso, [9] discute as propriedades do modelo definindo a esperança e a variância e apresentando a função de log-verossimilhança, a função *score* e o gradiente. Em seu trabalho, [9] apresenta uma condição relacionada aos parâmetros que garante a estacionaridade da série temporal de contagem. Além da vantagem citada anteriormente de poder admitir também covariância negativa o modelo também pode incluir covariáveis que resultam em coeficientes de regressão negativos, sem que a média do processo de Poisson se torne negativa. No entanto, uma limitação deste modelo se dá pelo fato de que ao assumir a forma log-linear com distribuição de Poisson assume-se também que os dados são equidispersos, ou seja, que este tenham média e variância iguais, o que reduz a abrangência e eficácia do modelo no que diz respeito a dados cuja variância excede a média.

O objetivo geral deste trabalho é ampliar a classe de modelos INGARCH, através uma extensão da proposta apresentada por [9], no qual a distribuição Poisson é substituída por outra distribuição pertencente à classe de distribuições Poisson mistas. O modelo denominado INGARCH log-linear Poisson misto, faz a substituição da Poisson por uma distribuião de probabilidade obtida quando a média do processo Poisson assume outra distribuição que é um membro da família exponencial. Neste texto, serão abordados especificamente duas propostas que são os modelos INGARCH log-linear Poisson Inversa Gaussiana (PIG) e INGARCH log-linear binomial negativo (BN). Com isso, abrem-se possibilidades para proposta de outros trabalhos que contemplem modelos além dos que aqui são apresentados e que utilizem outros membros que pertençam à família exponencial. A depender do comportamento da variável aleatória envolvida no processo, define-se a escolha de uma distribuição que melhor se adapte às suas características destes conjuntos de dados.

Para alcançar o objetivo geral, são exploradas as propriedades de máxima verossimilhança, verificando sua consistência quando o tamanho da amostra sofre acréscimo de observações, avaliando os estimadores através da convergência dos valores estimados com relação aos valores teóricos estudando este comportamento por meio da raiz do erro quadrático médio, do viés do estimador, além da análise gráfica

que permite avaliar a proximidade entre estes valores para os diferentes tamanhos de amostra. É possível verificar que o parâmetro de dispersão, que descreve o comportamento dos dados ao redor da média, não depende dos outros parâmetros que serão estimados para o modelo.

São avaliadas as estimações de parâmetros analisando o comportamento assintótico através de estudos de simulação que são desenvolvidas com o uso de linguagem R, que é uma linguagem de programação descrita por [11] como "totalmente funcional, adequada para aplicações científicas e programação em geral". Para melhor desenvolver a simulação, foi usada a IDE (ambiente de desenvolvimento integrado) Rstudio em que foram utilizados o pacote `tseries`[24] usado em análise de séries temporais e o pacote `gamlss.dist`[23], que é utilizado para modelos de distribuições discretas, como é o caso da PIG e da BN, ambos aplicados ao algoritmo EM (expectation–maximization). O algoritmo construído para o modelo apresenta comandos e funções que estão no repositório oficial do R, e seus detalhes são apresentados em [20].

Para melhor estruturar este trabalho, facilitando o desenvolvimento e a conexão da proposta a ser discutida, dividiu-se sua composição nas partes descritas a seguir.

O Capítulo 2 apresenta uma abordagem dos conceitos de séries temporais de contagem e dos modelos INGARCH linear descrito por [8] e INGARCH log-linear Poisson apresentado por [9], além de suas propriedades básicas. É feita ainda uma apresentação das distribuições de probabilidades mistas com membros da família exponencial, enunciando as principais propriedades e enfatizando as distribuições Inversa Gaussiana e Gama, que são usadas na definição do modelo INGARCH log-linear Poisson misto para os casos particulares estudados, que serão propostos no capítulo subsequente. São descritos ainda os conceitos e as principais características dos modelos mistos, como esperança e variância, além de apresentar ao final do capítulo, os conceitos relacionados aos estimadores, como a definição de erro quadrático médio e viés do estimador. Também são abordados os conceitos de AIC (*Akaike information criterion*) e BIC (*Bayesian information criterion*), que são critérios comparação entre modelos competidores.

No Capítulo 3, é definido o modelo log-linear Poisson misto de modo a se analisar os aspectos inferenciais por meio de verossimilhança. Esta abordagem é feita através do algoritmo EM, com o cálculo da esperança da log-verossimilhança inicialmente e, em seguida atualizar as esperanças condicionais. Para isso, é feito um estudo da função de log-verossimilhança completa, determinando as esperanças condicionais citadas anteriormente. O capítulo é finalizado com uma síntese da decomposição das distribuições Inversa Gaussiana e Gama como membros da família exponencial. O estudo de simulação também é apresentado no Capítulo 3, e consta

das tabelas com valores simulados, as raízes dos erros quadráticos médios e, em tabela separada, são apresentados os vieses dos estimadores. Este estudo de simulação, também apresenta em sua composição a análise dos boxplots gerados a partir dos dados de simulação, os gráficos das séries temporais de contagem geradas aleatoriamente na simulação e seus respectivos gráficos de autocorrelação e autocorrelação parcial, que permitem avaliar a estacionaridade dos dados obtidos.

O Capítulo 4 apresenta as aplicação do modelo INGARCH log-linear Poisson misto a três bancos de dados reais, compostos por dois grupos que contemplam o número de internações hospitalares por abuso de álcool nos estados da Paraíba e no segundo caso, do Piauí, respectivamente, e um terceiro conjunto que apresenta os dados de casos de infeção por campilobactérias na província de Quebec, no Canadá. Neste Capítulo ainda é feita uma análise do comportamento dos modelos log-linear PIG e BN que são propostos neste trabalho, comparando os resultados obtidos aos apresentados pelo modelo INGARCH log-linear Poisson proposto por [9], inicialmente através de uma análise gráfica e posteriormente utilizando critérios de avaliação de modelo para comparação entre eles.

O fechamento do trabalho se dá no Capítulo 5, fazendo um breve apanhado dos produtos obtidos no decorrer do processo através de simulação, aplicação e comparação dos modelos abordados, dos aspectos positivos e negativos observados na construção do processo de pesquisa, finalizando o trabalho com as considerações finais.

As rotinas computacionais em linguagem de programação R para a simulação da PIG e da BN apresentadas no trabalho encontram-se disponíveis nos apêndices, e os resultados aqui expostos podem ser reproduzidos usando para isso a semente `set.seed(10)`.

Capítulo 2

Revisão Bibliográfica

Neste capítulo, é apresentada uma abordagem geral dos principais conceitos utilizados na construção do modelo proposto neste trabalho, tomando como base os resultados estabelecidos na literatura de séries temporais de contagem que envolvem definições e propriedades das distribuições de probabilidades, de características das séries temporais, em especial as de contagem, a avaliação de valores estimados por meio de definição de raiz do erro quadrático médio e do viés de um estimador, além de apresentar modelos e contribuições relevantes da literatura atual e definir os critérios de escolha usados para avaliação de desempenho.

2.1 Breve abordagem sobre as séries temporais

Eventos que podem ser representados por séries temporais são encontrados frequentemente em diversas áreas das ciências. Exemplos de séries temporais na área de saúde como no trabalho de [1], avalia problemas relacionados aos coração. Um eletrocardiograma, que consta de uma série temporal que serve de exame inicial para verificar possíveis doenças relacionadas ao coração, avaliando comportamentos da variável que sugere irregularidades no ritmo de batimentos ao longo do tempo. Na Geofísica as séries temporais são obtidas por sismógrafos que captam a frequência de terremotos, em que podem ser analisadas, por exemplo, a amplitude, que avalia o tamanho do movimento sísmico, a magnitude, que mede a energia liberada no intervalo de tempo de duração do terremoto. Na economia, as séries temporais modelam, por exemplo, o comportamento do valor das ações no decorrer do tempo, permitindo que o investidor tenha uma ideia de quando comprar ou vender estas ações [3]. São muitas manifestações e aplicações das séries temporais, o que torna seu estudo de grande relevância já que este permite criar modelos de previsão que descrevam fenômenos e propiciam formular hipóteses de previsões futuras. Um conceito mais formal para as séries temporais é apresentado a seguir.

Definição 2.1.1 *Define-se por série temporal uma sequência de variáreis aleatórias* $\{Y_t, t \geq 0\}$ *que apresentam correlação temporal.*

A estrutura conceitual dada às séries temporais, que é apresentada na Definição 2.1.1 contempla variáveis aleatórias de modo geral. Como a proposta de abordagem a ser apresentada abrange apenas dados de contagem, é necessário que se restrinja a imagem do conjunto das variáveis aleatórias a valores discretos, o que é feito na definição a seguir.

Definição 2.1.2 *Se as observações* $Y_t = \{y_1, y_2, \cdots, y_t\}$ *são tais que* $y_i \in \mathbb{N}$*, então diz-se que a série temporal* $\{Y_t, t \geq 0\}$ *é uma série temporal de contagem.*

2.2 Estacionaridade de séries temporais

De maneira informal, uma série temporal é dita estacionária se as observações Y_t, $t = 1, \ldots$ mantem características estatísticas como média e variância aproximadamente constantes ao longo do tempo.

Basicamente, dois tipos principais de estacionaridade são definidos, a saber: estacionaridade forte, que apresenta média e variância constantes e a covariância entre pontos distintos não depende do tempo, além de também não mudar a sua estrutura de correlação; a estacionaridade fraca, que é quando a série temporal apresenta apenas média e variância constantes. De maneira mais formal, as seguintes definições de estacionaridade são fornecidas por [21].

Definição 2.2.1 *Uma série temporal estritamente estacionária é aquela para a qual o comportamento probabilístico de qualquer coleção de valores* $Y_{t_1}, Y_{t_2}, \cdots, Y_{t_k}$ *é idêntico ao do conjunto com deslocamento no tempo* $Y_{t_1+h}, Y_{t_2+h}, \cdots, Y_{t_k+h}$*. Isso significa que:*

$$P(Y_{t_1} \leq y_{t_1}, \cdots, Y_{t_k} \leq y_{t_k}) = P(Y_{t_1+h} \leq x_{t_1+h}, \cdots, Y_{t_k+h} \leq y_{t_k+h})$$

para todos os $k = 1, 2, \ldots$*, todos os pontos temporais* $t_1, t_2, \cdots, t_k$*, todos os números* $y_{t_1}, \cdots, y_{t_k}$*, e todos os deslocamentos no tempo* $h = 0, \pm 1, \pm 2, \ldots$.

Da Definição 2.2.1 tem-se que para qualquer deslocamento h no tempo, a série temporal mantém suas características estatísticas como a média, a variância e a covariância.

Definição 2.2.2 *Uma série temporal fracamente estacionária* Y_t*, é um processo de variância finita, tal que*

(i) a função de valor médio μ_t *é constante e não depende de* t*;*

(ii) a função de autocovariâncias, $\gamma(s, t)$*, depende apenas de* s *e* t *através da diferença entre eles,* $|s - t|$*.*

2.3 Modelos lineares e log-lineares

Na introdução deste texto foram mencionados alguns trabalhos que abordam modelos INGARCH e suas aplicações. No que se segue, serão definidos a seguir os modelos de maior relevância de modo a apresentar as suas principais propriedades. A primeira proposta de estrutura de modelo que será discutida, foi formulada por [8], que é denominado modelo linear, cuja variável aleatória segue uma distribuição Poisson. No artigo são discutidas as definições de valor esperado, de variância do modelo, além desenvolver um estudo por meio de verossimilhança. A definição do modelo linear de [8] é apresentada a seguir.

Definição 2.3.1 *Seja Y_t uma série de contagem e $\mathcal{F}_t$ é uma σ-álgebra gerada pelas variáveis aleatórias, $Y_1, \ldots, Y_n$, em que λ_t é a média condicional do processo de Poisson. Um processo INGARCH(p,q) é definido como um processo de valores inteiros $\{Y_t\}_{t\in\mathbb{Z}}$ tal que*

$$\begin{cases} Y_t|\mathcal{F}_{t-1} \sim Poisson(\lambda_t); \quad \forall\, t \in \mathbb{Z} \\ \lambda_t = \gamma_0 + \sum_{i=1}^{q} \gamma_i Y_{t-i} + \sum_{j=1}^{p} \delta_j \lambda_{t-j} \end{cases}$$

em que $\gamma_0 > 0, \gamma_i \geq 0, i = i, \cdots, q, \delta_0 \geq j$ e $j = 1, \cdots, p$.

Tomando então $p = q = 1$, o modelo INGARCH$(1,1)$ é dado por

$$\begin{cases} Y_t|\mathcal{F}_{t-1} \sim Poisson(\lambda_t) \\ \lambda_t = \gamma_0 + \gamma_1 Y_{t-1} + \delta_1 \lambda_{t-1} \end{cases} \tag{2.1}$$

em que $t \in \mathbb{Z}$.

O modelo linear Poisson descrito por [8] permite modelar séries temporais de contagem tais que a correlação entre as observações admite apenas valores positivos. Dada a simplicidade do modelo linear Poisson, a estimação dos parâmetros através do método de máxima verossimilhança não apresenta dificuldades, sendo este então um modelo de fácil implementação e que, como ressalta [9], apresenta comportamento assintoticamente normal para seus parâmetros sob determinadas condições. Uma conclusão sobre pontos negativos do modelo linear que é apresentada por [9] e diz respeito ao fato de que este não admite correlação negativa já que a e b são parâmetros que devem atender à restrição $|a + b| < 1$. Visando preencher a lacuna deixada pelo modelo linear proposto por [8], é que [9] introduz então o modelo INGARCH log-linear Poisson, cuja definição apresentamos a seguir.

Definição 2.3.2 *Suponha que $\mathcal{F}_t$ denote uma σ-álgebra gerada por $\{Y_0, \cdots, Y_t\}$. O*

modelo INGARCH log-linear Poisson é dado por

$$\begin{cases} Y_t|\mathcal{F}_{t-1} \sim Poisson(\lambda_t); \quad \forall\, t \in \mathbb{Z} \\ \nu_t = d + a\nu_{t-1} + b \cdot \log(Y_{t-1} + 1) \end{cases} \tag{2.2}$$

em que d, a *e* $b \in \mathbb{R}$, $\nu_t \equiv \log(\lambda_t)$, *com* $t > 0$.

Neste modelo, é possível a inserção de correlação negativa além da correlação positiva. Algumas vantagens desse modelo são citadas por [9], como é o caso da inserção de defasagem por meio do termo $\log(Y_{t-1} + 1)$ que permite o mapeamento de zeros. Entretanto, este modelo não é adequado para conjuntos de dados sobre-dispersos já que o modelo segue a distribuição Poisson, cujas média e a variância são iguais. Neste sentido, surge então a proposta do modelo INGARCH log-linear Poisson misto, que torna possível a inclusão de outros modelos probabilísticos que melhor se adequem à sobredispersão como média condicional do processo de Poisson, ajustando melhor o modelo para esse tipo de conjuntos de dados.

2.4 Distribuições de probabilidade mistas

As distribuições de probabilidade mistas são uma família de distribuições que vem sendo estudadas para elaboração de modelos capazes de capturar com maior precisão, características de conjuntos de dados sobredispersos. Um nome que se destaca no estudo deste tipo de distribuição de probabilidade é [12], que apresenta as principais definições e propriedades. As principais destas definições são apresentadas abaixo.

Definição 2.4.1 *Uma distribuição de probabilidade é considerada uma distribuição mista se sua função de distribuição* $F(\cdot)$ *puder ser escrita na forma:*

$$F(\cdot) = \int_{\Theta} f(\cdot|\lambda) dG(\lambda)$$

em que $f(\cdot|\lambda)$ *representa a função de distribuição dos compoentes considerados, que são indexados pelo parâmetro* λ *com função de distribuição* $G(\lambda), \lambda \in \Theta$.

A definição de função densidade de probabilidade também é apresentada por [12] e é dada a seguir:

$$f(y) = \int_{\Theta} F(y|\lambda) g_{\lambda}(\lambda)\, d\lambda,$$

em que a função mista é denotada por $f(y|\lambda) \underset{\lambda}{\wedge} g(\lambda)$, $g(\cdot)$ é chamada de densidade da mistura.

O valor esperado da função $h(Y)$, independentemente da forma de $f(y|\lambda)$, é dado por

$$E[h(Y)] = \int_{\Theta} E_{Y|\lambda}[h(Y)]g(\lambda)\, d\lambda.$$

Com isso, tem-se que a esperança e a variância são dadas em relação a distribuição codicional de Y por

$$E(X) = E[E_{Y|\lambda}(Y)] \quad \text{e} \quad V[Y] = V[E_{Y|\lambda}(Y)] + E[V_{Y|\lambda}(Y)].$$

2.5 Distribuições Poisson mistas

Um membro proeminente do grupo de distribuições misto é o caso da Poisson mista também apresentada por [12]. Este caso é fundamental para a construção do modelo a ser apresentado no próximo capítulo deste texto. Para a definição da mesma, assume-se que $F(\cdot)$ da Definição 2.4.1 segue uma distribuição de Poisson. Este tipo de distribuição apresenta-se como um tema largamente abordado na atualidade, dada a sua importância e seu potencial na construção de modelos para os mais variados tipos de conjuntos de dados de contagem. Sua definição mais rigorosa, apresentada por [12], é transcrita a seguir.

Definição 2.5.1 *Seja Y uma variável aleatória que segue uma distribuição Poisson mista que possui uma função de densidade g se sua função de probabilidade é dada por*

$$P(Y = y) = \int_0^{\infty} \frac{e^{-\lambda} \cdot \lambda^y}{y!} g(\lambda) d\lambda, \quad y = 1, 2, 3, \ldots. \tag{2.3}$$

Assim como assumido por [12], denominam-se por $F(\cdot\ |\lambda)$ as distribuições Poisson mistas com probabilidade como a Definição 2.5.1. Abaixo, a Proposição 2.5.1 apresenta propriedades das distribuições Poisson mistas que também foram discutidas por [12].

Proposição 2.5.1 *Seja Y uma variável aleatória cuja distribuição é uma distribuição de Poisson mista. Então, os seguintes resultados são válidos*

$$(a) \quad P(Y \leq y) = \int_0^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} G(\lambda)\, d\lambda$$

$$(b) \quad P(Y > y) = \int_0^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} [1 - G(\lambda)]\, d\lambda,$$

em que $G(\lambda) = \int_0^{\lambda} g(y)dy$ é função de distribuição de parâmetro λ.

Já são muitos os trabalhos publicados envolvendo modelos de distribuições Poisson mistas. Nestas propostas de trabalhos, a densidade de mistura $g(\lambda)$ assume diversos tipos de distribuição, cujo propósito é alcançar o melhor ajuste para tipos de variáveis aleatórias com características distintas. A Tabela 2.1 apresenta alguns modelos de distribuições Poisson mistas, cujo levantamento das mesmas foi feito por [12]. Ao avaliar esta tabela, é possível perceber a importância desse tipo de mistura, dada a quantidade de modelos que já foram propostos.

Tabela 2.1: Distribuições de probabilidade mistas de Poisson.

Distribuição Poisson Mista	**Distribuição de Mistura**	**Referência Importante**
Binomial Negativa	Gama Greenwood	Yule (1920)
Geométrica	Exponencial	Johnson *et. al.* (1992)
Família Poisson – Exponencial Linear	Família Exponencial Linear	Sankaran (1969)
Poisson – Lindley	Lindley	Sankaran (1970)
Poisson – Linear Exponencial	Linear exponencial	Kling & Goovaerts(1993)
Poisson – Lognormal	Loggnormal	Bulmer (1974)
Série Hipergeométrica Poisson – Confluente	Séries Hipergeométricas Confluentes	Bhattacharya (1966)
Poisson – Inversa Gaussiana Generalizada	Inversa Gaussiana Generalizada	Sichel (1974)
Sichel	Inversa Gaussiana	Sichel (1975)
Poisson – Inversa Gama	Inversa Gama	Willmot (1993)
Poisson – Normal Truncada	Normal Truncada	Patil (1964)
Guerra Generalizada	Proporção de Produto Gama	Irwin (1975)
Guerra Simples	Exponencial Beta	Pielou (1962)
Yule	Beta com Parâmetros de Valores Específicos	Simon (1955)
Poisson – Pareto Generalizada	Beta Pareto Generalizada Tipo I	Kempton (1975)
Poisson – Beta I	Beta Tipo I	Holla e Bhattacharya (1965)
Poisson – Beta II	Beta Tipo II	Gurland (1958)
Poisson – Beta Truncada II	Beta Truncada Tipo II	Willmot (1986)
Poisson – Uniforme	Uniforme	Bhattacharya (1966)
Poisson – Gama Truncada	Gama Truncada	Willmot (1993)
Poisson – Gama Generalizada	Gama Generalizada	Albrecht (1984)
Dellaporte	Gama Deslocada	Ruohonen (1988)
Poisson – Bessel Modificada do Terceiro Tipo	Bessel Modificada do Terceiro Tipo	Ong & Muthaloo (1995)
Poisson – Pareto	Pareto	Willmot (1993)
Poisson – Pareto Deslocada	Pareto Deslocada	Willmot (1993)
Família Poisson – Pearson	Família de Distribuições de Pearson	Albrecht (1982)
Poisson – Log – Student	Log – Student	Gaver & O'Muircheartaigh (1987)
Poisson – Função de Potência	Distribuição de Função Potência	Rai (1971)
Poisson – Lomax	Lomax	Al – Awadhi & Ghitany (2001)
Poisson – Poder de Variação	Família de Poder de Variação	Hougaard *et. al.* (1997)
Neyman	Poisson	Douglas (1980)
	Outras distribuições Discretas	Johnson *et. al.* (1992)

Fonte: Karlis e Xekalaki, 2005 [12]

Uma atualização para a Tabela 2.1 das distribuições Poisson mistas apresentadas por [12] é abordada por [18], que exprime algumas ramificações mais recentes de membros obtidos pela mistura.

2.6 Poisson mistas com família exponencial

Com o modelo Poisson misto, muitas possibilidades de distribuições de probabilidade podem ser propostas. Uma dessas possibilidades é a distribuição Poisson mista, que tem sua média condicional dada por uma variável aleatória que segue uma distribuição que seja membro da família exponencial tratada por [22], que trazem a Definição 2.6.1. Para melhor explorar as distribuições Poisson mistas em que a componente de mistura seja membro da família exponencial, são apresentadas as proposições que determinam a esperança e a variância para as distribuições que compõem esta família, de modo a permitir posteriormente uma melhor abordagem para os casos particulares tratados neste texto.

Definição 2.6.1 *Seja Z uma variável aleatória contínua e positiva com função de densidade dada por*

$$f(z) = \exp\{\phi[z\xi_0 - b(\xi_0)] + c(z;\phi)\}, \quad z > 0, \quad \phi > 0, \tag{2.4}$$

em que $b(\cdot)$ é uma função contínua e diferenciável três vezes, ξ_0 e tal que $b'(\xi_0) = 1$ e $c(\cdot,\cdot)$ é uma função de domínio $\mathbb{R}^+ \times \mathbb{R}^+ \subset \mathbb{R}$.

Diz-se então que Z pertence à família exponencial, denotada por $Z \sim EF(\phi)$.

Assumindo ainda que a função $c(\cdot\ ;\cdot)$ pode ser escrita como uma soma de funções da forma $c(z;\phi) = d(\phi) + \phi g(z) + h(z)$, com $d(\phi)$ três vezes diferenciável.

Para a família exponencial, tem-se que a função geradora de momentos é dada por

$$M_Z(t) = E(e^{tz}) = \int_0^\infty e^{tz} \exp\{\phi[z\xi_0 - b(\xi_0)] + c(z;\phi)\}dz. \tag{2.5}$$

Os resultados apresentados a seguir são obtidos através das derivadas da Equação (2.5), que, através de manipulações algébricas, possibilitam a dedução do valor esperado e da variância para um membro da família de distribuições exponenciais.

Proposição 2.6.1 *O valor esperado para um membro da família exponencial é dada por*

$$E(Z) = b'(\xi_0).$$

Proposição 2.6.2 *A variância para um membro qualquer da família exponencial é dada por*

$$Var(Z) = \frac{1}{\phi} b''(\xi_0).$$

2.7 Casos especiais: Poisson Inversa Gaussiana e Binomial Negativa

Dois casos especiais de distribuição Poisson mista com membros da família exponencial serão considerados aqui, já que são componentes do modelo proposto. O primeiro deles é a PIG, que é o misto da distribuição Poisson com a distribuição Inversa Gaussiana. O segundo caso é o da binomial negativa, que é obtida a partir da distribuição gama como membro da família exponecial.

A distribuição PIG é obtida através da mistura de Poisson com a distribuição Inversa Gaussiana através do parâmetro λ. É possível mostrar que esta pode ser escrita como membro da família exponencial, como é exposto a seguir. A função densidade de probabilidade Inversa Gaussiana é dada por

$$f(z;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi z^3}}\ \exp\left\{-\frac{\lambda(z-\mu)^2}{2\mu^2 z}\right\}.$$

Note que a expressão acima pode ser reescrita como expoente na base da constante de Euler de modo a assumir a forma da equação

$$f(z;\mu,\lambda) = \exp\left\{\frac{1}{2}\log\left(\frac{\lambda}{2\pi z^3}\right)\right\}\ \exp\left\{-\frac{\lambda(z-\mu)^2}{2\mu^2 z}\right\}.$$

Portanto,

$$\begin{aligned} f(z;\mu,\lambda) &= \exp\left\{\frac{1}{2}\log\left(\frac{\lambda}{2\pi z^3}\right) - \frac{\lambda z}{2\mu^2} + \frac{\lambda}{\mu} - \frac{\lambda}{2z}\right\} \\ &= \exp\left\{\lambda\left[-\frac{z}{2\mu^2} + \frac{1}{\mu}\right] + \frac{1}{2}\log\left(\frac{\lambda}{2\pi z^3}\right) - \frac{\lambda}{2z}\right\} \\ &= \exp\left\{\lambda\left[-\frac{z}{2\mu^2} + \frac{1}{\mu}\right] + \frac{1}{2}\log(\lambda) - \frac{1}{2}\log(2\pi z^3) - \frac{\lambda}{2z}\right\} \end{aligned}$$

Tomando $\lambda = \phi$, tem-se então $\xi_0 = -\frac{1}{2\mu^2}$, ou seja, $\mu = \sqrt{-\frac{1}{2\xi_0}}$. Substituíndo o valor obtido em $b(\xi_0)$, tem-se $b(\xi_0) = -\frac{1}{\mu} = -\frac{1}{\frac{1}{\sqrt{2\xi_0}}} = -\sqrt{-2\xi_0}$ e $c(z;\phi) = \frac{1}{2}\log(\phi) - \frac{1}{2}\log(2\pi z^3) - \frac{\phi}{2z}$ e então $f(z;\mu,\phi)$ assume a forma exponencial.

Como $c(z;\phi) = d(\phi) + \phi g(z) + h(z)$, é possível concluir então que as funções que compõem $c(\cdot;\cdot)$ são dadas por $d(\phi) = \frac{1}{2}\log\phi$, $g(z) = -\frac{1}{2z}$ e $h(z) = -\frac{1}{2}\log(2\pi z^3)$.

Deste modo, é possível notar que para o caso em que $G(\cdot)$ da equação (2.5) é a Inversa Gaussiana na forma exponencial, então a função de distribuição $F(\cdot)$ resultante da mistura é a PIG.

Uma definição mais refinada para a classe de distribuição de Poisson mista é

abordada de acordo com [2], já que esta definição, da forma como se apresenta, se adequa melhor aos casos que referem-se família exponencial que a definição sugerida por [12].

Definição 2.7.1 *Seja* $Y|Z = z \sim Poisson(\mu z)$, *com* $\mu > 0$ *e* $Z \sim EF(\phi)$. *Diz-se então que a variável aleatória* Y *pertence à classe de distribuições Poisson mista (MP), e esta é denotada por* $Y \sim MP(\mu, \phi)$.

Ao escolher Z como uma distribuição Inversa Gaussiana com média 1 e parâmetro de dispersão ϕ, obtem-se Y seguindo uma distribuição PIG (denotada por $PIG(\phi, \mu)$) com função de probabilidade dada por

$$P(Y = y) = \sqrt{\frac{2}{\pi}}[\phi(\phi + 2\mu)]^{-\frac{(y-1/2)}{z}} \frac{e^{\phi}(\mu\phi)^y}{y!} \mathcal{K}_{y-1/2}(\sqrt{\phi(\phi + 2\mu)}), \quad Y \in \mathbb{N},$$

em que

$$\mathcal{K}_{\lambda}(t) = \int_0^{\infty} u^{\lambda-1} \exp\left\{-\frac{t}{z}\left(u + \frac{1}{u}\right)\right\} du$$

é a função de Bessel modificada do terceiro tipo para $\lambda \in \mathbb{R}$ e $t > 0$.

De modo análogo ao que é feito com a distribuição Inversa Gaussiana para determinar a PIG, a distribuição Gama também pode ser escrita como membro da família exponencial. Para mostrar isso, basta observar então que

$$\begin{aligned} f(z; \alpha, \beta) &= \frac{\beta^{\alpha}}{\Gamma(\alpha)} z^{\alpha-1} \exp\{-\beta z\} \\ &= \exp\{\alpha \ \log \ \beta + (\alpha + 1) \log \ z - \beta z - \log \ \Gamma(\alpha)\} \\ &= \exp\left\{\alpha \left[-\frac{\beta}{\alpha} z + \log \ \beta\right] + (\alpha - 1) \log \ z - \log \ \Gamma(\alpha)\right\}. \end{aligned}$$

Fazendo $\phi = \alpha$ e $\xi_0 = -\frac{\beta}{\alpha}$, tem-se que $\beta = -\phi\xi_0$. Deste modo, tem-se

$$\begin{aligned} f(z) &= \exp\left\{\phi\left[\xi_0 z + \log\left(-\phi\xi_0\right)\right] + (\phi - 1) log \ z - \log \ \Gamma(\phi)\right\} \\ &= \exp\left\{\phi\left[\xi_0 z + \log\left(-\xi_0\right)\right] \phi \ \log \ \phi + (\phi - 1) \log \ z - \log \ \Gamma(\phi)\right\}. \end{aligned}$$

Basta então observar que $b(\xi_0) = -\log\ (-\xi_0)$ e que $c(z; \phi) = \phi \ \log \ \phi + (\phi - 1) \log \ z - \log \ \Gamma(\phi)$. Deste modo, a decomposição da função $c(\cdot; \cdot)$ tem como suas funções: $d(\phi) = \phi \ \log \ \phi - \log \ \Gamma(\phi)$, $g(z) = \log \ z$ e $h(z) = -\log \ z$.

Retomando a Definição 2.7.1, supor que Z segue uma distribuição gama com média 1 e parâmetro de dispersão ϕ, e deste modo, podemos obter Y seguindo a distribuição binomial negativa (denotada por $BN \sim (\phi, \mu)$), cuja função de proba-

bilidade é expressa por

$$P(Y=y)=\frac{\Gamma(y+\phi)}{y!\Gamma(\phi)}\left(\frac{\mu}{\mu+\phi}\right)^{y}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}$$

de modo que $\Gamma(\cdot)$ é a função gama e $y \in \mathbb{N}$.

2.8 Propriedades das Poisson mista

Tanto para o caso da PIG e da BN quanto para qualquer distribuição obtida por meio de uma mistura da Poisson com variável latente que segue uma distribuição de um membro da família exponencial, pela equação (2.7) da Seção 2.4, tem-se que $Y \sim MP(\phi,\mu)$:

$$E[h(y)]=\int_0^{\infty} E_{Y|Z}(h(y))\exp\{\phi[z\xi_0-b(\xi_0)]+c(z;\phi)\}dz$$

Portanto, a função geradora de momentos da distribuição MP, denotada por φ_y, é dada por:

$$\varphi_Y(t)=E(e^{ty})=\int_0^{\infty} E_{Y|Z}(e^{ty})\exp\{\phi[z\xi_0-b(\xi_0)]+c(z;\phi)\}dz$$

em que

$$E_{Y|Z}(e^{ty})=\exp\{\mu z(e^t-1)\}.$$

Assim, ao afetual as manipulações algébricas tem-se então que a função geradora de momentos condicional, denotada por $Y \sim$MP(μ,ϕ) tem a representação

$$\varphi_Y(t)=\exp\left\{\phi\left[b\left(\xi_0+\frac{\mu(e^t-1)}{\phi}\right)-b(\xi_0)\right]\right\},$$

com t pertencendo a algum intervalo que contém zero.

Com resultado obtido na equação (2.16), podemos determinar a esperança e a variância para a Poisson mista.

Proposição 2.8.1 *A esperança da distribuição Poisson mista $Y \sim MP(\phi,\mu)$ é dada por*

$$E(Y)=\mu,$$

desde que $b'(\xi_0)=1$.

Proposição 2.8.2 *A variância da distribuição Poisson mista $Y \sim MP(\phi,\mu)$ é dada por*

$$Var(Y)=\mu+\frac{\mu^2}{\phi}b''(\xi_0).$$

Todos os resultados apresentados acima compõem o suporte teórico para definição e estudo do modelo INGARCH log-linear Poisson misto, que é apresentado no capítulo subsequente.

2.9 Avaliação de viés de um estimador

Para avaliar as estimativas obtidas no processo de simulação, são utilizados o viés do estimador e a raiz do erro quadrático médio. O viés de um estimador pode ser descrito como a diferença entre o valor estimado e o valor do parâmetro. Uma definição mais formal é apresentada a seguir, adaptando à notação utilizada neste texto para a definição apresentada por [6].

Definição 2.9.1 *Um estimador $\hat{\theta}$ é não viesado para θ, se $E[\hat{\theta}] = \theta$, independentemente do valor de θ. A diferença $E[\hat{\theta}] - \theta$ é chamada de viés de $\hat{\theta}$.*

Quando $E[\hat{\theta}] - \theta \neq 0$, afirma-se então que $\hat{\theta}$ é um estimador viesado. Outra forma de avaliar a eficiência do valor estimado é através da raiz do erro quadrático médio, também apresentada por [6].

Definição 2.9.2 *Seja $\hat{\theta}$ um estimador para o parâmetro θ. Define-se como erro quadrático médio de $\hat{\theta}$ ao valor numérico dado por*

$$EQM(\hat{\theta}) = E[\hat{\theta} - \theta]^2.$$

O erro quadrático médio ainda pode ser expresso por

$$EQM(\hat{\theta}) = Var(\hat{\theta}) + (E[\hat{\theta}] - \theta)^2,$$

em que é possível então definir a raiz quadrada do erro quadrático médio por

$$REQM(\hat{\theta}) = \sqrt{Var(\hat{\theta}) + (E[\hat{\theta}] - \theta)^2}.$$

Além da avaliação dos valores estimados através do viés e da raiz do erro quadrático médio, também são analisados os boxplots, os gráficos de autocorrelação e autocorrelação parcial, para melhor compreensão do comportamento dos dados da série temporal de contagem obtida no processo de simulação.

2.10 Critérios de comparação de modelos

Para avaliar a eficiência de modelos de regressão, bem como para comparar dois ou mais modelos, basicamente dois métodos se destacam na literatura: o AIC

(critério de informação de Akaike) e o BIC (critério de informação Bayesiano). Esses critérios são utilizados para avaliar modelos de modo a indicar qual deles é o mais parcimonioso. Os critérios de informação foram estudados por [4], que apresenta a definição para o AIC pela equação

$$AIC = -2\log(L(\hat{\Theta})) + 2k, \tag{2.6}$$

em que $L(\Theta)$ é a função de verossimilhança aplicada aos valores maximizados para os parâmetros, e k é o número de parâmetros estimados pelo modelo de regressão. Já o BIC é dado em função dos mesmos parâmetros e do tamanho amostral n, e pode ser obtido pela equação

$$BIC = -2\log(L(\hat{\Theta})) + k\log(n). \tag{2.7}$$

Ao analisar os critérios de informação AIC e BIC, [4] ressalta que "os modelos, por definição, são apenas aproximações da realidade ou verdade desconhecidas; não existem modelos que reflitam perfeitamente toda realidade". Os modelos buscam aproximar fenômenos reais por meio de equações que capturem o máximo de informações sobre seu comportamento tornando possível a sua compreensão, pelo menos de forma geral. Nesta perspectiva, [4] ainda complementa ao afirmar que "efeitos menores muitas vezes só podem ser revelados à medida em que o tamanho da amostra aumenta".

Capítulo 3

Método Proposto

Neste capítulo é feita a introdução do modelo INGARCH log-linear Poisson mistos e suas propriedades, além do estudo das equações para a esperança condicional da função de log-verossimilhança para estimação dos parâmetros do modelo utilizando o algoritmo EM. Também serão abordados os estudos de simulação que apresentam o processo de estimação dos parâmetros por meio do método de Monte Carlo. A simulação contempla quatro cenários com quatro tamanhos de amostras para os dois casos que são abordados neste trabalho. São observados o gráfico da série temporal de contagem, de autocorrelação e autocorrelação parcial de dados gerados aleatoriamente, afim de observar a estacionaridade do processo. Para avaliar os estimadores são analisadas as raízes dos erros quadráticos médios, o viés dos estimadores e os boxplots.

3.1 Modelo INGARCH log-linear Poisson misto

A proposta a seguir tem como objetivo apresentar um modelo para séries temporais capaz de modelar conjuntos de dados sobredispersos. Para isso, define-se um modelo de misturas entre a distribuição Poisson com uma distribuição que possa assumir a forma de um membro da família exponencial de distribuições.

A partir desta definição, serão abordados dois casos de distribuições obtidas através desta mistura em que, no primeiro deles, a média do processo de Poisson indexa a distribuição Inversa Gaussiana, obtendo assim a distribuição PIG. No segundo caso, de modo análogo, a distribuição abordada é a distribuição Gama que permite a obtenção da BN como um membro de mistura. A seguir é apresentada a definição do modelo geral.

Definição 3.1.1 *Suponha que $\mathcal{F}_t$ denote uma σ-álgebra gerada por $\{Y_0, \cdots, Y_t\}$. O*

modelo INGARCH log-linear Poisson misto é dado por

$$\begin{cases} Y_t|\mathcal{F}_{t-1} \sim MP(\mu, \phi); \quad \forall\, t \in \mathbb{Z} \\ \nu_t = \alpha + \rho\nu_{t-1} + \eta \log(Y_{t-1} + 1) \end{cases} \tag{3.1}$$

em que α, ρ *e* $\eta \in \mathbb{R}$, $\nu_t \equiv \log(\mu_t)$, *com* $t > 0$.

3.2 Aspectos inferenciais

3.2.1 Algoritmo EM

Para realizar o processo de otimização, inicialmente foi feita uma tentativa através de um método de otimização direto, entretanto um problema de integração numérica relacionado à função de Bessel do terceiro tipo que se apresenta na distribuição PIG, inviabilizou esta abordagem. Este problema no processo de otimização é análogo ao relatado por [22] e que, a partir do mesmo caminho tomado no trabalho citado, uma abordagem alternativa foi explorada.

Para o processo de otimização e obtenção das estimativas dos parâmetros do modelo foi utilizado o algoritmo EM (*espectation maximization*), indicado para conjuntos de dados incompletos ou efeitos não observáveis. O algoritmo EM é um método iterativo composto de dois passos:

1. O passo E que calcula a esperança da função de log-verossimilhança com relação às variáveis latentes.

2. O passo M que atualiza os parâmetros maximizando a esperança condicional a partir da etapa anterior.

A estimação pelo algoritmo EM inicia-se ao se considerar um vetor de parâmetros arbitrários e este passa pelos dois passos do algoritmo de modo a devolver, ao final da iteração, os valores atualizados para estes parâmetros. De início, é estabelecido um conjunto de valores que é denominado "chute inicial", que é uma estimativa arbitrária para dar inicio ao processo de estimação. Ao final do processo são obtidos novos valores para os estimadores e verificada a tolerância para o erro estabelecido. Caso a tolerância estabelecida para o erro seja satisfeita, o processo se encerra, caso contrário, os valores atualizados vetor de parâmetros substitui os valores antigos e o processo se repete até que esta tolerância para o erro seja satisfeita.

Uma vantagem expressiva do uso do algoritmo EM como é observado por [22] está relacionado à obtenção do parâmetro de dispersão que se apresenta com grau satisfatório de eficácia.

Como foi ressaltado no Capítulo 2, ao se abordar uma distribuição que pertence à família exponencial, é assumido então que a função $c(\cdot;\cdot)$ pode ser expressa por meio da equação

$$c(z;\phi) = d(\phi) + \phi g(z) + h(z),$$

em que $d(\cdot)$ representa uma função três vezes diferenciável de modo que se possa garantir que suas três derivadas existam e que estas sejam contínuas.

Trabalhando com dados completos $(Y_1, Z_1), \cdots (Y_n, Z_n)$, em que $Y_1, \cdots, Y_n$ são contagens observadas e $Z_1, \cdots, Z_n$ são efeitos aleatórios não-observáveis.

A função completa de log-verossimilhança é dada por

$$l_c(\theta) = \sum_{t=1}^{n} \log\{P(Y_i = y_i, Z_i = z_i)\},$$

que equivale a

$$l_c(\theta) = \sum_{t=1}^{n} \log\{P(Y_i = y_i | Z_i = z_i) f_z(z_i)\}.$$

Portanto, temos então que a função completa de log-verossimilhança é dada por

$$l_c(\theta) = \sum_{t=1}^{n} \log\left\{ \frac{\mathrm{e}^{-\mu_t z_t}(\mu_t z_t)^{y_t}}{y_t!} \exp\{\phi[z_t\xi_0 - b(\xi_0)] + d(\phi) + \phi g(z_t) + h(z_t)\} \right\},$$

em que $\mu_t = \exp\{\nu_t\}$, com $\nu_t = \alpha + \rho\nu_{t-1} + \eta \log(Y_{t-1} + 1)$.

Atrávés de manipulações algébricas simples aplicando propriedades de exponenciais e logaritmos, podemos concluir que

$$l_c(\theta) = \sum_{t=1}^{n} \{-\mu_t z_t + y_t \log\ \mu_t + y_t \log\ z_t - \log\ y_t! + \phi z_t + \xi_0 - \phi b(\xi_0) + d(\phi) + \phi g(z_t) + h(z_t)\}.$$

Por fim, suprimindo os termos que não dependem de μ_t e ϕ, já que o vetor de parâmetros $\theta = (\alpha, \rho, \eta, \phi)$ a ser maximizado não depende destes termos, é possível encontrar uma equivalência para a função completa de log-verossimilhança dada por

$$l_c(\theta) \propto \sum_{t=1}^{n} \{y_t\nu_t - z_t \mathrm{e}^{\nu_t} + d(\phi) + \phi[z_t\xi_0 - b(\xi_0) + g(z_t)]\}. \tag{3.2}$$

No que se segue, será usado então o algotitmo EM para encontrar os estimadores de máxima verossimilhança $\hat{\theta} = (\hat{\alpha}, \hat{\rho}, \hat{\eta}, \hat{\phi})$ para o vetor de parâmetros $\theta = (\alpha, \rho, \eta, \phi)$.

Passo E (E-step)

No passo E, é utilizada a esperança condicional dentro da função completa de log-verossimilhança para estimar os efeitos aleatórios não-observáveis tornando possível a maximização dos parâmetros no passo seguinte. A esperança condicional para a função completa de log-verossimilhança é dada pela equação

$$Q(\theta;\theta^{(r)}) \equiv E(l_c(\theta)|Y;\theta^{(r)}).$$

Tomando então a função equivalente que é representada por

$$Q(\theta;\theta^{(r)}) \propto \sum_{t=1}^{n}\{y_t\nu_t - \gamma_t e^{\nu_t} + d(\phi) + \phi[\gamma_t\xi_0 - b(\xi_0)] + \psi_t\},$$

em que $\gamma_t = E(Z_t|Y;\theta^{(r)})$ representa a esperança condicional de Z_t dado Y e que $\psi_t = E(g(z_t)|Y;\theta^{(r)})$ expressa a esperança condicional de $g(z_t)$. De início, é possível determinar uma expressão para γ_t. Assim,

$$E(Z|Y=y) = \int_0^\infty zf(z|y)dz.$$

Usando as definições e propriedades de distribuição condicional, distribuição conjunta e distribuição marginal, além aplicação de manipulações algébricas simples, tem-se que

$$E(Z|Y=y) = \int_0^\infty z\frac{f(z,y)}{f_y(y)}dz = \frac{1}{p(y;\mu,\phi)}\int_0^\infty zf_z(z)f(y|z)dz.$$

Como Y tem distribuição Poisson mista (discreta), pode-se então ser adotada a mudança denotação de f para p. Deste modo tem-se que

$$\begin{aligned}
E(Z|Y=y) &= \frac{1}{p(y;\mu,\phi)}\int_0^\infty \frac{e^{-\mu z}(\mu z)^y}{y!}\exp\{\phi[z\xi_0 - b(\xi_0)]c(z;\phi)\}dz \\
&= \frac{\mu^y}{y!p(y;\mu,\phi)}\int_0^\infty e^{-\mu z}z^{y+1}\exp\{\phi[z\xi_0 - b(\xi_0)]c(z;\phi)\}dz \\
&= \frac{\mu\mu^y}{\mu y!p(y;\mu,\phi)}\frac{(y+1)!}{(y+1)!}\int_0^\infty e^{-\mu z}z^{y+1}\exp\{\phi[z\xi_0 - b(\xi_0)]c(z;\phi)\}dz \\
&= \frac{(y+1)}{\mu p(y;\mu,\phi)}\int_0^\infty \frac{e^{-\mu z}(\mu z)^{y+1}}{(y+1)!}\exp\{\phi[z\xi_0 - b(\xi_0)]c(z;\phi)\}dz
\end{aligned}$$

$$= \frac{(y+1)p(y+1;\mu,\phi)}{\mu p(y;\mu,\phi)}.$$

Portanto, a expressão que representa a esperança condicional de Z Y é dada por

$$\gamma_t = E(Z|Y=y) = \frac{(y+1)p(y+1;\mu,\phi)}{\mu p(y;\mu,\phi)}. \tag{3.3}$$

Após obter então uma expressão geral para γ_t, deve-se agora determinar a expressão geral para ψ_t usando a função geradora de momentos condicional. Novamente ao usar as definições e propriedades das distribuições conjuntas e marginais é possível determinar uma equação para γ_t. Assim, tem-se então que

$$E(e^{tg(z)}|Y=y) = \int_0^\infty e^{tg(z)} f(z|y)dz$$

$$= \int_0^\infty e^{tg(z)} \frac{f(z,y)}{f_y(y)}dz = \frac{1}{f_y(y)} \int_0^\infty e^{tg(z)} f(z|y) f_z(z)dz$$

$$= \frac{1}{p(y;\mu,\phi)} \int_0^\infty e^{tg(z)} \frac{e^{-\mu z}(\mu z)^y}{y!} f_z(z)dz$$

$$= \frac{1}{p(y;\mu,\phi)} \int_0^\infty e^{tg(z)} \frac{e^{-\mu z}(\mu z)^y}{y!} \quad \exp\{\phi[z\xi_0 - b(\xi_0)] + c(z;\phi)\}dz$$

$$= \frac{\mu^y}{y!p(y;\mu,\phi)} \int_0^\infty e^{-\mu z} z^y \quad \exp\{\phi[z\xi_0 - b(\xi_0)] + c(z;\phi)\}dz.$$

Portando, tem-se então

$$E(e^{tg(z)}|Y=y) = \frac{\mu^y}{y!p(y;\mu,\phi)} \int_0^\infty e^{-\mu z} z^y \ \exp\{\phi[z\xi_0 - b(\xi_0)] + tg(z) + d(\phi) + \phi g(z) + h(z)\}dz. \tag{3.4}$$

É fácil notar que, se for considerada então a parte da segunda exponencial da equação acima de modo que se possa colocar $g(z)$ em evidência, obtem-se então

$$\exp\{\phi[z\xi_0 - b(\xi_0)] + tg(z) + d(\phi) + \phi g(z) + h(z)\} = \exp\{\phi[z\xi_0 - b(\xi_0)] + (\phi+t)g(z) + d(\phi) + h(z)\}.$$

Multiplicamos então o termo $\phi[z\xi_0 - b(\xi_0)]$ por $\frac{(\phi+t)}{(\phi+t)}$, o que permite reescrever, sem perda de generalidade, a igualdade acima como

$$\exp\{\phi[z\xi_0 - b(\xi_0)] + tg(z) + d(\phi) + \phi g(z) + h(z)\} =$$

$$= \exp\left\{(\phi+t)\left[\frac{z\phi\xi_0}{\phi+t} - \frac{\phi b(\xi_0)}{\phi+t}\right] + (\phi+t)g(z) + d(\phi) + d(\phi+t) - d(\phi+t) + h(z)\right\}.$$

É possível ainda adicionar e subtrair o termo $(\phi)b\left(\frac{\phi\xi_0}{\phi+t}\right)$, de modo que novamente, com manipulações matemáticas simples, obtem-se a equação

$$\exp\{\phi[z\xi_0 - b(\xi_0)] + tg(z) + d(\phi) + \phi g(z) + h(z)\} =$$

$$= \exp\left\{(\phi+t)\left[\frac{z\phi\xi_0}{\phi+t} - b\left(\frac{\phi\xi_0}{\phi+t}\right)\right] + (\phi+t)b\left(\frac{\phi\xi_0}{\phi+t}\right) - \phi b(\xi_0)\right.$$

$$\left.+(\phi+t)g(z) + d(\phi) + d(\phi+t) - d(\phi+t) + h(z)\right\}$$

$$= \exp\left\{(\phi+t)b\left(\frac{\phi\xi_0}{\phi+t}\right) - \phi b(\xi_0) + d(\phi) - d(\phi+t)\right\}$$

$$\times \exp\left\{(\phi+t)\left[\frac{z\phi\xi_0}{\phi+t} - b\left(\frac{\phi\xi_0}{\phi+t}\right)\right] + d(\phi+t) + (\phi+t)g(z) + h(z)\right\}.$$

Pelas manipulações algébricas no termo acima, retornando então à equação (3.4), decorre então que

$$E(\mathrm{e}^{tg(z)}|Y=y) = \frac{1}{p(y;\mu;\phi)} \exp\left\{(\phi+t)b\left(\frac{\phi\xi_0}{\phi+t}\right) - \phi b(\xi_0) + d(\phi) - d(\phi+t)\right\}$$

$$\times \int_0^\infty \frac{e^{-\mu z}(\mu z)^y}{y!} \exp\left\{(\phi+t)\left[\frac{z\phi\xi_0}{\phi+t} - b\left(\frac{z\phi\xi_0}{\phi+t}\right)\right] + d(\phi+t) + (\phi+t)g(z) + h(z)\right\} dz.$$

Portanto, tomando $\mu^* = b'\left(\frac{\phi\xi_0}{\phi+t}\right)$, conclui-se então que

$$E(\mathrm{e}^{tg(z)}|Y=y) = \frac{p(y;\mu^*,\phi+t)}{p(y;\mu,\phi)} \exp\left\{(\phi+t)b\left(\frac{\phi\xi_0}{\phi+t}\right) - \phi b(\xi_0) - d(\phi) - d(\phi+t)\right\}. \tag{3.5}$$

Finalmente, derivando a equação (3.5) e aplicando em $t = 0$, obtem-se o primeiro momento, o que equivale a $E(g(Z)|Y) = \psi_t$. Assim, tem-se que

$$E(g(Z)|Y) = \frac{d}{dt}E(\mathrm{e}^{tg(z)}|Y)\Big|_{t=0}.$$

A derivada da função é dada por

$$\frac{d}{dt}E(\mathrm{e}^{tg(z)}|Y) = \frac{1}{p(y;\mu,\phi)}\left\{\frac{d}{dt}p(y;\mu^*,\phi+t) \exp\left[(\phi+t)b\left(\frac{\phi\xi_0}{\phi+t}\right) - \phi b(\xi_0) + d(\phi)\right.\right.$$

$$-d(\phi+t)+p(y;\mu^*,\phi+t)\left[-\frac{\phi^2\xi_0}{(\phi+t)^2}b'\left(\frac{\phi\xi_0}{\phi+t}\right)+b\left(\frac{\phi\xi_0}{\phi+t}\right)-\frac{t\phi\xi_0}{(\phi+t)^2}b'\left(\frac{\phi\xi_0}{\phi+t}\right)-d'(\phi+t)\right]$$

$$\times\exp\left[(\phi+t)b\left(\frac{\phi\xi_0}{\phi+t}\right)-\phi d(\phi)+d(\phi)-d(\phi+t)\right]\Bigg\}$$

$$=\frac{1}{p(y;\mu,\phi)}\left\{\frac{d}{dt}p(y;\mu^*,\phi+t)\bigg|_{t=0}+p(y;\mu^*,\phi)\left[b(\xi_0)-\xi_0b'(\xi_0)-d'(\phi)\right]\right\}.$$

Como $b'(\xi_0)=1$, então tem-se que

$$\psi_t=E(g(Z)|Y=y)=\frac{1}{p(y;\mu,\phi)}\frac{d}{dt}p(y;\mu^*,\phi)\bigg|_{t=0}+b(\xi_0)-d'(\phi). \tag{3.6}$$

Os resultados obtidos nas equações (3.3) e (3.6) são as expressões gerais para γ_t e ψ_t. Deste modo, é possível seguir então para o próximo passo do algoritmo EM, que consiste em maximizar a função completa de verossimilhança.

Passo M (M-step)

Neste passo serão tomadas as derivadas parciais da função completa de log-verossimilhança em relação aos parâmetros que compõem o vetor θ para obter os estimadores de máxima verossimilhança.

A função completa de verossimilhança é dada por

$$Q(\theta,\theta^{[r]})=\sum_{t=0}^{n}(Y_t\nu_t-\gamma_t\mathrm{e}^{\nu_t}+d(\phi)+\phi\{\gamma_t\xi_0-b(\xi_0)\}+\psi_t).$$

Para estimar os parâmetros α, ρ e η serão usadas as derivações da expressão equilalênte à função completa de log-verossimilhança em relação a ν_t de modo desconsideramos os termos que não estão relacionados a ν_t, a função equivalente é dada por

$$Q(\theta,\theta^{[r]})\propto\sum_{t=0}^{n}(Y_t\nu_t-\gamma_t\mathrm{e}^{\nu_t}).$$

Vamos considerar θ^* como sendo vetor de parâmetros composto pelas três primeiras componentes do vetor θ e $\nu_t=\alpha+\rho\nu_{t-1}+\eta\ log(Y_{t-1}+1)$. A função *score* é dada por

$$S(\theta^*)=\frac{\partial Q(\theta^*)}{\partial\theta^*}=\sum_{t=0}^{n}(Y_t-\gamma_t\mathrm{e}^{\nu_t})\frac{\partial\nu_t(\theta^*)}{\partial\theta^*}$$

em que $\dfrac{\partial\nu_t(\theta^*)}{\partial\theta^*}$ é um vetor tridimencional cujas componentes são dadas por

$$\frac{\partial\nu_t}{\partial\alpha}=1+\rho\cdot\frac{\partial\nu_{t-1}}{\partial\alpha};\quad\frac{\partial\nu_t}{\partial\rho}=\nu_{t-1}+\rho\cdot\frac{\partial\nu_{t-1}}{\partial\rho};\quad\frac{\partial\nu_t}{\partial\eta}=\log(Y_{t-1}+1)+\rho\cdot\frac{\partial\nu_{t-1}}{\partial\eta}$$

Agora serão feitas as manipulações algébricas para calcular ϕ considerando a função completa de log-verossimilhança sendo dada por

$$\frac{\partial Q}{\partial \phi} = \sum_{t=1}^{n} \{d'(\phi) + \gamma_t \xi_0 - b(\xi_0) + \psi_t\}$$

Pode-se então obter $\hat{\phi}$ através do seguinte cálculo:

$$\frac{\partial Q}{\partial \phi} = 0 \therefore \sum_{t=1}^{n} \{d'(\phi) + \gamma_t \xi_0 - b(\xi_0) + \psi_t\} = 0$$

Por isso,

$$nd'(\hat{\phi}) - nb(\xi_0) + \sum_{t=1}^{n} (\xi_0 \gamma_t + \psi_t) = 0 \therefore nd'(\hat{\phi}) = nb(\xi_0) - \sum_{t=1}^{n} (\xi_0 \gamma_t + \psi_t)$$

$$d'(\hat{\phi}) = b(\xi_0) - \frac{\sum_{t=1}^{n} (\xi_0 \gamma_t + \psi_t)}{n} \tag{3.7}$$

Escolhendo então um Z tal que esta segue distribuição inversa Gaussiana com média 1 e parâmetro de dispersão ϕ, tem-se então que $E(Z) = b'(\xi_0) = 1$, logo temos que $b'(\xi_0) = \frac{1}{\sqrt{-2\xi_0}}$ e então pode-se concluir que $\xi_0 = -\frac{1}{2}$. Com as retrições feitas acima, a decomposição da Inversa Gaussiana apresenta as seguintes parcelas:

i) $b(\xi_0) = b\left(-\frac{1}{2}\right) = -\sqrt{-2\left(-\frac{1}{2}\right)} = -1$;

ii) $d(\phi) = \frac{1}{2}\ \log\phi \Rightarrow d'(\phi) = \frac{1}{2\phi}$;

iii) $g(z) = -\frac{1}{2z}$;

iv) $h(z) = -\frac{1}{2}\ \log(2\pi z^3)$.

Agora é possível então determinar o estimador $\hat{\phi}$ substituíndo os resultados obtidos acima para a PIG na equação (3.7), de modo que

$$\begin{aligned} \frac{n}{2\hat{\phi}} &= -n - \sum_{t=1}^{n} \left(\psi_t - \frac{\gamma_t}{2}\right) \\ &= \sum_{t=1}^{n} \left(\frac{\gamma_t}{2} - \psi_t - 1\right). \end{aligned}$$

Tem-se então

$$\frac{1}{\hat{\phi}} = \frac{2\sum_{t=1}^{n} \left(\frac{\gamma_t}{2} - \psi_t - 1\right)}{n} \Rightarrow \hat{\phi} = \frac{n}{2\sum_{t=1}^{n} \left(\frac{\gamma_t}{2} - \psi_t - 1\right)},$$

em que $\hat{\phi}$ não depende de α, ρ e η.

Modo análogo ao que foi feito para a PIG, é repetido o processo agora para determinar as funções auxiliares para a BN ao escolher Z tal que este segue distribuição Gama com média 1 e parâmetro de dispersão ϕ. A esperança de Z é dada por $E(Z) = b'(\xi_0) = 1$, logo tem-se que $b'(\xi_0) = -\frac{1}{\xi_0}$ e então é possível concluir que $\xi_0 = -1$. As parcelas com as funções auxiliares são

i) $b(\xi_0) = -\log(-\xi_0) = -\log(1) = 0$;

ii) $d(\phi) = \phi\ \log\ \phi - \log(\Gamma(\phi)) \therefore d'(\phi) = log\ \phi + 1 - \frac{\Gamma'(\phi)}{\Gamma(\phi)}$;

iii) $g(z) = \log\ z$;

iv) $h(z) = -\log\ z$.

É possível notar pelas equações auxiliares através da expressão de $d'(\phi)$, que o parâmetro ϕ não apresenta uma forma fechada para a binomial negativa. Os resultados apresentados na Tabela 3.1 sintetizam as funções auxiliares usadas para a construção das simulações para os modelos PIG e BN e para a obteção do vetor de estimadores para o vetor parâmetros $\theta = (\alpha, \rho, \eta, \phi)$.

Tabela 3.1: Valores e funções auxiliares para as distribuições PIG e BN tomando $\xi_0 = -\frac{1}{2}$ e $\xi_0 = -1$ respectivamente

Distribuição	$b(\xi_0)$	$d(\phi)$	$d'(\phi)$	$g(z)$	$h(z)$
PIG	$-\sqrt{-2\xi_0}$	$\frac{1}{2}\log\ \phi$	$\frac{1}{2\phi}$	$-\frac{1}{2z}$	$-\frac{1}{2}\log(2\pi z^3)$
BN	$-\log(-\xi_0)$	$\phi\log\ \phi - \log(\Gamma(\phi))$	$\log\ (\phi) + 1 - \frac{\Gamma'(\phi)}{\Gamma(\phi)}$	$\log\ z$	$-\log\ z$

Através do conjunto de equações apresentados acima, é possível então efetuar a construção do algoritmo EM, que possibilita a obteção dos estimadores $\hat{\theta} = (\hat{\alpha}, \hat{\rho}, \hat{\eta}, \hat{\phi})$. Observa-se ainda que as esperanças condicionais γ_t e ψ_t são inseridas no algoritmo de simulação de acordo com o a variável latente Z.

3.3 Estudos de simulação

3.3.1 Cenários de simulação para os parâmetros

Com o procedimento de simulação feito pela estimação por verossimilhança através do algoritmo EM, foi realizada uma análise do comportamento dos valores estimados considerando as distribuições log-linear PIG e log-linear BN que são distribuições obtidas por meio de uma distribuição Poisson mista e que foram submetidas ao processo de simulação pelo método de Monte Carlo.

Por meio do algoritmo EM construído utilizando a linguagem de programação R, foram avaliados quatro cenários inicialmente estabelecidos aos quais foram denominados [A], [B], [C] e [D] para cada uma das duas distribuições consideradas. Cada um destes cenários foi composto por 4 tamanhos amostrais distintos, a saber 50, 100, 200 e 500 para os quais foram geradas 1000 réplicas de Monte Carlo para cada um dos casos propostos.

Os valores teóricos estabelecidos para os parâmetros considerando os casos dos modelos INGARCH log-linear PIG e BN, considerando os cenários [A], [B], [C] e [D] são apresentados, respectivamente, nas Tabelas 3.2 e 3.3.

Tabela 3.2: Valores teóricos dos parâmetros para os cenários de simulação da PIG pelo algoritmo EM.

Cenários para o modelo INGARCH Log-linear PIG				
Cenários	α	ρ	η	ϕ
[A]	2	0,5	$-0,3$	1
[B]	3	$-0,2$	0,4	1
[C]	1,5	0,4	$-0,2$	2
[D]	3,5	$-0,35$	0,45	2

Tabela 3.3: Valores teóricos dos parâmetros para os cenários de simulação da BN pelo algoritmo EM.

Cenários para o modelo INGARCH Log-linear BN				
Cenários	α	ρ	η	ϕ
[A]	2	0,5	$-0,3$	1
[B]	1,5	$-0,1$	0,2	1
[C]	2,5	0,3	$-0,4$	2
[D]	1,5	$-0,25$	0,5	2

A princípio seriam utilizados os mesmos cenários para as duas distribuições propostas, entretanto os valores propostos para o caso PIG apresentaram problemas para a distribuição BN. Tendo em vista que o caso da BN é mais complexo já que se tratar de uma distribuição cujas funções auxiliares apresentam forma mais complexa tendo em sua composição a função Gama e a sua derivada Digama, os cenários propostos tiveram que ser por vezes reelaborados por causa de erros durante o processo de simulação. Para os cenários propostos inicialmente para a PIG, apenas o cenário [A] foi possível de ser simulado para a BN.

Devido ao problema observado acima para um dos casos aqui abordados, vale ressaltar que para algumas distribuições Poisson mistas, a depender da escolha do membro da família exponencial, o processo de otimização pode se tornar muito

complicado ou até mesmo impossível de ser realizado através do algoritmo EM. Deste modo, é importante avaliar a distribuição Poisson mista a ser abordada antes de realizar a implementação por meio de algoritmo EM ou mesmo um método de otimização direto. Este cuidado pode evitar problemas com distribuições Poisson mistas que tenham expressões complicadas em sua composição.

Outra observação a ser feita acerca da montagem dos cenários é que existe uma alternância entre os sinais dos parâmetros ρ e η. Os valores teóricos para o parâmetro α também sofreram modificações para cada cenário, porém conservando-o como valor positivo. Já para o parâmetro de dispersão ϕ, foram considerados apenas os valores 1 para os cenários [A] e [B] e 2 para os cenários [C] e [D] para ambas as distribuições analisadas.

3.3.2 Simulações para o modelo INGARCH log-linear PIG

O primeiro grupo de simulações foi realizado para o modelo INGARCH log-linear PIG, contemplando os cenários sugeridos na Seção 3.3.1. A Figura 3.1 apresenta os gráficos da série temporal de contagem, da autocorrelação e da autocorrelação parcial dos 100 primeiros valores da geração aleatória obtida através da simulação do cenário [A]. A representação gráfica dos dados aleatórios da série temporal sugerem um comportamento estacionário enquanto os gráficos de autocorrelação e autocorrelação parcial sugerem correlação fraca entre as observações.

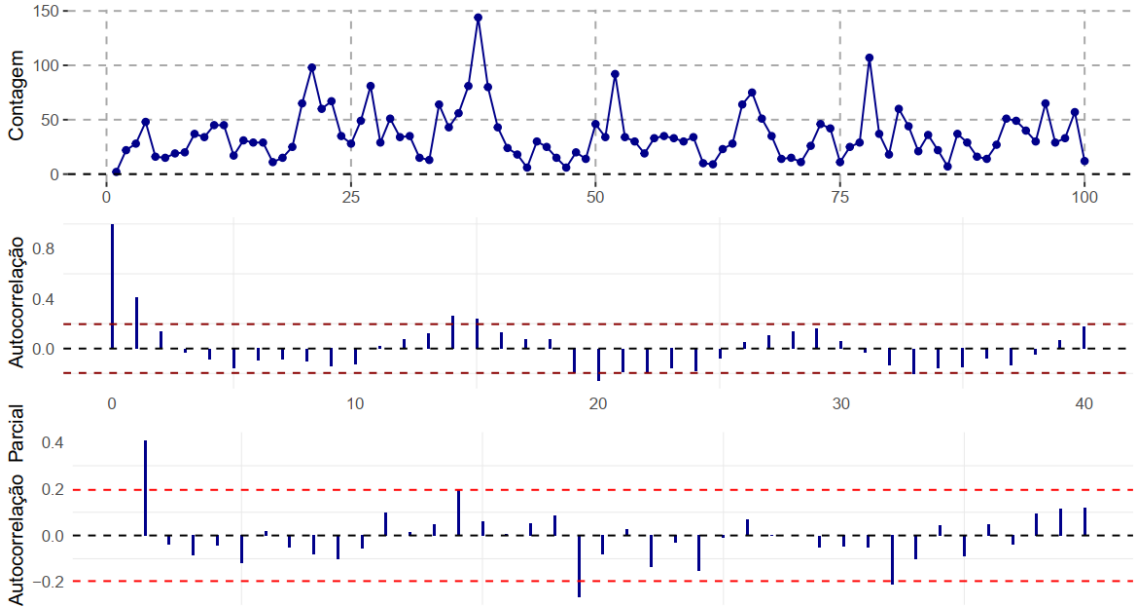

Figura 3.1: Série temporal de contagem, gráfico de autocorrelação e autocorrelação parcial para a PIG, cenário A, tamanho amostral 100.

Os valores obtidos nas simulações para o processo INGARCH log-linear PIG para a média estimada de cada parâmetro, bem como a raiz quadrada do erro

quadrático médio, destacada entre parênteses ao lado da estimativa correspondente, são apresentados na Tabela 3.4.

Tabela 3.4: Resultados das simulações para os cenários [A], [B], [C] e [D] para a distribuição PIG.

Cenário [A]: ($\alpha = 2$, $\rho = 0,5$, $\eta = -0,3$ e $\phi = 1$)				
Tamanho amostral	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	$\hat{\phi}$
50	1,988 (0,911)	0,636 (0,506)	$-0,448$ (0,293)	1,429 (0,764)
100	1,972 (0,772)	0,547 (0,327)	$-0,343$ (0,130)	1,123 (0,312)
200	1,985 (0,597)	0,516 (0,220)	$-0,314$ (0,078)	1,041 (0,171)
500	2,007 (0,367)	0,504 (0,131)	$-0,308$ (0,048)	1,014 (0,097)
Cenário [B]: ($\alpha = 3$, $\rho = -0,2$, $\eta = 0,4$ e $\phi = 1$)				
Tamanho amostral	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	$\hat{\phi}$
50	3,026 (0,940)	$-0,193$ (0,343)	0,373 (0,200)	1,173 (0,407)
100	3,062 (0,791)	$-0,216$ (0,242)	0,391 (0,114)	1,073 (0,237)
200	3,029 (0,624)	$-0,208$ (0,176)	0,398 (0,075)	1,027 (0,154)
500	3,024 (0,413)	$-0,206$ (0,110)	0,400 (0,047)	1,005 (0,092)
Cenário [C]: ($\alpha = 1,5$, $\rho = 0,4$, $\eta = -0,2$ e $\phi = 2$)				
Tamanho amostral	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	$\hat{\phi}$
50	1,473 (0,837)	0,547 (0,632)	$-0,342$ (0,311)	3,276 (2,811)
100	1,482 (0,789)	0,444 (0,471)	$-0,240$ (0,145)	2,331 (0,770)
200	1,484 (0,687)	0,414 (0,366)	$-0,208$ (0,085)	2,113 (0,412)
500	1,510 (0,469)	0,399 (0,242)	$-0,205$ (0,052)	2,040 (0,222)
Cenário [D]: ($\alpha = 3,5$, $\rho = -0,35$, $\eta = 0,45$ e $\phi = 2$)				
Tamanho amostral	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	$\hat{\phi}$
50	3,490 (0,744)	$-0,344$ (0,264)	0,440 (0,178)	2,302 (0,711)
100	3,516 (0,646)	$-0,354$ (0,188)	0,447 (0,107)	2,136 (0,425)
200	3,510 (0,522)	$-0,354$ (0,138)	0,450 (0,071)	2,056 (0,275)
500	3,513 (0,382)	$-0,354$ (0,092)	0,450 (0,047)	2,014 (0,171)

Fonte: Dados de simulações em linguagem R feitas pelo autor.

Através da análise dos valores obtidos nas simulações é possível perceber que os valores estimados para os diferentes tamanhos de amostras propostos tendem assintoticamente para os valores teóricos estabelecidos para os parâmetros. Este fato pode ser verificado também ao observar a raiz quadrada do erro quadrático médio que tende a diminuir para tamanhos amostrais maiores.

Uma complementação à análise das observações pode ser vista na Tabela 3.5 que apresenta os valores para o viés para cada estimador obtido na simulação e que, em alguns casos específicos, como pode ser observado para o parâmetro η no cenário

[B], chega a zero para o número de casas decimais consideradas por arredondamento para o tamanho amostral 500. O mesmo fato se repete para mesmo parâmetro no cenário [D] para os tamanhos amostrais 200 e 500.

Tabela 3.5: Viés dos estimadores para os cenários [A], [B], [C] e [D] para a PIG.

Cenário [A]: ($\alpha = 2$, $\rho = 0,5$, $\eta = -0,3$ e $\phi = 1$)				
Tamanho amostral	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	$\hat{\phi}$
50	0,011	$-0,136$	0,148	$-0,429$
100	0,028	$-0,047$	0,043	$-0,123$
200	0,015	$-0,016$	0,014	$-0,041$
500	$-0,007$	$-0,004$	0,008	$-0,014$
Cenário [B]: ($\alpha = 3$, $\rho = -0,2$, $\eta = 0,4$ e $\phi = 1$)				
Tamanho amostral	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	$\hat{\phi}$
50	$-0,026$	$-0,007$	0,027	$-0,173$
100	$-0,062$	$-0,016$	0,008	$-0,073$
200	$-0,029$	$-0,008$	0,002	$-0,027$
500	$-0,024$	$-0,006$	0,000	$-0,005$
Cenário [C]: ($\alpha = 1,5$, $\rho = 0,4$, $\eta = -0,2$ e $\phi = 2$)				
Tamanho amostral	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	$\hat{\phi}$
50	0,027	$-0,147$	0,142	$-1,276$
100	0,018	$-0,044$	0,040	$-0,332$
200	0,015	$-0,014$	0,008	$-0,113$
500	$-0,009$	0,001	0,005	$-0,040$
Cenário [D]: ($\alpha = 3,5$, $\rho = -0,35$, $\eta = 0,45$ e $\phi = 2$)				
Tamanho amostral	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	$\hat{\phi}$
50	0,010	$-0,006$	0,010	$-0,302$
100	$-0,165$	0,004	0,003	$-0,136$
200	$-0,010$	0,004	0,000	$-0,056$
500	$-0,013$	0,003	0,000	$-0,015$

Fonte: Dados de simulações em linguagem R feitas pelo autor.

Por meio da análise dos gráficos boxplot da Figura 3.2 é possível perceber que em todos os cenários observam-se muitos valores de *outliers*, porém para todos os casos fica evidente que o aumento do tamanho amostral diminui o intervalo entre o primeiro e o terceiro quartis indicando que há uma concentração cada vez maior em torno do valor teórico.

Nos boxplots também é possível notar que a mediana dos dados tende ao valor teórico dos parâmetros para todos os casos observados, enfatizando que em alguns deles, estas observações acontecem desde o menor tamanho amostral, como é

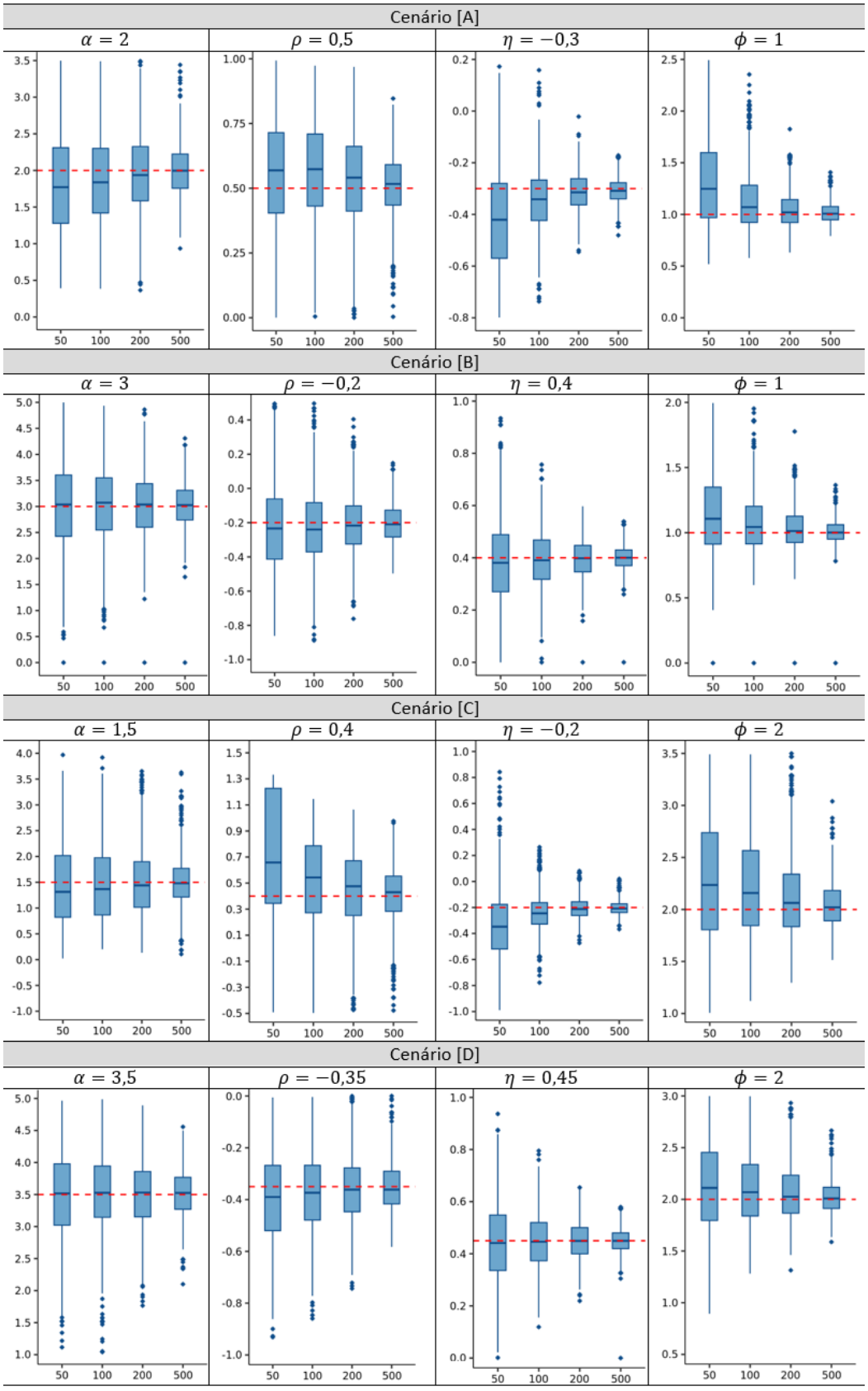

Figura 3.2: Boxplots das estimativas dos parâmetros α, ρ, η e ϕ para os cenários [A], [B], [C] e [D] - Modelo INGARCH log-linear PIG.

o caso por exemplo dos parâmetros α e η no cenário [D]. Deste modo, com a análise dos boxplots, a conclusão é a mesma em que se chega ao analisar a Tabelas 3.4 e a Tabela 3.5, ou seja, a média dos valores estimados tendem aos valores teóricos à medida em que o tamanho amostral sofre aumento.

3.3.3 Simulações para o modelo log-linear BN

Como já comentado anteriormente, para o modelo log-linear BN fez-se necessária uma reestruturação dos cenários propostos. Esta distribuição carrega em sua composição a função gama e, consequentemente retorna um fatorial, ocasionando muitos erros no processo de otimização. O uso da função `TryCatch` proporcionou o desenvolvimento do processo de obtenção dos estimadores, eliminando a possibilidade de perda do processo de simulação de cada cenário proposto. A função `TryCatch`, da forma como inserida no código, elimina a interação com erro retornando à interação anterior.

O gráfico da séries temporal, de autocorrelação e autocorrelação parcial apresentados na Figura 3.3.

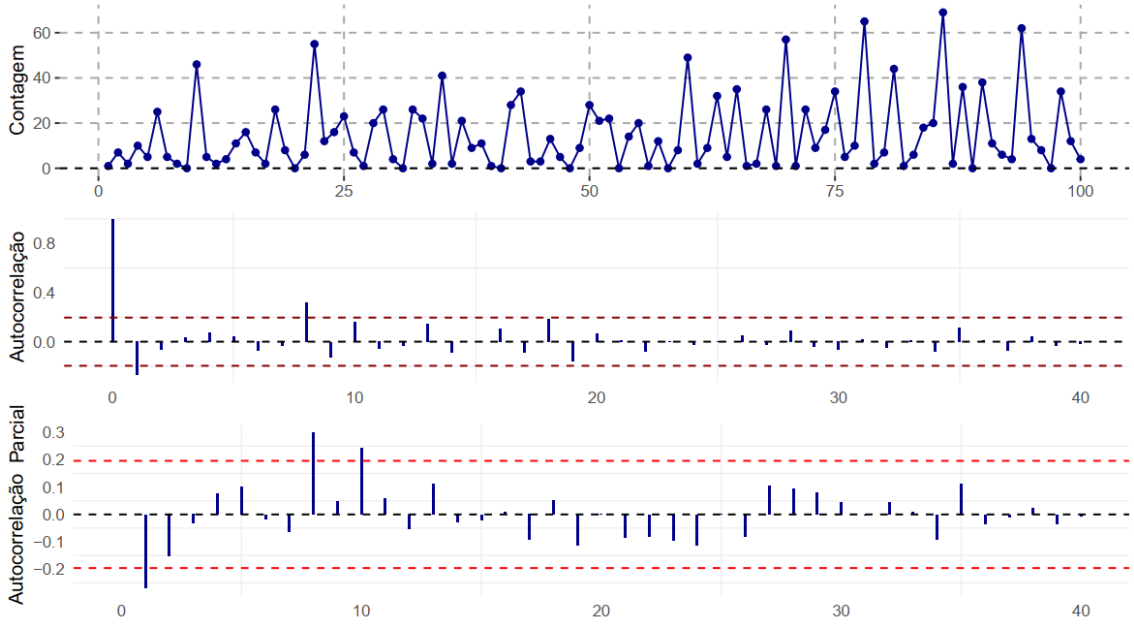

Figura 3.3: Série temporal de contagem, gráfico de autocorrelação e autocorrelação parcial para a BN, cenário A, tamanho amostral 100.

A exemplo do que foi feito para o caso da análise do caso anterior, também consideram a amostra gerada aleatoriamente, cujo tamanho amostral é $n = 100$ e, que considera os valores teóricos para os parâmestros que foram estabelecidos para o cenário [A]. É possível perceber que para o modelo INGARCH log-linear BN, os dados gerados também apresentam um comportamento que sugere a estacionaridade, além de que também se observa uma baixa correlação entre eles, repetindo aquilo

que foi observado para a distribuição anterior.

A Tabela 3.6 apresenta os valores obtidos no processo de otimização do modelo INGARCH log-linear BN.

Tabela 3.6: Resultados das simulações para os cenários [A], [B], [C] e [D] para a distribuição BN.

Cenário [A]: ($\alpha = 2$, $\rho = 0,5$, $\eta = -0,3$ e $\phi = 1$)				
Tamanho amostral	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	$\hat{\phi}$
50	1,923 (0,924)	0,606 (0,457)	$-0,394$ (0,234)	1,214 (0,413)
100	1,974 (0,768)	0,523 (0,305)	$-0,320$ (0,107)	1,069 (0,189)
200	1,981 (0,562)	0,506 (0,202)	$-0,300$ (0,071)	1,028 (0,112)
500	1,974 (0,347)	0,507 (0,119)	$-0,298$ (0,043)	1,014 (0,068)
Cenário [B]: ($\alpha = 1,5$, $\rho = -0,1$, $\eta = 0,2$ e $\phi = 1$)				
Tamanho amostral	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	$\hat{\phi}$
50	1,374 (1,849)	$-0,031$ (0,720)	0,172 (0,354)	1,277 (0,633)
100	1,421 (1,770)	$-0,056$ (0,565)	0,185 (0,261)	1,094 (0,247)
200	1,436 (1,698)	$-0,061$ (0,449)	0,192 (0,227)	1,031 (0,139)
500	1,492 (1,548)	$-0,094$ (0,320)	0,196 (0,212)	1,012 (0,083)
Cenário [C]: ($\alpha = 2,5$, $\rho = 0,3$, $\eta = -0,4$ e $\phi = 2$)				
Tamanho amostral	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	$\hat{\phi}$
50	2,262 (0,854)	0,458 (0,470)	$-0,463$ (0,213)	2,506 (1,051)
100	2,358 (0,671)	0,365 (0,299)	$-0,408$ (0,103)	2,156 (0,430)
200	2,451 (0,467)	0,320 (0,192)	$-0,401$ (0,070)	2,085 (0,269)
500	2,493 (0,467)	0,303 (0,109)	$-0,400$ (0,070)	2,031 (0,269)
Cenário [D]: ($\alpha = 1,5$, $\rho = -0,25$, $\eta = 0,5$ e $\phi = 2$)				
Tamanho amostral	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	$\hat{\phi}$
50	1,539 (0,625)	$-0,289$ (0,387)	0,496 (0,213)	2,423 (1,395)
100	1,543 (0,499)	0,268 (0,266)	0,487 (0,129)	2,129 (0,443)
200	1,534 (0,329)	$-0,267$ (0,162)	0,495 (0,078)	2,068 (0,286)
500	1,507 (0,201)	$-0,252$ (0,099)	0,497 (0,050)	2,025 (0,175)

Fonte: Dados de simulações em linguagem R feitas pelo autor.

Para o cenário [A], cujos valores foram os mesmos para ambas as distribuições analisadas, os resultados de simulação mostraram que a convergência dos parâmetros da BN foi melhor que da PIG para todos os tamanhos amostrais. Através dos resultados apresentados na Tabela 3.6 é possível observar que a raiz do erro quadrático médio apresentou valores menores em todos os casos observados para este cenário.

É importante ressaltar que a distribuição BN apresenta problemas para con-

Tabela 3.7: Viés dos estimadores para os cenários [A], [B], [C] e [D] para a BN.

Cenário [A]: ($\alpha = 2$, $\rho = 0,5$, $\eta = -0,3$ e $\phi = 1$)				
Tamanho amostral	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	$\hat{\phi}$
50	$-0,077$	0,106	$-0,094$	0,214
100	$-0,025$	0,023	$-0,020$	0,069
200	$-0,018$	0,006	0,000	0,028
500	$-0,026$	0,007	$-0,002$	0,014
Cenário [B]: ($\alpha = 3$, $\rho = -0,2$, $\eta = 0,4$ e $\phi = 1$)				
Tamanho amostral	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	$\hat{\phi}$
50	$-1,626$	0,169	$-0,228$	0,277
100	$-1,579$	0,144	$-0,214$	0,094
200	$-1,564$	0,139	$-0,208$	0,031
500	$-1,508$	0,106	$-0,204$	0,0118
Cenário [C]: ($\alpha = 2,5$, $\rho = 0,3$, $\eta = -0,4$ e $\phi = 2$)				
Tamanho amostral	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	$\hat{\phi}$
50	$-0,238$	0,158	$-0,063$	0,506
100	$-0,142$	0,065	$-0,008$	0,156
200	$-0,048$	0,020	$-0,001$	0,085
500	$-0,007$	0,003	0,000	0,031
Cenário [D]: ($\alpha = 1,5$, $\rho = -0,25$, $\eta = 0,5$ e $\phi = 2$)				
Tamanho amostral	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	$\hat{\phi}$
50	0,039	$-0,039$	$-0,004$	0,423
100	0,043	$-0,018$	$-0,013$	0,129
200	0,034	$-0,017$	$-0,005$	0,068
500	0,007	$-0,002$	$-0,003$	0,025

Fonte: Dados de simulações em linguagem R feitas pelo autor.

figuração de cenários de modo a não aceitar muitas configurações de valores teóricos para os parâmetros, mesmo sendo obedecida a condição estabelecida por [9] em que $|\rho - \eta| < 1$. Neste aspecto, a PIG se mostra mais eficiente já que ante ao processo de simulações não apresentou problemas a nenhum cenário proposto.

A Tabela 3.7 apresenta os valores dos viés obtidos para os cenários propostos para a BN. É possível observar que os valores estimados para α, ρ e η no cenário [A] apresentam um aumento, em módulo, para a variação de tamanho amostral de 200 para 500. Os demais cenários tiveram comportamento de valores esperados para a simulação.

A Figura 3.4 apresenta os gráficos boxplot referentes às simulações para os cenários da BN, sendo possível fazer uma observação análoga ao que já foi constatado

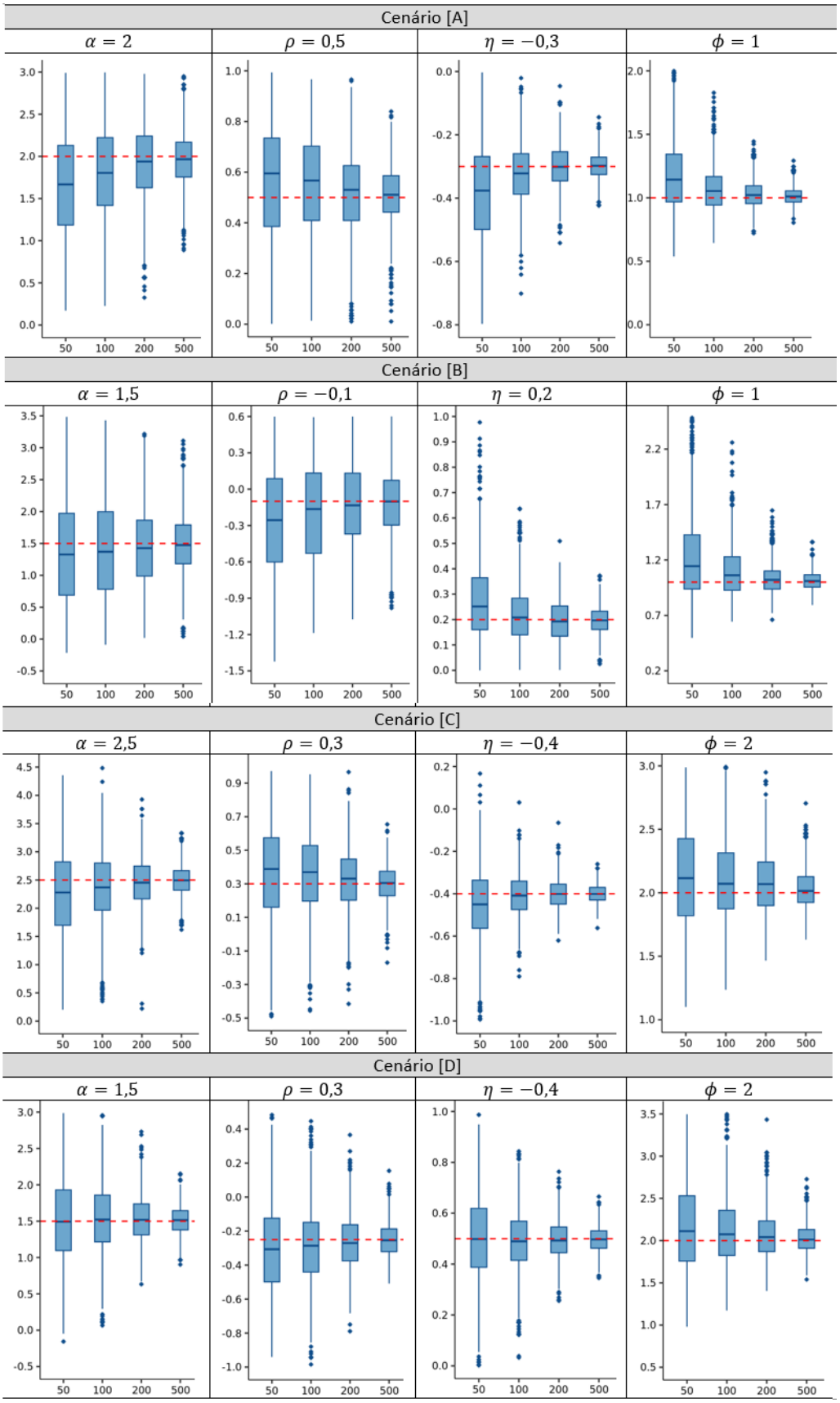

Figura 3.4: Boxplots das estimativas dos parâmetros α, ρ, η e ϕ para os cenários [A], [B], [C] e [D] - Modelo INGARCH log-linear BN.

ao analisar a Tabela 3.6, ou seja, à medida em que aumenta o tamanho amostral, os dados se concentram cada vez mais em torno da média e esta convergem para os valores teóricos dos parâmetros.

3.4 Custo computacional de simulação

Para obter os custos computacionais foi usada a função `sytem.time()` que é nativa da linguagem R e que retorna três valores que são, respectivamente, **usuário** que apresenta o tempo gasto pela CPU para a execução do código R, **sistema** que se refere ao tempo que o sistema operacional leva para executar o código R e **decorrido** que apresenta o tempo total do processo.

A Tabela 3.8 apresenta o tempo total de cada simulação para os modelos INGARCH log-linear PIG e BN já convertidos ao formato de hora, minuto e segundo. É possível notar através dos valores tabulados, que o processo de simulação para a distribuição BN demanda menos tempo que processos correspondentes para a PIG.

Tabela 3.8: Tempo total de simulação dos cenários [A], [B], [C] e [D] para as distribuições PIG e BN

	log-linear PIG				log-linear BN			
Cenários	[A]	[B]	[C]	[D]	[A]	[B]	[C]	[D]
50	2:45:00	4:53:09	1:53:56	3:39:27	0:57:38	0:52:42	0:50:44	0:23:11
100	3:00:31	8:50:14	1:36:41	6:41:06	2:09:25	1:17:51	1:06:43	0:40:06
200	5:05:33	17:21:37	1:54:15	12:04:39	3:10:20	2:32:54	2:00:29	1:31:38
500	11:56:43	32:06:25	3:54:56	30:33:39	7:41:05	5:57:18	4:44:16	2:50:38

Fonte: Dados de simulações em linguagem R feitas pelo autor.

De modo empírico foi possível observar que o desempenho da máquina durante o processo de simulação era afetado diretamente pela temperatura ambiente. Simulações executadas no período da noite apresentavam menor tempo de total de execução enquanto as que eram executadas durante o dia, quando a temperatura era mais elevada, demandava tempo maior de execução para o processo.

Capítulo 4

Aplicação do modelo a dados reais

Neste capítulo são apresentados os resultados da aplicação a dados reais visando avaliar o desempenho do modelo, além de comparar aos resultados deste modelo aos dados aplicados ao modelo INGARCH log-linear Poisson. A comparação entre os modelos análise gráfica e através da verificação dos critérios de informação AIC e BIC. Ao final de cada aplicação, ao escolher o modelo que melhor se ajusta, é feita a análise de resíduos afim de comprovar se o modelo de fato apresenta um bom ajuste.

4.1 Internações por abuso de álcool

O consumo de bebidas alcoólicas é um hábito que cresce no Brasil e, segundo o relatório do CISA (Centro de Informações sobre Saúde e Álcool) [5], o maior número de internações e óbitos é entre a população na faixa etária de mais de 55 anos, sendo que os homens os que mais sofrem pela dependência.

O relatório aponta que 76% dos óbitos são de homens. Além das internações que são atribuídas diretamente ao alcoolismo ou alguma complicação causada por ele, muitas outras internações são consequência de uso abusivo como é o caso de vítimas de acidente de trânsito que é causado por motoristas e condutores de motocicletas em estado de embriaguez. O relatório da OMS [26] aponta que em 2018 aproximadamente 2,348 bilhões de pessoas (43% da população) eram consumidores de álcool no mundo. Os dados refletem a importância de estudos e ações para o combate ao alcoolismo.

A CID (Classificação Internacional de Doenças) possui 7 classes de agravo a saúde atribuídos à ingestão de álcool. Deste modo, o alcoolismo é considerado pela OMS como um problema de saúde pública e que causa impacto nos principais índices relacionados a saúde da população nos mais diferentes eixos. Dada a importância de estudos relacionados ao alcoolismo e seus malefícios é que para a aplicação a dados reais para o modelo INGARCH log-linear PIG à série temporal de contagem com o

número de internação por problemas relacionados ao alcoolismo, de acordo com a classificação da CID-10.

Para avaliar o desempenho do modelo foram utilizados dois bancos de dados que apresentam a série temporal de contagem do número de casos de internação por abuso de álcool nos estados da Paraíba e do Piauí e apresentam os dados coletados entre janeiro de 2005 a dezembro de 2015. Os dados utilizados nas simulações encontram-se disponíveis no site do Ministério da Saúde - Sistema de Informação de Atenção Básica - SIAB no endereço eletrônico `http://tabnet.datasus.gov.br/cgi/deftohtm.exe?siab/cnv/SIABSpi.def`.

4.2 Dados de internação por abuso de álcool no estado da Paraíba

De início foram abordados os dados de internação referentes ao estado da Paraíba, sendo feita a análise dos principais aspectos que permitem avaliar o desempenho do modelo . Os dados são compostos por 132 observações que foram feitas entre os anos de 2005 e 2015, tendo estes uma média igual a 58,174 e variância igual a 746,893, ou seja, trata-se de um conjunto de dados que apresenta sobredispersão.

A Figura 4.1 apresenta o gráfico da série temporal, que sugere estacionaridade para os dados. os gráficos de autocorrelação e de autocorrelação parcial indicam que dados apresentam baixa correlação.

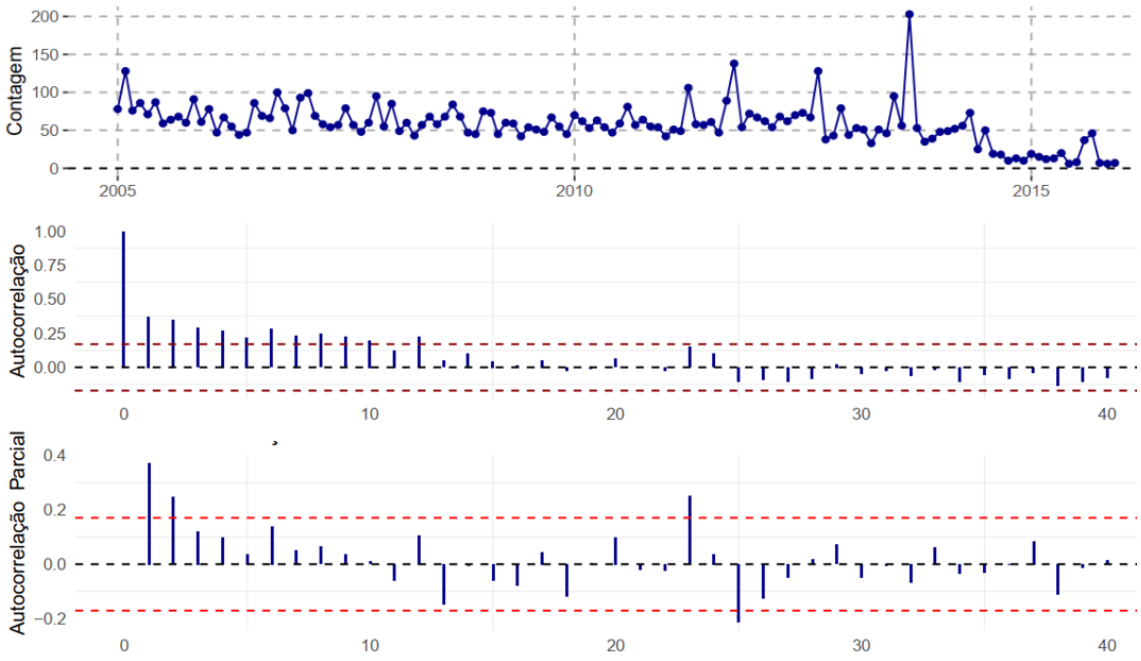

Figura 4.1: Série temporal de contagem do número de internações por abuso de álcool na Paraíba, autocorrelação e autocorrelação parcial

Os gráficos das Figuras 4.2 e 4.3 apresentam os ajustes para as séries temporais de contagem do número de internações por abuso de álcool no estado da Paraíba feitos respectivamente pelos modelos log-linear Poisson e PIG.

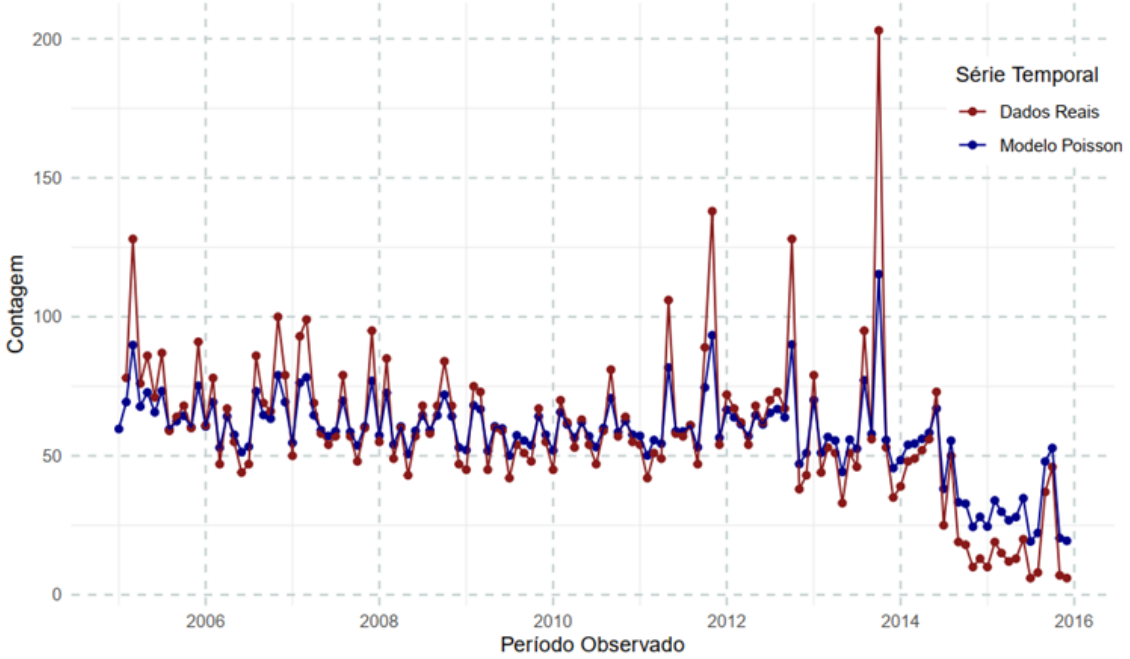

Figura 4.2: Ajuste dos dados pelo modelo log-linear Poisson para dados de internação na Paraíba.

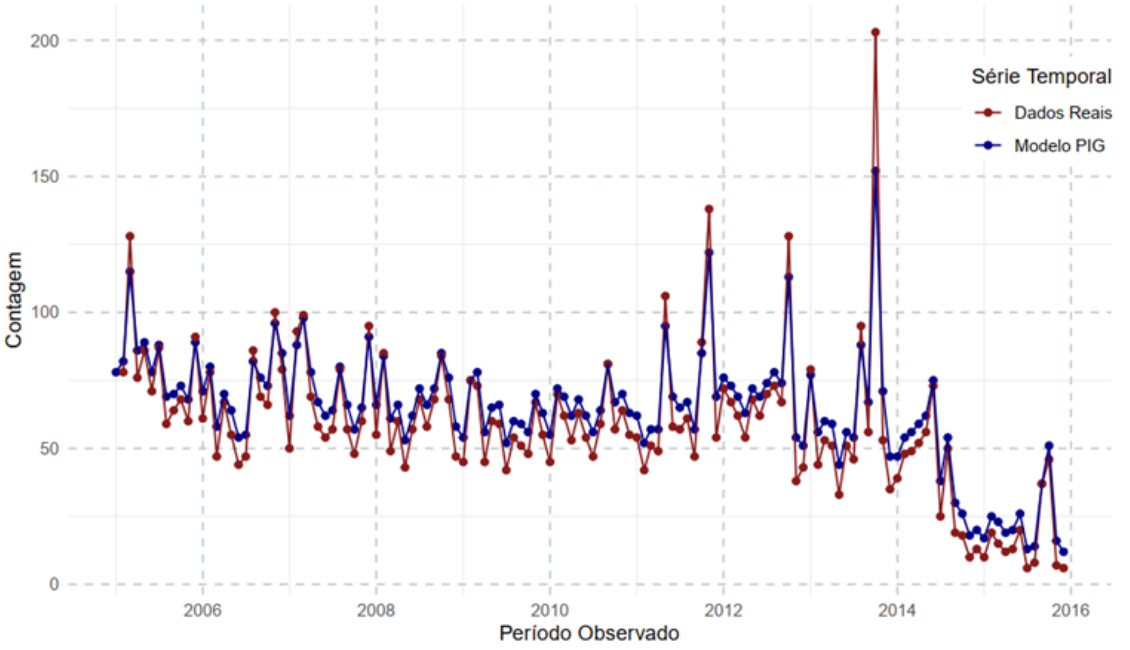

Figura 4.3: Ajuste dos dados pelo modelo log-linear PIG para dados de internação na Paraíba.

Através dos gráficos já é possível conjecturar visualmente que o melhor ajuste entre as duas distribuições abordadas é feito pela log-linear PIG ao observar os pontos de pico e principalmente, observando a disposição dos dados no decorrer do

ano de 2015 em que é visível uma melhor aproximação para o segundo modelo.

Os valores estimados pelo modelo log-linear Poisson apresentam média igual a 57,995 e variância igual a 217,755 enquanto o modelo INGARCH log-linear PIG apresenta média igual a 64,028 e variância igual a 473,309, ou seja, o modelo PIG captura a sobredispersão com maior eficácia .

Para confirmar a conjectura de que o modelo que melhor se ajusta ao conjunto de dados é o INGARCH log-linear PIG, basta observar os valores obtidos para o AIC, BIC e o EQM para os dados apresentado na Tabela 4.1 em que é possível confirmar então que a PIG apresenta os menores valores.

Tabela 4.1: Valores estimados para os parâmetros e critérios de comparação AIC, BIC e EQM para os dados de estado da Paraíba.

Modelo	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	AIC	BIC	EQM
Poisson	1,996	$-0,024$	0,536	1958,58	1967,23	1,581
PIG	0,683	0,170	0,682	**1205,16**	**1213,81**	**0,763**

Fonte: Dados de simulações em linguagem R feitas pelo autor.

O modelo escolhido é INGARCH log-linear PIG pois é o que melhor se ajustou aos dados. É possível ainda verificar se o modelo é adequado ao conjunto de dados através da análise de resíduos. Tomando os erros $e_i = y_i - \hat{\mu}_i$ em que y_i é a i-ésima observação, $\hat{\mu}_i$ é a i-ésima média obtida no processo de simulação e considerando a variância dos erros $var(e)$, então os resíduos podem ser obtidos fazendo

$$r_i = \frac{e_i}{var(e)},$$

em que r_i é o i-ésimo resíduo. A equação para os resíduos para a Poisson mista com membros da exponencial, como afirma [22] é dada por

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(1 + \hat{\mu}_i\hat{\phi}^{-1}b''(\xi_0))}}. \tag{4.1}$$

A Figura 4.4 apresenta os gráficos de autocorrelação e autocorrelação parcial, o gráfico da densidade dados e o QQ-plot. Através da analise dos gráficos fica claro que a correlação entre os resíduos é baixa e, pelos gráficos da densidade e QQ-plot, nota-se que estes indicam que os resíduos apresentam uma distribuição assintoticamente normal, o que aponta um bom ajuste do modelo log-linear PIG.

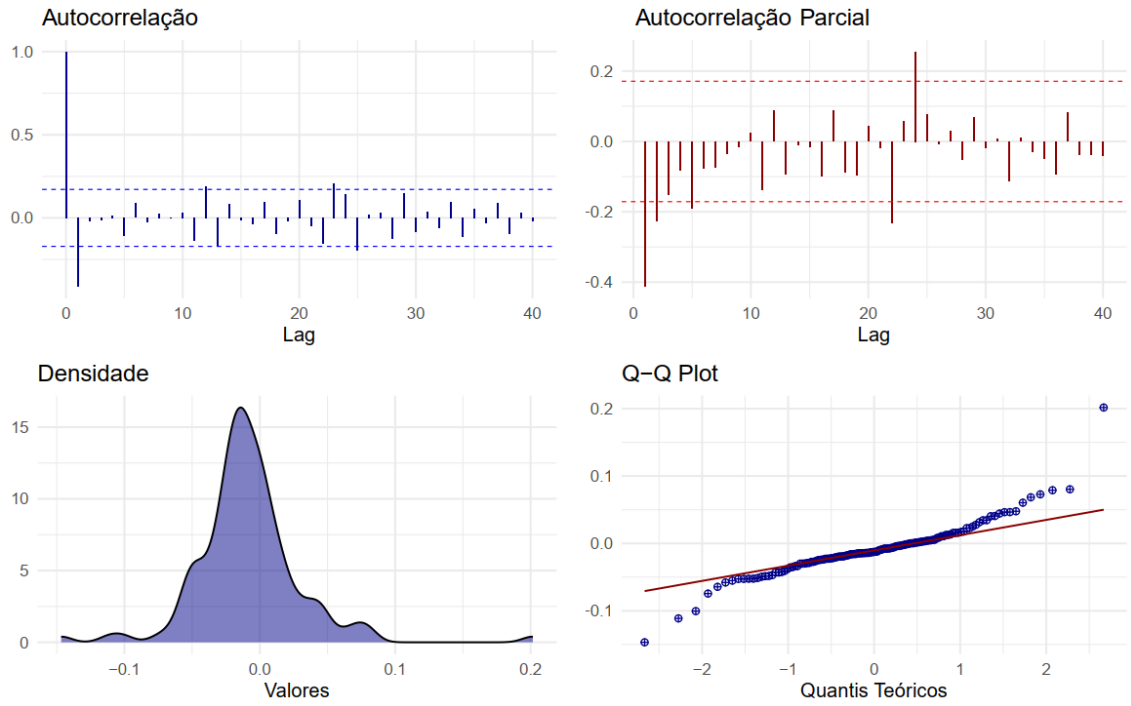

Figura 4.4: Resíduos para os dados de internação por abuso de álcool na Paraíba. ACF, PACF, densidade estimada e Q-Q plot.

4.3 Dados de internação por abuso de álcool no estado do Piauí

De modo análogo ao que se construiu para o conjunto de dados de internação para o estado da Paraíba é feito agora para os dados referentes às internações por abuso de álcool no estado do Piauí. Este conjunto de dados possui tamanho amostral de 132 observações feitas entre os anos de 2005 e 2015, com média igual a 51,470 e variância igual a 397,915, ou seja, trata-se de um conjunto que apresenta sobredispersão.

O conjunto de dados para o estado do Piauí são representados graficamente na Figura 4.5.

O comportamento temporal dos dados é explicitado no primeiro gráfico da figura e sugere estacionaridade, enquanto que os outros dois representam, respectivamente a autocorrelação e a autocorrelação parcial dos dados, sendo possível observar uma correlação baixa para as observações.

As figuras 4.6 e 4.7 apresentam os ajustes, respectivamente, pelos modelos INGARCH log-linear Poisson e o PIG. Como aconteceu na análise das simulações para o exemplo anterior, novamente é possível perceber visualmente que o segundo modelo se sobressai ao primeiro apresentando melhor ajuste. Isso pode ser notado principalmente nos pontos mais afastados no conjunto cuja proximidade dos pontos estimados pela PIG se mostra menor. Além disso, também é fácil perceber de forma

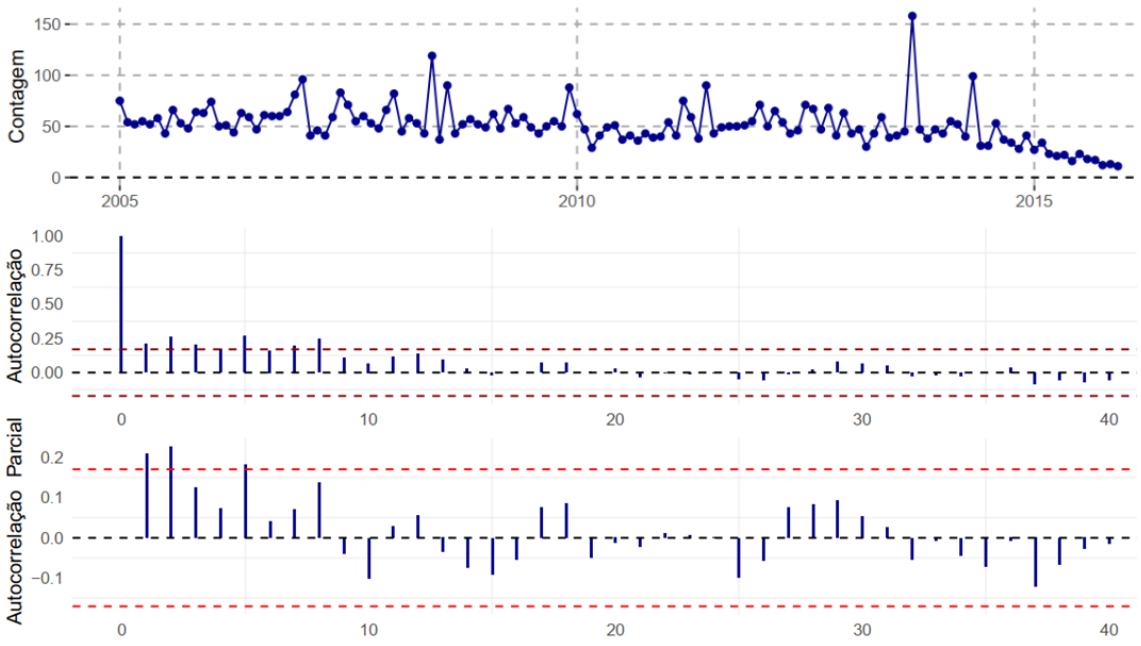

Figura 4.5: Série temporal de contagem do número de internações por abuso de álcool no Piauí, autocorrelação e autocorrelação parcial

visual que os dados estimados entre os anos de 2014 e 2016 se mostram bem mais afastados dos estimados pelo modelo Poisson que pelo modelo PIG.

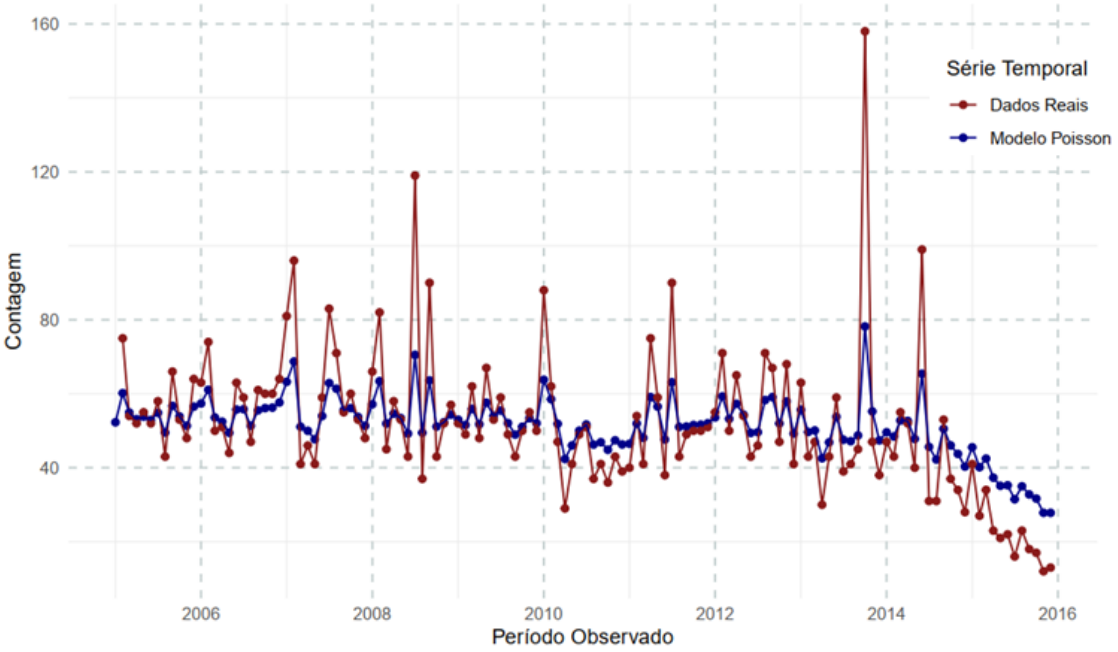

Figura 4.6: Ajuste dos dados pelo modelo log-linear Poisson para dados de internação no Piauí.

No ajuste feito pelo modelo log-linear Poisson, a média dos dados estimados é igual a 51,564 enquanto a variância é igual a 60,753. Já para o modelo INGARH log-linear PIG, a média é igual a 55,473 e a variância igual a 155,609, ou seja, a variância é três vezes maior que a média enquanto que o modelo log-linear Poisson

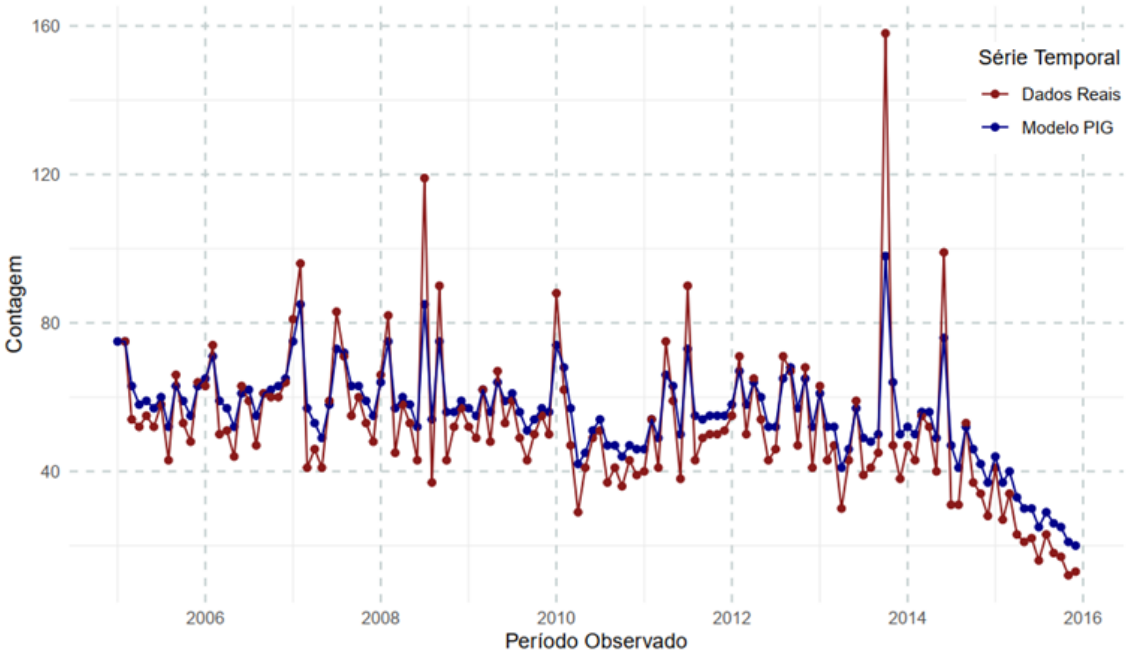

Figura 4.7: Ajuste dos dados pelo modelo log-linear PIG para dados de internação no Piauí.

apresenta média e variância bem próximas.

Como modo mais formal de avaliação e de comparação entre os modelos que foram utilizados, novamente foram utilizados os critérios AIC, BIC e EQM. Os valores obtidos são apresentados na Tabela 4.2 de modo que novamente os critérios apontam melhor ajuste é feito pelo modelo log-linear PIG que é novamente o modelo escolhido.

Tabela 4.2: Valores estimados para os parâmetros e critérios de comparação AIC, BIC e EQM para os dados de estado do Piauí.

Modelo	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	AIC	BIC	EQM
Poisson	1,616	0,216	0,376	1552,58	1561,23	1,676
PIG	0,652	0,308	0,539	**1136,42**	**1145,07**	**0,488**

Fonte: Dados de simulações em linguagem R feitas pelo autor.

A análise dos resíduos cujos gráficos encontram-se na Figura 4.8 apresentam função de autocorrelação e autocorrelação parcial que indicam a sugerindo baixa correlação entre os resíduos. Além disso, os gráficos da e a densidade e o QQ-plot mostram que os resíduos apresentam uma distribuição assintoticamente normal. Com a análise de resíduos fica claro que o modelo INGARCH log-linear PIG apresenta um bom ajuste aos dados.

Para os bancos de dados utilizados anteriormente, a BN apresentou problemas numéricos e não foi possível fazer as simulações usando esta distribuição. Deste modo, fez-se necessário a abordagem de outro conjunto de dados, cujos resultados dos ajustes são apresentados na seção a seguir.

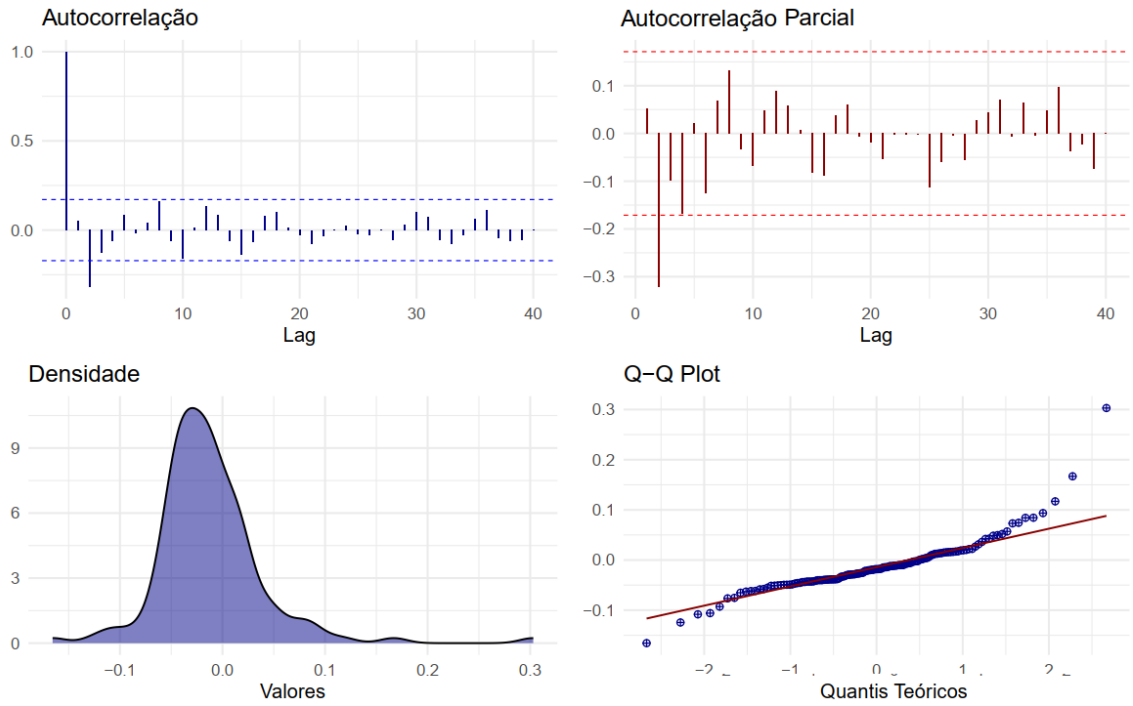

Figura 4.8: Resíduos para os dados de internação por abuso de álcool no Piauí. ACF, PACF, densidade estimada e Q-Q plot.

4.3.1 Casos de contaminação por campilobactérias

As infecções por bactérias são muito comuns e são elas as responsável por um grande número de mortalidade especialmente em crianças como ressalta [19]. O autor afirma que a infecção por campilobactéria é o segundo maior causador de doenças diarreicas no mundo. Seu contágio ocorre por ingestão de água ou alimentos contaminados. Seu efeito é, em muitos casos, fatal principalmente em crianças, pois como afirma [19], "a infecção é responsável por cerca de 25000 mortes de crianças de 0 a 5 anos anualmente".

O banco de dados denominado campy é parte integrante do pacote `tscount` cuja descrição completa é dada em [16] que também apresenta a descrição de outros bancos de dados disponíveis no pacote. Este banco de dados apresenta 140 observações feitas semanalmente entre o mês de janeiro de 1990 a outubro de 2000, que apresenta número de casos de infeções por campilobactérias no norte da província de Quebec no Canadá. Estes dados apresentam média igual a 11,543 e variância igual a 53,543, ou seja, são dados sobredispersos.

A Figura 4.9 apresenta o gráfico da série temporal de contagem, o gráfico de autocorrelação e autocorrelação parcial. Os dados não apresentam correlação sendo então estacionários.

O ajuste obtidos pelas simulações para os modelos log-linear Poisson, PIG e BN, respectivamente, são apresentados nas Figuras 4.10, 4.11 e 4.12.

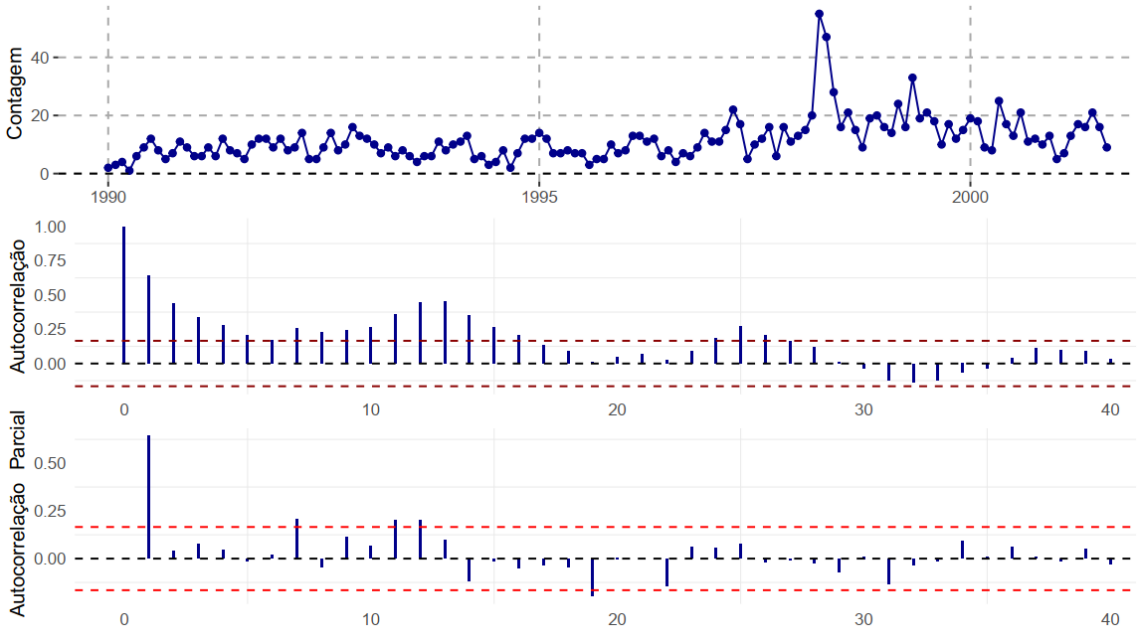

Figura 4.9: Série temporal de contagem do número de casos de infecções por campilobactérias, autocorrelação e autocorrelação parcial

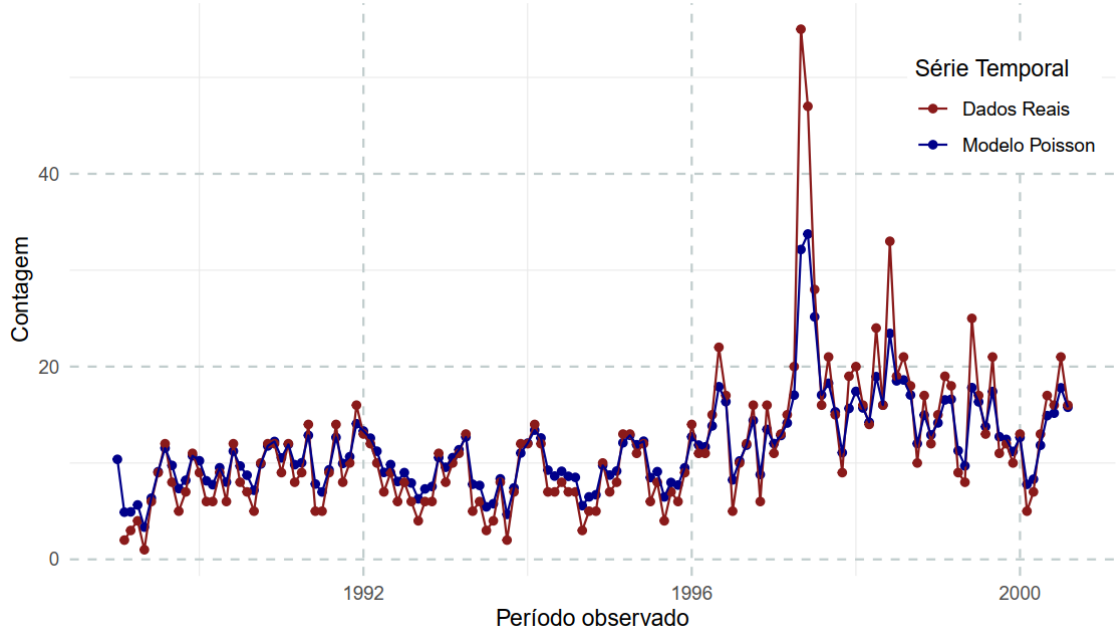

Figura 4.10: Ajuste dos dados pelo modelo log-linear Poisson para dados infecção por campilobactéria.

Para os dados de infecção por Campilobactéria, o modelo log-linear Poisson apresentou média e variâncias iguais, respectivamente, 11,608 e 21,171 enquanto que o modelo INGARCH log-linear PIG apresentou como média e variância, respectivamente, 11,510 e 20,101 e por fim, o modelo INGARCH BN apresentou média e variância dadas, respectivamente, por 11,521 e 21,086. Os valores para média e variância são bem próximos entre os três modelos.

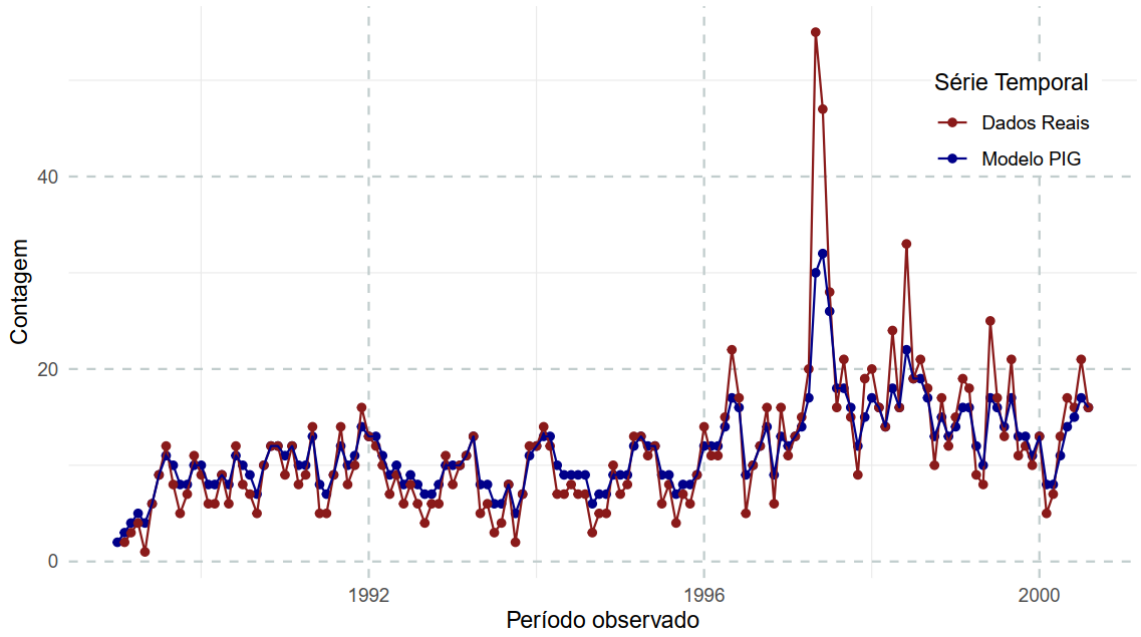

Figura 4.11: Ajuste dos dados pelo modelo INGARCH log-linear PIG para dados infecção por campilobactéria.

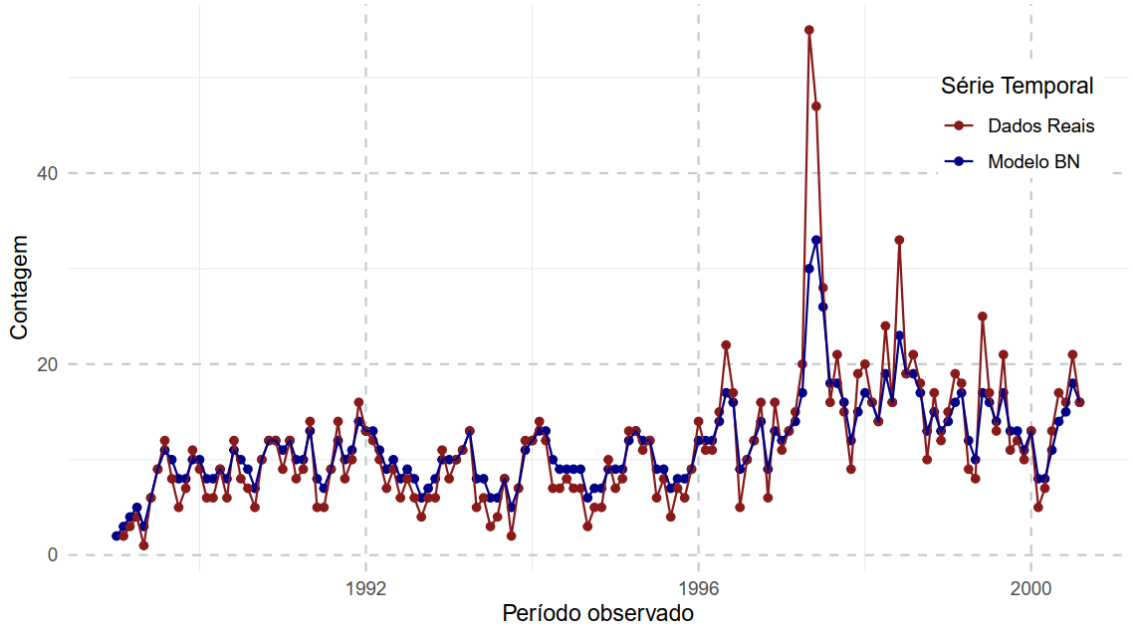

Figura 4.12: Ajuste dos dados pelo modelo INGARCH log-linear BN para dados infecção por campilobactéria.

Através de uma análise visual não foi possível conjecturar entre as três situações simuladas, qual o melhor ajuste, entretanto os valores obtidos para o AIC, BIC e EQM, que são apresentados na Tabela 4.9 deixam claro que o modelo INGARCH log-linear PIG é o que apresenta melhor resultado pelos critérios de escolha. É possível concluir, ainda pelos resultados da Tabela 4.9, que a BN apresenta melhores resultados que a Poisson para este conjunto de dados. Portanto, o modelo escolhido é o INGARCH log-linear PIG.

Tabela 4.3: Valores estimados para os parâmetros e critérios de comparação AIC, BIC e EQM para os dados de infecção por campilobactéria.

Modelo	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\eta}$	AIC	BIC	EQM
Poisson	0,401	0,224	0,605	879,21	888,03	0,557
PIG	0,380	0,307	0,532	**807,54**	**816,36**	**0,007**
BN	0,364	0,306	0,540	810,45	819,28	0,029

Fonte: Dados de simulações em linguagem R feitas pelo autor.

Com a escolha do modelo INGARCH log-linear PIG, a análise dos resíduos é mostrada nos gráficos da Figura 4.17 em que a autocorrelação e a autocorrelação parcial sugerem uma condição estacionária para os dados residuais enquanto que o gráfico da densidade e o QQ-plot apontam um comportamento assintoticamente normal para os resíduos.

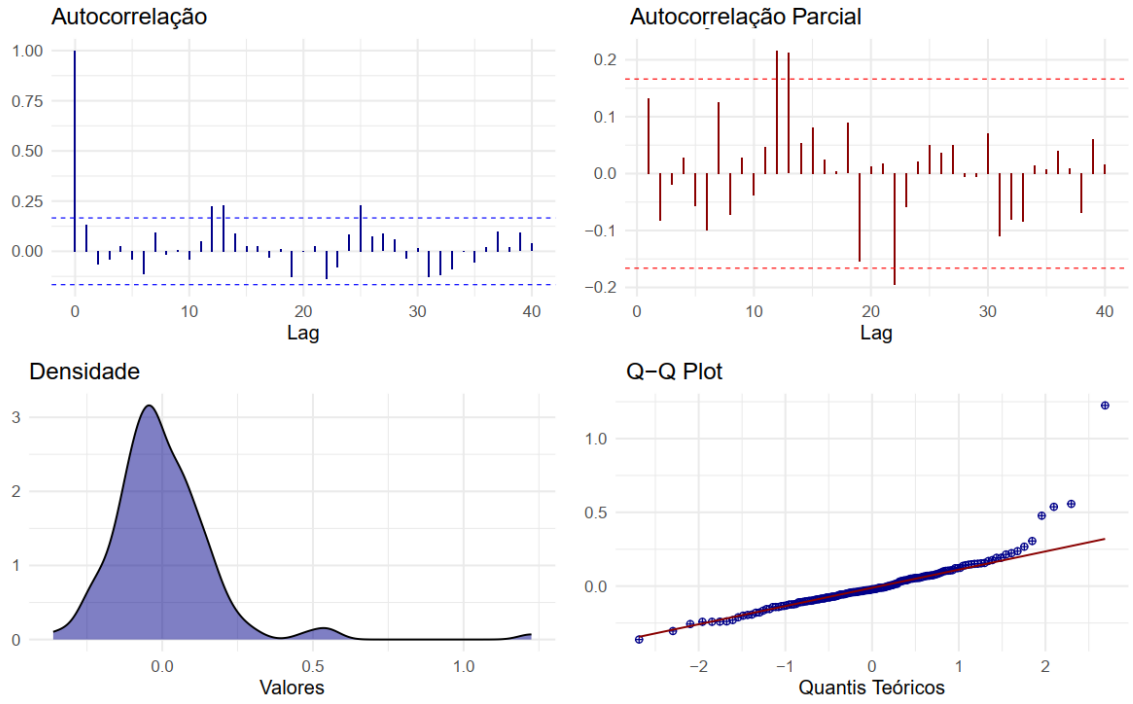

Figura 4.13: Resíduos para os dados de infecção por campilobactéria. ACF, PACF, densidade estimada e Q-Q plot.

Em todos os conjuntos de dados reais avaliados foi possível observar que o desempenho no ajuste foi melhor quando feito utilizando o modelo INGARCH log-linear PIG, membro do conjunto de distribuições INGARCH Poisson misto em que a média dos dados segue uma distribuição que integra a família exponencial.

Capítulo 5

Conclusões

Este trabalho teve como objetivo a proposta de uma nova classe de modelos para séries temporais, denominados INGARCH Poisson mistos, em que os efeitos não observáveis seguem uma distribuição pertencente à família exponencial. Para o desenvolvimento do estudo forma considerados dois casos particulares, aos quais denominou-se modelo INGARCH log-linear PIG e BN. Com o estudo de simulações para amostras finitas, foi possível avaliar de forma positiva a convergência dos parâmetros estimados para ambas as distribuições, ressaltando que o segundo caso abordado apresentou rejeição a alguns cenários de simulação propostos.

Algumas lacunas que foram deixadas pelo trabalho de referência para este texto ainda continuam em aberto, como é o caso das demonstrações das propriedades do modelo log-linear Poisson. Priorizou-se aqui a apresentação da proposta do modelo, abordando dois casos específicos através do algoritmo EM. Também não foram abordadas as propriedades assintóticas dos estimadores. Um estudo mais aprofundado sobre as propriedades supracitadas fica como sugestão para trabalhos futuros.

Para avaliação dos casos abordados, foram considerados três conjuntos de dados reais, verificando em todos eles um desempenho superior para o caso log-linear PIG com relação ao modelo de comparação, denominado log-linear Poisson. A avaliação do modelo escolhido nas aplicações se deu pela análise gráfica das séries temporais dos dados, pelos critérios de avaliação AIC e BIC, além da análise dos resíduos.

O comportamento do modelo PIG mostrou-se satisfatório, apresentando desempenho superior aos demais nas aplicações a dados reais e, deste modo, apresentando-se candidato forte a modelar séries temporais de contagem para os mais distintos contextos.

Vale a pena salientar que cada estudo relacionado a aplicação a dados de contagem no tempo tem suas particularidades, da mesma forma que cada proposta de distribuição de probabilidade visa melhor se ajustar ao comportamento da amostra.

Cabe ao pesquisador que irá fazer a utilização de um modelo, avaliar e escolher o processo que melhor retrate o fenômeno a ser estudado.

Um ponto crucial para justificar a importância de estudos relacionados a modelos de séries temporais de contagem pode ser observados nas aplicações apresentadas neste texto, em especial os primeiros a serem abordados. Os dados analisados retratam realidades de problemas de saúde pública que merecem a atenção das autoridades competentes no sentido de traçar estratégias de combate, conscientização, acompanhamento de casos e assistência a família. A possibilidade de estudo de uma projeção temporal do número de internações por abuso de álcool pode ajudar a entender muito sobre os fatores agravantes ao problema, além de oferecer a possibilidade de uma melhor programação de assistência a este tipo de problema.

Ainda em relação às aplicações de modelos de séries temporais aplicados a dados de saúde, estes ainda oferecem a possibilidade de uma análise periódica em que se possa traçar um panorama apresentando os períodos de maior ocorrência de casos, visando estabelecer conexão com o aumento ou diminuição do número de ocorrências. Esse foram só alguns exemplos do potencial de estudos relacionados às series temporais, em especial as de contagem.

Por fim é apresentado outro fruto importante gerado a partir da concepção apresentada neste trabalho, que é a possibilidade de novas propostas de modelos pertencentes à família INGARCH log-linear Poisson mista com membros da família exponencial, obviamente levando em consideração a complexidade da distribuição a ser abordada, bem como a viabilidade de execução do processo computacional. Esta colocação se dá exatamente pelo fato de que a ideia inicial de se trabalhar o com os modelos apresentados seriam abordadas a partir de um algoritmo mais simples de otimização direta por verossimilhança, mas que se mostrou inviável devido a funções complicadas de implementar.

Referências Bibliográficas

[1] ABBASI, W. A., 2004, *Time Series Analysis of Heart Rate Variability(HRV)Signals*. Tese de Doutorado, 12.

[2] BARRETO-SOUZA, W., SIMAS, A. B., 2016, "General mixed Poisson regression models with varying dispersion", *Statistics and Computing*, v. 26, pp. 1263–1280.

[3] BOX, G. E., JENKINS, G. M., REINSEL, G. C., et al., 2015, *Time series analysis: forecasting and control*. John Wiley & Sons.

[4] BURNHAM, K. P., ANDERSON, D. R., 2004, "Multimodel inference: understanding AIC and BIC in model selection", *Sociological methods & research*, v. 33, n. 2, pp. 261–304.

[5] CENTRO DE INFORMAÇÕES SOBRE SAÚDE E ÁLCOOL, 2023, *Álcool e a Saúde dos Brasileiros: PANORAMA 2023*. Relatório técnico, CISA. Disponível em: <https://cisa.org.br/images/upload/Panorama_Alcool_Saude_CISA2023.pdf?utm_source=sitecisa&utm_medium=cpc&utm_campaign=panorama_2023&utm_id=panorama2023&utm_term=panorama%2Bsaude%2Balcool&utm_content=btnlink>.

[6] DEKKING, F. M., KRAAIKAMP, C., LOPUHAÄ, H. P., et al., 2005, *A Modern Introduction to Probability and Statistics: Understanding why and how*, v. 488. Springer.

[7] EAS HEINEN, A., 2000, "Modeling Time Series Count Data: An Autoregressive Conditional Poisson Model", .

[8] FERLAND, R., LATOUR, A., ORAICHI, D., 2006, "Integer-valued GARCH process", *Journal of time series analysis*, v. 27, n. 6, pp. 923–942.

[9] FOKIANOS, K., TJØSTHEIM, D., 2011, "Log-linear Poisson autoregression", *Journal of multivariate analysis*, v. 102, n. 3, pp. 563–578.

[10] HAY, J. L., PETTITT, A. N., 2001, "Bayesian analysis of a time series of counts with covariates: an application to the control of an infectious disease", *Biostatistics*, v. 2, n. 4 (12), pp. 433–444. ISSN: 1465-4644. doi: 10.1093/biostatistics/2.4.433. Disponível em: <https://doi.org/10.1093/biostatistics/2.4.433>.

[11] JONES, O., MAILLARDET, R., ROBINSON, A., 2009, *Introduction to scientific programming and simulation using R*. Chapman and Hall/CRC.

[12] KARLIS, D., XEKALAKI, E., 2005, "Mixed poisson distributions", *International Statistical Review/Revue Internationale de Statistique*, pp. 35–58.

[13] KARMAKAR, S., GUPTA, R., CEPNI, O., et al., 2023, "Climate risks and predictability of the trading volume of gold: Evidence from an INGARCH model", *Resources Policy*, v. 82, pp. 103438.

[14] KIM, B., LEE, S., KIM, D., 2021, "Robust estimation for bivariate Poisson INGARCH models", *Entropy*, v. 23, n. 3, pp. 367.

[15] KIM, M., 2020, "Network traffic prediction based on INGARCH model", *Wireless Networks*, v. 26, n. 8, pp. 6189–6202.

[16] LIBOSCHIK, T., FRIED, R., FOKIANOS, K., et al., 2020, "Package 'tscount"', .

[17] MONTGOMERY, D. C., JENNINGS, C. L., KULAHCI, M., 2015, *Introduction to time series analysis and forecasting*. John Wiley & Sons.

[18] ONG, S.-H., LOW, Y.-C., TOH, K.-K., 2019, "Recent developments in mixed Poisson distributions", *Computational Statistics and Data Analysis*, v. 132.

[19] POLY, F., NOLL, A. J., RIDDLE, M. S., et al., 2019, "Update on Campylobacter vaccine development", *Human vaccines & immunotherapeutics*, v. 15, n. 6, pp. 1389–1400.

[20] R CORE TEAM, 2023, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <https://www.R-project.org/>.

[21] SHUMWAY, R. H., STOFFER, D. S., STOFFER, D. S., 2000, *Time series analysis and its applications*, v. 3. Springer.

[22] SILVA, R. B., BARRETO-SOUZA, W., 2019, "Flexible and robust mixed Poisson INGARCH models", *Journal of Time Series Analysis*, v. 40, n. 5, pp. 788–814.

[23] STASINOPOULOS, M., RIGBY, R., 2023, *gamlss.dist: Distributions for Generalized Additive Models for Location Scale and Shape*. Disponível em: <`https://CRAN.R-project.org/package=gamlss.dist`>. R package version 6.1-1.

[24] TRAPLETTI, A., HORNIK, K., 2023, *tseries: Time Series Analysis and Computational Finance*. Disponível em: <`https://CRAN.R-project.org/package=tseries`>. R package version 0.10-54.

[25] WEISS, C. H., 2018, *An introduction to discrete-valued time series*. John Wiley & Sons.

[26] WORLD HEALTH ORGANIZATION, 2018, *Global status report on alcohol and health 2018*. Geneva, Switzerland, World Health Organization.

[27] ZHU, F., 2011, "A negative binomial integer-valued GARCH model", *Journal of Time Series Analysis*, v. 32, n. 1, pp. 54–67.

[28] ZHU, F., 2012, "Modeling time series of counts with COM-Poisson INGARCH models", *Mathematical and Computer Modelling*, v. 56, n. 9-10, pp. 191–203.

Apêndices

Nos Apêndices são apresentadas as rotinas computacionais em linguagem R dos algoritmos EM para o modelo INGARCH log-linear PIG e BN, que forma nomeadas fitPIGINGARCH e fitBNINGARCH respectivamente. As funções recebem 3 entradas sendo estas `x`, que representa os dados temporais a serem modelados, `method` que é o método de otimização utilizado (no caso das simulações aqui apresentados foi utilizado o método `Nelder-Mead`) e `tol`, que representa a tolerância para o erro a que se deseja chegar.

Para facilitar a implementação, foram considerados $\alpha = d$, $\rho = a$ e $\eta = b$ na implementação. As funções denominadas auxiliares na implementação correspondem às decomposições dos membros da família exponencial utilizados para o modelo INGARCH log-linear PIG e BN respectivamente, apresentados para cada caso na Tabela 3.1.

Código R para o processo INGARCH log linear PIG

```r
# Codigo R para gerar autoregressao INGARCH log-linear PIG.
fitPIGINGARCH <- function(x, method, tol){
  # Bibliotecas R utilizadas
  pacman::p_load(tseries, gamlss.dist)

  # Numero de replicas de Monte Carlo
  N <- 1

  # Tamanho da amostra
  n <- length(x)

  # Funcoes auxiliares para o caso PIG
  # As funcoes sao obtidas a partir da decomposicao da Inversa
     Gaussiana como membro da familia exponencial
  qsi0 <- -1/2
  B <- function(x) {-sqrt((-2 * x))}
  d2B <- function(x) {(2 * sqrt(2) * (-x)^(3 / 2))^(-1)}
  D <- function(x) {log(x) / 2}
```

```
dD <- function(x) {1 / (2 * x)}
H <- function(x) {-log(2 * pi * x^3) / 2}
G <- function(x) {-1 / (2 * x)}

# Funcoes para as esperancas condicionais
Gamma_t <- function(x){
  y <- x[1:n]
  mu <- x[(n+1):(2*n)]
  phi <- x[2*n+1]
  (y + 1) * dPIG(y + 1, mu , 1 / phi) / (mu * dPIG(y, mu, 1 /
     phi))
}

Psi_t <- function(x){
  y <- x[1:n]
  mu <- x[(n+1):(2*n)]
  phi <- x[(2*n+1)]
  res <- (y == 0) * ((1 + sqrt(phi * (2 * mu + phi))) / phi) +
    (y > 0) * (mu * dPIG((y > 0) * (y - 1), mu, 1 / phi) /
                 (((y == 0) + y) * dPIG(y, mu, 1 / phi)))
  - .5 * res
}

# Vetor para receber os valores estimados
estimates <- array(0, c(N,4))

# Inicio do processo de Monte Carlo
k = 1
while(k <= N){
  x <- x

  # Inicio do processo de estimacao pelo algoritmo EM
  # Ajuste de tolerancia - criterio de parada.
  tol <- tol
  tolerance <- 1
  d_old <- .5
  a_old <- .5
  b_old <- .5
  phi_old <- .5

  while(tolerance > tol){
    # Vetor para receber as atualizacoes de nu_old
    nu_old <- NULL
    nu_old[1] <- 1
    for (i in 2:n){
      nu_old[i] <- d_old + a_old * nu_old[i-1] + b_old *
         log(x[i-1] + 1)
```

```
63      }
64
65      # Funcao e logverossimilhanca esperada
66      q.function <- function(z){
67        d <- z[1]
68        a <- z[2]
69        b <- z[3]
70        phi <- z[4]
71
72        # Vetor para receber os valores estimados para nu
73        nu <- NULL
74        nu[1] <- 1
75        for (i in 2:n){
76          nu[i] <- d + a * nu[i-1] + b * log(x[i-1] + 1)
77        }
78        sum(x * nu - exp(nu) * Gamma_t(c(x, exp(nu_old), phi_old))
             +
79              D(phi) + phi * (Gamma_t(c(x, exp(nu_old), phi_old))
                  * qsi0 -
80                                  B(qsi0) + Psi_t(c(x, exp(nu_old),
                                      phi_old))))
81      }
82
83      # Derivada de nu_t com relacao a d
84      Dnu_d <- function(z){
85        a <- z[1]
86        dnu <- array(0, c(n,1))
87        dnu[1] <- 0
88        dnu[2:n] <- 1 + a * dnu[1:(n-1)]
89        dnu
90      }
91
92      # Derivada de nu_t com relacao a a
93      Dnu_a <- function(z){
94        a <- z[1]
95        dnu <- array(0, c(n,1))
96        dnu[1] <- 0
97        dnu[2:n] <- nu[1:(n-1)] + a * dnu[1:(n-1)]
98        dnu
99      }
100
101     # Derivada de nu_t com relacao a b
102     Dnu_b <- function(z){
103       a <- z[1]
104       dnu <- array(0, c(n,1))
105       dnu[1] <- 0
106       dnu[2:n] <- log(x[1:(n-1)] + 1) + a * dnu[1:(n-1)]
```

```
107         dnu
108       }
109
110       # Vetor gradiente
111       g.function <- function(z){
112         d <- z[1]
113         a <- z[2]
114         b <- z[3]
115         phi <- z[4]
116
117         nu <- NULL
118         nu[1] <- 1
119         for (i in 2:n){
120           nu[i] <- d + a * nu[i-1] + b * log(x[i-1] + 1)
121         }
122
123         grad1 <- function(par){
124           x <- par[1:n]
125           nu <- par[(n+1):(2*n)]
126           nu_old <- par[(2*n+1):(3*n)]
127           sum((x[1:n] - Gamma_t(c(x, nu_old, phi_old))[1:n] *
                exp(nu[1:n])) * Dnu_d(a)[1:n])
128         }
129
130         grad2 <- function(par){
131           x <- par[1:n]
132           nu <- par[(n+1):(2*n)]
133           nu_old <- par[(2*n+1):(3*n)]
134           sum((x[1:n] - Gamma_t(c(x, nu_old, phi_old))[1:n] *
                exp(nu[1:n])) * Dnu_a(a)[1:n])
135         }
136
137         grad3 <- function(par){
138           x <- par[1:n]
139           nu <- par[(n+1):(2*n)]
140           nu_old <- par[(2*n+1):(3*n)]
141           sum((x[1:n] - Gamma_t(c(x, nu_old, phi_old))[1:n] *
                exp(nu[1:n])) * Dnu_b(a)[1:n])
142         }
143
144         grad4 <- function(par){
145           x <- par[1:n]
146           nu <- par[(n+1):(2*n)]
147           nu_old <- par[(2*n+1):(3*n)]
148           sum(dD(phi) + Gamma_t(c(x, nu_old, phi_old))[1:n] * qsi0
                - B(qsi0) + Psi_t(c(x, exp(nu_old), phi_old))[1:n])
149         }
```

```
150
151       c(grad1(c(x, nu, nu_old)), grad2(c(x, nu, nu_old)),
              grad3(c(x, nu, nu_old)), grad4(c(x, nu, nu_old)))
152     }
153
154     fit <- try(optim(par = c(d_old, a_old, b_old, phi_old), fn =
            q.function, gr = g.function,
155                      method = method , control = list(fnscale =
                             -1)))
156     d_new <- fit$par[1]
157     a_new <- fit$par[2]
158     b_new <- fit$par[3]
159     phi_new <- fit$par[4]
160
161     # Criterio de tolerancia
162     tolerance <- max(abs(c(d_new, a_new, b_new, phi_new) -
            c(d_old, a_old, b_old, phi_old)))
163
164     # Atualizacao dos valores dos parametros
165     d_old <- d_new
166     a_old <- a_new
167     b_old <- b_new
168     phi_old <- phi_new
169   }
170   estimates[k,] <- fit$par
171   print(k)
172   k <- k + 1
173 }
174 # Fim do processo de estimacao dos parametros
175
176 est <- apply(estimates, 2, mean)
177
178 # Inicio do processo de estima  o para as medias do modelo
179 dhat <- est[1]
180 ahat <- est[2]
181 bhat <- est[3]
182 phihat <- est[4]
183
184 # Geracao de observacoes para o modelo PIG
185 nu <- NULL
186 mu <- NULL
187
188 # Valores iniciais para nu[1] and mu[1]:
189 nu[1] <- log(x[1])
190 mu[1] <- exp(nu[1])
191 for (i in 2:n) {
192   nu[i] <- dhat + ahat*nu[i-1] + bhat * log(x[i-1]+1)
```

```
    mu[i] <- exp(nu[i])
  }

  # Calculo dos criterios de comparacao de modelos
  loglik <- sum(log(dPIG(x, mu, 1 / phihat)))
  # AIC: -2 * l(thetahat) + 2 * (p+q+1)
  aic <- -2 * loglik + (3) * 2
  # BIC: -2 * l(thetahat) + (p+q+1) * log(n)
  bic <- -2 * loglik + (3) * log(n)
}
```

Código R para o processo INGARCH log linear BN

```
# Codigo R para gerar autoregressao INGARCH log-linear BN.
# As entradas do modelo sao as mesmas do caso PIG
fitBNINGARCH <- function(x, method, tol){
  # Numero de replicas de Monte Carlo
  N <- 1

  # Tamanho da amostra
  n <- length(x)

  # Funcoes auxiliares: caso BN

  qsi0 <- -1
  B <- function(x) {-log(-x)}
  d2B <- function(x) {x^(-2)}
  D  <- function(x) {x * log(x) - log(gamma(x))}
  dD <- function(x) {1 + log(x) - digamma(x)}
  H <- function(x) {-log(x)}
  G <- function(x) {log(x)}

  # Esperancas condicionais

  Gamma_t <- function(x){
    y <- x[1:n]
    mu <- x[(n+1):(2*n)]
    phi <- x[2*n+1]

    (y + 1) * dnbinom(y + 1, size = phi, prob = phi/(mu + phi)) /
      (mu * dnbinom(y, size = phi, prob = phi/(mu + phi)))
  }

  Psi_t <- function(x){
    y <- x[1:n]
```

```
33      mu <- x[(n+1):(2*n)]
34      phi <- x[2*n+1]
35
36      res <- (mu^y / (factorial(y) * dnbinom(y, size = phi, prob =
            phi/(mu+phi)))) *
37        phi^phi * (gamma(y+phi)/gamma(phi)) *
38        ((digamma(y+phi)-log(mu+phi))/ (mu+phi)^(y+phi))
39
40      return(res)
41    }
42
43    estimates <- array(0, c(N,4))
44
45    # Inicio do processo de Monte Carlo
46    k = 1
47    while(k <= N){
48      x <- x
49      # Inicio do processo de estimacao pelo algoritmo EM
50      # Ajuste de tolerancia - criterio de parada.
51      tol <- tol
52      tolerance <- 1
53      d_old <- .5
54      a_old <- .5
55      b_old <- .5
56      phi_old <- .5
57      while(tolerance > tol){
58
59        # Valor inicial de nu
60        nu_old <- NULL
61        nu_old[1] <- 1
62        for (i in 2:n){
63          nu_old[i] <- d_old + a_old * nu_old[i-1] + b_old *
              log(x[i-1] + 1)
64        }
65        # Funcao de logverossimilhanca esperada
66        q.function <- function(z){
67          d <- z[1]
68          a <- z[2]
69          b <- z[3]
70          phi <- z[4]
71
72          # Atualizacao de nu
73          nu <- NULL
74          nu[1] <- 1
75
76          for (i in 2:n){
77            nu[i] <- d + a * nu[i-1] + b * log(x[i-1] + 1)
```

```
78          }
79
80          sum(x * nu - exp(nu) * Gamma_t(c(x, exp(nu_old), phi_old))
               +
81                D(phi) + phi * (Gamma_t(c(x, exp(nu_old), phi_old))
                     * qsi0 -
82                                  B(qsi0) + Psi_t(c(x, exp(nu_old),
                                       phi_old))))
83        }
84
85        # Derivada de nu_t com relacao a d
86        Dnu_d <- function(z){
87
88          a <- z[1]
89
90          dnu <- array(0, c(n,1))
91          dnu[1] <- 0
92
93          dnu[2:n] <- 1 + a * dnu[1:(n-1)]
94
95          dnu
96        }
97
98        # Derivada de nu_t com relacao a a
99        Dnu_a <- function(z){
100
101         a <- z[1]
102
103         dnu <- array(0, c(n,1))
104         dnu[1] <- 0
105
106         dnu[2:n] <- nu[1:(n-1)] + a * dnu[1:(n-1)]
107
108         dnu
109       }
110
111       # Derivada de nu_t com relacao a b
112       Dnu_b <- function(z){
113
114         a <- z[1]
115
116         dnu <- array(0, c(n,1))
117         dnu[1] <- 0
118
119         dnu[2:n] <- log(x[1:(n-1)] + 1) + a * dnu[1:(n-1)]
120
121         dnu
```

```
122       }
123
124       # Vetor gradiente
125       g.function <- function(z){
126
127         d <- z[1]
128         a <- z[2]
129         b <- z[3]
130         phi <- z[4]
131
132         # Aualizacao de nu
133         nu <- NULL
134         nu[1] <- 1
135
136         for (i in 2:n){
137           nu[i] <- d + a * nu[i-1] + b * log(x[i-1] + 1)
138         }
139
140         # Funcao score
141         grad1 <- function(par){
142
143           x <- par[1:n]
144           nu <- par[(n+1):(2*n)]
145           nu_old <- par[(2*n+1):(3*n)]
146
147           sum((x[1:n] - Gamma_t(c(x, nu_old, phi_old))[1:n] *
                  exp(nu[1:n])) * Dnu_d(a)[1:n])
148         }
149         grad2 <- function(par){
150
151           x <- par[1:n]
152           nu <- par[(n+1):(2*n)]
153           nu_old <- par[(2*n+1):(3*n)]
154
155           sum((x[1:n] - Gamma_t(c(x, nu_old, phi_old))[1:n] *
                  exp(nu[1:n])) * Dnu_a(a)[1:n])
156         }
157         grad3 <- function(par){
158
159           x <- par[1:n]
160           nu <- par[(n+1):(2*n)]
161           nu_old <- par[(2*n+1):(3*n)]
162
163           sum((x[1:n] - Gamma_t(c(x, nu_old, phi_old))[1:n] *
                  exp(nu[1:n])) * Dnu_b(a)[1:n])
164         }
165         grad4 <- function(par){
```

```
166
167           x <- par[1:n]
168           nu <- par[(n+1):(2*n)]
169           nu_old <- par[(2*n+1):(3*n)]
170
171           sum(dD(phi) + Gamma_t(c(x, nu_old, phi_old))[1:n] * qsi0
                  - B(qsi0) + Psi_t(c(x, exp(nu_old), phi_old))[1:n])
172         }
173
174         c(grad1(c(x, nu, nu_old)), grad2(c(x, nu, nu_old)),
              grad3(c(x, nu, nu_old)), grad4(c(x, nu, nu_old)))
175       }
176
177       fit <- try(optim(par = c(d_old, a_old, b_old, phi_old), fn =
            q.function, gr = g.function,
178                        method = method, control = list(fnscale =
                           -1)))
179
180       if(class(fit) != "try-error"){
181         d_new <- fit$par[1]
182         a_new <- fit$par[2]
183         b_new <- fit$par[3]
184         phi_new <- fit$par[4]
185
186         # Criterio de tolerancia
187         tolerance <- max(abs(c(d_new, a_new, b_new, phi_new) -
              c(d_old, a_old, b_old, phi_old)))
188
189         # Atualizacao dos valores dos parametros
190         d_old <- d_new
191         a_old <- a_new
192         b_old <- b_new
193         phi_old <- phi_new
194       }
195       if(class(fit) == "try-error"){
196         tolerance = tol
197       }
198
199     }
200     if(class(fit) != "try-error"){
201       estimates[k,] <- c(d_new, a_new, b_new, phi_new)
202
203       print(k)
204       k <- k + 1
205     }
206     else{
207       print("erro")
```

```
208         k <- k - 1
209       }
210     }
211     # Fim do processo de estimacao dos parametros
212
213     est <- apply(estimates, 2, mean);
214     # Inicio do processo de estimacao para as medias do modelo
215     dhat <- est[1]
216     ahat <- est[2]
217     bhat <- est[3]
218     phihat <- est[4]
219
220     # Geracao de observacoes para o modelo BN
221     nu <- NULL
222     mu <- NULL
223
224     # valores iniciais de nu e mu:
225     nu[1] <- log(x[1])
226     mu[1] <- exp(nu[1])
227
228     for (i in 2:n) {
229       nu[i] <- dhat + ahat*nu[i-1] + bhat * log(x[i-1]+1)
230       mu[i] <- exp(nu[i])
231     }
232     # Calculo dos critarios de comparacao de modelos
233     loglik <- sum(log(dnbinom(x, size = phihat, prob =
          phihat/(mu+phihat))))
234     # AIC: -2 * l(thetahat) + 2 * (p+q+1)
235     aic <- -2 * loglik + (3) * 2
236     # BIC: -2 * l(thetahat) + (p+q+1) * log(n)
237     bic <- -2 * loglik + (3) * log(n)
238 }
```