UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Comparação entre alguns modelos de regressão de Contagem

Lucas Akio Senaga Onuki

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Lucas Akio Senaga Onuki

Comparação entre alguns modelos de regressão de Contagem

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Jorge Luis Bazán Guzmán

USP – São Carlos
Junho de 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

O58c

Onuki, Lucas Akio Senaga
Comparação entre alguns modelos de regressão de Contagem / Lucas Akio Senaga Onuki; orientador Jorge Luis Bazán Guzmán. -- São Carlos, 2024.
78 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Modelos de Regressão. 2. Dados de Contagem. 3. Métodos de Estimação. 4. Inflacionamento de zeros. I. Guzmán, Jorge Luis Bazán, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Lucas Akio Senaga Onuki

Comparison between some Count regression models

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Master Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Jorge Luis Bazán Guzmán

USP – São Carlos
June 2024

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Lucas Akio Senaga Onuki, realizada em 27/09/2024.

Comissão Julgadora:

Prof. Dr. Jorge Luis Bazán Guzmán (USP)

Prof. Dr. Marcos Oliveira Prates (UFMG)

Prof. Dr. Alex de La Cruz Huayanay (PUC-Perú)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

AGRADECIMENTOS

Primeiramente agradeço a Deus, pela condução de minha jornada até aqui.

Sem menor prestígio, gostaria de agradecer ao meu orientador Jorge Luis Bazán Guzmán, por todo o suporte no meu aprendizado e realização desta dissertação.

Um agradecimento especial para os professores participantes das bancas de qualificação - Mário de Castro e Larissa Avila Matos e defesa desta dissertação - Alex de la Cruz, Marcos Prates. Com a colaboração de todos, o trabalho se tornou mais rico, organizado e detalhado. Muito obrigado.

Por fim, agradeço à minha família e amigos, que sempre estiveram presentes, apoiando minhas escolhas e acreditando em meu potencial.

RESUMO

ONUKI, L. A. S. **Comparação entre alguns modelos de regressão de Contagem**. 2024. 88 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Dados de contagem refletem o número de ocorrências de um comportamento de interesse em um período fixado de tempo (por exemplo, saldo de gols de um time no brasileirão). Um comportamento comum desse tipo de dado é a presença de muitos zeros observados, ie. inflacionamento de zeros, o que acaba viesando de certo modo as estimativas obtidas pelos modelos de Regressão Poisson e Binomial Negativa, usualmente utilizados para a modelagem desse tipo de dado. Com isso em mente, este trabalho propôs-se a estudar as variações desses modelos, seguindo em duas frentes: A primeira considerando modelos que conseguem comportar o possível excesso de zeros e uma segunda, que compara modelos da literatura recente para verificar se são boas alternativas em termos de estimativas e desempenho. No total, foram estudados sete modelos, sendo os dois mencionados acima, acrescidos de: Poisson-Tweedie, Bell, Zero-inflacionado Poisson, Zero-inflacionado Binomial Negativa e Zero-inflacionado Bell. Assim, diferentes cenários de simulação foram estudados computando métricas como média, desvio padrão, REQM e critérios de seleção de modelos, tais como AIC e BIC. Cabe ressaltar que tanto o método de estimação clássico, quanto a bayesiano, foram utilizados a critério comparativo das estimativas. Além dos estudos de simulação apresentamos duas aplicações a dados reais. Como resultado dos diversos cenários, entendemos que os modelos que possuem uma parte exclusiva para a acomodação de possíveis excessos de zeros possuíram maior aderência aos dados nas aplicações. Com relação aos modelos apresentados na literatura recente, podemos afirmar que há similaridade dos ajustes realizados, o que valida estudos anteriores e garante que são boas alternativas aos modelos Poisson e Binomial Negativa.

Palavras-chave: Modelos de Regressão, Dados de contagem, Métodos de estimação, Desempenho, Inflacionamento de zeros.

ABSTRACT

ONUKI, L. A. S. **Comparison between some Count regression models**. 2024. 88 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Count data reflects the number of occurrences of a behavior of interest in a given period of time (for example, a team's goals number in Brasileirão). A common behavior of this type of data is the presence of many zeros observed, i.e. zero-inflation, which ends up somewhat overturning the estimates obtained by the Poisson and Negative Binomial Regression models, usually used to model these type of data. With this in mind, this work set out to study the variations of these models, following two fronts: The first considering models that contain a possible excess of zeros and a second, which compares models from recent literature to check whether they are good alternatives in terms estimates and performance. In total, seven models were trained, the two mentioned above, plus: Poisson-Tweedie, Bell, Zero-inflated Poisson, Zero-inflated Negative Binomial and Zero-inflated Bell. Thus, different simulation scenarios were studied by computing metrics such as mean, standard deviation, REQM and model selection criteria, such as AIC and BIC. It is worth noting that both the classical and Bayesian study methods were used for comparative classification of estimates. In addition to the simulation studies, two applications to real data are presented. As a result of the different scenarios, we understand that the models that have an exclusive part to accommodate possible excesses of zeros had greater adherence to the data in applications. Regarding the models presented in recent literature, we can state that there is similarity in the adjustments made, which validates previous studies and guarantees that they are good alternatives to the Poisson and Negative Binomial models.

Keywords: Regression models, Counting data, Estimation methods, Performance, Zero-inflation.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AIC	Critério de Informação de Akaike
MLG's	Modelos Lineares Generalizados
pAIC	pseudo Critério de Informação de Akaike
PTw	Poisson-Tweedie
ZIBell	Zero-inflacionado Bell
ZIBN	Zero-inflacionado Binomial Negativa
ZIP	Zero-inflacionado Poisson

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

O passo inicial para adentrarmos o tema exposto neste trabalho é entendermos com que tipo de dado estamos lidando. Em qualquer análise com dados, esse passo é fundamental para sabermos qual técnica estatística é melhor aplicável ao contexto dos dados, dado que isso acaba direcionando a bons resultados e também evita erros. Nesse sentido, temos algumas características importantes sobre dados de contagem. O primeiro ponto válido a ressaltar é que os valores observados para uma variável de contagem são inteiros positivos, partindo do limite inferior zero, uma vez que não é possível observar contagens negativas. Podemos observar nesse tipo de dado certa assimetria positiva, evidenciando uma concentração de observações em valores menores. Ainda nesse sentido, quando há a observação de muitos zeros, sendo essa frequência consideravelmente superior a das demais observações, chamamos esse fenômeno de inflação de zeros. Essas características são suficientes para indicar que um modelo de contagem é aplicável. Podemos, assim, definir uma variável de contagem da seguinte forma: se Y é uma variável aleatória de contagem, os valores que ela pode assumir são $y = 0, 1, 2, 3, \ldots$ Na prática, podemos encontrar esse tipo de dado em cenários como competições esportivas, pesquisas médicas, educação e ciências sociais, por exemplo.

Comumente utilizamos para modelar esse tipo de dado os modelos Poisson e Binomial Negativa. O primeiro, considera a equidispersão dos dados, ou seja, o valor da média equivalente ao valor da variância. Isso acaba tornando-o um modelo mais simples, mais acessível, porém, um problema associado a ele é justamente a imposição de que a variância dos dados deva ser equivalente à média, algo que majoritariamente não é verdadeiro. O segundo modelo, comparando com o primeiro, pode ser descrito como um pouco mais flexível, dado ser possível acomodar o fenômeno de sobredispersão nos dados, ie. quando o valor da variância é maior que o valor da média. Esses modelos são apresentados no trabalho desenvolvido por Paula (2024), que descreve de forma didática não só esses modelos, mas também modelos utilizados para outros tipos de dados, todos pertencentes aos Modelos Lineares Generalizados, com aplicações no programa **R** (R Core Team (2024)), ponto que facilita a compreensão das metodologias e possíveis aplicações.

Cabe ressaltar que todas as análises apresentadas neste trabalho foram realizadas em **R**.

Ambos os modelos acabam com suas estimativas viesadas quando há a presença de muitos zeros nos dados. Com isso em mente, Yang *et al.* (2017) apresentam em seu trabalho as variantes zero-inflacionadas desses modelos aplicados à medicina, sendo os resultados discutidos para o acompanhamento comportamental de fumantes. Nesse sentido, elegeram uma comparação entre os modelos não inflacionados e os modelos inflacionados, comparando os ajustes via verificação da raiz do erro quadrático médio (REQM) e o AIC, tendo como conclusão que modelos inflacionados são mais adequados quando a frequência de zeros observados é consideravelmente superior aos demais valores amostrais.

Na literatura recente é possível encontrar modelos alternativos a Poisson e Binomial Negativa, dentre eles podemos citar três: Bell, Poisson-Tweedie e a variante zero-inflacionada do modelo Bell, podendo ser encontrados nos trabalhos discutidos a seguir. Castellares, Ferrari e Lemonte (2018) apresentam o modelo Bell destacando que para valores pequenos do parâmetro associado ao modelo encontramos similaridade com o modelo Poisson. Cabe ressaltar que apesar disso, trata-se de um modelo mais flexível, a medida que assim como a Binomial Negativa, o modelo Bell comporta a sobredispersão dos dados. Saha *et al.* (2020), em seu trabalho, descreve a mistura entre uma Poisson e uma Tweedie como o modelo de Poisson-Tweedie. Podemos definir mistura de modelos como sendo um modelo probabilístico que representa a presença de subpopulações dentro de uma população geral, sem exigir que um conjunto de dados observado deva identificar à qual subpopulação uma observação individual pertence. Um ponto forte apresentado pelo autor é a flexibilidade que a parte Tweedie garante à Poisson, de modo a tornar este, mais um bom modelo para acomodação de sobredispersão de dados. Ademais, vale salientar que apesar da flexibilidade e resultados similares ao modelo Binomial Negativa, esse modelo é mais complexo, dado que depende de uma aproximação para realização de sua estimação. Isso se deve ao fato desse modelo apresentar uma integral intratável. A aproximação baseia-se em suposições apenas de segundo momento, ou seja, média e variância. Essa aproximação garante acomodação de subdispersão nos dados, a partir de maior flexibilidade ao parâmetro de dispersão, a preço de não termos mais função massa de probabilidade nesse caso, como descrito por Jørgensen e Kokonendji (2015) e Bonat e Jørgensen (2016). Já o modelo Zero-inflacionado Bell, assim como os modelos estudados por Yang *et al.* (2017), consegue acomodar o comportamento de excesso de zeros nos dados. Uma possível vantagem comparando com os demais modelos inflacionados é o fato de que esse modelo é bem simples, em termos da sua função de distribuição, o que facilita a interpretação e pelo critério de parcimônia pode ser um diferencial, quando os resultados das estimações forem ao mínimo próximas as realizadas pelos modelos Zero-inflacionado Poisson e Zero-inflacionado Binomial Negativa.

Em termos de estimações de parâmetros, os trabalhos anteriores focaram em apresentar soluções seguindo a ótica clássica. Diferenciando-se do apresentado pelos autores, o presente trabalho propõe uma comparação entre a abordagem clássica e Bayesiana para os

modelos. Ressaltando que essa comparação foi realizada para os seguintes modelos: Poisson, Binomial Negativa, Zero-inflacionado Poisson e Zero-inflacionado Binomial Negativa. Bell e Zero-inflacionada Bell foi avaliada nos estudos de simulação sob olhar clássico, dado que em busca em referências bibliográficas Poisson-Tweedie não possui implementação Bayesiana porém, na aplicação, ambas foram avaliadas segundo as duas óticas de estimação.

Tendo essa contextualização em mente, tivemos por objetivo apresentar um panorama dos modelos comparando seus desempenhos, tanto em termos de abordagem - como mencionado anteriormente - quanto em termos do uso de diferentes pacotes do **R**, com o intuito de estudar os modelos em cenário simulado e verificar qual seria o modelo mais aderente em aplicação a dados reais. De mesmo modo, validar os estudos anteriores, que apontaram modelos inflacionados como melhores à tratativa de dados de contagem com excesso de zeros e os modelos Bell e Poisson-Tweedie, em termos de estimação, com resultados similares aos modelos Poisson e Binomial Negativa.

Sendo assim, este trabalho está organizado da seguinte maneira:

O Capítulo 2 apresenta os modelos regressão de contagem considerando a distribuição de Poisson (P) e a distribuição Binomial Negativa (BN), assim como aos respectivos modelos, zero-inflacionadas modelando à média (ZIP1 e ZIBN1) e modelando a média e a proporção de zeros (ZIP2 e ZIBN2), assim como o procedimento de estimação clássico e Bayesiano. Para estes modelos foram identificados diferentes pacotes (base, pscl, VGAM, mgcv, glmmTMB, GAMLSS - caso clássico e arm, JAGS, inla, MCMCglmm, glmmADMB, brms - caso Bayesiano) disponíveis em **R** (R Core Team (2024)) e, então, são desenvolvidos diferentes estudos de simulação para avaliar o desempenho deles em termos de precisão e tempo computacional.

O Capítulo 3 apresenta os modelos de regressão Bell e Poisson-Tweedie (PTw) e sua correspondente estimação clássica. Para estes modelos e desenvolvido um estudo de simulação que visa avaliar a capacidade destes modelos em comparação com o modelo Poisson e Binomial negativa.

No Capítulo 4 desenvolvemos uma aplicação em duas etapas dos modelos estudados considerando um conjunto de dados reais acerca do número de publicações de alunos de pós-graduação em bioquímica disponíveis no pacote Jackman (2024). Primeiro é desenvolvida uma avaliação do melhor modelo considerando os seis modelos estudados no Capítulo 2 considerado estimação clássica e Bayesiana. Num segundo momento considerando estimações clássica e Bayesiana para Bell e ZIBell, mas apenas clássica para PTw, é avaliado se os modelos Bell, ZIBell e PTw podem ser boas alternativas ao modelo ganhador da primeira etapa de análises.

Por fim o apêndice destina-se a apresentação dos códigos na linguagem **R** usados nos estudos de simulação e aplicação deste trabalho.

CAPÍTULO

2

MODELOS DE REGRESSÃO POISSON E BINOMIAL NEGATIVA INFLACIONADOS

2.1 Introdução - Modelos Inflacionados

Modelos de contagem compõem um tema importante na modelagem de dados, estando presentes em abundância na literatura estatística. Os modelos mais difundidos são o modelo de Regressão Poisson e o modelo de Regressão Binomial Negativa, ambos podendo ser encontrados, por exemplo, no trabalho desenvolvido por Paula (2024), o qual apresenta de forma didática esses modelos e outros pertencentes aos Modelos Lineares Generalizados (MLG's) (modelos que não possuem necessariamente variáveis respostas seguindo uma distribuição Normal), suas propriedades, a família exponencial linear, a qual esses modelos fazem parte, com algumas aplicações na linguagem de programação **R**. Outro trabalho que igualmente aborda esses modelos é o desenvolvido por Dunn e Smyth (2018).

Como comentado anteriormente, um fenômeno comum atrelado a dados de contagem é a frequência de muitos zeros em relação às outras observações. Esse fenômeno é chamado de inflacionamento de zeros. Tanto o modelo Poisson, quanto o modelo Binomial Negativa possuem dificuldade em acomodar de forma precisa esse fenômeno. Por consequência, podemos gerar subestimativas na hora da modelagem. Em outras palavras, não usar o método adequado de estimação pode resultar em estimativas ruins, à medida em que o viés das estimativas dos parâmetros será alto.

Alguns exemplos de situações nas quais encontramos o fenômeno podem ser: avaliação do número de gols marcados em função do investimento feito em determinado time, avaliação de notificação de uma determinada doença em função do IDH de uma localidade e verificação de fatores que influenciam na produção de artigos (em termos de quantidade) de alunos de pós-graduação.

Sabendo desse fenômeno comum que pode ocorrer ao modelar dados de contagem, os modelos zero-inflacionados são apresentados como alternativas mais interessantes, à medida em que são capazes de acomodar o possível excesso de zeros, visto que, possuem uma parte exclusiva para a modelagem dos zeros e outra para modelagem dos demais valores observados na variável resposta. Yang *et al.* (2017) em seu trabalho apresentam uma comparação entre as variantes zero-inflacionadas dos modelos Poisson e Binomial Negativa, os modelos Zero-inflacionado Poisson (ZIP) e Zero-inflacionado Binomial Negativa (ZIBN). Tanto nos estudos de simulação quanto na aplicação proposta pelos autores há o uso do Critério de Informação de Akaike (AIC) aliado ao teste de Vuong (vide Vuong (1989) e Castellares, Ferrari e Lemonte (2018)), ambos utilizados para seleção de modelos. O teste de Vuong pode ser descrito como uma generalização do teste de razão de verossimilhanças, no sentido de medir a distância entre os ajustes por meio do critério de informação de Kullback-Leibler. Como no teste de razão de verossimilhanças, consideramos as seguintes hipóteses - H0: a favor do modelo 1 e H1: a favor do modelo 2, ou seja, quando a razão possui o valor do denominador superior ao valor do numerador, temos evidências de que o modelo 2 realiza melhor ajuste que o modelo 1 e vice-versa. Como resultado, os autores constataram que os modelos ZIP e ZIBN apresentaram vantagem quando comparados aos demais modelos, pois garantem melhores estimativas sem o confundimento que a presença de zeros pode gerar nos modelos Poisson e Binomial Negativa.

Este capítulo tem por objetivo apresentar diferentes estudos de simulação e uma aplicação a dados reais para comparação dos modelos Poisson, Binomial Negativa, ZIP e ZIBN e metodologias de estimação (clássica e Bayesiana). Desse modo, o capítulo está organizado da seguinte forma, na Seção 2.2 introduzimos as distribuições de contagem, assim como os modelos de regressão para cada distribuição e as alternativas para tratativa dos casos inflacionados de zero. Há também uma explanação dos processos de estimação, abordando os processos metodológicos clássico e bayesiano. A Seção 2.4 apresenta os estudos de simulação, a partir dos pacotes presentes na Tabela 1. Por fim, a Seção 2.5 apresenta a conclusão para o capítulo.

2.2 Revisão de Conceitos - Modelos Inflacionados

Esta Seção destina-se a apresentar uma contextualização dos modelos de contagem estudados neste Capítulo.

2.2.1 Distribuições de Contagem Inflacionadas

Uma variável aleatória Y segue a distribuição de Poisson e escrevemos $Y \sim Poisson(\mu)$ se sua função de distribuição é dada por

$$P_Y(y;\mu) = \frac{exp\{-\mu\}\,\mu^y}{y!}, \quad y = 0,1,2,...,$$

com $E(Y) = \mu$ e $V(Y) = \mu$, $\mu > 0$.

Ao impor a equidispersão para os dados (valor da média igual ao da variância), essa distribuição se torna restritiva e sendo assim, comportamentos como subdispersão (variância menor que a média) de dados ou sobredispersão (variância maior do que a média) de dados não são bem acomodados, de modo a resultar em uma modelagem inadequada. Em contra partida, um ponto positivo é a simplicidade desta distribuição, dado que não possui expressões complicadas facilitando o entendimento.

Uma alternativa a Poisson é a distribuição Binomial Negativa. Uma variável aleatória Y segue a distribuição Binomial Negativa e escrevemos $Y \sim BN(\mu, \phi)$ se sua função de distribuição é dada por

$$P_Y(y;\mu,\phi) = \frac{\Gamma(\phi+y)}{y!\Gamma(\phi)}\Big(\frac{\mu}{\mu+\phi}\Big)^y\Big(\frac{\phi}{\mu+\phi}\Big)^\phi, \;\; y = 0,1,2,...$$

com $E(Y) = \mu$ e $V(Y) = \mu(1+\frac{\mu}{\phi})$, $\mu > 0$, $\phi > 0$.

Essa distribuição, diferentemente da Poisson, não impõe a equidispersão para os dados, de modo a ser mais aderente e igualmente mais flexível para dados em que a média não necessariamente é igual a variância.

Temos também as variantes zero-inflacionadas das duas distribuições anteriores, que visam uma aderência maior aos dados que apresentam o fenômeno de inflação de zeros, como mencionado anteriormente.

Seja $Y = 0$ com probabilidade ω e $Y \sim Poisson(\mu)$ com probabilidade $(1-\omega)$, então dizemos que Y segue uma distribuição Zero-inflacionada Poisson, denotada por $ZIP(\mu,\omega)$, se sua função de distribuição é dada por

$$P_Y(y;\mu,\omega) = \begin{cases} \omega + (1-\omega)exp\{-\mu\}, & y = 0 \\ (1-\omega)\dfrac{\mu^y exp\{-\mu\}}{y!}, & y = 1,2,..., \end{cases}$$

com $E(Y) = (1-\omega)\mu$ e $V(Y) = \mu(1-\omega)(1+\mu\omega)$, $\mu > 0$, $0 < \omega < 1$.

É importante ressaltar que essa variante garante a Poisson maior flexibilidade ao quebrar a imposição de equidispersão nos dados. Ou seja, é possível a acomodação de dados sobredispersos por essa distribuição.

Agora, seja $Y = 0$ com probabilidade ω e $Y \sim BN(\mu,\phi)$ com probabilidade $(1-\omega)$, então dizemos que Y segue uma distribuição Zero-inflacionada Binomial Negativa - $ZIBN(\mu,\phi,\omega)$. Uma das parametrizações desta distribuição é a que segue.

$$P_Y(\boldsymbol{y};\mu,\omega,\phi) = \begin{cases} \omega + (1-\omega)\left(\dfrac{\phi}{\phi+\mu}\right)^{\phi}, & y = 0 \\ (1-\omega)\dfrac{\Gamma(y+\phi)}{y!\Gamma(\phi)}\left(\dfrac{\mu}{\phi+\mu}\right)^{y}\left(\dfrac{\phi}{\phi+\mu}\right)^{\phi}, & y = 1,2,..., \end{cases}$$

com $E(Y) = (1-\omega)\mu$ e $V(Y) = (1-\omega)[1+(\phi+\omega)\mu]$, $\mu > 0$, $0 < \omega < 1$, $\phi > 0$, respectivamente.

2.2.2 Modelos de Regressão de Contagem Inflacionados

Para a utilização de MLG's, a variável resposta deve estar distribuída a partir da família exponencial linear. Considerando uma variável Y qualquer, podemos dizer que ela pertence à família exponencial se sua função de probabilidade possa ser escrita como

$$f(y) = exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\} \tag{2.1}$$

sendo ϕ o parâmetro de dispersão. A escolha de $b(\theta)$ determina qual distribuição a variável resposta segue.

Podemos ainda definir a média e a variância como sendo, respectivamente

$$E(Y) = b'(\theta) \;\; e \;\; V(Y) = \phi b''(\theta)$$

No caso do modelo Poisson, a título de exemplo, podemos escrever a distribuição como sendo parte da família exponencial escrevendo

$$f(y) = exp\left(\left(\frac{y\log(\mu) - \mu}{1}\right) + \log(y!)\right)$$

Rearranjando os termos de acordo com a equação 2.1, temos que para Poisson

$$\theta = \log(\mu); \;\; b(\theta) = \mu = exp\{\theta\}; \;\; \phi = 1; \;\; c(y,\theta) = -\log(y!)$$

Para mostrar como Binomial Negativa é parte da família exponencial, assim como as versões inflacionadas ver Paula (2024). Tanto para o modelo Poisson e Binomial Negativa, quanto para suas variantes zero-inflacionadas temos que a função de ligação canônica é dada por $\log(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$. Ou, de mesmo modo, $exp(\eta_i) = \mu_i$. Vale ressaltar que log(), aqui, é um logaritmo neperiano. Para os estudos de simulação e aplicação, essa foi a função de ligação considerada.

A seguir iremos propor os modelos de regressão para Poisson e Binomial Negativa considerando:

- $Y_1, \ldots, Y_n$ v.a independentes tais que $Y_i \sim Poisson(\mu_i)$ ou $BN(\mu_i, \phi)$;
- $g(\mu_i) = \log(\mu_i) = \eta_i$, em que $\log(\mu_i)$ é a função de ligação canônica;
- $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_1 + \beta_2 x_{i2} + ... + \beta_p x_{ip}$, i = 1,...,n é o preditor linear, em que $x_i^T = (x_{i1}, \cdots, x_{ip})^T$ é um vetor de covariáveis, sendo x_{i1} o intercepto.

Para propor modelos de regressão inflacionados devemos considerará-los como modelos de mistura, ou seja são a junção de um modelo de contagem com uma Regressão Logística, que atua como um preditor para dizer a qual grupo associar uma observação.

Consideremos, assim, que a parte não inflacionada segue um modelo Poisson ($Z_1 \sim Pois(\lambda_i)$), ou Binomial Negativo com $Z_1 \sim BN(\lambda_i, \phi)$ em que $\lambda_i = exp\left\{\boldsymbol{x}_i^T \boldsymbol{\beta}\right\}$. O exemplo a seguir, se baseia em $Z_1 \sim BN(\lambda_i, \phi)$. Para a parte de zeros consideremos um modelo Logístico para explicar a probabilidade de termos zero como nosso evento de sucesso, assim: $Z_2 \sim Bernoulli(\pi_i)$ e covariáveis $\boldsymbol{u}_i$ (que podem ser as mesmas utilizadas para a parte não inflacionada). Então, $\log(\frac{\pi_i}{1-\pi_i}) = \boldsymbol{u}_i^T \boldsymbol{\delta}$, e, portanto, obtemos a média do modelo Y_i, de forma que:

$$E(Y_i) = \mu_i = (1 - \pi_i)\lambda_i = \frac{exp\left\{\boldsymbol{x}_i^T \boldsymbol{\beta}\right\}}{1 + exp\left\{\boldsymbol{u}_i^T \boldsymbol{\delta}\right\}}, \tag{2.2}$$

e então, usando a ligação $\log()$, temos

$$\log(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} - \log(1 + exp\left\{\boldsymbol{u}_i^T \boldsymbol{\delta}\right\}).$$

Em suma, podemos dizer que ambos os modelos inflacionados seguem a seguinte estrutura:

- $Y_1, \ldots, Y_n$ v.a independentes tais que $Y_i \sim ZIP(\mu_i, \omega_i)$ ou $ZIBN(\mu_i, \phi, \omega_i)$;
- $g_1(\mu_i) = \log(\mu_i) = \eta_{1i}$ em que $\log(\mu_i)$ é uma função de ligação canônica;
- $\eta_{1i} = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_1 + \beta_2 x_{i2} + ... + \beta_p x_{ip}$, i = 1,...,n é o preditor linear para a taxa de resposta, em que em que $x_i^T = (x_{i1}, \cdots, x_{ip})^T$ é um vetor de covariáveis, sendo x_{i1} o intercepto;
- $g_2(\omega_i) = log(\omega_i / 1 - \omega_i) = \eta_{2i}$ em que $log(\omega_i / 1 - \omega_i)$ é uma função de ligação canônica;
- $\eta_{2i} = \boldsymbol{u}_i^T \boldsymbol{\delta} = \delta_1 + \delta_2 u_{i2} + ... + \delta_q u_{iq}$, i = 1,...,n é o preditor linear da média ou proporção de zeros, em que $u_i^T = (u_{i1}, \cdots, u_{iq})^T$ é um vetor de covariáveis, sendo u_{i1} o intercepto.

Ao consideramos as mesmas covariáveis na parte de contagem e na parte de zeros, fazemos $\boldsymbol{u}_i^T = \boldsymbol{x}_i^T$ e $q = p$ (caso considerado neste trabalho).

2.3 Estimação

Nesta Seção apresentamos os métodos de estimação utilizados neste trabalho, além de um panorama computacional dos pacotes utilizados nos estudos de simulação e aplicação deste Capítulo.

2.3.1 Método de Máxima Verossimilhança

Dentre diversos trabalhos que abordam o método de Máxima Verossimilhança, podemos citar o desenvolvido por Silvey (1975). Para definirmos o estimador de Máxima Verossimilhança, consideremos uma amostra aleatória $y_1, ..., y_n$ da distribuição de uma variável aleatória Y com função densidade $f(\boldsymbol{y};\boldsymbol{\theta})$, na qual θ é um escalar ou vetor dos parâmetros populacionais desconhecidos, caso expresso a seguir. Denominamos a função $L(\boldsymbol{\theta};\boldsymbol{y})$ definida por

$$L(\boldsymbol{\theta};\boldsymbol{y}) = \prod_{i=1}^{n} L(\boldsymbol{\theta};y_i) = \prod_{i=1}^{n} f(y_i;\theta), \quad \boldsymbol{\theta} \in \Theta \tag{2.3}$$

como Função de Verossimilhança. Aqui, os valores amostrais são conhecidos e, portanto, estão fixos. Θ é o espaço paramétrico, ou seja, é o conjunto de valores que o parâmetro pode assumir. Na regressão Poisson $\boldsymbol{\theta} = \boldsymbol{\beta}$ e na Binomial Negativa $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\delta}, \phi)$, com $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ e $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_q)$ Para efeitos de otimização é interessante utilizar o logaritmo neperiano da função expressa em (2.3). Desse modo, obtemos, que o estimador de Máxima Verossimilhança é dado por

$$l(\boldsymbol{\theta};\boldsymbol{y}) = log\ \ L(\boldsymbol{\theta};\boldsymbol{y}) = log \prod_{i=1}^{n} f(y_i;\boldsymbol{\theta}) = \sum_{i=1}^{n} log\ \ f(y_i;\boldsymbol{\theta}). \tag{2.4}$$

Muitas das vezes não é possível encontrar expressões relativamente simples, com forma fechada para as estimativas de Máxima Verossimilhança. De todo modo, usualmente é possível assumir que tais estimativas surgem como solução de equações de Verossimilhança, ou função escore, dadas por

$$\mathscr{U}(\boldsymbol{\theta}) = \left(\frac{\partial l(\boldsymbol{\theta})^T}{\partial \theta_1}, \ldots, \frac{\partial l(\boldsymbol{\theta})^T}{\partial \theta_p} \right)^T = \mathbf{0}. \tag{2.5}$$

Esse sistema de equações não lineares (2.5) pode ser resolvido de forma numérica. A entrada (i, j) da matriz $(p \times p)$ de informação de Fisher ($\mathscr{F}_{\boldsymbol{\theta}}$) para o vetor de parâmetros θ é dada pela seguinte expressão

$$\mathscr{F}_{\theta ij} = -E\left\{ \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right\}. \tag{2.6}$$

Tem-se que para resolver o sistema de equações $\mathscr{U}(\boldsymbol{\theta}) = \mathbf{0}$, utilizamos o algoritmo Newton-Raphson, encontrado de maneira abundante na literatura, podendo mencionar dentre tantos trabalhos, o desenvolvido por Ben-Israel (1966).

Para melhor elucidação dos modelos apresentados anteriormente, abaixo expressamos cada $l(\boldsymbol{\theta};\boldsymbol{y})$. Assim:

Modelo de regressão Poisson

Para o modelo Poisson, temos $\boldsymbol{\theta} = \boldsymbol{\beta}$. Nesse sentido:

$$L(\boldsymbol{\theta};\boldsymbol{y}) = \prod_{y_i=1}^{n} \frac{\mu_i^{y_i} \exp -\mu_i}{y_i!}$$

aplicando o $\log(\cdot)$ chegamos a

$$l(\boldsymbol{\theta};\boldsymbol{y}) = \log \mu_i \left(\sum_{i=1}^{n} y_i \right) - n\mu_i$$

Modelo de regressão Binomial Negativa

Agora para o modelo Binomial Negativa, temos $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)$. Assim:

$$L(\boldsymbol{\theta};\boldsymbol{y}) = \prod_{y_i=1}^{n} \frac{\Gamma(\phi + y)}{y!\Gamma(\phi)} \left(\frac{\mu_i}{\mu_i + \phi} \right)^{y} \left(\frac{\phi}{\mu_i + \phi} \right)^{\phi}$$

podemos escrever, assim, $\log(\cdot)$ como

$$l(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{i=1}^{n} \left[\log\Gamma(y_i + \phi) - \log\Gamma(\phi) - \log y_i! \right] + n\phi \log \phi - n\phi \log \phi - n\phi log(\phi + \mu_i) + \sum_{i=1}^{n} y_i$$

Modelo de regressão Zero-inflacionada Poisson

Nesse caso, temos que considerar nosso vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\delta})$. De modo que:

$$L(\boldsymbol{\theta};\boldsymbol{y}) = \prod_{y_i=0} \left(\omega_i + (1-\omega_i) exp\left\{-\mu_i\right\} \right) \prod_{y_i>0} \left((1-\omega_i) \frac{\mu_i^{y_i exp\{-\mu_i\}}}{y_i!} \right)$$

ao aplicar o $\log(\cdot)$ ficamos com

$$l(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{y_i=0} log\left(\omega_i + (1-\omega_i) exp\left\{-\mu_i\right\} \right) + \sum_{y_i>0} \left[\log(1-\omega_i) + log \left(\frac{\mu_i^{y_i exp\{-\mu_i\}}}{y_i!} \right) \right]$$

Modelo de regressão Zero-inflacionada Binomial Negativa

Para esse modelo, temos que considerar o seguinte vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi, \boldsymbol{\delta})$. De modo que:

$$L(\boldsymbol{\theta};\boldsymbol{y}) = \prod_{y_i=0} \omega_i + (1+\omega_i)\left(\frac{\phi}{\phi+\mu_i}\right)^{\phi} \prod_{y_i>0} (1-\omega_i)\frac{\Gamma(y_i+\phi)}{y_i!\Gamma(\phi)}\left(\frac{\phi}{\phi+\mu_i}\right)^{\phi}\left(\frac{\mu_i}{\phi+\mu_i}\right)^{y_i} \quad (2.7)$$

$\log(\cdot)$ pode ser escrito como

$$l(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{y_i=0} log\left(\omega_i + (1+\omega_i)\left(\frac{\phi}{\phi+\mu_i}\right)^{\phi}\right) + \sum_{y_i>0}\left(\log(1-\omega_i) + log\left(\frac{\Gamma(y_i+\phi)}{y_i!\Gamma(\phi)}\right) + \phi log\left(\frac{\phi}{\phi+\mu_i}\right) + y_i log\left(\frac{\mu_i}{\phi+\mu_i}\right)\right)$$

ressaltando que $\log(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ e $\log\left(\frac{\omega_i}{1-\omega_i}\right) = \boldsymbol{u}_i^T\boldsymbol{\delta}$.

Dentre os pacotes que utilizamos para as estimações desses modelos sob ótica clássica podemos citar base - R Core Team (2024), pscl - Jackman (2024), VGAM - Yee (2024), mgcv - Wood (2003), glmmTMB - Brooks *et al.* (2017) e GAMLSS - R. A. Rigby e D. M. Stasinopoulos (2005).

2.3.2 Método de Estimação Bayesiano

Para realizarmos a estimação sob ótica Bayesiana, além dos dados amostrais, precisamos de uma informação à *priori* sobre o(s) parâmetro(s) e o cálculo da distribuição à *posteriori* do(s) parâmetro(s). A informação à *priori* é dada pela densidade de probabilidade $p(\theta)$, que expressa o conhecimento do pesquisador sobre o(s) parâmetro(s) a ser(em) estimado(s). Quando não há conhecimento inicial suficiente para ser levado em conta, chamamos nossa *priori* de não-informativa. Um exemplo de priori não informativa é a priori de Jeffreys (Jeffreys, Kneale e David (1949)).

Consideremos o mesmo cenário do Método de Máxima Verossimilhança com uma variável aleatória Y, então, podemos considerar a expressão (2.3) como sendo equivalente a $L(yi|\theta)$. Desse modo, temos que pelo teorema de Bayes chegamos a seguinte expressão

$$p(\theta|y) = \frac{L(\boldsymbol{y}|\theta)p(\theta)}{\int L(\boldsymbol{y}|\theta)p(\theta)d\theta} \quad (2.8)$$

nossa distribuição à *posteriori*. O denominador é uma constante de integração, dado que só depende da amostra de dados. Assim, podemos reduzir a expressão 2.8 a apenas

$$p(\theta|\boldsymbol{y}) \propto p(\theta)p(\boldsymbol{y}|\theta), \quad (2.9)$$

em que $p(\boldsymbol{y}|\theta) = L(\theta|\boldsymbol{y})$.

Para melhor elucidação dos métodos, a seguir apresentamos um exemplo para o modelo ZIBN.

No final da Subseção 2.2.2 apresentamos uma breve explicação sobre o funcionamento dos modelos inflacionados. Partindo do fato de que neste trabalho consideramos as mesmas covariáveis para explicar os zeros observados e demais valores de Y, com $\boldsymbol{y} = (y_1, \cdots, y_N)^T$, podemos expressar μ como na expressão (2.2) e ω da seguinte forma:

$$\omega_i = \frac{exp\left\{\boldsymbol{x}_i^T \boldsymbol{\delta}\right\}}{1 + exp\left\{\boldsymbol{x}_i^T \boldsymbol{\delta}\right\}} \tag{2.10}$$

nessa expressão, temos que $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_q)$ expressa o(s) coeficiente(s) associado(s) à parte de zeros do modelo. Já $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ em (2.2) o(s) coeficiente(s) associado(s) à parte de contagem do modelo. Tomemos como exemplo o modelo ZIBN, assim, o primeiro passo para aplicação do método Bayesiano é a descrição da Verossimilhança do modelo. Baseando-se em (2.2) e (2.10), temos que:

$$\mathrm{p}(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{yi=0} P_Y(y_i|\boldsymbol{\theta}) \;\; (I) \times \prod_{y_i>0} P_Y(y_i|\boldsymbol{\theta}) \;\; (II) \tag{2.11}$$

em que (I) pode ser expresso como

$$\prod_{y_i=0} \left[\frac{\exp\left\{x_i^T \delta_i\right\}}{1+\exp\left\{x_i^T \delta_i\right\}} + \left(\frac{1}{1+\exp\left\{x_i^T \delta_i\right\}} \right) + \left(\frac{\phi\left(1+\exp\left\{x_i^T \delta_i\right\}\right)}{\phi_i\left(1+\exp\left\{x_i^T \delta_i\right\}\right)+\exp\left\{x_i^T \beta_i\right\}} \right)^{\phi} \right] \tag{2.12}$$

e (II) é dado por

$$\prod_{y_i>0} \left[\left(\frac{1}{1+\exp\left\{x_i^T \delta_i\right\}} \right) \frac{\Gamma(y_i+\phi)}{y_i!\Gamma(\phi)} \left(\frac{\exp\left\{x_i^T \beta_i\right\}}{\phi[1+\exp\left\{x_i^T \delta_i\right\}]+\exp\left\{x_i^T \beta_i\right\}} \right)^{y_i} \left(\frac{\phi\left(1+\exp\left\{x_i^T \delta_i\right\}\right)}{\phi\left(1+\exp\left\{x_i^T \delta_i\right\}\right)+\exp\left\{x_i^T \beta_i\right\}} \right)^{\phi} \right] \tag{2.13}$$

Distribuição a Priori para os Parâmetros

Quando não temos informações históricas com relação aos dados ou experimentos prévios, devemos utilizar prioris não informativas para os parâmetros, de modo a não trazer nenhum viés da escolha do pesquisador às distribuições posterioris dos parâmetros. Não é

diferente no nosso caso, de modo que apresentaremos uma priori conjugada não informativa. Desse modo, temos que a priori conjunta para o modelo ZIBN é, tal que

$$\begin{aligned} p(\boldsymbol{\theta}) = &\prod_{i=1}^{p}\left[\frac{1}{\sigma_{\beta_i}\sqrt{2\pi}}\exp\left\{\frac{-(\beta_i-\mu_{\beta_i})^2}{2\sigma_{\beta_i}^2}\right\}\right]\times \\ &\prod_{j=1}^{q}\left[\frac{1}{\sigma_{\delta_j}\sqrt{2\pi}}\exp\left\{\frac{-(\delta_j-\mu_{\delta_j})^2}{2\sigma_{\delta_j}^2}\right\}\right]\times \\ &\frac{1}{b^a\Gamma(a)}\phi^{a-1}\exp\{-\phi/b\}, \end{aligned} \tag{2.14}$$

em que consideramos $\beta_i \sim N(\mu_{\beta_i}, \sigma^2_{\beta_i})$, $\delta_j \sim N(\mu_{\delta_i}, \sigma^2_{\delta_i})$ e $\phi \sim Gamma(a,b)$. Podemos indicar as especificações das prioris como $\beta_i \sim N(0,1000)$, $\delta_j \sim N(0,1000)$ e $\phi \sim Gamma(0{,}001, 0{,}001)$, por exemplo, ressaltando que como são independentes entre si, para obtermos a priori conjunta de todos os parâmetros, basta multiplicá-las. Garantindo, assim, que não tenhamos interferências de suposições anteriores nos dados observados.

Distribuição a Posteriori do vetor de parâmetros $\boldsymbol{\theta}$

Agora, combinando a priori com a Verossimilhança, temos como resultante a posteriori para os parâmetros $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ e ϕ, como expresso em (2.9), sendo a que segue.

$$\begin{aligned} &\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{p}\left[\frac{1}{\sqrt{2\pi(1000)}}exp\left\{-\frac{\beta_i^2}{2(1000)}\right\}\right]\times\prod_{j=1}^{q}\left[\frac{1}{\sqrt{2\pi(1000)}}exp\left\{-\frac{\delta_i^2}{2(1000)}\right\}\right] \\ &\times\frac{1}{0{,}001^{0{,}001}\Gamma(0{,}001)}\phi^{-0{,}999}exp\{-\phi/0{,}001\} \\ &\times\prod_{y_i=0}\left[\frac{\exp\{x_i^T\delta_i\}}{1+\exp\{x_i^T\delta_i\}}+\left(\frac{1}{1+\exp\{x_i^T\delta_i\}}\right)+\left(\frac{\phi\left(1+\exp\{x_i^T\delta_i\}\right)}{\phi_i\left(1+\exp\{x_i^T\delta_i\}\right)+\exp\{x_i^T\beta_i\}}\right)^{\phi}\right] \\ &\times\prod_{y_i>0}\left[\left(\frac{1}{1+\exp\{x_i^T\delta_i\}}\right)\frac{\Gamma(y_i+\phi)}{y_i!\Gamma(\phi)}\left(\frac{\exp\{x_i^T\beta_i\}}{\phi[1+\exp\{x_i^T\delta_i\}]+\exp\{x_i^T\beta_i\}}\right)^{y_i}\left(\frac{\phi\left(1+\exp\{x_i^T\delta_i\}\right)}{\phi\left(1+\exp\{x_i^T\delta_i\}\right)+\exp\{x_i^T\beta_i\}}\right)^{\phi}\right]. \end{aligned} \tag{2.15}$$

Ao obtermos a distribuição a posteriori, normalmente não é possível resolvê-la analiticamente - é o caso aqui. Isso se deve pela dificuldade que a expressão pode evidenciar. Contudo, um método numérico de simulação via cadeias de Markov Monte Carlo-Gibbs sampling pode ser utilizado para a atualização dos valores iniciais conhecidos dos parâmetros de interesse, e igualmente, as amostras dos parâmetros conhecidos acabam por convergentes, assim como expressa o estudo de Shafira, Abdullah e Lestari (2020).

Dentre os pacotes que utilizamos para as estimações segundo a ótica Bayesiana podemos citar arm - Gelman e Su (2024) - Funções Bayesianas para modelagem linear generalizada com distribuição a priori independente normal, t ou Cauchy para os coeficientes; JAGS - Plummer (2023) - Usado para criar um objeto que representa um modelo gráfico Bayesiano, especificado com uma descrição da distribuição a priori na linguagem BUGS e um conjunto de dados; inla - Lindgren e Rue (2015) - É um método para inferência Bayesiana aproximada. Nos últimos anos, estabeleceu-se como uma alternativa a outros métodos, como o Monte Carlo via Cadeias de Markov (MCMC), devido à sua velocidade e facilidade de uso; MCMCglmm - Hadfield (2010) - Amostrador de Monte Carlo via Cadeias de Markov para Modelos Lineares Generalizados Mistos Multivariados, com ênfase especial em efeitos aleatórios correlacionados decorrentes de genealogias e filogenias; glmmADMB - Skaug *et al.* (2016) - Construído com base no motor de ajuste não linear de código aberto AD Model Builder, para ajuste de modelos lineares mistos generalizados e suas extensões; brms - Bürkner (2021) - Ajusta modelos Bayesianos multivariados multinível generalizados (não-)lineares usando Stan para inferência Bayesiana completa.

2.3.3 Implementação Computacional dos Modelos Inflacionados

Tabela 1 – Pacotes para regressão de contagem utilizados no Capítulo.

Pacotes - R	
clássicos	
base	https://stat.ethz.ch/R-manual/R-devel/library/base/html/base-package.html
pscl	https://cran.r-project.org/web/packages/pscl/index.html
VGAM	https://cran.r-project.org/web/packages/VGAM/index.html
mgcv	https://cran.r-project.org/web/packages/mgcv/index.html
glmmTMB	https://cran.r-project.org/web/packages/glmmTMB/index.html
GAMLSS	https://cran.r-project.org/web/packages/gamlss/index.html
Bayesianos	
arm	https://cran.r-project.org/web/packages/arm/index.html
JAGS	https://cran.r-project.org/web/packages/rjags/index.html
inla	https://www.r-inla.org
MCMCglmm	https://cran.r-project.org/web/packages/MCMCglmm/index.html
glmmADMB	https://glmmadmb.r-forge.r-project.org
brms	https://cran.r-project.org/web/packages/brms/index.html

Como primeiro passo para realizarmos os ajustes dos modelos listamos vários pacotes do software **R** que podem ser utilizados para as estimações dos parâmetros, dentre os estudados neste Capítulo podemos listar os pacotes presentes na Tabela 1.

Sendo 6 pacotes que estimam segundo abordagem clássica e outros 6 para a abordagem Bayesiana. Cabe ressaltar que conseguimos aplicar os pacotes para ajustes inflacionados, porém, não foi possível em alguns casos. Um ponto a salientar é que para os modelos inflacionados

há pacotes em que é possível associar parâmetros à parte de zeros (ao invés de somente a probabilidade de zeros), assim como à parte de contagem, média. Separamos, então, esses modelos em dois casos cada. O sufixo '1', indica os casos em que somente a probabilidade de zeros pode ser associada à parte de zeros do modelo. Já o sufixo '2', indica que ambas as partes do modelo podem ter coeficientes regressivos associados às covariáveis aplicadas em cada parte. Dentre os pacotes testados e implementações bem-sucedidas por parte do autor, temos os pacotes expressos na Tabela 2.

Tabela 2 – Aplicabilidade dos pacotes para os modelos estudados neste Capítulo.

	Pacotes	Modelos					
		Poisson	Binomial Negativa	ZIP1	ZIP2	ZIBN1	ZIBN2
clássicos	base	X	X				
	pscl				X		X
	VGAM			X	X		
	mgcv	X	X	X			
	glmmTMB	**X**	**X**	**X**	**X**	**X**	**X**
	GAMLSS	**X**	**X**	**X**	**X**	**X**	**X**
Bayesianos	arm	X					
	JAGS	**X**	**X**	**X**	**X**	**X**	**X**
	inla	X	X	X		X	
	MCMCglmm	X		X	X		
	glmmADMB	X	X	X		X	
	brms	**X**	**X**	**X**	**X**	**X**	**X**

Como nota adicional, como foi comentado por um membro da banca, inla seria aplicável a todos os modelos, porém um dos desafios encontrados neste trabalho foi justamente a implementação dos modelos inflacionados e suas respectivas estimações. Ressaltasse então, que as simulações e aplicações realizadas neste trabalho foram feitas com base nas implementações já disponíveis nos pacotes expressos na Tabela 1, não adentrando em implementação própria por parte do autor. Nesse sentido, não foi possível aplicar o pacote inla para ZIP2 e ZIBN2. Adicionalmente, uma das métricas coletada nos estudos de simulação do Capítulo foi o tempo computacional, dessa forma, cabe ressaltar que a máquina na qual as simulações e aplicação foram rodadas tem como configuração um i5 de décima geração e 8gb de RAM.

2.4 Estudos de Simulação - Modelos Inflacionados

Esta Seção destinasse a apresentar três estudos de simulação realizados com os modelos Poisson, Binomial Negativa, ZIP e ZIBN, em termos de performance de estimação com mensuração de diferentes métricas comparativas e desempenho em termos de tempo demandado para processamento.

2.4.1 Estudo 1 - Avaliação de pacotes

Partindo da Tabela 1, que contém os pacotes utilizados para a realização dos ajustes dos modelos na linguagem **R**, desenvolvemos três estudos de simulação para verificar a qualidade dos ajustes realizados e respectivos tempos computacionais em segundos.

Sendo assim, para este primeiro estudo consideramos uma amostra de tamanho 500, que foi gerada para cada uma das distribuições estudadas. Temos ao todo um total de 6 amostras, sendo que são uma para Poisson, uma para Binomial Negativa, duas para ZIP e mais duas para ZIBN. Compondo os conjuntos de dados, definimos duas covariáveis independentes, aleatoriamente, utilizando-se de uma Binomial ($B(500,0,5)$) e uma Uniforme ($U_{[-1,1]}$) para ajustes não inflacionados e Normais padrão ($N(0,1)$), para os ajustes inflacionados.

O vetor de médias foi definido por $\boldsymbol{\mu} = \exp\{\beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2\}$, sendo utilizado para gerar os valores y da variável resposta Y. Foram considerados os seguintes valores de parâmetros:

$$\beta_0 = 0,5,\ \beta_1 = 1,0 \text{ e } \beta_2 = -1,0.$$

Agora, para os ajustes inflacionados, Zero-inflacionado Poisson (ZIP) e Zero-inflacionado Binomial Negativa (ZIBN), há o uso dos mesmos $\boldsymbol{\beta}$'s para a parte de contagem do modelo (parte não inflacionada). Antes de prosseguir, há uma particularidade a ressaltar dos ajustes zero-inflacionados. Ao observar o comportamento de estimação dos pacotes, podemos dividi-lo em dois, como já comentado anteriormente. A primeira divisão se dá pelos ajustes em que é possível a estimação apenas da probabilidade de zeros, não sendo possível a estimativa de coeficientes de regressão para a parte inflacionada. Um exemplo é o pacote mgcv. Já a segunda divisão se dá justamente pelos casos em que isso é possível. Com isso, passamos a considerar a divisão entre ZIP1 e ZIP2 e, igualmente, ZIBN1 e ZIBN2. Nesse sentido, para apenas uma parte, foram considerados os seguintes valores de parâmetros: ZIP1 - $\pi = 0,4$ e ZIBN1 - $\pi = 0,2$. Agora, nos modelos ZIP2 e ZIBN2 para a parte de zeros do modelo fixaram-se os seguintes valores de parâmetros:

$$\delta_0 = 0,4,\ \delta_1 = -0,5 \text{ e } \delta_2 = 0,9,$$

e

$$\delta_0 = -4,0,\ \delta_1 = 0,5 \text{ e } \delta_2 = -3,5,$$

respectivamente. A probabilidade de zeros para esses casos foi obtida a partir da função *plogis* nas combinações lineares de $\boldsymbol{\delta}$'s. Como levamos em conta as mesmas covariáveis para a parte inflacionada dos modelos, então temos um cenário equivalente ao realizado para formulação do vetor de médias. Assim, temos π_{ZIP} e $\pi_{ZIBN} = \exp\{\delta_0 + \delta_1 \times x_1 + \delta_2 \times x_2\}$. Cabe ressaltar que a escolha dos parâmetros garante que tenhamos muitos zeros observados, justamente o comportamento que gostaríamos de investigar. Nesse sentido, o gráfico 1, exemplifica como fica a distribuição ZIBN2 para os parâmetros escolhidos.

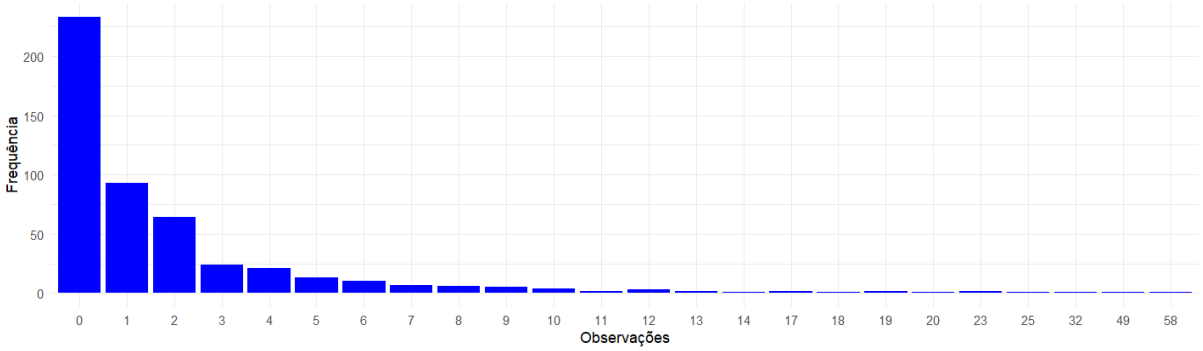

Figura 1 – Variável resposta segundo ZIBN2 simulada a partir dos parâmetros fixados.

Como medida de afastamento foi utilizada a seguinte expressão (2.4.1), considerando a raiz da soma da diferença ao quadrado entre as estimativas coletadas e o valor real (fixado) dos parâmetros ao quadrado sobre p (quantidade de parâmetros, 3 no nosso caso). Medida utilizada em conjunto com as demais para mensurar a precisão dos ajustes realizados.

$$\frac{\sqrt{\Sigma_{j=1}^{p}(\hat{\beta}_j - \beta_j)^2}}{p}. \tag{2.16}$$

Sendo assim, foram computadas as seguintes quantidades resultantes de cada um dos ajustes: a estimativa média dos parâmetros, erro padrão das estimativas, o afastamento 2.4.1 para os coeficientes de regressão e probabilidades de zero, tempo de execução (em segundos), critério de seleção de modelos, como AIC (Akaike (1974)) e a versão Bayesiana desse critério, EAIC (presente em Spiegelhalter *et al.* (2002) e Anyosa (2017)), que pode ser entendida como uma extensão do critério proposto por Akaike em seu trabalho. Podemos então descrevê-lo da seguinte forma, como apresentado no trabalho de Spiegelhalter *et al.* (2002):

$$EAIC = D(\overline{\boldsymbol{\theta}}) + 2p$$

em que p o número efetivo de parâmetros estimados e $D(\overline{\boldsymbol{\theta}})$ expressa a desviância aplicada na média à posteriori dos parâmetros.

O objetivo deste primeiro estudo foi avaliar quais pacotes seriam mais indicados para cada metodologia de estimação e baseado nessa escolha seguir para uma análise mais aprofundada. As Tabelas 3, 4, 5, 6, 7, 8 apresentam os resultados obtidos para cada modelo de contagem.

Notamos similaridade entre os ajustes, independentemente da distribuição simulada. Ressaltamos aqui que o afastamento das estimativas é baixo e próximo entre os ajustes, não sendo critério para decidir a favor de um ou de outro. Cabe somente pontuar que o pacote JAGS, para casos inflacionados apresenta afastamentos maiores. Considerando o cenário clássico, o pacote base do **R** de maneira geral é capaz de realizar bons ajustes, comparando-o com o pacote GAMLSS, porém há limitações de seu uso para modelos inflacionados. Caso o estudo conduzido

utilize apenas Poisson ou Binomial Negativa, ele é sim uma boa opção para a estimação de coeficientes. O pacote glmmTMB é igualmente similar ao GAMLSS, sendo igualmente aplicável a todos os modelos estudados neste Capítulo.

Com relação aos pacotes que realizam os ajustes segundo ótica Bayesiana, como mencionado anteriormente a respeito dos afastamentos, entendemos que JAGS, apesar de ter bons resultados para a parte não inflacionada do modelo, apresenta certa dificuldade em estimar os coeficientes da parte inflacionada. Quando lidamos apenas com a probabilidade de zeros o cenário é parelho aos demais pacotes. Além disso, temos que na maioria dos casos, esse pacote foi o mais demorado em termos de execução, o que é um ponto a se levar em conta na hora da escolha de um pacote para realização das estimações. JAGS é aplicável a todos os modelos aqui estudados, entretanto entendemos que o pacote brms seria mais interessante, tanto em relação a tempo de execução, quanto com relação às suas estimativas melhores.

Tabela 3 – Medidas coletadas para o ajuste Poisson.

Estimativas - Poisson							
Pacotes	Medidas	β_0	β_1	β_2	Afastamento	AIC/EAIC	Tempo (seg)
base	Média	0,46	0,83	-0,97	0,06	1303,40	0,03
	Erro Padrão	0,05	0,08	0,09			
mgcv	Média	0,46	0,83	-0,97	0,06	1303,38	0,31
	Erro Padrão	0,05	0,08	0,09			
glmmTMB	Média	0,46	0,83	-0,97	0,06	1303,40	0,23
	Erro Padrão	0,05	0,08	0,09			
GAMLSS	Média	0,46	0,83	-0,97	0,06	1303,38	0,06
	Erro Padrão	0,05	0,08	0,09			
arm	Média	0,46	0,83	-0,96	0,06	1303,40	0,03
	Erro Padrão	0,05	0,08	0,09			
jags	Média	0,46	0,83	-0,97	0,06	1304,00	63,06
	Erro Padrão	0,05	0,08	0,09			
inla	Média	0,46	0,83	-0,97	0,06	1303,40	1,69
	Erro Padrão	0,05	0,08	0,09			
MCMCglm	Média	0,42	0,86	-0,93	0,06	1304,20	1,45
	Erro Padrão	0,12	0,16	-0,17			
glmmADMB	Média	0,46	0,83	-0,97	0,06	1304,40	0,58
	Erro Padrão	0,05	0,08	0,09			
brms	Média	0,46	0,83	-0,97	0,06	1303,38	59,79
	Erro Padrão	0,05	0,07	0,09			

A partir desse primeiro estudo foi possível ter um panorama inicial das estimativas geradas pelos diferentes pacotes na modelagem das distribuições apresentadas neste Capítulo. Optamos em seguir com GAMLSS, glmmTMB, JAGS e brms dado que foram os pacotes que conseguimos realizar as simulações para todas as distribuições. Ressaltasse a similaridade entre os dois primeiros e considerável vantagem do quarto sobre o terceiro, com relação a tempo computacional.

2.4.2 Estudo 2 - Desempenho dos modelos utilizando réplicas e apenas pacotes mais versáteis

Para este novo estudo de simulação, foram mantidos os mesmos parâmetros considerados no primeiro estudo, de modo que temos $\beta_0 = 0,5$, $\beta_1 = 1,0$, $\beta_2 = -1,0$, $\delta_0 = 0,4$, $\delta_1 = -0,5$,

Tabela 4 – Medidas coletadas para o ajuste Binomial Negativa.

Estimativas - Binomial Negativa								
Pacotes	Medidas	β_0	β_1	β_2	ϕ	Afastamento	AIC/EAIC	Tempo (seg)
base	Média Erro Padrão	0,40 0,07	1,01 0,10	-0,95 0,11	2,33	0,04	1371,90	0,06
mgcv	Média Erro Padrão	0,40 0,07	1,01 0,10	-0,95 0,11	2,27	0,04	1371,87	0,05
glmmTMB	Média Erro Padrão	0,42 0,06	0,95 0,09	-0,91 0,11	2,58	0,04	1386,30	0,53
GAMLSS	Média Erro Padrão	0,40 0,07	1,01 0,10	-0,94 0,11	2,62	0,04	1373,57	0,36
jags	Média Erro Padrão	0,40 0,07	1,01 0,10	-0,95 0,11	2,62	0,04	1374,50	178,31
inla	Média Erro Padrão	0,40 0,07	1,01 0,10	-0,95 0,11	2,60	0,04	1376,85	1,43
glmmADMB	Média Erro Padrão	0,40 0,07	1,01 0,10	-0,95 0,11	2,33	0,04	1375,54	0,75
brms	Média Erro Padrão	0,4 0,07	1,01 0,1	-0,95 0,11	2,33	0,04	1372,19	72,39

Tabela 5 – Medidas coletadas para o ajuste ZIP1.

Estimativas - ZIP1								
Pacotes	Medidas	β_0	β_1	β_2	π	Afastamento	AIC/EAIC	Tempo (seg)
VGAM	Média Erro Padrão	0,48 0,07	1,05 0,04	-0,98 0,04	0,26 0,05	0,02	1129,32	0,33
mgcv	Média Erro Padrão	0,49 0,07	1,05 0,04	-0,97 0,04	0,34 0,06	0,02	1145,48	1,18
glmmTMB	Média Erro Padrão	0,48 0,07	1,05 0,04	-0,98 0,05	0,26 0,05	0,02	1129,30	2,60
GAMLSS	Média Erro Padrão	0,48 0,07	1,05 0,04	-0,98 0,05	0,26 0,05	0,02	1129,32	0,80
jags	Média Erro Padrão	0,47 0,07	1,05 0,04	-0,98 0,05	0,26 0,05	0,02	1129,00	88,00
inla	Média Erro Padrão	0,49 0,07	1,05 0,04	-0,96 0,05	0,69 0,08	0,02	1180,55	2,06
MCMCglm	Média Erro Padrão	0,07 0,18	1,17 0,07	-1,08 0,11	0,12 0,10	0,02	1235,89	2,75
glmmADMB	Média Erro Padrão	0,48 0,07	1,05 0,04	-0,98 0,05	0,56 0,11	0,02	1132,35	1,30
brms	Média Erro Padrão	0,48 0,07	1,05 0,04	-0,98 0,05	0,57 0,11	0,02	1129,70	97,89

Tabela 6 – Medidas coletadas para o ajuste ZIP2.

Estimativas - ZIP2											
Pacotes	Medidas	β_0	β_1	β_2	δ_0	δ_1	δ_2	Afastamento β	Afastamento δ	AIC/EAIC	Tempo (seg)
pscl	Média Erro Padrão	0,59 0,07	0,98 0,03	-0,93 0,04	0,38 0,09	-0,47 0,14	0,87 0,08	0,04	0,02	1264,33	0,12
VGAM	Média Erro Padrão	0,59 0,07	0,98 0,03	-0,93 0,04	0,38 0,11	-0,47 0,15	0,87 0,08	0,04	0,02	1264,33	0,31
glmmTMB	Média Erro Padrão	0,59 0,07	0,98 0,03	-0,93 0,04	0,38 0,07	-0,47 0,10	0,87 0,11	0,04	0,02	1264,30	0,64
GAMLSS	Média Erro Padrão	0,59 0,07	0,98 0,03	-0,93 0,04	0,38 0,08	-0,47 0,07	0,87 0,14	0,04	0,02	1264,33	0,23
jags	Média Erro Padrão	0,59 0,06	0,98 0,03	-0,93 0,04	0,38 0,14	-0,47 0,14	0,88 0,15	0,04	0,01	1175,55	197,24
MCMCglm	Média Erro Padrão	0,56 0,09	0,98 0,04	-0,95 0,07	0,29 0,07	-0,64 0,11	0,93 0,16	0,03	0,06	1267,78	2,88
brms	Média Erro Padrão	0,58 0,06	0,98 0,03	-0,93 0,04	0,37 0,14	-0,47 0,14	0,88 0,15	0,04	0,02	1266,70	86,10

Tabela 7 – Medidas coletadas para o ajuste ZIBN1.

Estimativas - ZIBN1									
Pacotes	Medidas	β_0	β_1	β_2	π	ϕ	Afastamento	AIC/EAIC	Tempo (seg)
glmmTMB	Média	0,44	0,99	-1,04	0,19	2,46	0,03	1713,00	0,76
	Erro Padrão	0,07	0,05	0,06	0,04				
GAMLSS	Média	0,43	0,99	-1,04	0,19	2,46	0,03	1702,98	0,36
	Erro Padrão	0,07	0,05	0,06	0,05				
jags	Média	0,22	0,99	-1,06	0,33	1,98	0,10	1745,00	211,64
	Erro Padrão	0,07	0,07	0,07	0,11				
inla	Média	0,45	0,99	-1,02	0,47	2,59	0,02	1825,38	5,31
	Erro Padrão	0,09	0,07	0,07	0,08				
glmmADMB	Média	0,44	0,99	-1,04	0,19	2,46	0,03	1727,00	1,74
	Erro Padrão	0,07	0,05	0,06	0,04				
brms	Média	0,44	0,99	-1,03	0,19	2,43	0,02	1712,70	203,06
	Erro Padrão	0,07	0,05	0,06	0,04				

Tabela 8 – Medidas coletadas para o ajuste ZIBN2.

Estimativas - ZIBN2												
Pacotes	Medidas	β_0	β_1	β_2	δ_0	δ_1	δ_2	ϕ	Afastamento β	Afastamento δ	AIC/EAIC	Tempo (seg)
pscl	Média	0,43	0,97	-1,01	-4,21	0,79	-3,91	2,74	0,03	0,18	1489,19	0,09
	Erro Padrão	0,06	0,06	0,08	0,16	0,16	0,28					
glmmTMB	Média	0,43	0,97	-1,01	-4,21	0,79	-3,91	2,74	0,03	0,18	1495,20	0,72
	Erro Padrão	0,06	0,06	0,08	0,30	0,21	0,53					
GAMLSS	Média	0,43	0,97	-1,01	-4,21	0,79	-3,91	2,74	0,03	0,18	1489,19	0,36
	Erro Padrão	0,06	0,06	0,08	0,66	0,43	0,75					
jags	Média	0,43	0,98	-1,01	4,52	-0,87	4,14	2,72	0,03	3,84	1574,20	415,74
	Erro Padrão	0,06	0,06	0,08	0,50	0,27	0,44					
brms	Média	0,44	0,97	-1,00	-4,00	0,72	-3,73	2,80	0,02	0,11	1493,11	112,27
	Erro Padrão	0,06	0,06	0,08	0,75	0,31	0,64					

$\delta_2 = 0,9$, para o modelo ZIP2 e $\delta_0 = -4,0$, $\delta_1 = 0,5$, $\delta_2 = -3,5$, para o modelo ZIBN2. Agora, ZIP1 e ZIBN1, na parte inflacionada, tem como probabilidade de zeros $\pi = 0,4$ e $\pi = 0,2$, respectivamente. O objetivo deste segundo estudo foi validar, com o uso de 100 réplicas, a qualidade dos pacotes pacotes GAMLSS, glmmTMB, JAGS e brms.

Assim, 100 conjuntos de 500 observações foram criados para cada modelo. Ao todo temos 6 modelos avaliados, dada a partição dos ajustes inflacionados em dois cada (ZIP1, ZIP2 e ZIBN1, ZIBN2) e os modelos Poisson e Binomial Negativa.

Aqui, coletamos o tempo computacional, média e erro padrão das estimativas, viés e REQM, ou raiz do erro quadrático médio. Calculamos também os respectivos intervalos de confiança sob olhar assintótico, partindo da estatística de Wald, dada por $\hat{\beta}_j \pm z_{1-\alpha/2} se(\hat{\beta}_j)$ (encontrada por exemplo no trabalho de Gouriéroux, Holly e Monfort (1982)), para as médias das estimativas. Cabe pontuar que trouxemos como resultado as probabilidades de cobertura dos intervalos realizados para cada modelo, assim podemos ver a variação para cada pacote. Podemos ressaltar tempos superiores para ajustes Bayesianos, algo já constatado anteriormente.

As Tabelas 9, 10, 11, 12, 13, 14 apresentam os resultados do estudo.

Algo a ressaltar é a especificação dos modelos Bayesianos. para JAGS consideramos 3 cadeias, com 4000 de tamanho total, além de 1000 iterações para etapa de burn-in. Esse cenário é aplicável para quase todas as distribuições, com exceção da ZIBN2, em que para atingir a

Tabela 9 – Medidas coletadas - Modelo Poisson considerando 100 réplicas.

Poisson				
Medidas	β_0	β_1	β_2	Tempo médio (seg)
Pacote GAMLSS				
Média	0,506	0,994	-1,019	0,024
Desvio padrão	0,053	0,091	0,084	
Viés	0,006	-0,006	-0,019	
REQM	0,053	0,091	0,086	
Pacote glmmTMB				
Média	0,506	0,994	-1,019	0,277
Desvio padrão	0,053	0,091	0,084	
Viés	0,006	-0,006	-0,019	
REQM	0,053	0,091	0,086	
Pacote JAGS				
Média	0,504	0,994	-1,021	61,248
Desvio padrão	0,053	0,091	0,085	
Viés	0,004	-0,006	-0,021	
REQM	0,053	0,091	0,087	
Pacote brms				
Média	0,504	0,995	-1,021	111,237
Desvio padrão	0,053	0,091	0,085	
Viés	0,004	-0,005	-0,021	
REQM	0,053	0,091	0,087	

Tabela 10 – Medidas coletadas - Modelo Binomial Negativa considerando 100 réplicas.

Binomial Negativa					
Medidas	β_0	β_1	β_2	ϕ	Tempo médio (seg)
Pacote GAMLSS					
Média	0,498	1,005	-1,011	2,254	0,481
Desvio padrão	0,071	0,097	0,116	0,981	
Viés	-0,002	0,005	-0,011	0,254	
REQM	0,071	0,097	0,116	1,009	
Pacote glmmTMB					
Média	0,499	0,926	-0,931	1,389	1,038
Desvio padrão	0,073	0,098	0,111	0,295	
Viés	-0,001	-0,074	0,069	-0,611	
REQM	0,072	0,123	0,131	0,678	
Pacote JAGS					
Média	0,497	1,009	-1,013	1,843	232,542
Desvio padrão	0,072	0,097	0,116	0,622	
Viés	-0,003	0,009	-0,013	-0,157	
REQM	0,071	0,097	0,116	0,638	
Pacote brms					
Média	0,497	1,009	-1,013	2,120	550,926
Desvio padrão	0,072	0,098	0,116	0,434	
Viés	-0,003	0,009	-0,013	0,120	
REQM	0,071	0,097	0,116	0,448	

convergência foi necessário ampliar o tamanho das cadeias para 10000. De todo modo, cabe pontuar, que para todas as distribuições JAGS convergiu, apresentando um potencial de escala de redução (R-hat), próximos a 1. Para brms, para todos os casos foi considerado o cenário de 4 cadeias, com um burn-in de 1000 e tamanho total das cadeias de 6000. A tabela 15 apresenta os valores dos R-hats para cada pacote e modelo.

Para ajustes não inflacionados, brms apresentou tempo maior do que JAGS, porém esse cenário se inverte conforme passamos a aumentar a complexidade do modelo. Seguindo sob ótica Bayesiana, há casos em que JAGS apresenta valores de REQM menores, porém para o modelo ZIBN2, além dos valores serem bem elevados, temos que para esse ajuste, o pacote se perdeu na parte inflacionada, gerando estimativas ruins, longes dos verdadeiros valores dos parâmetros.

Tabela 11 – Medidas coletadas - Modelo ZIP1 considerando 100 réplicas.

ZIP1					
Medidas	β_0	β_1	β_2	π	Tempo médio (seg)
Pacote GAMLSS					
Média	0,497	0,995	-1,006	0,405	
Desvio padrão	0,068	0,043	0,047	0,119	
Viés	-0,003	-0,005	-0,006	0,005	0,160
REQM	0,068	0,043	0,047	0,118	
Pacote glmmTMB					
Média	0,497	0,995	-1,006	0,405	
Desvio padrão	0,068	0,043	0,047	0,119	
Viés	-0,003	-0,005	-0,006	0,005	0,628
REQM	0,068	0,043	0,047	0,118	
Pacote JAGS					
Média	0,520	0,984	-0,993	0,542	
Desvio padrão	0,002	0,001	0,001	0,002	
Viés	0,020	-0,016	0,007	0,142	82,914
REQM	0,020	0,016	0,007	0,142	
Pacote brms					
Média	0,524	0,984	-1,000	0,631	
Desvio padrão	0,053	0,036	0,035	0,000	
Viés	0,024	-0,016	0,000	0,231	119,516
REQM	0,058	0,039	0,035	0,231	

Tabela 12 – Medidas coletadas - Modelo ZIP2 considerando 100 réplicas.

ZIP2							
Medidas	β_0	β_1	β_2	δ_0	δ_1	δ_2	Tempo médio (seg)
Pacote GAMLSS							
Média	0,508	0,995	-0,998	0,401	-0,510	0,919	
Desvio padrão	0,061	0,034	0,031	0,145	0,127	0,133	
Viés	0,008	-0,005	0,002	0,001	-0,010	0,019	0,127
REQM	0,061	0,035	0,031	0,144	0,127	0,134	
Pacote glmmTMB							
Média	0,508	0,995	-0,998	0,401	-0,510	0,919	
Desvio padrão	0,061	0,034	0,031	0,145	0,127	0,133	
Viés	0,008	-0,005	0,002	0,001	-0,010	0,019	0,841
REQM	0,061	0,035	0,031	0,144	0,127	0,134	
Pacote JAGS							
Média	0,505	0,996	-0,998	0,399	-0,512	0,929	
Desvio padrão	0,061	0,035	0,031	0,148	0,130	0,136	
Viés	0,005	-0,004	0,002	-0,001	-0,012	0,029	222,095
REQM	0,061	0,035	0,031	0,147	0,130	0,138	
Pacote brms							
Média	0,503	0,996	-0,998	0,394	-0,510	0,927	
Desvio padrão	0,062	0,035	0,031	0,147	0,130	0,136	
Viés	0,003	-0,004	0,002	-0,006	-0,010	0,027	148,520
REQM	0,061	0,035	0,031	0,146	0,130	0,138	

Em cenário clássico, temos equivalência em quase todos os ajustes. Cabe ressaltar que para ajustes inflacionados com uma parte Binomial Negativa, GAMLSS acaba apresentando maior variabilidade de resultados, apresentando tanto resultados adequados, bem próximos dos valores reais dos parâmetros, porém também um número considerável de estimativas inadequadas. Isso influencia no valor do REQM e média maiores. De todo modo, cabe pontuar que para ZIBN2, a parte inflacionada é mais aderente para GAMLSS, observando viés e REQM menores, além das médias mais próximas.

Para todos os casos consideramos um nível de confiança de 95%, o que podemos considerar como probabilidade nominal. Em muitos dos casos podemos ver que as probabilidades de cobertura ultrapassam esse valor fixado. Considerando o cenário estudado, podemos ver na

Tabela 13 – Medidas coletadas - Modelo ZIBN1 considerando 100 réplicas.

ZIBN1						
Medidas	β_0	β_1	β_2	π	ϕ	Tempo médio (seg)
Pacote GAMLSS						
Média	0,575	0,995	-1,006	0,228	1,877	
Desvio padrão	0,442	0,729	0,678	0,220	1,147	
Viés	0,075	-0,005	-0,006	0,028	-0,123	1,649
REQM	0,446	0,726	0,675	0,221	1,148	
Pacote glmmTMB						
Média	0,489	0,997	-1,012	0,197	2,025	
Desvio padrão	0,073	0,055	0,056	0,034	0,408	
Viés	-0,011	-0,003	-0,012	-0,003	0,025	0,811
REQM	0,074	0,055	0,057	0,034	0,407	
Pacote JAGS						
Média	0,261	1,008	-1,014	0,334	2,323	
Desvio padrão	0,067	0,065	0,073	0,004	0,256	
Viés	-0,239	0,008	-0,014	0,134	0,323	287,653
REQM	0,248	0,065	0,074	0,134	0,412	
Pacote brms						
Média	0,458	1,010	-1,012	0,200	2,050	
Desvio padrão	0,340	0,275	0,389	0,033	0,355	
Viés	-0,042	0,010	-0,012	0,000	0,050	168,371
REQM	0,341	0,273	0,388	0,032	0,387	

Tabela 14 – Medidas coletadas - Modelo ZIBN2 considerando 100 réplicas.

ZIBN2								
Medidas	β_0	β_1	β_2	δ_0	δ_1	δ_2	ϕ	Tempo médio (seg)
Pacote GAMLSS								
Média	0,652	0,847	-0,782	-4,151	0,511	-3,628	2,061	
Desvio padrão	0,233	0,154	0,131	1,103	0,354	0,836	0,537	
Viés	0,152	-0,153	0,218	-0,151	0,011	-0,128	0,061	2,124
REQM	0,277	0,216	0,254	1,108	0,352	0,842	0,538	
Pacote glmmTMB								
Média	0,491	1,004	-1,027	-3,273	0,432	-3,094	2,090	
Desvio padrão	0,062	0,067	0,086	2,528	0,424	1,579	0,438	
Viés	-0,009	0,004	-0,027	0,727	-0,068	0,406	0,090	0,403
REQM	0,062	0,067	0,090	2,618	0,428	1,622	0,445	
Pacote JAGS								
Média	0,444	1,012	-1,024	2,639	-0,097	0,118	2,143	
Desvio padrão	0,098	0,068	0,075	2,291	0,500	0,667	0,507	
Viés	-0,056	0,012	-0,024	6,639	-0,597	3,618	0,143	1550,941
REQM	0,112	0,069	0,078	7,020	0,777	3,679	0,524	
Pacote brms								
Média	0,497	1,002	-0,999	-3,870	0,447	-3,407	2,157	
Desvio padrão	0,061	0,066	0,066	0,694	0,288	0,556	0,435	
Viés	-0,003	0,002	0,001	0,130	-0,053	0,093	0,157	253,815
REQM	0,061	0,065	0,066	0,703	0,292	0,561	0,461	

Tabela 15 – R-hats dos ajustes Bayesianos.

Distribuição / Pacotes	**JAGS**	**brms**
Poisson	1,0001	1,00
BN	1,0002	1,00
ZIP1	1,0004	1,00
ZIP2	1,0004	1,00
ZIBN1	1,0010	1,00
ZIBN2	1,0017	1,00

Tabela 16 – Probabilidade de cobertura dos Intervalos de Confiança assintóticos clássicos (GAMLSS; glmmTMB) e Bayesianos (jags; brms).

Poisson							
Pacotes/medidas	β_0	β_1	β_2	π	δ_0	δ_1	δ_2
GAMLSS	90,00%	90,00%	97,00%				
glmmTMB	90,00%	90,00%	97,00%				
JAGS	90,00%	90,00%	96,00%			-	
brms	90,00%	90,00%	97,00%				
Binomial Negativa							
GAMLSS	95,00%	97,00%	95,00%				
glmmTMB	94,00%	85,00%	90,00%				
JAGS	96,00%	95,00%	95,00%			-	
brms	95,00%	97,00%	95,00%				
ZIP1							
GAMLSS	96,00%	95,00%	95,00%	95,00%			
glmmTMB	89,00%	95,00%	97,00%	95,00%			
JAGS	94,00%	92,00%	95,00%	93,00%		-	
brms	94,00%	95,00%	95,00%	96,00%			
ZIP2							
GAMLSS	96,00%	94,00%	94,00%		92,00%	96,00%	94,00%
glmmTMB	94,00%	90,00%	96,00%		93,00%	96,00%	94,00%
JAGS	94,00%	92,00%	95,00%	-	90,00%	98,00%	92,00%
brms	96,00%	94,00%	96,00%		98,00%	95,00%	100,00%
ZIBN1							
GAMLSS	84,00%	80,00%	87,00%	87,00%			
glmmTMB	94,00%	95,00%	90,00%	90,00%			
JAGS	86,00%	100,00%	95,00%	90,00%		-	
brms	95,00%	96,00%	91,00%	100,00%			
ZIBN2							
GAMLSS	84,00%	80,00%	82,00%		97,00%	96,00%	97,00%
glmmTMB	99,00%	91,00%	90,00%		85,00%	90,00%	90,00%
JAGS	100,00%	91,00%	93,00%	-	10,00%	70,00%	12,00%
brms	100,00%	94,00%	92,00%		97,00%	98,00%	100,00%

Tabela 16 que a estatística intervalar concorda com a estimativa pontual, sendo que o comentário para GAMLSS possuir maior REQM por conta de uma variabilidade maior nos resultados bons e ruins, acabou impactando negativamente na probabilidade de cobertura dos intervalos de confiança assintóticos realizados. Isso observando todos os coeficientes das partes não inflacionadas. Cabe ressaltar que para coeficientes da parte inflacionada, os resultados são favoráveis a GAMLSS. Para o modelo ZIBN2, ainda pontua-se que JAGS não conseguiu estimar adequadamente os coeficientes para a parte inflacionada do modelo, algo evidenciado nas estatísticas pontuais e consequentemente intervalares. De maneira geral brms apresenta intervalos consistentes, de certo modo superando na maioria dos casos a probabilidade nominal, a qual os intervalos foram construídos. GAMLSS e glmmTMB empatam, tendo cenários mais favoráveis a GAMLSS e outros a glmmTMB.

É interessante pontuar que de maneira geral, conforme a complexidade do modelo vai aumentando, no caso Bayesiano, o pacote brms tende a ser mais preciso e ter tempos menores do que JAGS, ao mesmo passo que as estimativas de JAGS começam a piorar e os tempos aumentar. Há similaridade quase que em todos os casos para os pacotes clássicos, porém há vantagem de tempo para GAMLSS.

A partir das Tabelas das medidas, entende-se que, de fato, há a validação de que o pacote brms possui vantagem sobre o JAGS, principalmente observando a qualidade das estimativas. Para a parte inflacionada do ZIBN2, vemos que há um descolamento do pacote JAGS. Sendo

que para esse caso, o REQM fica bem elevado. Por esse motivo, entendemos que brms seria a indicação para realização dos ajustes sob ótica Bayesiana.

Agora, com relação ao GAMLSS e glmmTMB, a maior diferença entre eles se dá no tempo computacional, que apesar de próximo, acaba sendo menor na maioria dos casos para GAMLSS. Um ponto a destacar é que os erros começam a aumentar de acordo com a complexidade do modelo, nos dois casos. glmmTMB tende a errar mais a parte inflacionada e para algumas amostras GAMLSS acaba errando a parte não inflacionada, algo que fica claro para ZIBN2. Como nos modelos Poisson, Binomial Negativa, ZIP1 e ZIP2, houve empate, sendo o critério de desempate o tempo, então, para ajustes clássicos, GAMLSS fica como nossa indicação para ajustes clássicos. Ressaltando sempre o cenário aqui simulado.

2.4.3 Estudo 3 - Desempenho do modelo Zero-inflacionado Binomial Negativo de duas partes utilizando réplicas

Em sequência ao estudo de simulação anterior, um novo estudo foi proposto, partindo do resultado obtido, que validou que para o caso clássico a melhor alternativa é oferecida pelo pacote GAMLSS e de maneira similar para o caso Bayesiano temos o pacote brms. Ressaltamos que os parâmetros mais uma vez foram mantidos, dada a necessidade que temos em observar um volume de zeros superior às demais observações (vide Figura 1).

Deste modo, o proposto para este novo estudo foi uma avaliação, agora dos métodos de estimação, porém restritos ao uso de apenas o modelo mais complexo: Zero-inflacionado Binomial Negativo de duas partes. Propomos, desta forma, uma comparação de desempenho dos pacotes GAMLSS e brms considerando ZIBN2 para avaliar o tempo computacional em escala logarítmica e a precisão dos estimadores utilizando o REQM, utilizado no estudo de Morris, White e Crowter (2019) para comparação dos modelos estudados. A escala logarítmica é utilizada dada a discrepância observada entre os tempos computacionais dos dois pacotes nos estudos anteriores. Adicionalmente, foram considerados três diferentes tamanhos de amostra: 200, 500 e 1000, sendo que para cada tamanho amostral estamos considerando 50 réplicas. Ou seja, para cada pacote, teremos 150 conjuntos de dados diferentes a serem utilizados nos ajustes de cada metodologia, sendo 50 com tamanho amostral 200, 50 com tamanho amostral 500 e 50 com tamanho amostral 1000. Totalizando assim 300 conjuntos. Cada conjunto conta com 3 variáveis, seguindo os moldes do primeiro estudo, ou seja, temos uma variável resposta que segue uma Zero-inflacionada Binomial Negativa com vetor de médias dado por $\boldsymbol{\mu} = \exp\{\beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2\}$, mais uma vez sendo utilizado para simular os valores y de Y. Aqui X_1 e X_2 são duas normais padrão ($N(0,1)$).

Cada ajuste dispôs de informações que foram coletadas, tais como o tempo de execução, o parâmetro de dispersão ϕ de valor verdadeiro igual a 2 e as estimativas dos parâmetros. Os parâmetros tiveram valores fixados, considerando $\beta_0 = 0,5$, $\beta_1 = 1,0$, $\beta_2 = -1,0$ e, para a parte

não inflacionada $\delta_0 = 4,0$, $\delta_1 = 0,5$, $\delta_2 = -3,5$ para a parte inflacionada. Esses valores foram escolhidos levando em conta o experimento conduzido anteriormente, que fez uso dos mesmos valores fixados para os parâmetros. De maneira sumarizada obteve-se as informações relativas a média das 150 estimativas clássicas/Bayesianas, desvios padrões, viés - $B(\hat{\boldsymbol{\theta}}) = E(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}$ e a raiz do Erro Quadrático Médio, REQM. Podemos dizer que tal medida é obtida a partir da soma da diferença das estimativas $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\delta}}$ considerando as 50 réplicas e o valor real do parâmetro, elevada ao quadrado sobre o número de simulações (no nosso caso repetições). A expressão dessa medida no nosso caso fica: $\sqrt{1/n_{simul}\sum_{i=1}^{n_{simul}}(\hat{\beta}i - \beta_i)^2}$ e $\sqrt{1/n_{simul}\sum_{i=1}^{n_{simul}}(\hat{\delta}i - \delta_i)^2}$. Os resultados da simulação se encontram na Tabela 17.

Ressaltamos que o n_{simul} é igual a 50, dado que esse é o número de repetições que consideramos por limitações computacionais e que essa medida e as demais foram mensuradas para cada um dos tamanhos amostrais estudados (200, 500, 1000).

Pode-se observar na Tabela 17, que para a parte inflacionada, a estimação Bayesiana é mais precisa, tendo REQMs mais baixos que comparados com a estimação clássica. Isso se torna mais evidente com tamanho amostral menor, mas não é diferente para os demais tamanhos. Cada cenário pode ser observado nas Figuras 3, 4, 5, 6, 7, 8. Especificamente para o parâmetro β_1 o cenário se inverte, sendo que o REQM para a amostra de tamanho 200, sob uma abordagem clássica se torna menor que quando comparado à abordagem Bayesiana.

Com tamanho amostral 1000 as diferenças entre os métodos é quase irrisória (sendo equivalente no caso de β_2) e desse modo quanto maior o tamanho amostral entende-se que os métodos se tornam equivalentes e para ambos os casos o REQM diminui. A estimação clássica melhora conforme o tamanho da amostra cresce, dado que a diferença entre as estimações e os valores de referência diminui.

Tabela 17 – Sumarização dos resultados do modelo ZIBN2 considerando 50 réplicas por tamanho amostral.

Estimativas								
			Parte Binomial Negativa			Parte zero-inflacionada		
Medidas	ϕ	Tempo	β_0	β_1	β_2	δ_0	δ_1	δ_2
			0,5	1	-1	-4	0,5	-3,5
Tamanho amostral 200								
Pacote GAMLSS								
Média	2,111	0,255	0,482	1,003	-1,016	-4,383	0,739	-3,726
Desvio Padrão	0,709	0,239	0,110	0,096	0,146	1,649	1,103	2,632
Viés	-	-	-0,018	0,003	-0,016	-0,390	0,244	-0,230
MSE	-	-	0,012	0,009	0,021	2,810	1,249	6,838
REQM	-	-	0,111	0,095	0,146	1,676	1,117	2,615
Pacote BRMS								
Média	2,301	46,137	0,507	0,993	-1,029	-3,778	0,439	-3,324
Desvio Padrão	0,662	18,049	0,108	0,096	0,107	1,021	0,462	0,833
Viés	-	-	0,007	-0,007	-0,030	0,227	-0,062	0,180
MSE	-	-	0,011	0,009	0,012	1,072	0,213	0,711
REQM	-	-	0,107	0,095	0,110	1,035	0,462	0,843
Tamanho amostral 500								
Pacote GAMLSS								
Média	2,136	0,466	0,494	0,999	-1,005	-4,342	0,638	-3,827
Desvio Padrão	0,433	0,120	0,082	0,052	0,056	1,073	0,381	0,771
Viés	-	-	-0,006	-0,001	-0,005	-0,349	0,141	-0,334
MSE	-	-	0,007	0,003	0,003	1,245	0,161	0,689
REQM	-	-	0,081	0,052	0,056	1,116	0,401	0,830
Pacote BRMS								
Média	2,256	109,317	0,500	0,994	-1,002	-4,030	0,562	-3,580
Desvio Padrão	0,523	35,631	0,081	0,052	0,055	0,808	0,327	0,585
Viés	-	-	0,000	-0,006	-0,002	-0,031	0,063	-0,081
MSE	-	-	0,006	0,003	0,003	0,641	0,109	0,341
REQM	-	-	0,080	0,052	0,054	0,801	0,329	0,584
Tamanho amostral 1000								
Pacote GAMLSS								
Média	2,030	0,626	0,507	0,992	-0,998	-4,031	0,456	-3,564
Desvio Padrão	0,223	0,210	0,037	0,048	0,057	0,573	0,177	0,504
Viés	-	-	0,007	-0,008	0,002	-0,032	-0,045	-0,065
MSE	-	-	0,001	0,002	0,003	0,323	0,033	0,253
REQM	-	-	0,037	0,048	0,056	0,568	0,180	0,503
Pacote BRMS								
Média	2,048	188,912	0,510	0,990	-0,996	-3,924	0,432	-3,478
Desvio Padrão	0,222	33,046	0,036	0,057	0,515	0,048	0,169	0,462
Viés	-	-	0,011	-0,010	0,078	-0,070	0,004	0,023
MSE	-	-	0,001	0,002	0,265	0,033	0,003	0,210
REQM	-	-	0,038	0,048	0,515	0,181	0,056	0,458

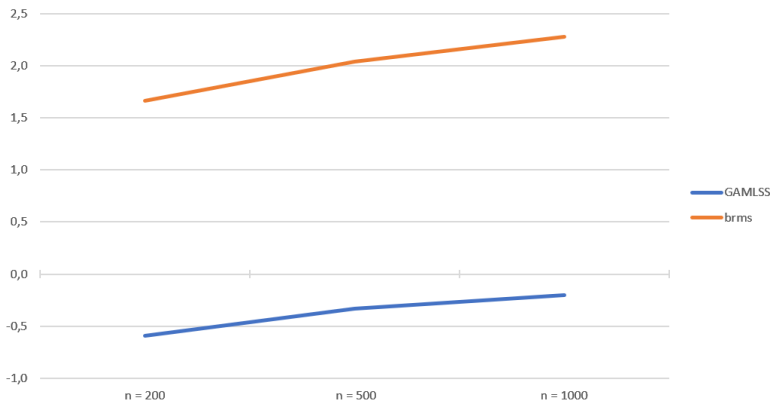

Figura 2 – Tempo computacional médio em logaritmo de segundos de acordo com cada tamanho amostral e metodologia.

A Figura 2 mostra os tempos em escala logarítmica (para melhor comparabilidade).

Pontua-se aqui que era esperado que conforme o tamanho amostral aumentasse, o tempo computacional atrelado seria maior. Outro ponto é o de que a abordagem Bayesiana possui maior complexidade e por isso, temos essa discrepância em relação aos tempos desempenhados sob abordagem clássica. No sentido de que cada cadeia de Markov possui parte das iterações que são chamadas de *burn-in*, passo de calibração da cadeia, de modo a descartar valores iniciais até que a cadeia atinja um estado estacionário. Cada etapa conta com cadeias de tamanho 1000, 6000, respectivamente, assim como já comentado para o estudo anterior. Por esse motivo, em escala logarítmica vemos que todos os tempos Bayesianos (laranja) se encontram muito acima dos tempos clássicos (azul).

Realizamos um levantamento das probabilidades de cobertura, observando os parâmetros simulados. Expressos assim na Tabela 18.

Tabela 18 – Probabilidades de cobertura dos Intervalos de Confiança assintóticos clássico (GAMLSS) e Bayesiano (brms).

ZIBN2							
Pacotes/medidas	β_0	β_1	β_2	π	δ_0	δ_1	δ_2
n = 200							
GAMLSS	90,00%	89,00%	84,00%		86,00%	89,00%	87,00%
brms	92,00%	90,00%	90,00%	-	94,00%	92,00%	93,00%
n = 500							
GAMLSS	93,00%	92,00%	91,00%		95,00%	92,00%	90,00%
brms	95,00%	93,00%	93,00%	-	97,00%	95,00%	93,00%
n = 1000							
GAMLSS	96,00%	95,00%	96,00%		98,00%	95,00%	97,00%
brms	98,00%	96,00%	95,00%	-	100,00%	96,00%	96,00%

Observamos para as novas amostras simuladas, que houve um resultado diferente com relação primeiramente ao REQM, consideravelmente menor aqui para GAMLSS, comparando com os resultados do último estudo e isso favoreceu positivamente nas estimativas intervalares. Isso para todos os tamanhos de amostras. Novamente consideramos uma probabilidade nominal de 95%. O resultado aqui seguiu o que vemos nas Figuras 3, 4, 5, 6, 7 e 8. Conforme o REQM foi diminuindo para o caso clássico, principalmente na parte inflacionada do modelo, as estimativas intervalares foram convergindo para uma probabilidade similar para ambas as metodologias.

Como conclusão, ambas as metodologias se mostram equivalentes para estimativa de parâmetros a medida que o tamanho amostral aumenta. Um ponto é que isso acarreta em maiores tempos computacionais, que começa a ter um peso maior do lado Bayesiano. Com relação aos resultados das estimações, podemos dizer que o método Bayesiano de estimação é mais preciso, principalmente para amostras de tamanho pequeno, porém ao preço de tempo de processamento muito superior ao observado para o método clássico. De todo modo, as estimativas intervalares, apontam ligeira vantagem do método Bayesiano, dado que mais intervalos contém o valor verdadeiro dos parâmetros simulados. Porém, entende-se que a metodologia clássica se sobrepõe neste caso e então é a mais indicada para modelagem de dados de contagem.

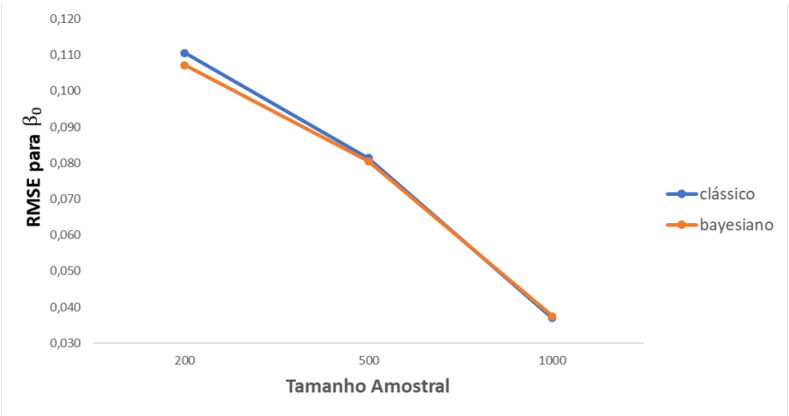

Figura 3 – Raiz do MSE da estimativa de β_0 dado os diferentes tamanhos amostrais (200, 500 e 1000, da esquerda para direita), sob 50 réplicas para o modelo ZIBN.

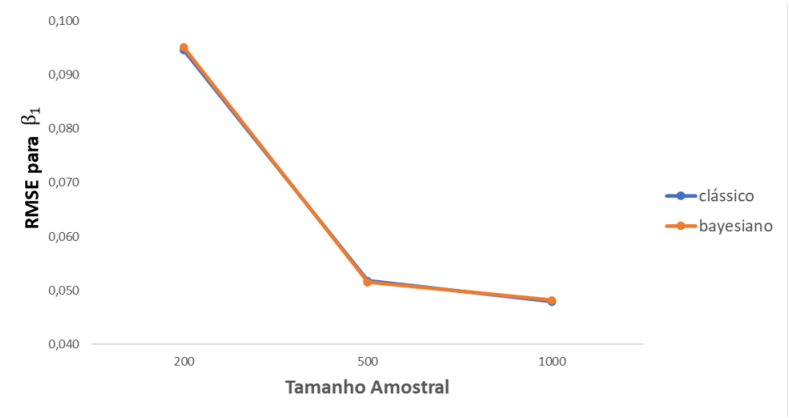

Figura 4 – Raiz do MSE da estimativa de β_1 dado os diferentes tamanhos amostrais (200, 500 e 1000, da esquerda para direita), sob 50 réplicas para o modelo ZIBN.

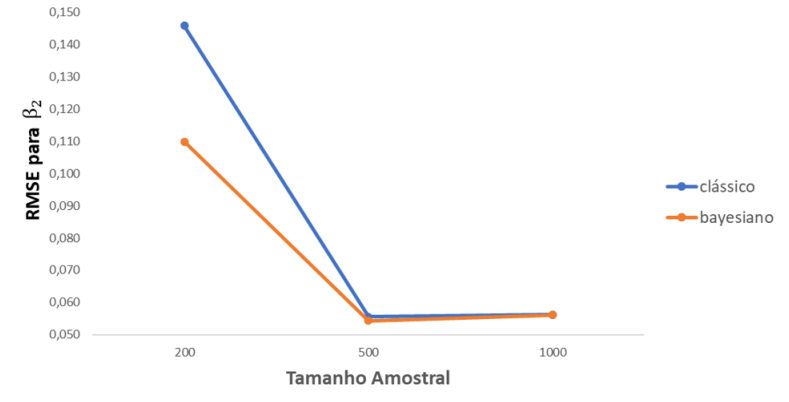

Figura 5 – Raiz do MSE da estimativa de β_2 dado os diferentes tamanhos amostrais (200, 500 e 1000, da esquerda para direita), sob 50 réplicas para o modelo ZIBN.

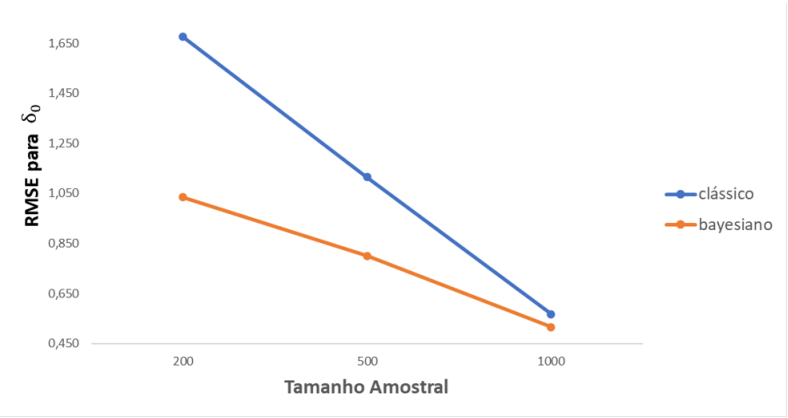

Figura 6 – Raiz do MSE da estimativa de δ_0 dado os diferentes tamanhos amostrais (200, 500 e 1000, da esquerda para direita), sob 50 réplicas para o modelo ZIBN.

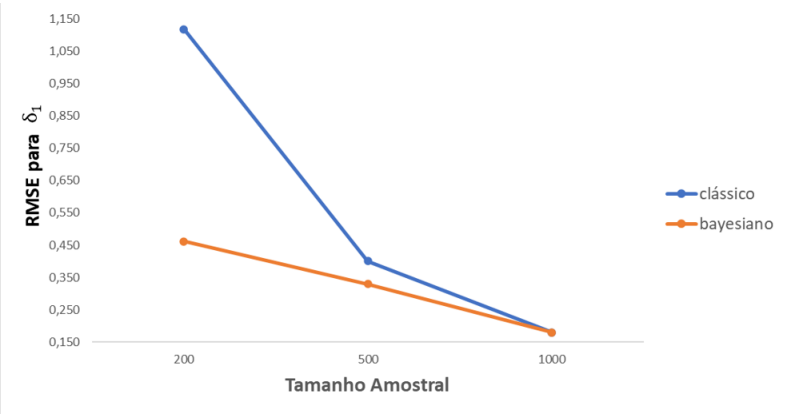

Figura 7 – Raiz do MSE da estimativa de δ_1 dado os diferentes tamanhos amostrais (200, 500 e 1000, da esquerda para direita), sob 50 réplicas para o modelo ZIBN.

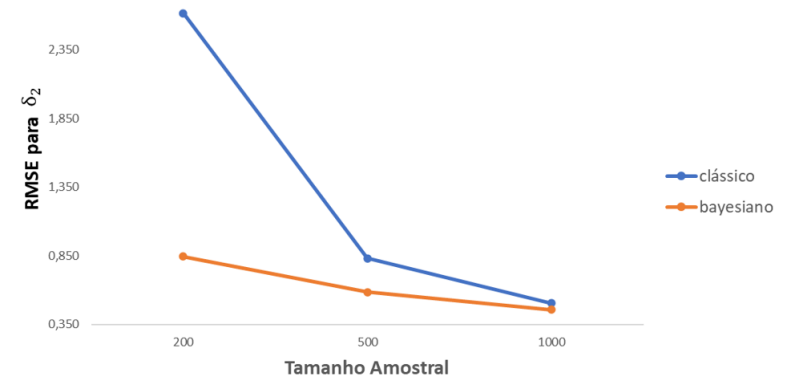

Figura 8 – Raiz do MSE da estimativa de δ_2 dado os diferentes tamanhos amostrais (200, 500 e 1000, da esquerda para direita), sob 50 réplicas para o modelo ZIBN.

2.5 Considerações do Capítulo - Modelos Inflacionados

Este Capítulo apresentou os modelos Poisson e Binomial Negativa e suas variantes zero-inflacionadas. Através dos estudos de simulação pontua-se que dois pacotes se destacaram: GAMLSS e brms, sendo o primeiro para a modelagem via metodologia clássica e o segundo via metodologia Bayesiana de estimação para os parâmetros. Comparando os resultados comparativos entre metodologias, entendemos que as implementações em **R** tendem a ter melhores resultados para abordagem Bayesiana, porém a preço de um tempo computacional consideravelmente maior. A depender da velocidade necessária para uma análise, seria interessante o uso da metodologia clássica para estimação dos parâmetros. Outro ponto a destacar é que segundo o cenário simulado, quanto maior o número da amostra, há uma convergência entre os resultados observados, ou seja, independentemente da abordagem escolhida para modelagem, os resultados se tornam de fato muito próximos, ficando a critério do pesquisador escolher a metodologia que melhor lhe atende.

Por fim, os códigos dos estudos de simulação estão presentes no Apêndice A.

CAPÍTULO

3

MODELOS DE REGRESSÃO BELL E POISSON-TWEEDIE

3.1 Introdução

No Capítulo anterior apresentamos os modelos Poisson, Binomial Negativa e duas alternativas para a tratativa de inflacionamento de zeros, o ZIP e o ZIBN. Na literatura recente, diferentes autores se propuseram a estudar outras alternativas. Abordamos neste Capítulo duas distribuições que vem ganhando destaque: Bell e Poisson-Tweedie (PTw).

Primeiramente, Castellares, Ferrari e Lemonte (2018) propõem em seu estudo uma nova distribuição que não é obtida a partir do processo usual de discretização, mas sim, das expansões da função de Bell, propostas em Bell (1934), que acabam nomeando-a. A distribuição de Bell possui um único parâmetro - θ e à medida em que sua função massa de probabilidade não apresenta expressões complicadas, se torna bem simples de lidar. Algumas de suas propriedades mais importantes são:

- Esta é uma distribuição que faz parte da família exponencial uniparamétrica;
- Apesar da Poisson não ser englobada na família da distribuição Bell, há uma aproximação entre as distribuições quando o parâmetro da distribuição Bell assume valores pequenos;
- Esta distribuição é infinitamente divisível;
- Pode-se afirmar que essa distribuição é um caso especial de processos de Poisson múltiplo (como afirma Daniel e Seminars (2008) em seu trabalho).

Outra distribuição que ganhou destaque na literatura recente para a modelagem de dados de contagem é a PTw, apresentada por Bonat *et al.* (2018). Como o nome pode sugerir, PTw é uma

mistura entre uma parte Poisson e outra Tweedie. Uma observação importante é que modelos de mistura que possuem uma parte Poisson normalmente são aplicáveis quando há falta de heterogeneidade nos dados, o que acaba implicando em variabilidade extra e consequentemente dados sobredispersos. É necessária, assim, a inclusão de efeitos aleatórios a nível observacional para a tratativa dos dados.

Um ponto forte apresentado pelo autor é a flexibilidade que a Tweedie garante à Poisson, de modo que, assim como acontece com ZIP por exemplo, garante que o modelo proveniente de uma PTw seja capaz de acomodar dados sobredispersos. Cabe ressaltar que o modelo PTw é consideravelmente mais complexo que os modelos Poisson e Binomial Negativa, pois possui uma integral intratável como função de distribuição e assim, há a necessidade de uma aproximação para a realização da estimação de parâmetros. Essa aproximação se dá considerando a modelagem em suposições apenas para momentos de segunda ordem. Isso acaba por garantir ainda mais flexibilidade do parâmetro de dispersão, dado que há a possibilidade de acomodar tanto subdispersão, quanto sobredispersão de dados, sendo que inicialmente somente sobredispersão era suportada. Entretanto, passamos a não ter mais função massa de probabilidade, como descrito por Jørgensen e Kokonendji (2015) e Bonat e Jørgensen (2016).

Os modelos inflacionados para essas distribuições ainda se encontram pouco explorados pelos pesquisadores. Porém, podemos citar aqui dois trabalhos, o primeiro proposto por Saha *et al.* (2020), que apresenta o modelo PTw, testando-o com dados inflacionados e sobredispersos. Como resultado, o autor afirma que o modelo PTw oferece uma estrutura unificada para modelar dados de contagem sobredispersos, subdispersos, zero-inflacionados, espaciais e longitudinais. E desse modo, no seu entendimento, não haveria a necessidade de criar um modelo com mais uma distribuição para tratativa dos zeros. O segundo trabalho, proposto por AJ Moreno-Arenas G (2020) apresenta o modelo Zero-inflacionado Bell (ZIBell). Assim como Yang *et al.* (2017), os autores concluem que a variante zero-inflacionada é bem-vinda para a tratativa de excesso de zeros amostrais, de modo a obter melhores estimativas.

Cabe pontuar que todos os trabalhos para essas variáveis se baseiam na ótica clássica para estimação de parâmetros. De mesmo modo, as estimações realizadas para os estudos de simulação e aplicação deste Capítulo foram elaboradas sob mesma ótica.

Este Capítulo está organizado da seguinte forma, na Seção 3.2 introduzimos as distribuições de contagem, assim como os modelos de regressão para cada distribuição. A Seção 3.3 apresenta o estudo de simulação que compara a assertividade dos modelos Poisson, Binomial Negativa, Bell e PTw quando a distribuição origem dos dados varia. Por fim, a Seção 3.4 apresenta a conclusão para o Capítulo.

3.2 Revisão de Conceitos - Modelos Alternativos

Esta seção destina-se a apresentar uma contextualização dos modelos de contagem estudados neste capítulo.

3.2.1 Distribuições de Contagem Alternativas

Como apresentado por Castellares, Ferrari e Lemonte (2018), podemos dizer que $Y \sim Bell(\theta)$ quando sua função de distribuição é dada por

$$P_Y(y;\mu) = \frac{\theta^y e^{-e^\theta+1} B_y}{y!}, \quad y = 0,1,2,..., \tag{3.1}$$

com $E(Y) = \theta e^\theta$ e $V(Y) = \theta(1+\theta)e^\theta$, $\theta > 0$.

Dados os valores da média e variância, podemos dizer que $\frac{V(Y)}{E(Y)} = 1+\theta > 1$. Essa razão é chamada de Índice de Dispersão. A interpretação dessa medida é feita a partir do resultado da razão ser menor, igual ou maior do que 1. Sendo menor do que 1, há indício de que o modelo proveniente da distribuição em questão será bom para a tratativa de dados subdispersos. Sendo igual a 1, podemos dizer que estamos lidando com o modelo de Poisson, a medida em que média e variância, por conta da equidispersão imposta aos dados, devem possuir valores equivalentes. Agora, sendo maior que 1, temos indício de que teremos bons resultados para dados sobredispersos. Dado isso, a medida para distribuição Bell é maior do que 1, então caímos no terceiro caso e de mesmo modo, isso é aplicável para o modelo Binomial Negativa, por exemplo.

Outro ponto a ressaltar é o de que os B_y presentes (3.1) são os números de Bell, os quais são expressos como

$$B_y = \frac{1}{e}\sum_{k=0}^{\infty} \frac{k^y}{k!}, \quad k = 0,1,2,3,... \tag{3.2}$$

Podemos ainda relacionar os números de Bell à distribuição de Poisson. Dado que representam o n-ésimo momento de uma Poisson com parâmetro λ igual a 1 (vide Castellares, Ferrari e Lemonte (2018)).

Assim como nas variantes zero-inflacionadas para Poisson e Binomial Negativa, a distribuição Bell também pode ser associada a uma distribuição degenerada em zero para a tratativa de muitos zeros. Assim, segundo AJ Moreno-Arenas G (2020) podemos dizer que $Y \sim ZIBell(\mu,\omega)$ quando sua função de distribuição é dada por

$$P_Y(y;\mu,\omega) = \begin{cases} \omega + (1-\omega)\exp\left\{1-e^{W(\mu)}\right\}, & y = 0 \\ (1-\omega)\exp\left\{1-e^{W(\mu)}\right\}\dfrac{W(\mu)^y B_y}{y!}, & y = 1,2,..., \end{cases} \tag{3.3}$$

com $E(Y) = (1-\omega)\mu$ e $V(Y) = (1-\omega)\mu[1-W(\mu)+\mu\omega]$, $\mu > 0$, $0 < \omega < 1$.

Aqui, o índice de dispersão é dado por

$$I_{ZIBell} = \frac{V(Y)-E(Y)}{E(Y)} = W(\mu)+\mu\omega, \quad \mu > 0, \quad \omega \in (0,1).$$

Os autores pontuam a simplicidade da função massa de probabilidade dessa distribuição, dado que assim como Bell, não há a presença de funções complexas. Isso é um ponto positivo principalmente comparando com a distribuição ZIBN, que acaba necessitando de um parâmetro a mais (o parâmetro de dispersão) e possui uma função massa de probabilidade mais complexa. Caso o modelo ZIBell apresente resultados similares, mesmo que piores, por conta do critério de parcimônia seria interessante considerá-lo ao invés do modelo ZIBN.

A distribuição Poisson-Tweedie como o nome sugere é uma mistura entre uma parte Poisson e outra Tweedie. Aqui a parte Tweedie garante adicionar efeitos aleatórios no nível observacional de variáveis aleatórias Poisson. Assim, podemos partir da explanação da parte Tweedie, que neste caso se baseia no seguinte modelo de dispersão exponencial:

$$f_Z(z;\mu,\phi,b) = a(z,\phi,b)\exp\{(z\psi - k_b(\psi))/\phi\}. \tag{3.4}$$

com $E(Z) = \mu = k_b'(\psi)$ e $V(Z) = \phi V(\mu)$. Temos ainda, $\phi > 0$, que é o parâmetro de dispersão, ψ, chamado de parâmetro canônico e $k_b(\psi)$, a função cumulativa. Além disso, $V(\mu) = k"_b(\psi)$, que podemos chamar de função de variância.

No caso da Tweedie podemos caracterizá-la por funções potenciais de variância, na forma $V(\mu) = \mu^b$, em que $b \in (-\infty, 0] \cup [1, \infty)$, sendo b o indicador determinante da distribuição. Nesse sentido, há a dependência do suporte da distribuição no valor desse parâmetro potencial. Quando $b \geq 2$ temos suporte correspondente aos valores positivos, $1 < b < 2$, não negativos e $b = 0$, valores reais. Agora, quando o valor de b é negativo, então o suporte se mantém nos números reais, com esperança μ positiva. No nosso caso, com $b \geq 1$ temos $Tw_b(\mu,\phi)$ não negativa.

Jørgensen e Kokonendji (2015) em seu trabalho discutem sobre a função $a(z,\phi,b)$, que não possui forma fechada para ser expressa. Nesse sentido, apresentam alguns casos em que isso se torna possível, fixando o valor dos parâmetros. Assim: $b = 0$ corresponde a uma Gaussiana, $\phi = 1$ e $b = 1$, corresponde a Poisson, $b = 3/2$, corresponde a uma Gama não centralizada, $b = 2$, corresponde a uma Gama e finalmente $b = 3$, correspondendo a uma Inversa Gaussiana. Outra distribuição discutida por Bonat *et al.* (2018) é a Composta Poisson, obtida a partir de $1 < b < 2$. Essa distribuição é interessante para o cenário em que precisemos lidar com dados não negativos, com função massa de probabilidade zero e com grande assimetria à direita.

Ademais dos casos especiais que a Poisson-Tweedie contempla, cabe ressaltar que Bonat *et al.* (2018) e Saha *et al.* (2020) afirmam que a Poisson-Tweedie é uma distribuição muito

flexível, no sentido de ser possível acomodar o inflacionamento de zeros sem a necessidade de uma parte logística, como os demais modelos zero-inflacionados.

A partir desse panorama inicial, podemos dizer que $Y \sim PTw_b(\mu, \phi)$ Poisson-Tweedie quando sua função distribuição é dada pela seguinte especificação hierárquica:

$$\begin{aligned} Y|Z &\sim Poisson(Z), \\ Z &\sim Tw_b(\mu, \phi), \end{aligned} \tag{3.5}$$

com $b \geq 1$, garantindo que Z seja não-negativo. Com essa condição, podemos dizer que Poisson-Tweedie é um modelo de dispersão fatorial sobredisperso (Jørgensen e Kokonendji (2015)). Assim, a função massa de probabilidade para $Y \sim PTw_b(\mu, \phi)$, sendo $b > 1$, é dada por

$$P_Y(y; \mu, \phi, b) = \int_0^\infty \frac{z^y \exp\{-z\}}{y!} a(z, \phi, b) \exp\{(z\psi - k_b(\psi))/\phi\} dz. \tag{3.6}$$

Possuímos forma fechada para essa integral apenas nos casos em que $b = 2$, retornando uma Binomial Negativa. $b = 1$, caso em que a integral é substituível por uma soma, resultando em uma Neyman tipo A. Ademais, casos especiais incluem a composta Poisson ($1 < b < 2$), distribuição Stable fatorial positiva discreta ($b > 2$) e Poisson-inversa Gaussiana ($b = 3$), distribuições apresentadas por Bonat *et al.* (2018) e Kokonendji, Demétrio e Gbete (2004) em seus respectivos estudos.

Baseando-se em (3.4), mais especificamente na função cumulativa, temos que esperança e variância dessa distribuição, se resumem a

$$\begin{aligned} E(Y) &= \mu \\ V(Y) &= \mu + \phi \mu^b \end{aligned} \tag{3.7}$$

Bonat *et al.* (2018), propõe a avaliação numérica da integral, utilizando-se do método de Monte Carlo. Ao utilizarmos o método de Monte Carlo, é necessária a especificação de uma distribuição de proposta, a partir da qual amostras serão consideradas para cálculo da integral como expectativas. No caso da Poisson-Tweedie é interessante o uso da Tweedie como distribuição proposta. Assim, Bonat *et al.* (2018) indica que a vantagem do método numérico ao método via função massa de probabilidade é a reutilização dos valores para todas as avaliações da função massa de probabilidade, já que precisamos simular os valores apenas uma vez. Os códigos do processo de simulação da Poisson-Tweedie são encontrados no apêndice A.

3.2.2 Modelos de Regressão de Contagem Alternativos

Como mencionado anteriormente, a distribuição de Bell faz parte da família exponencial uniparamétrica. Um ponto a ressaltar é que a parametrização apresentada em (3.1) precisa ser adaptada para que possamos ficar com um aspecto mais familiar. Assim, seja $\mu = \theta e^{\theta}$, com $\theta = W_0(\mu)$, em que $W_0(\cdot)$ sendo a função Lambert. Logo, podemos reescrever a distribuição Bell da seguinte forma:

$$P_Y(y;\mu) = \exp(1 - e^{W_0(\mu)})\frac{W_0(\mu)^y B_y}{y!}, y = 0,1,2,..., \tag{3.8}$$

com $E(Y) = \mu$ e $V(Y) = \mu[1 + W_0(\mu)]$, $\mu > 0$ e B_y são os números de Bell, descritos em (3.2). Tomando uma amostra aleatória de $Y_1,...,Y_n$, em que $Y_i \sim Bell(\mu_i)$, com distribuição dada por (3.8), assim, devemos satisfazer

$$g(\mu_i) = \eta_i = \boldsymbol{x_i^T}\boldsymbol{\beta}, \;\; i = 1,...,n, \tag{3.9}$$

nosso preditor linear. No caso de Bell, algumas opções de função de ligação são as que seguem.

- logarítmica $g(\mu) = log(\mu)$;
- raiz quadrática $g(\mu) = \sqrt{\mu}$;
- identidade $g(\mu) = \mu$ (com especial atenção a positividade das estimativas).

Canonicamente, utiliza-se a função de ligação logarítmica e é baseado nela que os estudos de simulação e aplicação deste capítulo foram conduzidos.

Em suma esse modelo segue a seguinte estrutura:

- $Y_1...Y_n$ v.a independentes tais que $Y_i \sim Bell(\mu_i)$;
- $g(\mu_i) = log(\mu_i) = \eta_i$, em que $log(\mu_i)$ é a função de ligação canônica;
- $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta} = \beta_1 + \beta_2 x_{2i} + ... + \beta_p x_{pi}$, i = 1,...,n é o preditor linear, em que $x_i^T = (x_{i1}, \cdots, x_{ip})^T$ é um vetor de covariáveis, sendo x_{i1} o intercepto.

Para definirmos o modelo ZIBell, consideremos uma amostra aleatória $Y_1, \ldots, Y_n$ independentes, em que cada $Y_i \sim ZIBell(\mu_i, \omega_i)$, para $i = 1, \ldots, n$, que satisfaça as seguintes relações funcionais:

$$\begin{aligned} g_1(\mu_i) = log(\mu_i) = \eta_{1i} = \boldsymbol{x}_i^T\boldsymbol{\beta}, \\ g_2(\omega_i) = log(\tfrac{\omega_i}{1-\omega_i}) = \eta_{2i} = \boldsymbol{u}_i^T\boldsymbol{\delta}, \end{aligned} \tag{3.10}$$

Assim como para os demais modelos inflacionados, podemos assumir diferentes funções de ligações, tanto para a parte de zeros, quanto para a parte de contagem. Entretanto, optamos em utilizar as funções canônicas.

Em suma, podemos dizer que esse modelo segue a seguinte estrutura:

- $Y_1 ... Y_n$ v.a independentes tais que $Y_i \sim ZIBell(\mu_i, \omega_i)$;
- $g_1(\mu_i) = log(\mu_i) = \eta_{1i}$ em que $log(\mu_i)$ é uma função de ligação canônica;
- $\eta_{1i} = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_1 + \beta_2 x_{2i} + ... + \beta_p x_{pi}$, i = 1,...,n é o preditor linear para a taxa de resposta, em que $x_i^T = (x_{i1}, \cdots, x_{ip})^T$ é um vetor de covariáveis, sendo x_{i1} o intercepto;
- $g_2(\omega_i) = log(\omega_i / 1 - \omega_i) = \eta_{2i}$ em que $log(\omega_i / 1 - \omega_i)$ é uma função de ligação canônica;
- $\eta_{2i} = \boldsymbol{u}_i^T \boldsymbol{\delta} = \delta_1 + \delta_2 u_{2i} + ... + \delta_q u_{qi}$, i = 1,...,n é o preditor linear da média ou proporção de zeros, em que $u_i^T = (u_{i1}, \cdots, u_{iq})^T$ é um vetor de covariáveis, sendo u_{i1} o intercepto..

Como apresentado anteriormente, a função distribuição de uma Poisson-Tweedie é dada pela integral apresentada em (3.6). Um ponto negativo é o de que ela não possui forma fechada, se tornando intratável, a não ser pelos casos particulares. De todo modo, média e variância são fáceis de serem obtidos. Por conta disso, Bonat *et al.* (2018) especificaram um modelo baseado apenas em suposições de primeiro e segundo momentos. Chamamos esse modelo de Modelo de Regressão Poisson-Tweedie estendido. Considerando $Y_1, \ldots, Y_n$ independentes, temos:

$$Y_i \sim PTw_b(\mu_i, \phi), \text{ com } g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

de mesmo modo como para Poisson, Binomial Negativa e Bell, aqui a função ligação canônica é a logarítmica. Podemos escrever a esperança e a variância e, assim, o modelo Poisson-Tweedie estendido pode ser definido por

$$\begin{aligned} E(Y_i) = &\quad \mu_i = g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta}) \\ V(Y_i) = &\quad \mu_i + \phi \mu_i^b = C_i, \end{aligned} \tag{3.11}$$

logo, a parametrização desse modelo é dada por $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\lambda}^T)^T$, com $\boldsymbol{\lambda} = (\phi, b)$. Um ponto importante é que com base nas suposições de primeiro e segundo momentos, a única restrição para ter um modelo adequado é $V(Y_i) > 0$, de modo que

$$\begin{aligned} V(Y_i) &> 0 \\ \mu_i + \phi \mu_i^b &> 0 \\ \phi &> -\mu_i^{(1-b)}, \end{aligned} \tag{3.12}$$

de certa forma, valores negativos são permitidos para o parâmetro de dispersão. Isso possibilita a extensão do modelo Poisson-Tweedie para a tratativa de dados subdispersos, no entanto, não existe neste caso função massa de probabilidade associada. Porém, como temos interesse apenas nos valores dos coeficientes de Regressão, não há nenhuma perda interpretativa ou de aplicabilidade nesse caso.

É importante ressaltar que, neste caso, a relação entre média e variância é proporcional ao parâmetro de dispersão ϕ. Assim, esperamos que com $\phi < 0$ e $b = 1$ o modelo Poisson-Tweedie apresente resultados em muito próximos aos encontrados em outros modelos como Gama e Conway-Maxwell-Poisson, presentes nos trabalhos de Zeviani *et al.* (2014), Sellers, Borle e Shmueli (2012) e Bonat, Zeviani e Jr (2017), respectivamente.

Assim, o modelo Poisson-Tweedie estendido segue a seguinte estrutura:

- $g(\mu_i) = log(\mu_i) = \eta_i$, em que $log(\mu_i)$ é a função de ligação canônica;
- $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_1 + \beta_2 x_{2i} + ... + \beta_p x_{pi}$, i = 1,...,n é o preditor linear, em que $x_i^T = (x_{i1}, \cdots, x_{ip})^T$ é um vetor de covariáveis, sendo x_{i1} o intercepto.

3.2.3 Estimação dos Modelos de Regressão Alternativos

Baseando-se no método de Máxima Verossimilhança (vide 2.3.1), apresentamos aqui o processo de estimação dos modelos estudados neste capítulo.

3.2.3.1 Regressão Bell

Segundo Castellares, Ferrari e Lemonte (2018), a partir do método de Máxima Verossimilhança, escrevemos $l(\beta)$ com a nova parametrização adotada (vide 3.2.2), obtendo:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \ \log(W_0(\mu_i)) - e^{W_0(\mu_i)}],$$

em que $\mu_i = g^{-1}(\eta_i)$ é uma função de $\boldsymbol{\beta}$. A função escore é dada por

$$U(\boldsymbol{\beta}) = X^T \boldsymbol{W}^{1/2} \boldsymbol{V}^{-1/2} (\boldsymbol{y} - \boldsymbol{\mu}),$$

com X de posto completo, $\boldsymbol{W}$ e $\boldsymbol{V}$, vetores diagonais, com elementos dados por

$$wi = \frac{(d\mu_i / d_{\eta_i})^2}{V_i}, \ \ V_i = \mu_i[1 + W_0(\mu_i)], \ \ i = 1, 2, ..., n,$$

sendo V_i é função de variância de Y_i. A informação de Fisher, nesse caso pode ser expressa da seguinte forma $K(\boldsymbol{\beta}) = X^T \boldsymbol{W} X$. Para obtermos o estimador de Máxima Verossimilhança para o vetor de $\boldsymbol{\beta}$'s, basta igualar a função escore a um vetor de zeros, de modo que encontramos

estimativas para cada $\boldsymbol{\beta}$. Infelizmente esse estimador não apresenta forma fechada, sendo necessário o uso de métodos numéricos para a determinação, tal como Newton-Raphson. Aqui, cabe pontuar que o método de Escore de Fisher pode ser utilizado para estimar $\boldsymbol{\beta}$ a partir da solução iterativa da equação que segue:

$$\beta^{(m+1)} = (X^T \boldsymbol{W}^{(m)} X)^{-1} X^T \boldsymbol{W}^{(m)} \boldsymbol{z}^{(m)}, \tag{3.13}$$

em que $m = 0, 1, \ldots$ é um contador das iterações, $\boldsymbol{z} = \boldsymbol{\eta} + \boldsymbol{W}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu})$, tem o papel de variável modificada em (3.13).

A desviância para esse modelo é dada por $D = \sum_{i=1}^{n} d_i^2(y_i, \mu_i)$. Agora, para cada componente temos:

$$d_i^2(y_i, \hat{\mu}_i) = 2 \begin{cases} \exp\left\{1 - e^{W_0(\hat{\mu}_i)}\right\}, & y_i = 0, \\ e^{W_0(\hat{\mu}_i)} + e^{W_0(y_i)} + y_i \log\left(\dfrac{W_0(y_i)}{W_0(\hat{\mu}_i)}\right), & y_i > 0, \end{cases}$$

em que $\hat{\mu}_i = g^{-1}(\boldsymbol{x_i^T}\hat{\boldsymbol{\beta}})$, estimativa para μ_i. Sob condições de regularidade, pode-se afirmar que $D \sim^a \chi^2_{n-p}$. Tem-se que D é uma ótima candidata a ser utilizada como medida de bondade de ajuste para o modelo de regressão aplicado a dados reais. De modo que quanto menor o valor de D, melhor o nosso ajuste. O processo de estimação se encontra implementado no pacote bellreg (vide Tabela 19).

3.2.3.2 Regressão ZIBell

Considerando a parametrização $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\delta}^T)^T$, temos por objetivo estimar esse vetor. Desse modo podemos partir da função log-verossimilhança, que a menos de termos constantes, é dada por

$$l(\boldsymbol{\theta}) = \sum_{y_i : y_i = 0} \log\left[e^{\eta_{2i}} + \exp\left(1 - e^{W(\mu_i)}\right)\right] - \sum_{i=1}^{n} \log(1 - e^{\eta_{2i}}) + \sum_{y_i : y_i > 0} y_i \log[W(\mu_i)] - \sum_{y_i : y_i > 0} e^{W(\mu_i)},$$

com $\mu_i = e^{\eta_{1i}} = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$ para $i = 1, \ldots, n$. Agora, a estimativa de Máxima Verossimilhança $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}^T}, \hat{\boldsymbol{\delta}^T})^T$ de $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\delta}^T)^T$, em que $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ e $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, \ldots, \hat{\delta}_q)^T$, é obtida pela maximização da função log-verossimilhança $l(\boldsymbol{\theta})$ com respeito a $\boldsymbol{\theta}$.

A função escore, obtida pela diferenciação da função de log-verossimilhança com respeito aos parâmetros desconhecidos, é dada pelo vetor $(p+q)$ $U(\boldsymbol{\beta}, \boldsymbol{\delta}) = (U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\delta})^T, U_{\boldsymbol{\delta}}(\boldsymbol{\beta}, \boldsymbol{\delta})^T)^T$, em que $U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\delta}) = \boldsymbol{X}^T \boldsymbol{\varepsilon}$, $U_{\boldsymbol{\delta}}(\boldsymbol{\beta}, \boldsymbol{\delta}) = \boldsymbol{S}^T \boldsymbol{\gamma}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ e $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^T$, com

$$\varepsilon_i = \begin{cases} -\dfrac{\exp(1+\eta_{1i}-e^{W(\mu_i)})}{[e^{\eta_{2i}}+\exp(1-e^{W(\mu_i)})][1+W(\mu_i)]} & ,y_i=0, \\ \dfrac{y_i-\mu_i}{1+W(\mu_i)} & ,y_i>0, \end{cases}$$

$$\gamma_i = \frac{e^{\eta_{2i}} I(y_i=0)}{e^{\eta_{2i}}+\exp(1-e^{W(\mu_i)})} - \frac{e^{\eta_{2i}}}{1+e^{\eta_{2i}}},$$

em que $I(\cdot)$ representa a função indicadora. As estimativas de MV $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1,\ldots,\hat{\beta}_p)^T$ e $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1,\ldots,\hat{\delta}_q)^T$ podem ser obtidas também, como AJ Moreno-Arenas G (2020) propõem, pela solução do sistema não linear de equações $U_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}}) = \mathbf{0_p}$ e $U_{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}}) = \mathbf{0_q}$, em que $\mathbf{0_k}$ denota um vetor k-dimensional de zeros. Nesse sistema, as estimativas de MV precisam ser obtidas através de uma maximização numérica, usando algoritmos de otimização não linear, como é o caso do Newton-Raphson. Esse algoritmo, por ser iterativo, demanda especificação de valores iniciais. AJ Moreno-Arenas G (2020) sugerem utilizar como chutes iniciais estimativas obtidas a partir de um modelo ZIP.

Para o modelo ZIBell, valem as propriedades assintóticas, de modo que os estimadores via MV são assintoticamente normais, não viesados e possuem matriz de variância e covariância dada pelo inverso da matriz da informação de Fisher. Seja $K(\boldsymbol{\beta},\boldsymbol{\delta})$ a matriz de informação de Fisher $((p+q)\times(p+q))$ para $(\boldsymbol{\beta},\boldsymbol{\delta})$. Com n grande e sob as condições de regularidade, temos que

$$\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{\delta}}\end{pmatrix} \underset{\sim}{a} N_{p+q}\left(\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\delta}\end{pmatrix}, K(\boldsymbol{\beta},\boldsymbol{\delta})^{-1}\right),$$

em que $'\underset{\sim}{a}'$ significa aproximadamente distribuído. Podemos expressar $K(\boldsymbol{\beta},\boldsymbol{\delta})$, portanto, da seguinte maneira:

$$K(\boldsymbol{\beta},\boldsymbol{\delta}) = \begin{bmatrix} X^T\boldsymbol{W}_1X & X^T\boldsymbol{W}_2S \\ S^T\boldsymbol{W}_2X & S^T\boldsymbol{W}_3S \end{bmatrix},$$

com $\boldsymbol{W}_1 = diag\{w_{1i}\}$, $\boldsymbol{W}_2 = diag\{w_{2i}\}$, $\boldsymbol{W}_3 = diag\{w_{3i}\}$ e $diag\{a_i\}$, que representa uma matriz diagonal com elemento típico a_i $(i=1,\ldots,n)$. Seja P a matriz diagonal $(2n\times(p+q))$

$$P = \begin{bmatrix} X & 0_{n,q} \\ 0_{n,p} & S \end{bmatrix},$$

em que $0_{l,c}$ denota uma matriz $(l\times c)$ de zeros. Consideremos M uma matriz $(2n\times 2n)$

$$M = \begin{bmatrix} W_1 & W_2 \\ W_2 & W_3 \end{bmatrix}.$$

Desse modo, podemos expressar $K(\boldsymbol{\beta},\boldsymbol{\delta}) = P^T MP$. A distribuição normal assintótica acima pode ser usada para construção de intervalos de confiança aproximados para os parâmetros. Considerando $\beta_r (r = 1,\ldots,p)$ e $\delta_j (j = 1,\ldots,q)$ os componentes r-ésimo e j-ésimo de $\boldsymbol{\beta}$ e $\boldsymbol{\delta}$. Para $0 < \alpha < 1/2$, temos intervalos assintóticos, tais como $\hat{\beta}_r \pm z_{1-\alpha/2} se(\beta_r)$ e $\hat{\delta}_j \pm z_{1-\alpha/2} se(\hat{\delta}_j)$, ambos com cobertura assintótica de $100(1-\alpha)\%$. Cabe ressaltar que $se(\cdot)$ é a raiz quadrada do elemento diagonal de $K(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}})^{-1}$ correspondente a cada parâmetro, ie. o erro padrão assintótico, e $z_{1-\alpha/2}$ denota o $(1-\alpha/2)$-ésimo quantil da distribuição normal padrão.

3.2.3.3 Regressão Poisson-Tweedie

Dada a função massa de probabilidade presente em 3.4, temos que a função log-verossimilhança para uma variável Tweedie, pode ser escrita da forma:

$$l(\phi, p) = \sum_{i=1}^{n} \log f_Z(z; \mu, \phi, b)$$

Agora, uma vez que o modelo apresentado em (3.11) baseia-se unicamente em suposições de primeiro e segundo momento, o método de Máxima Verossimilhança não pode ser implementado. Bonat e Jørgensen (2016) baseados em ideias de Kokonendji, Demétrio e Gbete (2004) combinam, dessa forma, duas funções para obtenção das estimativas dos parâmetros, elas: quase-escore e Pearson. Desse modo, temos que a função quase-escore proposta tem a seguinte forma

$$\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta},\boldsymbol{\lambda}) = \left(\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_1} C_i^{-1} (y_i - \mu_i)^T, \ldots, \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_Q} C_i^{-1} (y_i - \mu_i)^T \right)^T,$$

em que $\partial \mu_i / \partial \beta_j = \mu_i x_{ij}$ para $j = 1,\ldots,Q$. Cabe reforçar que de (3.11), temos $C_i = \mu_i + \phi \mu_i^b$. Segundo Bonat *et al.* (2018), a matriz de sensibilidade é definida como a esperança da primeira derivada da função de estimação com respeito aos parâmetros do modelo. Assim, uma entrada (j,k) de uma matriz de sensibilidade $Q \times Q$ para ψ_β é dada por

$$S_{\beta_{jk}} = E\left(\frac{\partial}{\partial \beta_k} \psi_{\beta_j}(\beta,\lambda) \right) = -\sum_{i=1}^{n} \mu_i x_{ij} C_i^{-1} x_{ik} \mu_i. \tag{3.14}$$

Similarmente, a matriz de variabilidade é definida pela variância da função de estimação. Em particular, para a função quase-escore a entrada (j,k) de uma matriz de variabilidade $Q \times Q$ é dada por

$$V_{\beta_{jk}} = V(\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta},\boldsymbol{\lambda})) = \sum_{i=1}^{n} \mu_i x_{ij} C^{-1} x_{ik} \mu_i.$$

A função de estimação de Pearson para os parâmetros de dispersão fica da seguinte forma

$$\psi_{\boldsymbol{\lambda}}(\boldsymbol{\lambda},\boldsymbol{\beta}) = \left(\sum_{i=1}^{n} W_{i\phi}[(y_i-\mu_i)^2 - C_i]^T, \sum_{i=1}^{n} W_{ib}[(y_i-\mu_i)^2 - C_i]^T \right)^T,$$

sendo $W_{i\phi} = -\partial C_i^{-1}/\partial\phi$ e $W_{ib} = -\partial C_i^{-1}/\partial b$. Note que, de acordo com Bonat *et al.* (2018), as funções de estimação de Pearson são não viesadas para estimar $\boldsymbol{\lambda}$ baseando-se nos resíduos quadráticos $(y_i-\mu_i)^2$ com valor esperado C_i. Agora, uma entrada (j,k) de uma matriz de sensibilidade para os parâmetros de dispersão é dada por

$$S_{\lambda_{jk}} = E\left(\frac{\partial}{\partial\lambda_k}\psi_{\lambda_j}(\boldsymbol{\lambda},\boldsymbol{\beta})\right) = -\sum_{i=1}^{n} W_{i\lambda_j} C_i W_{i\lambda_k} C_i, \tag{3.15}$$

com λ_1 e λ_2 denotando tanto ϕ ou b. De maneira equivalente, entradas cruzadas para a matriz de sensibilidade são dadas por

$$S_{\beta_j\lambda_k} = E\Big(\frac{\partial}{\partial\lambda_k}\psi_{\beta_j}(\boldsymbol{\beta},\boldsymbol{\lambda})\Big) = 0 \tag{3.16}$$

e

$$S_{\lambda_j\beta_k} = E\left(\frac{\partial}{\partial\beta_k}\psi_{\lambda_j}(\boldsymbol{\lambda},\boldsymbol{\beta})\right) = -\sum_{i=1}^{n} W_{i\lambda_j} C_i W_{i\beta_k} C_i. \tag{3.17}$$

com $W_{i\beta_k} = \partial C_i^{-1}/\partial\beta_k$. Por fim, a matriz de sensibilidade para o vetor de parâmetro $\boldsymbol{\theta}$ é dada por

$$S_{\boldsymbol{\theta}} = \begin{pmatrix} S_{\boldsymbol{\beta}} & \mathbf{0} \\ S_{\boldsymbol{\lambda\beta}} & S_{\boldsymbol{\lambda}} \end{pmatrix},$$

cujas entradas estão definidas pelas equações (3.14), (3.15), (3.16) e (3.17). Ainda de acordo com Bonat *et al.* (2018) é possível calcularmos a variância assintótica dos estimadores $\hat{\boldsymbol{\theta}}$, obtidos da inversa da matriz de informação de Godambe, cuja forma geral para o parâmetro $\boldsymbol{\theta}$ é $J_{\boldsymbol{\theta}}^{-1} = S_{\boldsymbol{\theta}}^{-1} V_{\boldsymbol{\theta}} S_{\boldsymbol{\theta}}^{-T}$, em que $-T$ denota uma transposição inversa. Assim, a matriz de variabilidade para θ pode ser escrita como

$$V_{\boldsymbol{\theta}} = \begin{pmatrix} V_{\boldsymbol{\beta}} & V_{\boldsymbol{\beta\lambda}} \\ V_{\boldsymbol{\lambda\beta}} & V_{\boldsymbol{\lambda}} \end{pmatrix}, \tag{3.18}$$

em que $V_{\boldsymbol{\lambda\beta}} = V_{\boldsymbol{\beta\lambda}}^T$ e $V_{\boldsymbol{\lambda}}$ depende do terceiro e quarto momentos de Y_i. Para evitar esta dependência para momentos de maior grandeza, utilizamos as versões empíricas de $V_{\boldsymbol{\lambda}}$ e $V_{\boldsymbol{\lambda\beta}}$, sendo estas

$$\tilde{V}_{\lambda_{jk}} = \sum_{i=1}^{n} \psi_{\lambda_j}(\boldsymbol{\lambda},\boldsymbol{\beta})_i \psi_{\lambda_k}(\boldsymbol{\lambda},\boldsymbol{\beta})_i, \quad \tilde{V}_{\lambda_j\beta_k} = \sum_{i=1}^{n} \psi_{\lambda_j}(\boldsymbol{\lambda},\boldsymbol{\beta})_i \psi_{\beta_k}(\boldsymbol{\lambda},\boldsymbol{\beta})_i.$$

Jørgensen e Knudsen (2004) descrevem em seu trabalho a distribuição assintótica de $(\hat{\boldsymbol{\theta}})$, dada por

$$(\hat{\boldsymbol{\theta}}) \sim N(\boldsymbol{\theta}, J_{\boldsymbol{\theta}}^{-1}), \text{ em que } J_{\boldsymbol{\theta}}^{-1} = S_{\boldsymbol{\theta}}^{-1} V_{\boldsymbol{\theta}} S_{\boldsymbol{\theta}}^{-T}.$$

Para resolver o sistema de equações $\psi_{\boldsymbol{\beta}} = \mathbf{0}$ e $\psi_{\boldsymbol{\lambda}} = \mathbf{0}$, Jørgensen e Knudsen (2004) propuseram o algoritmo de busca modificado.

Algoritmo 1 – Método de busca modificado proposto por Bonat *et al.* (2018)

1: $\beta^{(i+1)} = \beta^{(i)} - S_{\beta}^{-1} \psi_{\beta}(\beta^{(i)}, \lambda^{(i)})$;
2 : $\lambda^{(i+1)} = \lambda^{(i)} - \alpha \ S_{\lambda}^{-1} \psi_{\lambda}(\beta^{(i+1)}, \lambda^{(i)})$.

Esse algoritmo (Algoritmo 1) de busca modificado utiliza da propriedade de insensibilidade (3.16), que nos permite usar duas equações separadas para atualização de β e λ. α aqui serve como um controle para o comprimento do passo. Assim, a estimativa para o modelo Poisson-Tweedie estendido é facilmente implementada através do pacote mcglm presente em 3.2.4.

3.2.4 Implementação Computacional dos Modelos Alternativos

Tabela 19 – Pacotes para regressão de contagem utilizados.

Pacotes - R	
Clássicos	
GAMLSS	https://cran.r-project.org/web/packages/gamlss/index.html
bellreg	https://cran.r-project.org/web/packages/bellreg/index.html
mcglm	https://cran.r-project.org/web/packages/mcglm/index.html

A Tabela 19 mostra os pacotes disponíveis em linguagem R, utilizados para estimar os modelos Bell, ZIBell (bellreg) e Poisson Tweedie (mcglm), respectivamente. Apresentamos na Tabela o pacote GAMLSS, dado que para o estudo de simulação o utilizamos para realizar estimações para Poisson e Binomial Negativa.

Como apresentado na subseção 3.2.2, temos aqui o uso do modelo estendido, já que ele é mais flexível em termos de aceitabilidade de valores para o parâmetro de dispersão e igualmente já está presente implementado em linguagem **R**. Pontua-se que as estimações desses pacotes se baseiam na ótica clássica, como já discutido anteriormente, desse modo, tanto o estudo de simulação, comparativo com os modelos Poisson e Binomial Negativa (aqui não levando em conta suas variantes inflacionadas), quanto a aplicações, que levam em conta o modelo mais bem ajustado - ZIBN - serão baseadas somente nessa abordagem para o modelo PTw, Bell e ZIBell serão avaliados segundo ambas as metodologias.

3.3 Estudos de Simulação - Modelos Alternativos

Esta Seção destinasse a apresentar o estudo de simulação comparativo entre a similaridade das estimativas produzidas pelos pacotes GAMLSS, bellreg e mcglm para os modelos Poisson, Binomial Negativa, Bell e Poisson-Tweedie, respectivamente.

3.3.1 Estudo 1 - Similaridade dos ajustes de diferentes pacotes utilizando réplicas

Considerando os pacotes GAMLSS, bellreg e mcglm, presentes na Tabela 19, conduzimos um estudo de simulação investigando se a partir de um conjunto de dados de um modelo, utilizando outro, conseguimos recuperá-lo. Nesse sentido, simulamos amostras de tamanho 500 para as distribuições Poisson, Binomial Negativa, Bell e Poisson-Tweedie, considerando 100 réplicas. Ao todo foram gerados 400 conjuntos de dados diferentes, sendo 100 para cada distribuição. Foram mantidas aqui as premissas dos valores dos parâmetros do primeiro estudo do capítulo anterior (vide 3.3.1), de modo que $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_2 = -1$. Binomial Negativa e Poisson-Tweedie apresentam parâmetros próprios, ϕ - parâmetro de dispersão. Sendo assim, para ambas, esse parâmetro foi fixado. Para a primeira distribuição em 2 e para segunda, 1.

Cabe pontuar que este estudo foi realizado com olhar clássico, pois no momento de construção dessa análise, não foi identificada abordagem Bayesiana para o modelo PTw. Assim, preferiu-se comparar os modelos a partir do olhar clássico.

Em todos os casos foram consideradas duas variáveis explicativas, variando entre Normal Padrão ($N(0,1)$), Uniformes ($U_{[-1,1]}$ e $U_{[0,1]}$), Binomial ($B(500,0,5)$) e no caso da Poisson-Tweedie, uma Tweedie ($TW_b(10,1)$). Então, para ajustes Poisson e Binomial Negativa as variáveis explicativas consideradas foram $B(500,0,5)$ e $U_{[-1,1]}$. Para o ajuste Bell foram consideradas $N(0,1)$ e $U_{[0,1]}$. Por fim, para o ajuste PTw foram consideradas $N(0,1)$ e $TW_b(10,1)$.

Realizamos os ajustes dos respectivos modelos de regressão para Poisson, Binomial Negativa, Bell e Poisson-Tweedie, totalizando de 1600 ajustes, sendo 400 para cada distribuição, 100 para cada modelo. Foram coletadas as médias, desvios padrões, REQM das réplicas de cada distribuição, medidas expressas na Tabela 20.

Para dados gerados de uma Poisson, temos similaridade de todos os ajustes realizados. Observando os REQM coeficiente a coeficiente, entendemos que o modelo Binomial Negativa acaba tendo um erro maior no β_1, mas um erro também mais alto no β_2, sendo superado apenas pelo modelo Poisson-Tweedie. De todo modo, observando exclusivamente as médias, não temos destoamentos do verdadeiro valor dos parâmetros. O que já é um indicativo que qualquer modelo utilizado para dados de origem Poisson, poderiam ser bons substitutos, porém chamando atenção para Bell, que tem equivalência teórica com a Poisson para valores pequenos de θ. Observando a cobertura dos intervalos assintóticos (Tabela 16) e percebemos que Binomial Negativa apresenta

Tabela 20 – Resultados das simulações de dados Poisson, Binomial Negativa (GAMLSS), Bell (bellreg) e Poisson-Tweedie (mcglm).

Dados Poisson														
Distribuição	Poisson			Binomial Negativa				Bell			Poisson-Tweedie			
Medidas	β_0	β_1	β_2	β_0	β_1	β_2	ϕ	β_0	β_1	β_2	β_0	β_1	β_2	ϕ
Média	0,50	1,00	-1,01	0,50	1,00	-1,00	1,64	0,50	1,00	-1,01	0,50	1,00	-1,01	0,00
Desvio Padrão	0,06	0,07	0,09	0,06	0,11	0,12	0,16	0,06	0,07	0,09	0,06	0,07	0,09	0,08
REQM	0,06	0,07	0,09	0,05	0,10	0,12	-	0,05	0,07	0,09	0,05	0,07	0,13	-
Dados Binomial Negativa														
Distribuição	Poisson			Binomial Negativa				Bell			Poisson-Tweedie			
Medidas	β_0	β_1	β_2	β_0	β_1	β_2	ϕ	β_0	β_1	β_2	β_0	β_1	β_2	ϕ
Média	0,50	1,00	-0,99	0,50	1,00	-0,99	2,17	0,50	0,99	-0,99	0,50	0,99	-0,99	0,49
Desvio Padrão	0,07	0,10	0,11	0,07	0,09	0,11	0,63	0,07	0,09	0,11	0,07	0,09	0,11	0,16
REQM	0,07	0,10	0,11	0,07	0,09	0,11	-	0,07	0,09	0,11	0,07	0,09	0,11	-
Dados Bell														
Distribuição	Poisson			Binomial Negativa				Bell			Poisson-Tweedie			
Medidas	β_0	β_1	β_2	β_0	β_1	β_2	ϕ	β_0	β_1	β_2	β_0	β_1	β_2	ϕ
Média	0,50	1,01	-1,02	0,50	1,01	-1,03	1,78	0,50	1,01	-1,03	0,50	1,01	-1,03	0,54
Desvio Padrão	0,10	0,07	0,18	0,10	0,06	0,18	0,44	0,10	0,06	0,18	0,10	0,06	0,18	0,13
REQM	0,10	0,07	0,18	0,10	0,06	0,18	-	0,10	0,07	0,18	0,10	0,07	0,18	-
Dados Poisson-Tweedie														
Distribuição	Poisson			Binomial Negativa				Bell			Poisson-Tweedie			
Medidas	β_0	β_1	β_2	β_0	β_1	β_2	ϕ	β_0	β_1	β_2	β_0	β_1	β_2	ϕ
Média	0,49	1,01	-1,00	0,50	0,94	-0,92	1,95	0,50	0,94	-0,35	0,50	1,01	-1,00	2,04
Desvio Padrão	0,09	0,06	0,05	0,07	1,15	0,74	0,53	0,06	0,15	0,07	0,08	0,06	0,05	0,53
REQM	0,09	0,06	0,05	0,07	1,15	0,74	-	0,06	0,16	0,65	0,08	0,06	0,05	-

Tabela 21 – Probabilidades de cobertura dos Intervalos de Confiança assintóticos.

Dados - Poisson			
Pacotes/medidas	β_0	β_1	β_2
Poisson - GAMLSS	94,00%	95,00%	97,00%
BN - GAMLSS	87,00%	89,00%	87,00%
Bell - bellreg	95,00%	95,00%	87,00%
PTw - mcglm	90,00%	99,00%	100,00%
Dados - Binomial Negativa			
Poisson - GAMLSS	98,00%	97,00%	99,00%
BN - GAMLSS	94,00%	98,00%	95,00%
Bell - bellreg	89,00%	100,00%	88,00%
PTw - mcglm	95,00%	93,00%	92,00%
Dados - Bell			
Poisson - GAMLSS	99,00%	88,00%	85,00%
BN - GAMLSS	97,00%	80,00%	88,00%
Bell - bellreg	99,00%	92,00%	95,00%
PTw - mcglm	89,00%	83,00%	80,00%
Dados - Poisson-Tweedie			
Poisson - GAMLSS	92,00%	87,00%	89,00%
BN - GAMLSS	84,00%	78,00%	83,00%
Bell - bellreg	91,00%	82,00%	66,00%
PTw - mcglm	96,00%	98,00%	97,00%

menor probabilidade de conter o verdadeiro valor dos parâmetros simulados nos intervalos calculados (vide Tabela 21). Ressaltando que novamente, como nos demais estudos, o nível de confiança nominal foi de 95%.

Ao observarmos os resultados para dados segundo Binomial Negativa, permanece a similaridade entre todos os modelos estudados. Podemos apenas destacar um erro ligeiramente maior no coeficiente β_1, utilizando o modelo Poisson. Bell e Poisson-Tweedie tiveram resultados equivalentes. Em conformidade com os estudos anteriores podemos indicar esses dois modelos como boas alternativas aos modelos Poisson e Binomial Negativa. Aqui, qualquer um dos modelos teve probabilidades de cobertura se não maiores, muito próximas ao nível nominal, o que é bom indicativo de que ademais da escolha, eles comportarão os valores verdadeiros dos parâmetros, ou seja é possível a recuperação dos dados por qualquer modelo utilizado (vide Tabela 21).

Considerando resultados mostrados na Tabela 20, quando dados foram gerados do modelo de regressão Bell, se ajustamos os dados com a regressão Poisson, Binomial Negativa e Poisson-Tweddie os coeficientes de regressão nestes modelos são muito próximos dos verdadeiros valores correspondentes ao modelo de regressão Bell, sinalizando que os coeficientes de regressão não são afetados por causa do modelo de contagem, mas certamente os modelos são diferentes, já que a Binomial negativa e o modelo Poisson-Tweddie estimam o parâmetro ϕ que não e parte do modelo de regressão Bell. No caso das probabilidades de cobertura (Tabela 21), Poisson-Tweedie acaba apresentando valores menores, indicando que o modelo Poisson e Binomial Negativa podem apresentar resultados mais precisos e próximos aos do modelo originário.

Para dados segundo uma Poisson-Tweedie, podemos ver que o modelo mais próximo foi o modelo Poisson. Bell acaba com estimativas ruins para o parâmetro β_2, podendo ser comprovado tanto pela média, quanto pelo erro. Apesar das estimativas do modelo Binomial Negativa não ficarem longe das esperadas em termos de erro, há um destoamento considerável dos demais modelos. Em algumas das repetições os valores para os parâmetros β_1 e β_2 foram inadequados e isso acabou influenciando negativamente para a média final das repetições e igualmente as estimativas intervalares, resultando em probabilidades de cobertura menores (Tabela 21). Como majoritariamente a natureza dos dados não é conhecida, entendemos que somente a Poisson teve resultados bem equivalentes aos realizados pelo modelo Poisson-Tweedie.

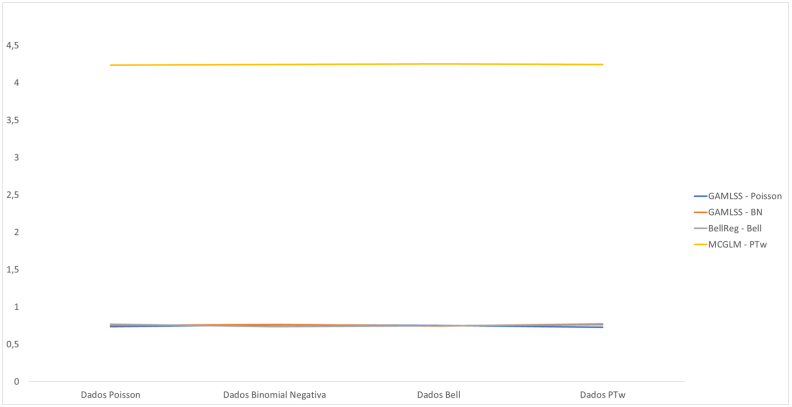

Figura 9 – Tempos médios em segundos considerando os ajustes realizados para cada origem de dados.

A Figura 9 apresenta os tempos computacionais médios em segundos de acordo com cada ajuste e origem de dados. Temos a evidência de que ademais da origem dos dados, os tempos computacionais são equivalentes para a realização dos ajustes, assim não sendo um fator predominante para justificar o uso de um modelo ou de outro.

3.4 Considerações do Capítulo - Modelos Alternativos

Este Capítulo apresentou duas alternativas aos modelos Poisson e Binomial Negativa, Bell e PTw, além da variante zero-inflacionada do modelo Bell. O estudo de simulação apresentado foi conduzido sob ótica clássica. Comparamos os modelos Poisson, Binomial Negativa, Bell e PTw através do pacote determinado como melhor para a abordagem clássica, no Capítulo anterior, o GAMLSS, e os pacotes bellreg e mcglm, utilizando-se de réplicas para mensurar média, desvios padrões e REQM's das estimativas dos parâmetros simulados. Além dos tempos computacionais de cada ajuste. Observamos também as probabilidades de cobertura assintóticas para cada um dos parâmetros simulados de cada modelo. Como conclusão, os modelos Bell e PTw se confirmaram boas alternativas aos modelos Poisson e Binomial Negativa, o que valida os estudos anteriores. Porém, o caminho inverso nem sempre é razoável, pontuando que apenas se mostrou válido para o modelo Bell.

Por fim, o código do estudo de simulação está presente no Apêndice A.

CAPÍTULO

4

APLICAÇÃO

Retomando os resultados obtidos nos estudos de simulação dos capítulos passados, temos que para abordagem clássica, o pacote GAMLSS foi o mais indicado como pacote com melhores estimativas e para abordagem Bayesiana, temos o brms. No que se refere aos modelos Bell e Poisson-Tweedie, confirmamos que são, sim, boas alternativas aos modelos Poisson e Binomial Negativa.

Nesse sentido, num primeiro momento, comparamos os modelos do capítulo 2 e num segundo momento os modelos do capítulo 3.

4.1 Etapa 1 - Modelos Inflacionados

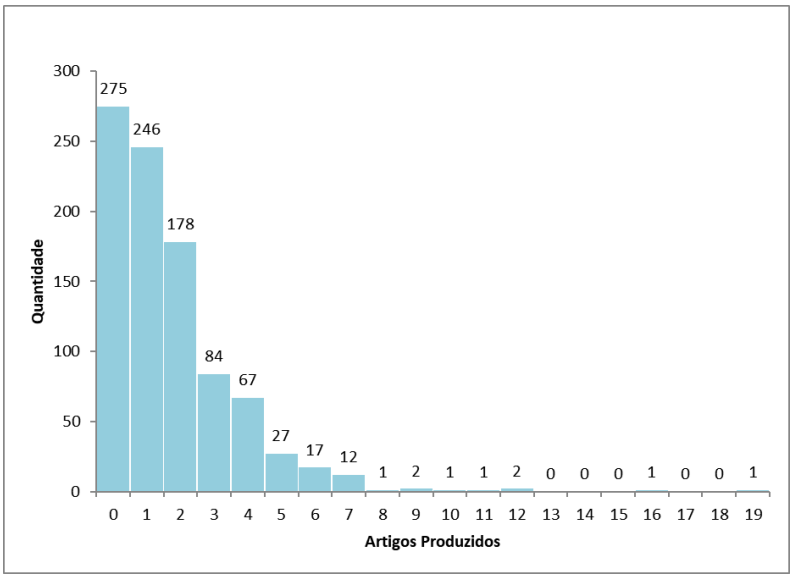

Figura 10 – Histograma referente ao número de publicações dos alunos de pós-graduação em bioquímica.

Com os resultados dos cenários simulados em mente, escolhemos um conjunto de dados para analisar. O conjunto selecionado foi o bioChemists, no qual 915 estudantes de pós-graduação em Bioquímica foram escolhidos aleatoriamente com o intuito de coletar as seguintes informações: sexo (*fem*), número de filhos (*kid5*), estado civil (*mar*), prestígio do instituto (*phd*), o qual o aluno faz parte, o número de artigos produzidos pelo seu orientador (*ment*) e por fim a produção de artigos dos próprios alunos (*art*). Long (1990) em seu estudo estava interessado em entender quais fatores influenciam na quantidade de artigos produzidos dos alunos. E de mesmo modo, o exposto neste capítulo segue as mesmas diretrizes.

Podemos ver na Figura 10 a variável resposta. Há a predominância de zeros na amostra, o que acaba representando aproximadamente 30% da volumetria total. Essa informação é importante, pois, já evidencia que provavelmente os modelos inflacionados terão melhores resultados comparando-os com os modelos não inflacionados. Outro ponto importante é a presença de sobredispersão nessa variável (variância = $3,71$ e média = $1,69$). Nesse sentido, o modelo Poisson já não é o mais indicado, dado que para melhores resultados, as quantidades devem ser equivalentes, ou ao menos similares.

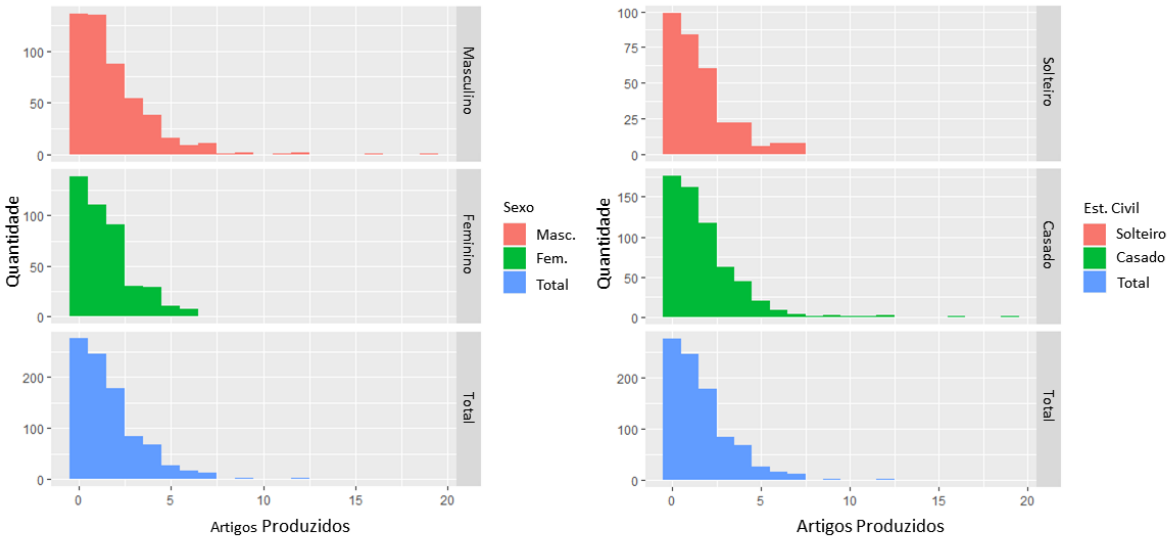

Figura 11 – Histogramas de sexo (à esquerda) e estado civil (à direita) relacionados com a variável resposta.

Na Figura 11 apresentamos aberturas para o sexo e o estado civil dos alunos. Visualmente parece haver diferença entre os níveis das duas variáveis. Indicando possibilidade de homens e de casados serem públicos que mais publicam artigos.

A partir da Figura 12, verificamos se de fato há diferença entre os níveis das variáveis sexo e estado civil. Os boxplots ajustados são gráficos interessantes para essa verificação, à medida em

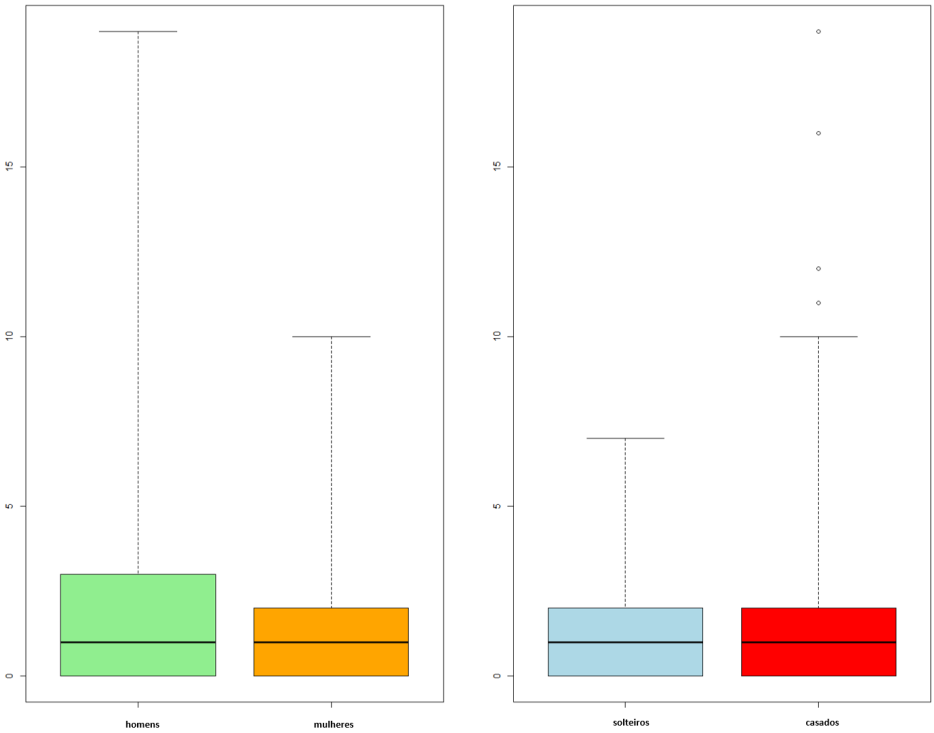

Figura 12 – Boxplots ajustados dos níveis das variáveis sexo e estado civil de acordo com a variável resposta.

que já estão removidos quaisquer outlier do conjunto. Como podemos ver, o indício inicialmente a favor de maior produtividade de homens e casados não representa diferença estatística com mulheres e solteiros, dado que as médias são muito próximas em todos os casos.

Na Figura 13 temos as aberturas do número de filhos dos alunos. Há indício de que quanto maior o número de filhos menos o aluno produzirá.

Com relação ao prestígio do instituto do aluno e quantidade de artigos produzidos pelos orientadores, calculamos a correlação linear de Pearson existente entre essas variáveis e a variável resposta. Como conclusão, obtivemos que para a primeira variável há uma correlação de 7% e para a segunda variável, uma correlação de 30%. Desse modo, entendemos que a correlação entre a produtividade dos alunos e o prestígio do instituto é baixa, indicando que essa variável não seria interessante para ser covariável nos modelos a serem ajustados. Já a produtividade do orientador é mais relevante, apresentando uma correlação moderada com a variável resposta e, portanto, indicando que a produtividade do orientador influencia na produtividade final do aluno.

Como passo seguinte, realizamos um ajuste inicial utilizando sexo, estado civil, número de filhos e a produtividade do orientador como covariáveis. Foram obtidos os valores de AIC e EAIC presentes na Tabela 22.

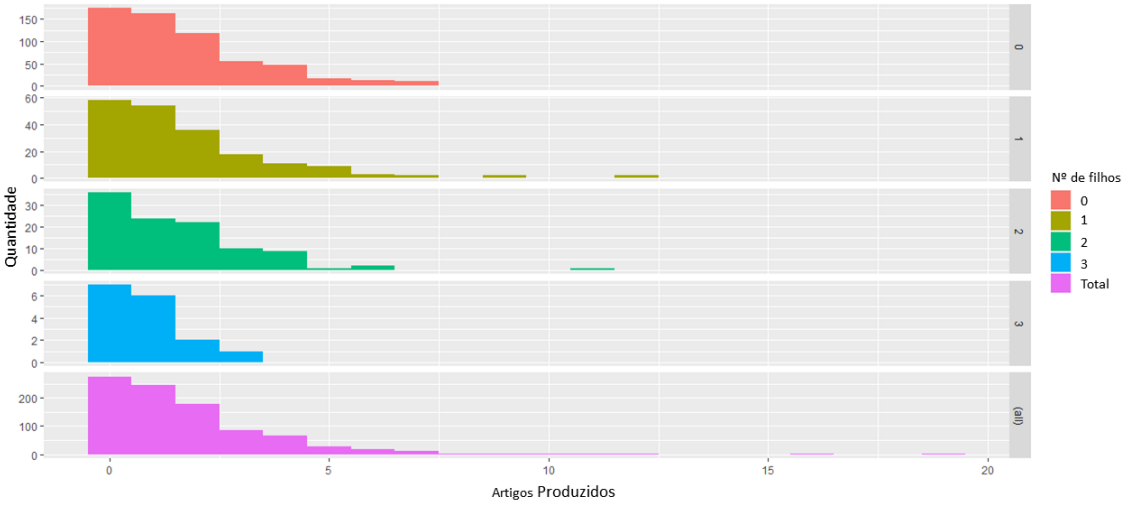

Figura 13 – Histogramas do número de filhos relacionados com a variável resposta.

Tabela 22 – AIC e EAIC para os modelos estudados a partir de GAMLSS e brms na aplicação.

Modelos	GAMLSS	brms
	AIC	EAIC
Poisson	3312,3	3328,6
Binomial Negativo	3136,1	3144,2
ZI-Poisson 1 parte	3253,5	3268,7
ZI-Poisson 2 partes	3229,5	3246,0
ZI-Binomial Negativo 1 parte	3253,5	3136,3
ZI-Binomial Negativo 2 partes	3123,5	3127,2

Assim, os menores AIC's foram obtidos pelos modelos que possuem a Binomial Negativa em sua composição. Como modelo mais aderente, temos portanto o modelo ZIBN2. Comparando entre metodologias de estimação, temos ainda que o melhor resultado foi realizado sob ótica clássica. Nesse sentido, trazemos na Tabela 23 o resumo do ajuste clássico.

Tabela 23 – Modelo ZIBN2 via estimação clássica - pacote GAMLSS

ZIBN2 - sob ótica clássica (GAMLSS)				
	Função de ligação μ: log			
Coeficientes μ:	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
(Intercepto)	$0,41$	$0,09$	$4,74$	$< 0,05$
feemWomen	$-0,19$	$0,07$	$-2,58$	$< 0,05$
marMarried	$0,10$	$0,08$	$1,17$	$> 0,05$
kid5	$-0,15$	$0,05$	$-2,79$	$< 0,05$
ment	$0,02$	$3,40 \times 10^{-3}$	$7,27$	$< 0,05$
	Função de ligação ν: logit			
Coeficientes ν:	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
(Intercepto)	$-0,32$	$0,83$	4,74	$> 0,05$
feemWomen	$0,66$	$0,82$	-2,58	$> 0,05$
marMarried	$-1,47$	$0,89$	1,17	$> 0,05$
kid5	$0,62$	$0,44$	-2,79	$> 0,05$
ment	$-0,88$	$0,31$	7,27	$< 0,05$
ϕ		$2,66$		

A nível de significância de 5% temos que para a parte não inflacionada do modelo todas as covariáveis com exceção do estado civil são significativas na explicação da produção dos

alunos. Agora, para a parte inflacionada, temos que apenas a produção do orientador é relevante para explicabilidade dos zeros. Para validar os argumentos expostos acima, realizamos um processo de seleção de variáveis, utilizando o método Stepwise. Assim, foram ajustados dois novos modelos, um sob cada ótica de estimação, obtendo assim novos valores de AIC e EAIC, removendo estado civil da parte não inflacionada do modelo e sexo, estado civil e número de filhos da parte inflacionada do modelo. Assim obtivemos: 3124,5 e 3139,4, respectivamente. Escolhemos pelo modelo clássico, dado o valor menor do critério de seleção de modelos.

4.2 Etapa 2 - Modelos Bell e Poisson-Tweedie

Partindo do resultado anterior, em que o modelo ZIBN2 foi escolhido como modelo final e a abordagem clássica como metodologia de estimação, agora estamos interessados em comparar Bell, ZIBell e Poisson-Tweedie, partindo do modelo com sexo, número de filhos e produção do orientador para a parte não inflacionada e somente a produção do orientador para a parte inflacionada do modelo, assim como para o modelo ZIBN2.

Aqui, trazemos um comparativo dos critérios de seleção para a abordagem clássica e Bayesiana para os modelos Bell e ZIBell e apenas clássica para PTw. Assim, os ajustes obtiveram os valores de AIC e EAIC presentes na Tabela 24.

Tabela 24 – AIC e EAIC para os modelos estudados na aplicação.

Modelos	AIC/EAIC
ZIBN2	3124,5
Bell - clássico	3149,5
Bell - Bayesiano	3151,6
ZIBell - clássico	3144,9
ZIBell - Bayesiano	3153,3
PTw	3573,1

O modelo PTw apresenta como critério de informação a medida pseudo Critério de Informação de Akaike (pAIC), que segundo Bonat (2018) é uma medida equivalente ao AIC tradicional, porém baseia-se em pseudo-verossimilhança, à medida em que como discutido a respeito do modelo Poisson-Tweedie estendido (Subseção 3.2.2), não é possível tratar a integral associada a função massa de probabilidade dessa distribuição, de modo que o modelo estendido é uma solução aproximada, porém sem perda de interpretabilidade ou aplicabilidade. Mesmo sendo um modelo mais flexível, que pode abranger diferentes comportamentos dos dados, o modelo PTw fica atrás dos modelos inflacionados, nesse caso, e acaba apresentando o maior AIC. Agora, realizamos os ajustes segundo ótica clássica e Bayesiana para Bell e ZIBell, no caso desse conjunto de dados, acabamos com vantagem, segundo o critério de seleção de modelos para os modelos clássicos, cabe ressaltar que o modelo que acaba mais próximo de ZIBN2 é o ZIBell.

Por esse critério de seleção de modelos, temos que o modelo ZIBN2 (lembrando que 2 indica o modelo que considera os coeficientes associados às partes inflacionada e não inflacio-

nada, vide Subseção 3.3.1). Então as conclusões apontadas na etapa anterior não se alteraram. Entretanto, um ponto a ressaltar é o fato do valor do AIC do modelo ZIBell ser muito próximo ao obtido para o modelo ZIBN2. Dado isso, apesar de não termos menor valor de AIC para esse modelo, o critério de parcimônia pode ser utilizado como critério de escolha. Com isso em mente, o modelo ZIBell indica vantagem sobre o ZIBN2, dado que aparentemente é possível ter bons resultados com um modelo mais simples. Para validar a qualidade dos ajustes, na Figura 14, trazemos lado a lado ambos os gráficos de envelopes de ZIBell e ZIBN2.

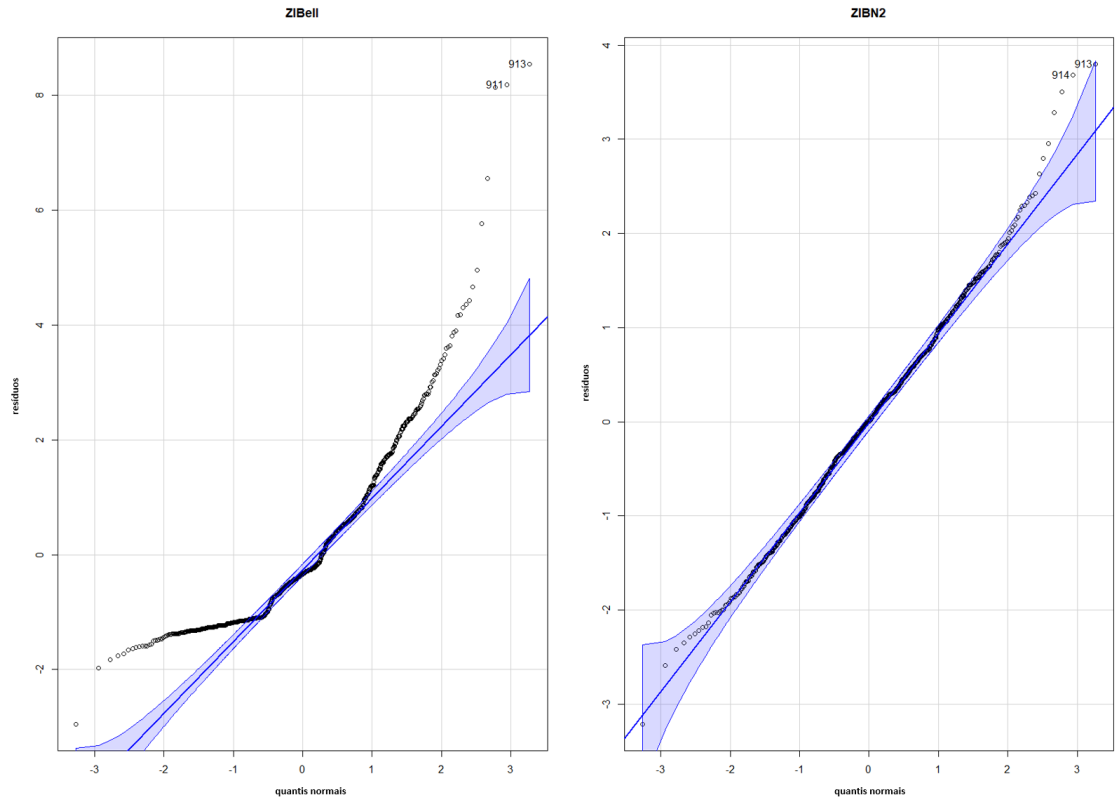

Figura 14 – Envelopes comparativos entre os modelos finais ZIBell e ZIBN2.

Podemos ver que para ZIBell apesar da proximidade em valor de AIC, para os dados estudados nas aplicações, não é um modelo adequado. ZIBN2 em contrapartida se mostra um modelo mais aderente, à medida em que a leitura desse gráfico se baseia na permanência dos resíduos no interior das bandas.

Como o modelo ZIBN2 é o modelo final escolhido, para ele trazemos uma análise dos seus resíduos. Assim, trazemos alguns gráficos na Figura 15. Não observamos padrões que indiquem má qualidade do ajuste, tais como concentrações acima ou abaixo da referência, zero, nem tendências fortes. Cabe ressaltar que há alguns outliers e por esse motivo, quando vemos o gráficos de resíduos quantílicos contra índices, há uma subida nos valores dos resíduos. Apesar de não ser pressuposto para aplicabilidade em muitos MLG's, temos que os resíduos seguem uma distribuição normal padrão $N(0,1)$, por conta de estar bem distribuída ao redor de zero e não possuir descolamentos no gráfico de quantis teóricos contra quantis amostrais. Com o

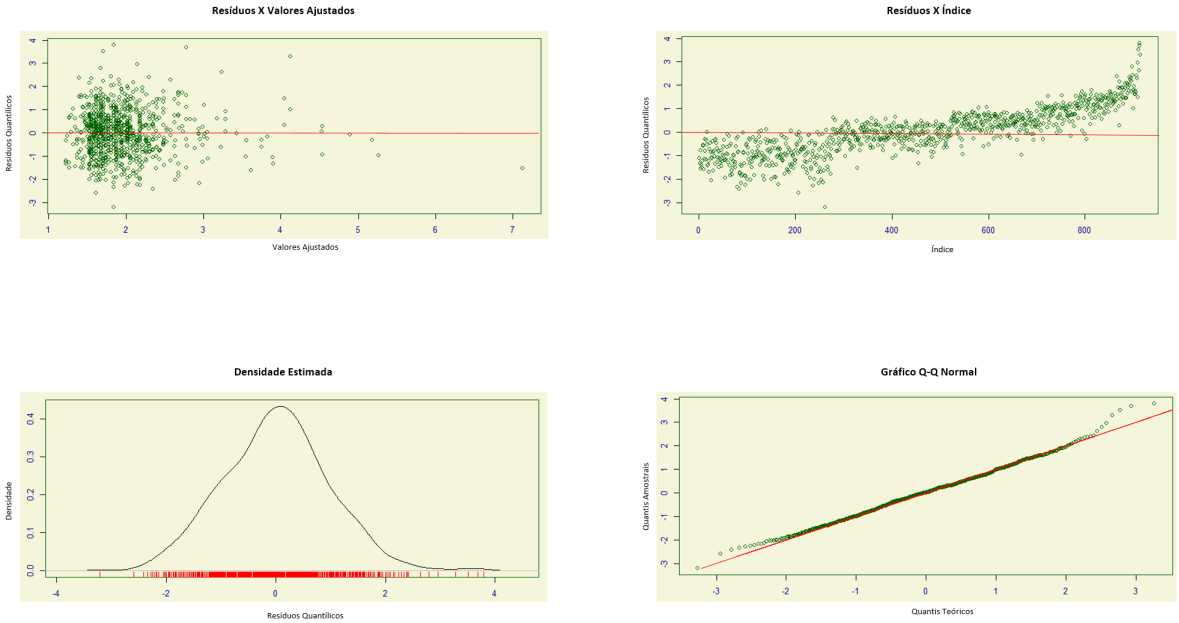

Figura 15 – Gráficos de resíduos para o modelo ZIBN2.

modelo final em mãos, podemos apresentar o resumo do ajuste mais aderente.

Tabela 25 – Modelo final via estimação Clássica - pacote GAMLSS.

ZIBN2 final - sob ótica clássica (GAMLSS)				
		Função de ligação μ: log		
Coeficientes μ:	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
(Intercepto)	$0,49$	$0,07$	$4,74$	$< 0,05$
feemWomen	$-0,23$	$0,07$	$-2,58$	$< 0,05$
kid5	$-0,13$	$0,05$	$-2,79$	$< 0,05$
ment	$0,02$	$3,46 \times 10^{-3}$	$7,27$	$< 0,05$
		Função de ligação ν: logit		
Coeficientes ν:	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
(Intercepto)	$-0,79$	$0,35$	$-2,27$	$< 0,05$
ment	$-0,62$	$0,25$	$-2,48$	$< 0,05$
ϕ		$2,70$		

Na Tabela 25, podemos observar alguns pontos importantes. Ser mulher no meio acadêmico acaba contribuindo negativamente à produtividade de artigos, isso aparenta estar diretamente relacionado com a quantidade de filhos (que equivalentemente, contribui negativamente à medida em que vai se tornando maior). Algo que contribui positivamente às publicações dos alunos é a quantidade de artigos do seu orientador. De mesmo modo, um orientador que não possui publicações, contribui negativamente para que seus orientandos possuam publicações.

Assim, como conclusão, temos que a metodologia clássica foi preferível à abordagem Bayesiana para esse conjunto de dados. Tendo como modelo mais aderente aos dados o modelo ZIBN2. Como comentado anteriormente nos estudos de simulação, quando há uma quantidade de dados considerável, há convergência das abordagens, cabendo ao pesquisador a escolha entre as metodologias. Cabe ressaltar que para esta aplicação o tempo de processamento do pacote GAMLSS foi de $0,53$ segundos contra $110,45$ segundos gastos pelo pacote brms. Realizando uma razão simples entre os tempos temos que o tempo do pacote Bayesiano foi mais de 200

vezes maior. Não diferente do cenário simulado. De modo que a escolha se torne tendenciosa para a abordagem clássica, mesmo quando o AIC for ligeiramente maior. Agora, por conta dos modelos Poisson-Tweedie não possuir implementações Bayesianas, ao menos não foi encontrado material nesse sentido. Então, até o presente momento não foi possível realizar a comparação entre metodologias de estimação, ficando limitados à abordagem clássica apenas para este modelo. O modelo que de fato mais se aproximou dos resultados obtidos, observando AIC foi o ZIBell, porém a análise de diagnóstico indicou vantagem ao uso do modelo ZIBN2. Isso acaba validando o estudo de Yang *et al.* (2017) que comparou modelos inflacionados com modelos não inflacionados e constatou vantagem ao uso de modelos inflacionados, quando nos dados há a presença desse fenômeno.

4.3 Considerações do Capítulo - Aplicação

Em cenário de aplicação, identificamos que o modelo mais aderente aos dados foi o ZIBN2, que apesar de mais complexo, apresentou menor valor de AIC para os dados. Num segundo momento, foi possível entender que o modelo ZIBN2 ainda se manteve como modelo de melhor ajuste aos dados bioChemists, sendo que ao que se refere a excesso de zeros, Bell e PTw não acomodaram tão bem este comportamento, porém com valores de AIC até que próximos entre si. ZIBell acabou entregando resultados muito similares aos do ZIBN2, ao menos na questão do valor do AIC. Comparando os valores dos critérios de seleção de modelos clássico e Bayesiano de Bell e ZIBell, ainda entendemos, ao menos para esses dados, que a abordagem clássica teve vantagem, tendo em mente tempos computacionais menores e equivalência das abordagens quando o tamanho da amostra é grande, assim como foi discutido nos estudos de simulação. Entendemos também, que de fato, possuir uma parte específica para a tratativa de zeros é benéfica, o que valida o estudo de Yang *et al.* (2017). Em termos de parcimônia, dada a similaridade dos resultados, podemos apontar vantagem para o modelo ZIBell. Porém, ao observar o gráfico de envelopes, percebemos que o modelo ZIBN2 é de fato mais aderente, para o qual também analisamos seus resíduos quantílicos. Sendo, portanto a escolha final do cenário de aplicação. Assim, o critério de parcimônia não se aplicou aqui.

Por fim, os códigos das aplicações estão presentes no Apêndice A.

CAPÍTULO

5

DISCUSSÃO E CONCLUSÕES

Sumarizamos a seguir as contribuições do presente trabalho, permeando sobre o que é apresentado em cada Capítulo e discutindo sobre possíveis perspectivas futuras. Salientamos que o objetivo do presente trabalho foi realizar simulações e aplicação com foco em apresentar um leque de distribuições e seus respectivos modelos ademais do que é visto na graduação em Estatística. Nesse sentido, tomamos como foco avaliar pacotes existentes que implementaram os modelos, destacando que é possível, ainda, o desenvolvimento de códigos próprios para os modelos explorados no presente trabalho. Como é o caso comentado sobre o inla em 2.2.

Disponibilizamos um repositório com códigos para facilitar replicabilidade e expansão. Ressaltando variação das funções de ligação, uma análise mais aprofundada dos resíduos dos modelos, abordagem de mais distribuições como Gama de contagem e COM-Poisson, presentes no trabalho de Bonat, Zeviani e Jr (2017), pontos não abordados no presente trabalho.

5.1 Contribuições

Neste trabalho nos orientamos a estudar os dois principais modelos de contagem: Poisson e Binomial Negativa, uma alternativa para cenário em que existe superdispersão, ie. variância da variável resposta e maior do que a média. Estudamos as versões inflacionadas de zeros para ambos os modelos.

Para os quatro modelos P, BN, ZIP e ZIBN consideramos variáveis explicativas à média. Ainda discutimos ZIP2 e ZIBN2, casos em que inclui-se covariáveis na proporção de zeros. Identificamos pacotes estatísticos disponíveis em **R** que implementam a estimação clássica e a estimação Bayesiana. Identificamos que os seis modelos podem ser ajustados nos pacotes base, pscl, VGAM, mgcv, glmmTMB e GAMLSS (estimação clássica) e arm, JAGS, inla, MCMCglmm, glmmADMB, brms (estimação Bayesiana).

Desenvolvemos três estudos de simulação sucessivos com e sem réplicas, que permitiram

identificar os pacotes GAMLSS na estimação clássica e brms na estimação Bayesiana obtiveram melhor desempenho em termos de precisão na recuperação de parâmetros e tempo computacional. Observando que em general os tempos computacionais dos pacotes Bayesianos são demorados, com exceção de inla que não foi possível usar em todos os cenários, mais especificamente para os modelos ZIP2 e ZIBN2.

Como alternativos ao modelo BN foram estudados dois novos modelos de regressão considerando as distribuições Bell e Poisson-Tweedie. Embora seja possível obter a estimação bayesiana do modelo Bell (referência do pacote) para propósitos de comparação, somente consideramos a estimação clássica de ambos modelos já que a estimação Bayesiana do modelo Poisson-Twedie não estava disponível. Uma versão zero-inflacionada foi considerada somente para a Bell já que para o modelo PTw não esta disponível.

Desenvolvemos um estudo de simulação, comparando a possibilidade de recuperação de um conjunto de dados originado por um certo modelo (P, BN, Bell e PTw) utilizando outro modelo. Esse estudo permitiu constatar que Bell e PTw são boas alternativas a P e BN. Ressalta-se que dados PTw não foram recuperados com os demais modelos, sendo interessante, nesse caso, o uso exclusivo deste modelo.

A aplicação a dados reais comparou primeiramente os modelos inflacionados (ZIP e ZIBN), P e BN, considerando também a metodologia clássica e Bayesiana para estimação dos parâmetros. Constatamos que há benefício ao uso de modelos inflacionados. A abordagem clássica foi mais aderente. Ambos os pontos anteriores resultando em valores de AIC menores, comparando com os de EAIC. Num segundo momento, comparamos Bell, ZIBell e PTw ao modelo ZIBN2, ZIBN2 se manteve com AIC menor, com proximidade de ZIBell. A análise de diagnóstico indicou ZIBN2 como modelo mais aderente aos dados.

Este trabalho abordou, portanto, sete diferentes distribuições de contagem e seus respectivos modelos (não pontuando aqui ZIP2 e ZIBN2, que acresceriam mais dois modelos). Temos que o modelo Poisson, apesar de mais difundido na literatura acadêmica e aplicações, apresenta uma limitação importante, que é a imposição de equidispersão aos dados, fenômeno pouco característico referente a esse tipo de dado. Binomial Negativa e demais modelos inflacionados (ZIP, ZIBN) e alternativos (Bell, ZIBell, PTw) adicionam uma maior flexibilidade à análise, dado que acomodam o comportamento de superdispersão. Como pontuado, o inflacionamento de zeros é um fenômeno comum observado em dados de contagem e de fato há vantagem no uso dos modelos inflacionados aos não inflacionados, uma vez que tratam separadamente o excesso de zeros amostrais. E ademais, Bell e PTw são boas alternativas a dados Poisson e Binomial Negativa, validando os estudos anteriores.

5.2 Próximos Passos

Como sugestão para passos futuros, entendemos que seria interessante estudar o método Bayesiano para o modelo Poisson-Tweedie, ainda pouco discutido na literatura e também implementado, ao menos ao que buscamos na literatura. Assim será possível realizar a verificação de desempenho perante os demais modelos alterando a abordagem de estimação dos parâmetros. Outro ponto, já comentado acima, seria variar as funções de ligação para ver a influência da escolha no resultado das estimações dos modelos.

Associado a isso, entendemos que caberia uma análise de cada modelo estudado observando seus resíduos, especialmente os quantílicos, para validar de mais uma forma a qualidade dos modelos estudados neste trabalho.

REFERÊNCIAS

AJ MORENO-ARENAS G, C. F. L. Zero-inflated bell regression models for count data. **Journal of applied statistics**, v. 47, n. 2, p. 265–286, 2020. Disponível em: <https://www.jstatsoft.org/index.php/jss/article/view/v084i04>. Citado nas páginas 54, 55 e 62.

AKAIKE, H. A new look at the statistical model identification. **IEEE Trans. on Automatic Control**, IEEE, v. 19, p. 716–723, 12 1974. Citado na página 38.

ANYOSA, S. A. C. Regressão binária usando ligações potência e reversa de potência. Universidade Federal de São Carlos - UFSCar, 2017. Citado na página 38.

BELL, E. T. Exponential polynomials. Annals of Mathematics, v. 35, p. 258–277, 04 1934. Citado na página 53.

BEN-ISRAEL, A. A newton-raphson method for the solution of systems of equations. **Journal of Mathematical Analysis and Applications**, v. 15, n. 2, p. 243–252, 1966. ISSN 0022-247X. Disponível em: <https://www.sciencedirect.com/science/article/pii/0022247X66901156>. Citado na página 31.

BONAT, W. H. Multiple response variables regression models in r: The mcglm package. **Journal of Statistical Software**, v. 84, n. 4, p. 1–30, 2018. Disponível em: <https://www.jstatsoft.org/index.php/jss/article/view/v084i04>. Citado na página 75.

BONAT, W. H.; JøRGENSEN, B. Multivariate covariance generalized linear models. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, [Wiley, Royal Statistical Society], v. 65, n. 5, p. 649–675, 2016. ISSN 00359254, 14679876. Disponível em: <http://www.jstor.org/stable/44681850>. Citado nas páginas 22, 54 e 63.

BONAT, W. H.; JøRGENSEN, B.; KOKONENDJI, C. C.; HINDE, J.; DEMéTRIO, C. G. B. Extended poisson–tweedie: Properties and regression models for count data. **Statistical Modelling**, v. 18, n. 1, p. 24–49, 2018. Disponível em: <https://doi.org/10.1177/1471082X17715718>. Citado nas páginas 53, 56, 57, 59, 63, 64 e 65.

BONAT, W. H.; ZEVIANI, W. M.; JR, E. E. R. **Regression Models for Count Data: beyond the Poisson model**. Universidade Federal do Paraná, 2017. Acessado em 16 de novembro de 2024. Disponível em: <https://cursos.leg.ufpr.br/rmcd/introduction.html>. Citado nas páginas 60 e 79.

BROOKS, M. E.; KRISTENSEN, K.; van Benthem, K. J.; MAGNUSSON, A.; BERG, C. W.; NIELSEN, A.; SKAUG, H. J.; MAECHLER, M.; BOLKER, B. M. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. **The R Journal**, v. 9, n. 2, p. 378–400, 2017. Citado na página 32.

BüRKNER, P.-C. Bayesian item response modeling in R with brms and Stan. **Journal of Statistical Software**, v. 100, n. 5, p. 1–54, 2021. Citado na página 35.

CASTELLARES, F.; FERRARI, S. L.; LEMONTE, A. J. On the bell distribution and its associated regression model for count data. **Applied Mathematical Modelling**, v. 56, p. 172–185, 2018. ISSN 0307-904X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0307904X17307448>. Citado nas páginas 22, 26, 53, 55 e 60.

DANIEL, J. W.; SEMINARS, A. A. **Poisson processes (and mixture distributions)**. [S.l.]: CAS - Casualty Actuarial Society, 2008. Citado na página 53.

DUNN, P. K.; SMYTH, G. K. **Generalized Linear Models with Examples in R**. [S.l.]: Springer Science & Business Media, 2018. Citado na página 25.

GELMAN, A.; SU, Y.-S. **arm: Data Analysis Using Regression and Multilevel/Hierarchical Models**. [S.l.], 2024. R package version 1.14-4. Disponível em: <https://CRAN.R-project.org/package=arm>. Citado na página 35.

GOURIéROUX, C.; HOLLY, A.; MONFORT, A. Likelihood ratio test, wald test, and kuhn-tucker test in linear models with inequality constraints on the regression parameters. **Econometrica**, [Wiley, Econometric Society], v. 50, n. 1, p. 63–80, 1982. ISSN 00129682, 14680262. Disponível em: <http://www.jstor.org/stable/1912529>. Citado na página 41.

HADFIELD, J. D. Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package. **Journal of Statistical Software**, v. 33, n. 2, p. 1–22, 2010. Disponível em: <https://www.jstatsoft.org/v33/i02/>. Citado na página 35.

JACKMAN, S. **pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory**. Sydney, Australia, 2024. R package version 1.5.9. Disponível em: <https://github.com/atahk/pscl/>. Citado nas páginas 23 e 32.

JEFFREYS, H.; KNEALE, W.; DAVID, F. N. Theory of probability. **Journal of the Institute of Actuaries (1886-1994)**, Cambridge University Press, v. 75, n. 2, p. 262–264, 1949. ISSN 00202681. Disponível em: <http://www.jstor.org/stable/41138851>. Citado na página 32.

JøRGENSEN, B.; KNUDSEN, S. J. Parameter orthogonality and bias adjustment for estimating functions. Scandinavian Journal of Statistics, v. 31, p. 93–114, 02 2004. Citado na página 65.

JøRGENSEN, B.; KOKONENDJI, C. C. Discrete dispersion models and their tweedie asymptotics. AStA Advances in Statistical Analysis, v. 100, p. 43–78, 04 2015. Citado nas páginas 22, 54, 56 e 57.

KOKONENDJI, C.; DEMéTRIO, C.; GBETE, S. D. Overdispersion and poisson-tweedie exponential dispersion models. **VIII Journées Zaragoza-Pau de Mathématiques Appliquées et de Statistiques :Jaca, Spain, September 15-17, 2003, 2003-01-01, ISBN 84-7733-720-9, pags. 365-374**, v. 31, 01 2004. Citado nas páginas 57 e 63.

LINDGREN, F.; RUE, H. Bayesian spatial modelling with R-INLA. **Journal of Statistical Software**, v. 63, n. 19, p. 1–25, 2015. Disponível em: <http://www.jstatsoft.org/v63/i19/>. Citado na página 35.

LONG, J. S. The origins of sex differences in science. Oxford University Press, v. 68, p. 1297–1316, 06 1990. Citado na página 72.

MORRIS, T. P.; WHITE, I. R.; CROWTER, M. J. Using simulation studies to evaluate statistical methods. Statistics in Medicine, v. 38, p. 2074–2102, 05 2019. Citado na página 46.

PAULA, G. A. **Modelos de Regressão com apoio computacional**. [S.l.]: IME-USP São Paulo, 2024. Citado nas páginas 21, 25 e 28.

PLUMMER, M. **rjags: Bayesian Graphical Models using MCMC**. [S.l.], 2023. R package version 4-15. Disponível em: <https://CRAN.R-project.org/package=rjags>. Citado na página 35.

R. A. Rigby; D. M. Stasinopoulos. Generalized additive models for location, scale and shape,(with discussion). **Applied Statistics**, v. 54, p. 507–554, 2005. Citado na página 32.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2024. Disponível em: <https://www.R-project.org/>. Citado nas páginas 21, 23 e 32.

SAHA, D.; ALLURI, P.; DUMBAUGH, E.; GAN, A. Application of the poisson-tweedie distribution in analyzing crash frequency data. **Accident Analysis Prevention**, v. 137, p. 105456, 2020. ISSN 0001-4575. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0001457519315258>. Citado nas páginas 22, 54 e 56.

SELLERS, K. F.; BORLE, S.; SHMUELI, G. The com-poisson model for count data: a survey of methods and applications. **Applied Stochastic Models in Business and Industry**, v. 28, n. 2, p. 104–116, 2012. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.918>. Citado na página 60.

SHAFIRA, S.; ABDULLAH, S.; LESTARI, D. Bayesian zero inflated negative binomial regression model for the parkinson data. EAI, 1 2020. Citado na página 34.

SILVEY, S. D. **Statistical Inference**. [S.l.]: CRC Press, 1975. Citado na página 30.

SKAUG, H.; FOURNIER, D.; BOLKER, B.; MAGNUSSON, A.; NIELSEN, A. **Generalized Linear Mixed Models using 'AD Model Builder'**. [S.l.], 2016. R package version 0.8.3.3. Citado na página 35.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. V. D. Bayesian Measures of Model Complexity and Fit. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, v. 64, n. 4, p. 583–639, 10 2002. ISSN 1369-7412. Disponível em: <https://doi.org/10.1111/1467-9868.00353>. Citado na página 38.

VUONG, Q. H. A new look at the statistical model identification. **Econometria**, The Econometric Society, v. 57, p. 307–333, 03 1989. Citado na página 26.

WOOD, S. N. Thin-plate regression splines. **Journal of the Royal Statistical Society (B)**, v. 65, n. 1, p. 95–114, 2003. Citado na página 32.

YANG, S.; HARLOW, L.; PUGGIONI, G.; REDDING, C. A comparison of different methods of zero-inflated data analysis and an application in health surveys. **Journal of Modern Applied Statistical Methods**, v. 16, p. 518–543, 05 2017. Citado nas páginas 22, 26, 54 e 78.

YEE, T. W. **VGAM: Vector Generalized Linear and Additive Models**. [S.l.], 2024. R package version 1.1-12. Disponível em: <https://CRAN.R-project.org/package=VGAM>. Citado na página 32.

ZEVIANI, W. M.; JR, P. J. R.; BONAT, W. H.; SHIMAKURA, S. E.; MUNIZ, J. A. The gamma-count distribution in the analysis of experimental underdispersed data. **Journal of Applied Statistics**, Taylor Francis, v. 41, n. 12, p. 2616–2626, 2014. Disponível em: <https://doi.org/10.1080/02664763.2014.922168>. Citado na página 60.

APÊNDICE

A

REPOSITÓRIO DE CÓDIGOS

Dada a quantidade de códigos utilizados não é possível transcrevê-los neste documento, para que não fique muito extenso. Sendo assim, um repositório no *GitHub* é disponibilizado para que seja possível o acesso a todo material computacional utilizado neste trabalho.

Link:<https://github.com/Lucas-Onuki/codigos_desempenho_modelos_contagem>

Neste repositório você leitor encontrará os códigos organizados em pastas, separados em capítulos, estudo ou aplicação respectivos. Abaixo, trazemos os códigos 1 e 2, mencionados na seção 3.2, em que exemplificamos o processo de simulação da Poisson-Tweedie.

Código-fonte 1 – Funções para utilizar o método de Monte Carlo de Poisson-Tweedie

```
# Integrando Poisson X Tweedie
integrand <- function(x, y, mu, phi, power) {
    int = dpois(y, lambda = x)*dtweedie(x, mu = mu,
                                        phi = phi, power =
    power)
    return(int)
}

# Computando o pmf utilizando Monte Carlo
dptw <- function(y, mu, phi, power, control_sample) {
    pts <- control_sample$pts
    norma <- control_sample$norma
    integral <- mean(integrand(pts, y = y, mu = mu, phi = phi,
                               power = power)/norma)
    return(integral)
}
dptw <- Vectorize(dptw, vectorize.args = "y")
```

Código-fonte 2 – Geração de valores de Poisson Tweedie

```
require(tweedie)
set.seed(123)
pts <- rtweedie(n = 1000, mu = 10, phi = 1, power = 2)
norma <- dtweedie(pts, mu = 10, phi = 1, power = 2)
control_sample <- list("pts" = pts, "norma" = norma)
dptw(y = c(0, 5, 10, 15), mu = 10, phi = 1, power = 2,
     control_sample = control_sample)
```


ICMC
USP
SÃO CARLOS