Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

Lucas Ferrari Pereira

O modelo Poisson-Exponencial do ponto de vista Bayesiano

Maringá - PR
30 de março de 2021

Lucas Ferrari Pereira

O modelo Poisson-Exponencial do ponto de vista Bayesiano

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito para obtenção do título de Mestre em Bioestatística.

Universidade Estadual de Maringá – UEM

Departamento de Estatística

Programa de Pós-Graduação em Bioestatística

Maringá - PR

30 de março de 2021

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

Pereira, Lucas Ferrari

P436m O modelo poisson-exponencial do ponto de vista bayesiano / Lucas Ferrari Pereira. -- Maringá, PR, 2021.
66 f.: il., tabs.

Orientador: Prof. Dr. Carlos Aparecido dos Santos.
Coorientador: Prof. Dr. Willian Luís de Oliveira.
Dissertação (Mestrado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Estatística, Programa de Pós-Graduação em Bioestatística, 2021.

1. Modelos bivariados . 2. Bayesiano - Método de estimação. 3. Bayesiano - Diagnóstico. 4. Poisson-exponencial - Modelo. I. Aparecido dos Santos, Carlos, orient. II. Oliveira, Willian Luís de, coorient. III. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Estatística. Programa de Pós-Graduação em Bioestatística. IV. Título.

CDD 23.ed. 519.5

Elaine Cristina Soares Lira - CRB-9/1202

LUCAS FERRARI PEREIRA

O Modelo Poisson-Exponencial do ponto de vista Bayesiano

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de Mestre em Bioestatística.

BANCA EXAMINADORA

Prof. Dr. Willian Luís de Oliveira

Universidade Estadual de Maringá – PBE/UEM

Prof. Dr. Brian Alvarez Ribeiro de Melo

Universidade Estadual de Maringá – PBE/UEM

Prof. Dr. Vanderly Janeiro

Universidade Estadual de Maringá – PBE/UEM

Maringá, 30 de abril de 2021.

Este trabalho é dedicado aos meus pais, minhas irmãs e ao meu sobrinho.

Agradecimentos

A concretização deste trabalho se deu pelo apoio, auxílio, compreensão e dedicação de várias pessoas. Meu agradecimento a todos que, de alguma forma, contribuíram para a conclusão deste trabalho e, de uma maneira especial, agradeço:

- a Deus, pelo dom da vida e por todas as bênçãos recebidas;
- aos meus pais, Simone e Jorge, por me disponibilizarem esta formação, por todo amor, paciência e dedicação. Quero um dia retribuir tudo o que já fizeram por mim;
- às minhas irmãs Carolina e Gabriela, e ao meu sobrinho João Vitor, por todo amor, carinho, compreensão e incentivo para seguir em frente.
- aos meus amigos, os de longe e os de perto, e aqueles que de alguma forma fazem parte da minha vida e que são essenciais para eu ser, a cada dia dessa longa jornada, uma pessoa melhor;
- à minha grande amiga Patrícia por todo o carinho, compreensão, incentivo e apoio. Você foi, literalmente, a peça chave para que essa dissertação saísse e sem teus conselhos e ajudas eu jamais chegaria aqui. Agradeço a Deus por colocar uma pessoa tão incrível na minha vida, saiba que estarei sempre aqui;
- ao professor orientador deste trabalho, professor Carlos, e em especial ao professor coorientador, professor Willian, pela dedicação, auxilio, compreensão, paciência e pelas palavras motivadoras na elaboração e condução deste trabalho. Os meus mais sinceros votos de felicidades e bênçãos de Deus em suas vidas;
- à banca examinadora deste trabalho, o professore Dr. Brian A. Ribeiro de Melo e o professor Dr. Vanderly Janeiro pela participação na banca examinadora, pela ajuda e sugestões para o enriquecimento do trabalho;
- aos professores do Programa de Pós Graduação em Bioestatística, pela contribuição ao meu aprendizado, pelos ensinamentos e experiências transmitidas;

- à Universidade Estadual de Maringá, pela acolhida, infraestrutura e pela oportunidade de realização deste trabalho;
- à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro;

*“Não importa o que aconteça, continue a **nadar**.*

*(WALTERS, GRAHAM; **Procurando Nemo**, 2003)*

Resumo

Este trabalho propõe uma metodologia Bayesiana para o ajuste e diagnóstico do modelo bivariado Poisson-Exponencial em que a resposta discreta segue uma distribuição Poisson e a resposta contínua, condicionada à discreta, segue uma distribuição Exponencial. É estabelecida uma estrutura de dependência entre as variáveis respostas, que depende das médias marginais, da variável resposta discreta e de uma medida de associação entre as variáveis respostas, e são assumidos conjuntos de covariáveis que se relacionam às médias marginais através de funções de ligação. A estimação dos parâmetros é realizada sob a perspectiva da inferência Bayesiana, adotando distribuições *a priori* não informativas para os parâmetros do modelo e propõe-se o uso do algoritmo de Metropolis-Hastings para a obtenção de amostras Monte Carlo via Cadeias de Markov (MCMC), necessárias para a obtenção de estimativas *a posteriori* para os parâmetros. Para a análise de resíduos, são considerados os resíduos baseados na densidade Preditiva Condicional Ordinária (CPO) e os resíduos baseados na distribuição *a posteriori* do modelo. Por fim, é realizado um estudo com dados simulados para verificar as propriedades inferenciais dos estimadores e a performance dos resíduos à alguns cenários pré-determinados. Quanto a verificação de possíveis observações influentes, realiza-se algumas perturbações em algum dos conjuntos de dados simulados, a fim de verificar as modificações ocorridas. Os resultados foram bastante satisfatórios, os valores estimados estão bem próximos dos valores reais, além disso, foi observado que as cadeias convergiram, através do Teste de Geweke.

Palavras-chave: Diagnóstico Bayesiano. Método de Estimação Bayesiano. Modelos Bivariados. Variáveis Respostas Discretas e Contínuas.

Abstract

This paper proposes a Bayesian methodology for the adjustment and diagnosis of the bivariate Poisson-Exponential model in which the discrete response follows a Poisson distribution and the continuous response, conditioned to the discrete, follows an Exponential distribution. A dependency structure is established between the response variables, which depends on the marginal means, the discrete response variable and a measure of association between the response variables, and sets of covariables are assumed that relate to the marginal means through link functions. For Bayesian estimation, non-informative *a priori* distributions are adopted for the model parameters and it is proposed to use the Metropolis-Hastings algorithm to obtain Monte Carlo samples via Markov Chains (MCMC), necessary to obtain estimates *a posteriori* for the parameters. For the analysis of residues, residues based on Ordinary Conditional Predictive density (CPO) and residues based on the *a posteriori* distribution of the model are considered. Finally, a study is carried out with simulated data to verify the inferential properties of the estimates and the performance of the residues in some predetermined scenarios. As for the verification of possible influential observations, some disturbances are carried out in any of the simulated data sets, in order to verify the modifications that have occurred.

Keywords: Bayesian diagnosis. Bayesian Estimation Method. Bivariate models. Variables Discrete and Continuous Answers. Bayesian Estimation Method. Bayesian diagnosis.

Sumário

Introdução

Há muito tempo, a estatística vem sendo utilizada em várias áreas do conhecimento, tais como: Engenharias, Economia, Saúde, Biologia, entre outras, a fim de explicar problemas e situações do cotidiano. Para tal fato, a estatística se apropria de expressões matemáticas com o intuito de associar as fórmulas e expressões aos problemas encontrados no dia a dia, e traduz os resultados da forma mais simplificada para um maior entendimento das pessoas. Uma dessas ferramentas, que é bastante estudada por grandes pesquisadores e que utiliza a matemática para resolver os problemas encontrados em muitos estudos, é a modelagem estatística.

Existem muitos tipos de modelos estatísticos na literatura e cada vez mais surgem novos modelos estatísticos para serem utilizados nas mais diversas situações. Um exemplo de modelagem estatística encontrado na literatura são os modelos apropriados para dados de seleção genômica. Esse estudo de seleção genômica tem como objetivo selecionar a característica da área de olho de lombo (AOL) em ovinos da raça Santa Inês, que é um trabalho desenvolvido por Oliveira (2019). O autor utilizou a metodologia estatística, dentro do contexto Bayesiano, para resolver o problema de selecionar um grupo de ovinos da raça Santa Inês que tinham uma determinada característica ocular.

Outro exemplo de modelagem estatística é o trabalho Mello e Silva (2009), que investigaram a precipitação mensal e anual das chuvas no estado de Minas Gerais, a fim de auxiliar estudos estratégicos associados ao planejamento do meio ambiente, geração de energia e manejo da agricultura. O objetivo do trabalho foi ajustar modelos lineares, por meio de uma regressão múltipla, para predição da precipitação média mensal e anual (durante o período úmido e seco), baseados nas coordenadas geográficas (latitude, longitude e altitude) para o Estado de Minas Gerais.

Os exemplos mencionados trazem uma ideia da importância da modelagem estatística no que diz respeito a respostas de questões sérias dentro de várias linhas de pesquisa. Podemos notar, nesses exemplos, que a variável de interesse é somente uma resposta (característica da AOL ou a precipitação da chuva). Mas, além da modelagem

estatística para uma única variável resposta de interesse do pesquisador (podendo ser ela discreta ou contínua), existem também situações que é desejável modelar mais de uma resposta para o problema de interesse, uma vez em que as observações acontecem simultaneamente.

Esses modelos que admitem duas variáveis respostas são chamados modelos bivariados. Na literatura, encontramos modelos bivariados cujas variáveis respostas são ambas discretas, ambas contínuas ou ainda modelos cujas variáveis respostas são mistas, ou seja, uma discreta e outra contínua.

Um exemplo de trabalho que envolva a modelagem estatística para mais de uma resposta é o trabalho de Cunha et al. (2018), que decidiu modelar a variável dependente da equação de rendimento minceriana de forma separada, ou seja, modelar o rendimento-hora em duas partes, rendimento e horas trabalhadas. Essa equação é muito utilizada na economia para a determinação de salários com base no rendimento e na hora trabalhada. Os autores utilizaram modelos de regressão bivariada baseados nas distribuições normal, t e Birnbaum-Saunders, para analisar dados do mercado de trabalho.

Também ilustra uma modelagem para duas variáveis respostas o trabalho de Song, Barnhart e Lyles (2004), que estima a correlação entre as variáveis respostas de uma pesquisa sobre HIV, em mulheres grávidas, através de equações de estimação generalizadas. Para a realização do estudo, foram utilizadas as contagens de células CD4+ em 36 semanas de gestação e a carga viral de HIV na lavagem cérvico-vaginal com 38 semanas de gestação. E, ainda, um quarto exemplo é o trabalho de Oliveira, Diniz e Durbán (2019), que busca modelar a relação entre o gasto obtido e uso do centro cirúrgico com outras variáveis coletadas.

Ainda, para fins de exemplificação envolvendo modelos bivariados com respostas discretas é o trabalho de Jung e Winkelmann (1993), que utiliza um modelo de Poisson Bivariado para o estudo da distinção de dois tipos de mobilidade de trabalho: emprego direto a mudanças de emprego (que são assumidas como voluntárias) e mudanças de emprego depois de passar por um período de desemprego (assumido como involuntário). Outro exemplo está na investigação de Khafri, Kazemnejad e Eskandari (2008), o qual apresenta uma análise de dados de fertilização *in vitro* em casais inférteis, utilizando um modelo de Poisson bivariado e para a estimação dos parâmetros, utilizaram os métodos Bayesianos.

Para os modelos bivariados com respostas contínuas, podemos citar a pesquisa de Scollnik (2002) que utiliza dois modelos de regressão, Pareto-Gama e o Pareto Bivariado, para analisar a relação de seguros contra acidentes, considerando valores de perda e valores de despesas alocadas com ajuste do sinistro.

Já em relação aos modelos bivariados com resposta mista, citamos os trabalhos de

Olkin, Tate et al. (1961) que, motivados por estudos no campo da psicologia, trabalharam com um modelo Bernoulli-Normal e estenderam para uma versão multivariada. Também, podemos citar o trabalho de Catalano e Ryan (1992), o qual empregou uma estrutura de variável latente para obter a distribuição conjunta de um modelo misto com uma resposta contínua e uma resposta discreta. Além disso, temos a investigação de Fitzmaurice e Laird (1995), que apresentam uma extensão do modelo de Olkin, Tate et al. (1961) com duas vantagens, em que a primeira se refere aos parâmetros terem interpretação marginal e a segunda se volta ao fato de que as estimativas de máxima verossimilhança dos parâmetros são consistentes, mesmo que a associação entre as variáveis respostas não seja especificada.

Ainda, dentro desse contexto de modelos bivariados, Stulp (2019) propôs um modelo bivariado com resposta mista, com dependência entre as variáveis respostas. Em seu trabalho, a autora expõe que variável resposta discreta provinha da distribuição Poisson e a variável resposta contínua era condicionada à variável resposta discreta, e essa resposta condicionada seguia uma distribuição Exponencial. Além disso, Stulp (2019) realizou a estimação dos parâmetros de interesse via inferência frequentista.

Pensando nas possíveis situações em que o modelo bivariado Poisson-Exponencial com abordagem frequentista não é uma boa alternativa para ser utilizado, o presente trabalho tem por objetivo apresentar a estimação do modelo bivariado Poisson-Exponencial, proposto por Stulp (2019), agora dentro do contexto Bayesiano, no qual os parâmetros são tratados como variáveis aleatórias. Para isso, este trabalho está organizado em três capítulos.

Dessa forma, além desta introdução, este trabalho está organizado em mais três capítulos, que são descritos na sequência. No capítulo 1, abordamos o referencial teórico, no qual discutimos os principais trabalhos que norteiam esta pesquisa, além de expormos o modelo proposto. As definições do modelo constam no capítulo 2, no qual são evidenciados, também, os processos de estimação, critérios de comparação, análises de resíduos, simulações e os resultados. E, por fim, no capítulo 3, são abordadas as considerações finais e as propostas futuras deste trabalho.

Capítulo 1

Referêncial Teórico

1.1 Inferência Bayesiana

Segundo Bolfarine e Sandoval (2001), a inferência estatística é um conjunto de técnicas e métodos utilizados na estimação de um ou mais valores para um parâmetro desconhecido $\theta \in \Theta$ associado a uma variável aleatória X com função de densidade $f(x|\theta)$. Há duas abordagens para interpretar a inferência estatística: a frequentista e a Bayesiana.

No método clássico, conhecido também como o método frequentista, a ideia de probabilidade é envolver uma sequência de repetições para um determinado evento, tratado como um subconjunto de Θ. Ou seja, o ponto fundamental é que o parâmetro θ, que ainda é desconhecido, além de ser tratado como constante ao invés de aleatório, é dado. Portanto, a interpretação correta do intervalo é a de que, se o procedimento for aplicado "muitas vezes", então, "em uma longa sequência", os intervalos que serão construídos vão conter o verdadeiro valor de θ (COLES; JR, 2016). Além disso, o estimador $\hat{\theta}$ pode assumir um valor fixo de acordo com o método de estimação adotado.

O método Bayesiano é semelhante ao anterior: há um parâmetro populacional θ em que se deseja fazer inferências, porém θ é tratado como uma quantidade aleatória, ou seja, a inferência vai ser baseada em $f(\theta|y)$, isto é, a probabilidade do parâmetro condicional aos dados obtidos (COLES; JR, 2016). A principal vantagem dessa técnica é que o estimador segue uma distribuição conhecida, descreve o comportamento do parâmetro que pode ser controlada e ajustada de acordo com o conhecimento que se tem sobre θ (OLIVEIRA, 2019). São chamadas de distribuições *a priori*.

As distribuições *a priori* $p(\theta_i)$ são combinadas com a função de verossimilhança $f(y|\Theta)$, que sintetiza a informação amostral do vetor de observações **y** condicionada aos parâmetros do modelo Θ (OLIVEIRA, 2019). Diante disso, e utilizando o teorema de Bayes,

é possível obter a distribuição *a posteriori*, que retrata a distribuição de probabilidade dos parâmetros associada à informação de **y** (FARIA et al., 2007).

Essa estratégia é bastante conveniente, já que a distribuição *a posteriori* fornece estimativas considerando a informação amostral e as distribuições *a priori*, que podem acrescentar a subjetividade acerca dos parâmetros, além de trazer vantagens em relação aos métodos frequentistas (OLIVEIRA, 2019).

Em algumas situações, dependendo da complexidade dos modelos e das distribuições *a priori* adotadas, não é possível obter a distribuição *a posteriori* de forma analítica. Nesses casos, são usados métodos iterativos como o de Monte Carlo via Cadeias de Markov (MCMC), um dos mais utilizados quando não é possível gerar valores diretamente da posteriori, ou seja, quando não possui forma fechada. Dentro do MCMC, os dois algoritmos mais adotados para obtenção das estimativas são o amostrador de Gibbs e o Metropolis de Hastings (OLIVEIRA, 2019).

1.2 Modelos com abordagem Bayesiana

A metodologia bayesiana está presente em diversas aplicações, nos mais variados contextos. Os trabalhos descritos na sequência são alguns exemplos de modelos univariados e bivariados no contexto bayesiano. Sumariamente o trabalho de Bastos (2018) refere-se a um modelo hierárquico bayesiano, ao passo que o estudo realizado por Oliveira (2019) é referente a modelos Bayesianos de seleção genômica ampla e, por fim, a pesquisa de Ribeiro (2017) apresenta uma abordagem bayesiana no contexto de dados bivariados de sobrevivência.

1.2.1 O Modelo de Bastos (2018)

Com o aumento de casos de Zika Virus na cidade do Rio de Janeiro, nos anos de 2015 e 2016, Bastos (2018) desenvolveu um estudo com o objetivo de estimar o tamanho total e os parâmetros de transmissão da epidemia de Zika, no ano de 2016. Como as arboviroses transmitidas pelo Aedes Aegypti se manifestam de forma sazonal, foi estabelecido um período de estudo através da estimativa de marcos temporais que representassem o início (τ_1) e o término (τ_2) do período epidemiológico de 2016.

Para realizar essas estimativas, o autor utilizou uma modelagem Bayesiana hierárquica adaptada ao modelo epidemiológico SIR (Suscetíveis - Infectados - Removidos). Os dados foram fornecidos pelo Sistema de Informação de Agravos de Notificação (SINAN) e do Sistema de Informações sobre Nascidos Vivos (SINASC).

As variáveis de estudo são os homens ($\mathbf{D}_m$) e as mulheres, que foram divididas em categorias: mulheres na idade fértil ($\mathbf{D}_{ffa}$ - nascidas entre os anos de 1968 e 2002),

mulheres fora da idade fértil ($\mathbf{D}_{fnfa}$) e o total de mulheres ($\mathbf{D}_f$). Para representar o número semanal de abortos causados pelo vírus da Zika, foi produzida a matriz $\mathbf{D}_{mz}$.

Inicialmente, Bastos (2018) realizou uma inferência utilizando a tabela $\mathbf{D}_{ffa}$ e $\mathbf{D}_{mz}$, pois ambas estão intimamente ligadas e os parâmetros estimados comuns as demais frações populacionais foram utilizadas como uma variável observável. Esse procedimento foi necessário, pois a presença de mais dados $\mathbf{D}_{mz}$ são responsáveis pelas estimativas mais precisas desses parâmetros comuns, diminuindo a variância das distribuições dos parâmetros epidemiológicos dos demais modelos, por essa razão, foram descritos apenas uma vez.

Como os dados observados são apenas uma fração da população total de afetados pela epidemia, Bastos (2018) definiu que as observações $\mathbf{D}_{ffa}$ (modelo feminino dentro da idade fértil) são governadas pela variável observada $\mathbf{Y}_o^{ffa}$ que possui distribuição binomial com parâmetros $\mathbf{Y}_{ffa}$ (total de mulheres infectadas na idade fértil) e $\mathbf{po}_{ffa}$ que representa a probabilidade de uma mulher na idade fértil e infectada ser observada.

Seja $\mathbf{Y}_p$ a variável aleatória com o número de mulheres grávidas e infectadas com o vírus da Zika, considerando que mz é a probabilidade de uma dessas mulheres sofrer aborto por conta do vírus, então as entradas na tabela $\mathbf{D}_{mz}$ podem ser modeladas pela variável observada $M \sim Bin(Y_p, mz)$.

Para os demais modelos (masculino, feminino e feminino fora da idade fértil), a verossimilhança referente aos dados ocorreu de forma análoga à utilização no modelo feminino dentro da idade fértil. Portanto, para $x \in \{m, f, fnfa\}$, $Y_o^x \sim Bin(Y_x, po_x)$ é a parcela da população que foi observada o vírus da Zika, Y_x é o total de infectados e po_x é a probabilidade de um infectado ser observado.

Após todos os processos de inferência dos modelos probababilísticos referentes às mulheres dentro da idade fértil, às mulheres fora da idade fértil, à população total de mulheres e à população total de homens, o autor concluiu que para entender o impacto total de uma doença tanto em termos de morbidade quanto nas perdas econômicas, é importante conhecer o tamanho total de uma epidemia.

No estudo de Bastos (2018), foi verificado uma assimetria na disseminação da doença causada pela transmissão sexual de homens para mulheres em idade fértil. Além disso, o impacto da Zika no aborto precoce transformou o registro de nascidos vivos em uma fonte importante de informações, permitindo a estimativa das taxas de subnotificação.

Compreender o tamanho real da epidemia de Zika, na cidade do Rio de Janeiro, em 2016, é um resultado muito relevante se quisermos entender completamente os riscos de ter uma segunda onda de Zika, na cidade, em um futuro próximo (BASTOS, 2018).

1.2.2 O Modelo de Oliveira (2019)

Como o avanço nas pesquisas em genética molecular favoreceram o conhecimento do genoma de algumas especies, Oliveira (2019) aplicou e avaliou modelos de seleção genômica ampla, usando os modelos Bayesianos robustos para a característica área de olho de lombo (AOL) medida por ultrassonografia, em ovinos da raça Santa Inês.

A amostra foi composta por 389 ovinos da raça Santa Inês criados em rebanhos localizados nos estados do Piauí e Maranhão e todos devidamente registrados. Para avaliar a característica AOL, foi realizado um procedimento por meio de imagens ultrassonográficas do corte transversal do músculo *Longissimus dorsi* entre a 12ª e 13ª vértebras lombares e foi medida em cm^2. Nos modelos de Seleção Genômica Ampla, a variável AOL, não negativa, foi utilizada como variável resposta.

Oliveira (2019) utilizou modelos Bayesianos de seleção genômica ampla, empregando a distribuição t (Bayes t) e a dupla-exponencial (Bayes DE) para a variável resposta. As distribuições *a priori* são equivalentes aos modelos propostos, isto é, normal para os efeitos sistemáticos e para o efeito dos SNPs e qui-quadrado invertida para os componentes de variância.

Para o ajuste dos modelos, foi realizada uma implementação mediante a com a inclusão de dados fenotípicos e genotípicos. Desta forma, têm-se o modelo aditivo:

$$y_i = \mu \; + \; \boldsymbol{x}'_{r_i}\boldsymbol{\beta_r} \; + \; \boldsymbol{x}'_{l_j}\boldsymbol{\beta_l} \; + \; \varepsilon_i, \;\; \forall\, i \; = \; 1, 2, \ldots, n; \; j \; = 1, 2, \ldots, m, \tag{1.1}$$

sendo que μ é o intercepto; $\boldsymbol{\beta_r}$ é o vetor de parâmetros associado aos efeitos sistemáticos; $\boldsymbol{\beta_l}$ é o vetor associado ao efeito dos marcadores genéticos; $\boldsymbol{x}'_{r_i}$ e $\boldsymbol{x}'_{l_j}$ são matrizes de incidência dos efeitos sistemáticos e aleatórios dos SNPs, respectivamente; e ε_i é o resíduo do modelo, considerado independente e identicamente distribuído. A função de verossimilhança, foi escrita como:

$$p(y \mid \mu, \boldsymbol{\beta_r}, \boldsymbol{\beta_l}, \sigma^2_\varepsilon) = \prod_{i=1}^{n} N(y_i \mid \mu \; + \; \boldsymbol{x}'_{r_i}\boldsymbol{\beta_r} \; + \; \boldsymbol{x}'_{l_i}\boldsymbol{\beta_l}, \sigma^2_\varepsilon) \tag{1.2}$$

As distribuições *a priori* foram descritas como:

$$\begin{aligned} p(\mu, \boldsymbol{\beta_r}, \boldsymbol{\beta_l}, u, \sigma^2_\varepsilon, \sigma^2_l) \;\; &= \;\; N(\mu|0, \sigma^2_\mu) N(\boldsymbol{\beta_r}|\mathbf{0}, \boldsymbol{I}\sigma^2_r) \\ &\times \;\; \chi^{-2}(\sigma^2_\varepsilon \mid S_\varepsilon, d.f._\varepsilon)\chi^{-2}(\sigma^2_l \mid S_l, d.f._l) \\ &\times \;\; N(\boldsymbol{\beta_l} \mid \mathbf{0}, \boldsymbol{I}\sigma^2_l) \; \forall i = 1, 2, \ldots, n; j = 1, 2, \ldots, m \end{aligned} \tag{1.3}$$

em que σ^2_μ e σ^2_r são as variâncias de μ e β_{r_i}, respectivamente; $d.f.$ e S são o grau de liberdade e o parâmetro de escala correspondentes às distribuições χ^{-2} de σ^2_ε e σ^2_l; e o índice i se refere ao número de animais e j à quantidade de marcadores.

Portanto, pôde-se representar hierarquicamente o modelo por:

$$y_i|\mu, \beta_r, \beta_l, \sigma^2_l, \sigma^2_l \sim N(\mu + \boldsymbol{x}'_{r_i}\boldsymbol{\beta_r} + \boldsymbol{x}'_{l_i}\boldsymbol{\beta_l}, \sigma^2_\varepsilon) \tag{1.4}$$

$$
\begin{aligned}
\mu \sim N(0, \sigma_\mu^2 \quad &: \quad \text{Intercepto} \\
\boldsymbol{\beta_r} \sim N_p(\mathbf{0}, \boldsymbol{I}\sigma_r^2) \quad &: \quad \text{Efeitos sistemáticos do modelo} \\
\boldsymbol{\beta_{l_j}} \sim N_p(\mathbf{0}, \boldsymbol{I}\sigma_l^2) \; \forall \; j = 1, 2, 3, \ldots, m \quad &: \quad \text{Efeitos do marcador genético} \\
\sigma_\varepsilon^2 \sim \chi^{-2}(\sigma_\varepsilon^2 | S_\varepsilon, d.f._\varepsilon) \quad &: \quad \text{Variância residual} \\
\sigma_l^2 \sim \chi^{-2}(\sigma_l^2 | S = S_l, d.f. = d.f._l) \quad &: \quad \text{Variância associada a } \boldsymbol{\beta_l}
\end{aligned} \tag{1.5}
$$

Alguns estudos apontam que pode ocorrer conflitos de informações, como *outliers*. Por essa razão de Oliveira (2019), ajustou o modelo, cuja variável resposta do modelo (y_i) tem distribuição t-Student e dupla-exponencial, uma vez que a distribuição t com baixos graus de liberdade é mais robusta comparada com a distribuição normal.

Desta forma, foi adotada a distribuição $t_{(v)}$ com v graus de liberdade e a dupla-exponencial com média μ e desvio-padrão σ para a variável resposta y_i do modelo, que continua a ser especificado da mesma forma de 1.1. Contudo as verossimilhanças passam a ser escritas da seguinte maneira:

$$
p(y|\mu, \boldsymbol{\beta_r}, \boldsymbol{\beta_l}, \sigma_\varepsilon^2) \;=\; \prod_{i=1}^{n} t(y_i|\mu + \boldsymbol{x'_{r_i}}\boldsymbol{\beta_r} + \boldsymbol{x'_{l_i}}\boldsymbol{\beta_l}, \sigma_\varepsilon^2, d.f.) \tag{1.6}
$$

para a distribuição t-Student;

$$
p(y|\mu, \boldsymbol{\beta_r}, \boldsymbol{\beta_l}, \sigma_\varepsilon^2) \;=\; \prod_{i=1}^{n} DE(y_i|\mu + \boldsymbol{x'_{r_i}}\boldsymbol{\beta_r} + \boldsymbol{x'_{l_i}}\boldsymbol{\beta_l}, \sigma_\varepsilon^2) \tag{1.7}
$$

para a distribuição dupla-exponencial.

Com relação às distribuições *a priori*, tem-se a mesma estrutura de 1.3. Portanto, pode-se representar hierarquicamente os modelos por:

$$
y_i|\mu, \beta_r, \beta_l, u, \sigma_l^2, \sigma_u^2, \sigma_e^2 \sim t_v(\mu + \boldsymbol{x'_{r_i}}\boldsymbol{\beta_r} + \boldsymbol{x'_{l_i}}\boldsymbol{\beta_l} + u_i, \sigma_\varepsilon^2) \tag{1.8}
$$

$$
y_i|\mu, \beta_r, \beta_l, u, \sigma_l^2, \sigma_u^2, \sigma_e^2 \sim DE(\mu + \boldsymbol{x'_{r_i}}\boldsymbol{\beta_r} + \boldsymbol{x'_{l_i}}\boldsymbol{\beta_l} + u_i, \sigma_\varepsilon^2), \tag{1.9}
$$

de modo que as distribuições *a priori* são semelhantes a 1.5 e as condicionais completas são:

$$
\begin{aligned}
p(\boldsymbol{\beta_r}, \boldsymbol{\beta_l}, \sigma_l^2, \sigma_\varepsilon^2) \;&=\; p(\boldsymbol{\beta_r}|\sigma_\varepsilon^2)p(\sigma_\varepsilon^2)p(\boldsymbol{\beta_r}|\sigma_l^2)p(\sigma_l^2) \\
&=\; \prod_{i=1}^{n} N(\beta_{r_i}, \sigma_\varepsilon^2) \times \chi^{-2}(\sigma_\varepsilon^2|d.f., S) \\
&\times\; \prod_{j=1}^{m} N(\beta_{l_j}, \sigma_m^2) \times \chi^{-2}(\sigma_l^2|d.f., S_l)
\end{aligned} \tag{1.10}
$$

Em sua análise, Oliveira (2019) comparou os modelos Bayes t, Bayes DE e RRBLUP (Regressão de Cumeeira do tipo BLUP), que já é consolidado na literatura e que sugere a

distribuição normal para y_i. De modo geral, foi constatada diferença na implementação dos ajustes com diferentes distribuições para a variável resposta AOL do modelo de seleção genômica ampla.

Ao verificar os resíduos, o modelo de Bayes t se mostrou melhor, mais concentrado em torno 0 e menos disperso, enquanto o RRBLUP teve sua média mais distante de 0 e apresentou uma maior variabilidade. Com excessão do DIC, todos os critérios adotados apontaram que os modelos propostos foram melhores do que o RRBLUP, o que pode ter sido reflexo da presença de *outliers* e pelo tamanho da amostra, indicando uma alternativa viável para cenários com poucos animais genotipados e com dados com *outliers*.

1.2.3 O Modelo de Ribeiro (2017)

Com o objetivo de modelar dados de sobrevivência bivariados, Ribeiro (2017) propõe dois modelos derivados das cópulas de Ali-Mikhail-Haq (AMH) e de Frank para modelar a dependência de dados bivariados na presença de covariáveis e observações censuradas. Para isso, é realizado uma abordagem Bayesiana usando métodos de Monte Carlo em Cadeias de Markov (MCMC).

Definição: *Uma cópula é uma distribuição multivariada cujas marginais são Uniforme* $(0,1)$*. Considere o vetor aleatório* $U = (U_1, \ldots, U_n) \in I^n$ *com cópula n-dimensional C, temos:*

$$C(u_1, \ldots, u_n; \phi) \;=\; P(U_1 \leq u_1, \ldots, U_n \leq u_n; \phi), \tag{1.11}$$

em que ϕ *é o parâmetro associado à função cópula.*

Atualmente, a classe de cópulas Arquimedianas é a mais utilizada na prática, pois sua representação permite reduzir o estudo de cópula multivariada ao estudo de uma função univariada φ, comumente chamada de gerador de uma cópula Arquimediana.

Um exemplo de cópula Arquimediana é a Cópula de Frank. Essa função cópula assume a seguinte forma:

$$C_\phi(u,v) = log_\phi\Big(1 + \frac{(\phi^u - 1)(\phi^v - 1)}{\phi - 1}\Big), \phi \in (0,1). \tag{1.12}$$

Para ϕ tendendo a 1 temos $C_\phi(u,v) = uv$ denotando independência. Além disso, a sua função geradora é dada por:

$$\varphi(t) = -ln\Big(\frac{1-\phi^t}{1-\phi}\Big). \tag{1.13}$$

e sua medida de concordância Tau de Kendal é $\tau_\phi = 1 + \frac{4}{ln(\phi)}\Big(\frac{1}{ln(\phi)}\int_0^{-ln(\phi)} \frac{t}{e^t - 1} dt + 1\Big)$.

Outro exemplo de cópula Arquimediana é a Cópula de Ali-Mikhail-Haq. Essa função cópula assume a seguinte forma:

$$C_\phi(u,v) = \frac{uv}{1 - \phi(1-u)(1-v)}, -1 \leq \phi \leq 1. \tag{1.14}$$

Para ϕ tendendo a 0 temos $C_\phi(u,v) = uv$ denotando independência. Além disso, a sua função geradora é dada por:

$$\varphi(t) = ln\frac{1-\phi(1-t)}{t}. \tag{1.15}$$

e sua medida de concordância Tau de Kendal é $\tau_\phi = 1 - \frac{2(\phi\ +\ (1-\phi)^2\ log(1-\phi))}{3\phi^2}$.

Para a construção da função de verossimilhança, a função de sobrevivência conjunta dada pela cópula AMH, tomando as funções de sobrevivência $u = S_1(t_1)$ e $v = S_2(t_2)$, é dada por:

$$S(t_1,t_2) = C_\phi(S_1(t_1), S_2(t_2)) = \frac{S_1(t_1)S_2(t_2)}{1-\phi(1-S_1(t_1))(1-S_2(t_2))} = \frac{S_1(t_1)S_2(t_2)}{1-\phi F_1(t_1)F_2(t_2)} \tag{1.16}$$

Para a análise bayesiana, foram consideradas as seguintes distribuições *a priori* independentes $\alpha_j \sim Gama(0,1;0,01)$ e $\beta_{ij} \sim N(0,10^3)$, tanto para o caso da distribuição Weibull quanto para a distribuição Exponencial Generalizada, $i = 0,1$ e $j = 1,2$. Para o parâmetro da cópula AMH, foi considerada $\phi \sim U(-1,1)$. Além disso foram realizadas as densidades *a posteriori*.

Para a construção da função de verossimilhança, a função de sobrevivência conjunta dada pela cópula de Frank, tomando as funções de sobrevivência $u = S_1(t_1)$ e $v = S_2(t_2)$, é dada por:

$$S(t_1,t_2) = C_\phi(S_1(t_1), S_2(t_2)) = log_\phi\Big(1 + \frac{(\phi^{S_1(t_1)}-1)(\phi^{S_2(t_2)}-1}{\phi-1}\Big), \phi \in (0,1). \tag{1.17}$$

Para realizarmos a análise bayesiana para o modelo da cópula de Frank, foram consideradas as seguintes distribuições *a priori* independentes $\alpha_j \sim Gama(0,1;0,01)$ e $\beta_{ij} \sim N(0,10^3)$ para o caso das duas distribuições, $i = 0,1$ e $j = 1,2$. Assumimos $\phi \sim U(0,1)$ para o parâmetro da cópula de Frank. Foram realizadas as densidades *a posteriori*.

Ribeiro (2017) realizou um estudo de simulação para verificar o bom comportamento das estimativas Bayesianas, com base na média *a posteriori*, além de realizar a comparação dos modelos por meio das medidas EAIC, EBIC, DIC e LPML. Ainda, ao final do estudo, faz uma comparação quanto ao desempenho dos modelos de sobrevivências apresentados. Os modelos das Cópulas Arquimedianas de Ali-Mikhail-Haq e Frank com distribuições marginais Weibull foram os que revelaram um melhor desempenho, de acordo com os critérios de comparção de modelos.

Além da construção de modelos bivariados utilizando cópulas, como adotado por Ribeiro (2017), existem outras maneiras de construir modelos bivariados, por exemplo, utilizando fatoração, o método da transformação marginal, o método de construção via cópulas, o método de mistura e combinação, entre outros. Alguns trabalhos com algumas dessas abordagens foram encontrados na literatura e se encontram na sequência.

1.3 Modelos Bivariados

Dentre os tipos de modelos bivariados, encontram-se modelos cujas respostas são ambas contínuas, ambas discretas ou ainda modelos cujas respostas são mistas, uma discreta e outra contínua.

1.3.1 Modelos Bivariados com Respostas Discretas

Nos modelos bivariados com respostas discretas, temos o trabalho de Khafri, Kazemnejad e Eskandari (2008) que expõe uma análise de dados de fertilização *in vitro* em casais inférteis, tomando como variáveis resposta o número de embriões obtidos e o número de óvulos maduros. O objetivo da pesquisa é encontrar o efeito de diferentes fatores clínicos e demográficos nos dois resultados, simultaneamente. Nesse trabalho, os autores utilizaram o modelo bivariado de Poisson.

O interesse do trabalho está em apresentar modelos nos quais as observações são correlacionadas entre as respostas, entretanto não correlacionadas entre os indivíduos. Assim, seja Y_{ij}, variáveis aleatórias, com $i = 1, ..., n_j$ e $j = 1, 2$, para o indivíduo i e a resposta j, em que $\mathbf{Y}_i = (Y_{i1}; Y_{i2})$ é o vetor de contagem para o indivíduo i em relação as duas respostas. Dessa forma,

$$Cov(Y_{ij}, Y_{kl}) \begin{cases} = 0 & para \quad i \neq k \\ \neq 0 & para \quad i = k; j \neq l \end{cases}$$

Com isso, Y_{ij} pode assumir como uma variável aleatória uma distribuição Poisson, com parâmetro λ_{ij}, ou seja,

$$(Y_{ij}|Z_{ij}, b_{ij}) \sim \text{Poisson}(\lambda_{ij})$$

onde, $\lambda_{ij} = exp(z'_{ij}\beta_j + b_{ij})$, com $i = 1, 2, ..., n$ e $j = 1, 2, z_{ij}$ representando a variável explicativa, β_j um vetor $k \times 1$ de parâmetros e b_{i1} e b_{i2} representando os componentes que modelam a dependência entre Y_{i1} e Y_{i2}. Para construirmos a função densidade, vamos considerar $\mathbf{b}_i = (b_{i1}, b_{i2})$, sendo que $\mathbf{b}_i \sim N_2(0, \sum)$, em que $\sum$ é uma estrutura de variância e covariância para comportar a correlação entre b_{i1} e b_{i2}. Então, a função densidade de probabilidade para o vetor de contagem $\mathbf{Y}_i = (Y_{i1}; Y_{i2})$ é dada por:

$$f(y_{i1}, y_{i2} \mid z_i, \beta_j, \Sigma) = \int \prod_{j=1}^{2} f(y_{ij} \mid \beta_j, b_{ij}) \phi_j(\mathbf{b}_i \mid 0, \Sigma) d\mathbf{b}_i. \tag{1.18}$$

A integral em 1.18 é intratável analiticamente e, por este motivo, o método MCMC, em um contexto Bayesiano, pode ser utilizado. Nesse sentido, Khafri, Kazemnejad e Eskandari (2008) optaram por utilizar os modelos Bayesianos hierárquicos (BHM).

Segundo Wikle et al. (2013), a ideia principal dos modelos Bayesianos hierárquicos é considerar o modelo do conjunto de dados, processos e parâmetros como três componentes gerais relacionados ao modelo, ou seja, os dados condicionados ao processo e aos parâmetros, o processo condicionando aos parâmetros e os parâmetros.

As distribuições a priori para os parâmetros do modelo proposto por Khafri, Kazemnejad e Eskandari (2008) foram, para o primeiro nível de hierarquia, $\beta \sim N_k(\mu_\beta, V_\beta^{-1})$ e $\Sigma^{-1} \sim Wishart(\mu_\Sigma, V_\Sigma)$ e, para o segundo nível de hierarquia, $\mu_\beta \sim N_k(\beta_0, \sigma_0)$, $V_\beta^{-1} \sim Wishart(\mu_{0_\beta}, V_{0_\beta})$ e $V_\beta^{-1} \sim Wishart(\mu_{0_\Sigma}, V_{0_\Sigma})$, em que β_0, σ_0, μ_{0_β}, μ_{0_Σ}, V_{0_β} e V_{0_Σ} são conhecidos, supondo que β_0 e σ_0 são independentes. Por construção, $\boldsymbol{\lambda}_i = (\lambda_{i1}, \lambda_{i2})$ tem uma distribuição lognormal bivariada e $\mathbf{Y}_i = (Y_{i1}, Y_{i2})$ tem uma distribuição poisson-lognormal bivariada.

Conhecida as observações, a distribuição a posteriori conjunta dos parâmetros do modelo foi obtida combinando a função de verossimilhança e as distribuições a priori, via teorema de Bayes. Como a distribuição a posteriori é intratável analiticamente foi utilizado, para a estimação dos parâmetros do modelo, um método de *Gibbs Sampler* para modelos hierárquicos.

Nos resultados encontrados pelos autores Khafri, Kazemnejad e Eskandari (2008), foi observado que todas as distribuições a posteriores são simétricas em relação a suas médias e que as estimativas dos parâmetros sugerem que as informações femininas desempenham um papel importante na previsão do número de óvulos maduros e embriões obtidos. Além disso, mostrou a falta de influência dos parâmetros do sexo masculino no número de embriões obtidos.

Em relação às técnicas de reprodução assistida em dois estágios de fertilização *in vitro*, realizados em diferentes épocas do ano, o modelo apresentado não apontou diferenças significativas. Ademais, é esperado que o modelo de regressão Lognormal Poisson Bivariado renda um modelo de predição superior, dado que o número de óvulos maduros e de embriões obtidos são considerados correlacionados em um mesmo caso.

1.3.2 Modelos Bivariados com Respostas Contínuas

Em relação aos modelos bivariados com respostas contínuas, temos o trabalho de Scollnik (2002), no qual o autor faz uma análise em relação à apólice de seguro contra acidentes. O objetivo foi descobrir se as despesas administrativas em um determinado acidente etão diretamente relacionadas ao pagamento da própria despesa. Neste estudo, o estudioso apresenta dois modelos para o ajuste de dados de despesa, um modelo Pareto-Gama e um modelo Pareto-Pareto.

No modelo Pareto-Gama, a variável X segue uma distribuição Pareto, com parâmetros α e θ e a variável Y segue uma distribuição Gama com parametros δ e λ_x, sendo

$\lambda_x = \exp(\gamma + \beta \log x)$.

A função densidade conjunta para esse modelo é dada por:

$$f(x, y \mid \Psi) = \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}} \frac{\lambda_x^\alpha y^{\delta-1}}{\Gamma(\delta) exp(y\lambda_x)} \tag{1.19}$$

em que $\Psi = (\alpha, \theta, \delta, \gamma, \beta)$, em que α, θ e δ dever assumir valores positivos e γ e β podem assumir tanto valores positivos quanto negativos.

No modelo Pareto-Pareto, a variável X segue uma distribuição Pareto, com parâmetros α e θ e a variável Y também segue uma distribuição Pareto, com parametros δ e λ_x, onde $\lambda_x = exp(\gamma + \beta logx)$.

A função densidade conjunta para este tal modelo é dada por:

$$f(x, y \mid \Psi) = \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}} \frac{\delta\lambda_x^\delta}{(y+\lambda_x)^{\delta+1}} \tag{1.20}$$

em que $\Psi = (\alpha, \theta, \delta, \gamma, \beta)$, em que α, θ e δ dever assumir valores positivos e γ e β podem assumir tanto valores positivos quanto negativos.

Para ambos os modelos, é possível observar a esperança condicional $E(Y \mid X = x, \Psi)$. A esperança do modelo Pareto-Gama sempre existe e tem valor $\frac{\delta}{\lambda_x}$. Já para o modelo Pareto-Pareto, só existe quando $\delta > 1$ e tem valor $\frac{\lambda_x}{\delta-1}$.

Ao considerar a variável Y, uma distribuição condicional gama, no modelo Pareto-Gama, e, pareto, no modelo Pareto-Pareto, com parâmetros δ e $\tilde{\lambda}_x$, com $\tilde{\lambda}_x = exp(\gamma + \beta[log(x) - k])$, em que k é igual o valor médio observado de $log(x)$, Scollnik (2002) obteve uma versão centralizada da função de densidade dos modelos. Para encontrar as estimativas de máxima verossimilhança, $\hat{\alpha}$, foi utilizado os métodos padrões. Os valores para essas estimativas, em ambos os modelos, sugere que os dois primeiros momentos de X existem.

Nesse estudo, Scollnik (2002) realizou uma análise Bayesiana, pois ao realizar um teste de hipótese, observou uma implicação de que as inferências preditivas condicionadas ao valor de $\hat{\alpha}$ podem estar sujeitas a uma variabilidade significamente menor. Consequentemente, a perda total prevista pode ser subestimada. Portanto, tal análise é adequada para explicar a incerteza inerente do parâmetro.

Para a estimação dos parâmetros, foi considerada uma abordagem Bayesiana, via MCMC, implementada utilizando o pacote WinBUGS. Esse processo resultou em distribuições preditivas significativamente mais dispersas porque incorpora a incerteza do parâmetro que é efetivamente ignorada pela análise de máxima verossimilhança.

Os resultados obtidos por Scollnik (2002) mostraram que a distribuição preditiva Bayesiana para a variável perda, em cada caso de cobertura, é mais dispersa do que a

distribuição preditiva correspondente às estimativas de máxima verossimilhança.

1.3.3 Modelos Bivariados com Respostas Mistas

Para os modelos bivariados com respostas mistas, apresentamos os trabalhos desenvolvidos por Catalano e Ryan (1992), Fitzmaurice e Laird (1995) e Oliveira, Diniz e Durbán (2019).

1.3.3.1 Modelo de Catalano e Ryan (1992)

O modelo introduzido por Catalano e Ryan (1992) foi motivado por estudos de toxicidade em ratas grávidas, considerando má formação fetal como variável resposta discreta e peso fetal como variável resposta contínua. Para este tal modelo, as respostas contínuas e discretas podem assumir, primeiramente, a independência entre as observações em diferentes fetos e depois a depedência.

O modelo bivariado sugerido, considerando d_i a dose administrada na i-ésima rata gravida, é dado por:

$$\begin{aligned} Y_{1ij} &= \alpha_0 + \alpha_1 d_i + \epsilon_{1ij} \\ Y_{2ij} &= \beta_0 + \beta_1 d_i + \epsilon_{2ij} \end{aligned}$$

sendo que a variável Y_{1ij} é relacionada ao peso fetal e a variável Y_{2ij} é uma variável latente não observada correspondente à má formação para o feto j da ninhada i, com $j = 1, ..., n_i$ e $i = 1, ..., I$. Os parâmetros α e β correspondem ao peso fetal e à má formação, respectivamente.

Para um modelo bivariado independente, suponha que os fetos na mesma ninhada são independentes e que

$$\epsilon_{ij} = \begin{pmatrix} \epsilon_{1ij} \\ \epsilon_{2ij} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \tau\sigma_1\sigma_2 \\ \tau\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right)$$

em que os resíduos ϵ_{ij} são independentes para todos os i e j.

Para um mesmo feto, as variáveis têm correlação constante τ para todo i e j. A variável indicadora observada para má formação, denotada por Y^*_{2ij}, é determinada pela variável latente Y_{2ij}, para o feto j da ninhada i, ou seja,

$$Y^*_{2ij} = \begin{cases} 1 & \text{se} \quad Y_{2ij} > 0 \\ 0 & \text{se} \quad Y_{2ij} \leq 0 \end{cases}$$

em que $Y^*_{2ij} = 1$ indica ocorrência de má formação.

Baseado no modelo descrito inicialmente $\mathbf{Y}^*_2 = (Y_{2i1}, Y_{2i2}, ..., Y_{2in_i})$, segue um modelo probito, isto é:

$$P(Y^*_{2ij} = 1) = \Phi\left(\frac{\beta_0 + \beta_1 d_i}{\sigma_2} \right).$$

A distribuição conjunta de Y_{1ij} e Y^*_{2ij} é obtida pelo produto da distribuição marginal e da distribuição condicional. $Y_{2ij} \mid Y_{1ij}$ segue uma distribuição normal com uma média que depende do resíduo do modelo para: $\mathbf{Y}_1$.

$$Y_{2ij} \mid Y_{1ij} \sim N(\mu_1, \sigma_2^2(1-\tau^2)), \tag{1.21}$$

em que $\mu_1 = \beta_0 + \beta_1 d_i + \left(\frac{\sigma_2}{\sigma_1}\right)\tau e_{1ij}$, em que $e_{1ij} = Y_{1ij} - (\alpha_0 + \alpha_1 d_i)$ é o resíduo do modelo para o peso.

A partir da distribuição 1.21, obtêm-se a distribuição condicional de $Y^*_{2ij} \mid Y_{1ij}$ que também é um modelo probito, isto é:

$$P(Y^*_{2ij} = 1 \mid Y_{1ij}) = \Phi\left(\frac{\mu_1}{\sqrt{\sigma_2^2(1-\tau^2)}}\right),$$

que pode ser reparametrizada, ou seja:

$$P(Y^*_{2ij} = 1 \mid Y_{1ij}) = \Phi(\beta^*_0, \beta^*_1 d_i + \beta^*_2 e_{1ij}).$$

Assim, com altos valores da variável latente Y_{2ij} resultam em uma má formação, espera-se que Y_{1ij} e Y_{2ij} sejam negativamente correlacionadas, ou seja, $\tau < 0$.

Para o um modelo dependente, seja τ a correlação constante entre as observações de um mesmo feto e, ρ_1 e ρ_2 as correlações separadas entre as observações em diferentes fetos na mesma ninhada, relacionadas a peso e a má formação, respectivamente.

Para a ninhada i, considere que $\mathbf{Y}_{1i} = (Y_{1i1}, ..., Y_{1in_i})'$ seja o vetor para o peso e $\mathbf{Y}_{2i} = (Y_{2i1}, ..., Y_{2in_i})'$ seja o vetor para a variável latente. Tomando $\mathbf{Y}_i = ((\mathbf{Y}_{1i_1}, \mathbf{Y}_{2i_1}), (\mathbf{Y}_{1i_2}, \mathbf{Y}_{2i_2}), ..., (\mathbf{Y}_{1i_{n_i}}, \mathbf{Y}_{2i_{n_i}}))$ como vetor de observações bivariadas para a ninhada i, $\mathbf{Y}_i$ tem distribuição normal multivariada com média:

$$E(\mathbf{Y}_i) = E\begin{pmatrix} \mathbf{Y}_{1i} \\ \mathbf{Y}_{2i} \end{pmatrix} = \begin{pmatrix} 1 & d_i 1 & 0 & 0 \\ 0 & 0 & 1 & d_i 1 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \beta_0 \\ \beta_1 \end{pmatrix}$$

e matriz de covariância:

$$Var(\mathbf{Y}_i) = \begin{pmatrix} \sigma_1^2[(1-\rho_1)\mathbf{I} + \rho_1\mathbf{J}] & \sigma_1\sigma_2[(\tau-\rho_{12})\mathbf{I} + \rho_{12}\mathbf{J}] \\ \sigma_1\sigma_2[(\tau-\rho_{12})\mathbf{I} + \rho_{12}\mathbf{J}] & \sigma_2^2[(1-\rho_2)\mathbf{I} + \rho_2\mathbf{J}] \end{pmatrix}$$

em que $\mathbf{I}$ é a matriz identidade e $\mathbf{J}$ é uma matriz de "uns".

A distribuição condicional da variável latente, dado o vetor do peso fetal, é normal $\mathbf{Y}_{2i} \mid \mathbf{Y}_{1i} \sim N_{n_i}(\mu_{2i}, \sigma_2^2\Sigma_i)$, em que o j-ésimo elemento da média condicional é dado por:

$$\mu_{2ij} = \beta_0 + \beta_1 d_i + \frac{\sigma_2}{\sigma_1}\left(\frac{\tau + (n_i-1)\rho_{12}}{1 + (n_i-1)\rho_1}\right)\bar{e}_{1i} + \frac{\sigma_2}{\sigma_1}\left(\frac{\tau - \rho_{12}}{1 - \rho 1}\right)(e_{1ij} - \bar{e}_{1i})$$

em que $e_{1ij} = Y_{1ij} - (\alpha_0 + \alpha_1 d_i)$, sendo o resíduo do modelo para Y, $\bar{e}_{1i} = \bar{Y}_{1i} - (\alpha_0 + \alpha_1 d_i)$, sendo a média do resíduo e_{1ij} e

$$\Sigma_i = \left[(1-\rho_2) - \frac{(\tau-\rho_{12})^2}{1-\rho_1}\right]\mathbf{I} + \left[\rho_2 - \frac{(1-\rho_1)(\tau^2+(n_i-1)\rho_{12}^2) - (\tau-\rho_{12})^2}{(1-\rho_1)(1+(n_i-1)\rho_1}\right]\mathbf{J}.$$

A distribuição condicional para o indicador de má formação observável, dado peso fetal, segue um modelo probito correlacionado com:

$$E(Y_{2ij}^* \mid Y_{1ij}) = \Phi\left(\frac{\mu_{2ij}}{\sigma_3}\right)$$

e

$$Var(Y_{2ij}^* \mid Y_{1ij}) = \Phi\left(\frac{\mu_{2ij}}{\sigma_{3i}}\right)\left[1 - \Phi\left(\frac{\mu_{2ij}}{\sigma_{3i}}\right)\right]$$

em que

$$\sigma_{3i}^2 = \sigma_2^2\left[1 - \frac{\tau^2(1-\rho_1) + (n_i-1)[\rho_1(\rho_{12}-\tau)^2 + (1-\rho_1)\rho_{12}^2}{(1-\rho_1)(1+(n_i-1)\rho_1)}\right]$$

Para o processo de estimação, Catalano e Ryan (1992) apresentaram um método baseado na fatoração da distribuição conjunta em dois componentes de regressão, a distribuição marginal de $\mathbf{Y}_1$ e a distribuição condicional de $\mathbf{Y}_2^* \mid \mathbf{Y}_1$, e, em cada componente, aplicaram uma abordagem de equações de estimações generalizadas (GEE).

Para o desenvolvimento dessa metodologia, as estimações foram realizadas em duas etapas. A primeira foi ajustar uma regressão correlacionada ao peso $\mathbf{Y}_1$ sobre a dose d e tamanho da ninhada $(n_i - \bar{n})$. A partir dessa etapa, foram obtidas as estimativas e variâncias dos parâmetros do modelo, α_0, α_1 e α_2 bem como a estimativa de ρ_1.

A segunda etapa consistiu em ajustar uma regressão probito correlacionada de $\mathbf{Y}_2^* \mid \mathbf{Y}_1$, usando a dose, assim como resíduos do peso fetal médio e individual e covariáveis do tamanho da ninhada como variáveis explicativas e um parâmetro de correlação constante para explicar o efeito da ninhada.

Na sequência, apresentamos alguns trabalhos que também utilizam a técnica da fatoração na construção da distribuição conjunta, em que a distribuição bivariada é dada pelo produto entre uma distribuição marginal e uma distribuição condicional.

1.3.3.2 Modelo de Fitzmaurice e Laird (1995)

Fitzmaurice e Laird (1995) desenvolveram um modelo bivariado com resposta binária a partir da análise de dados de um experimento de peso fetal e toxicidade epitelial$(*)$ por má formação do etileno glicol em camundongos.

No modelo, a variável resposta discreta é uma distribuição Bernoulli e representa a má formação, enquanto que a variável resposta contínua segue uma distribuição Normal e é representada pela variável peso, condicionada à má formação.

Seja Y_i uma variável discreta tal que $Y_i \sim Bernoulli(\mu_{iY})$, em que μ_{iY} é a média de Y_i, com $i = 1, ..., n$. A distribuição marginal de Y_i é Bernoulli e pertence à familia exponencial. A variância é dada por $\mathbb{V}ar(Y_i) = \mu_{iY}(1 - \mu_{iY})$.

Considerando p um conjunto de covariáveis $z_{i1}, z_{i2}, ..., z_{ip}$ e que esteja disponível para indicar Y_i, então:

$$\eta_{iY} = \beta_0 + \beta_1 z_{i1} + ... + \beta_p z_{ip} = Z_{i1} B_1$$

em que $\beta_i, i = 1, ..., p$ são os coeficientes de regressão.

Além disso, considerando que $g_1(\cdot)$ é uma função de ligação canônica para o modelo Bernoulli, ou seja,

$$g_1(\mu_{iY}) = log\left(\frac{\mu_{iY}}{1 - \mu_{iY}}\right) = \eta_{iY}.$$

tem-se

$$\mu_{iY} = g_1^{-1}(\eta_{iY}) = \frac{exp(\eta_{iY})}{1 + exp(\eta_{iY})}.$$

Seja X_i a variável resposta contínua que condicionada à variável resposta $Y_i = y_i$, segue uma distribuição normal, ou seja, $X_i \mid Y_i = y_i \sim N(\alpha_i, \sigma^2)$, em que α_i é a média e σ^2 é a variância de $X_i \mid Y_i = y_i, i = 1, ..., n$. A função densidade de $X_i \mid Y_i = y_i$ também é escrita na forma da família exponencial.

Considerando que o mesmo conjunto p de covariáveis $z_{i1}, z_{i2}, ..., z_{ip}$, esteja disponível para predizer X_i, e denotando por $\mu_{iX} = E(X_i)$ a média marginal da variável resposta contínua X_i, relaciona-se μ_{iX} às covariáveis disponíveis por $g_2(\mu_{iX}) = \eta_{iX}$, em que $g_2(\cdot)$ é uma função de ligação monótona e diferenciável e η_{iX} é o componente sistemático dado por:

$$\eta_{iX} = \delta_0 + \delta_1 z_{i1} + ... + \delta_p z_{ip} = Z_{i1} \Delta_1$$

A estrutura de dependência entre as variáveis respostas discretas e contínuas é inserida no modelo pela média condicional da variável resposta contínua X_i dada a variável resposta discreta Y_i, isto é,

$$\alpha_i = h(y_1, \mu_{iY}, \mu_{iX}, \gamma) = \mu_{iX} + \gamma(y_i - \mu_{iY}),$$

em que $h(\cdot)$ é uma função linear ou não linear de y_i, μ_{iY}, μ_{iX} e γ, um parâmetro desconhecido. Portanto, $\alpha_i = \mu_{iX} + \gamma(y_i - \mu_{iX})$.

É possível notar que $\mathbb{E}(\alpha_i) = \mathbb{E}[\mu_{iX} + \gamma(y_i - \mu_{iX})] = \mathbb{E}[\mu_{iX}] = \mu_{iX}$, mostrando que a escolha da função $h(\cdot)$ está de acordo com as especificações do modelo, como mencionado em Oliveira, Diniz e Durbán (2019).

1.3.3.3 Modelo de Oliveira, Diniz e Durbán (2019)

Oliveira, Diniz e Durbán (2019) propuseram uma classe de modelos bivariados de resposta mista, a partir da aplicação de um conjunto de dados concedido por uma operadora de plano de saúde. O objetivo era determinar a relação entre o gasto obtido e a utilização do centro cirúrgico com outras variáveis coletadas.

Como caso particular, para variável resposta discreta, os autores se apropriaram de uma distribuição Bernoulli, pois exige apenas duas possibilidades, a utilização ou não do centro cirúrgico. Já para a variável resposta contínua, condicionada à variável resposta discreta, apropriaram-se da distribuição Exponencial.

Com isso, Y_i é uma variável resposta discreta com parâmetros μ_{iY}, que está associada a um conjunto de p covariáveis $z_{i1}, z_{i2}, ..., z_{ip}$, através da relação $g_1(\mu_{iY}) = \eta_{iY}$ com η_{iY} e $g_1(\cdot)$, conforme exposições na subseção 2.3.3.2, $i = 1, ..., n$, e X_i é uma variável resposta contínua na qual a distribuição condicional de X_i, dado a variável resposta discreta $Y_i = y_i$ é Exponencial com parâmetro $\frac{1}{\alpha_i}$, na qual $\alpha_i = \mathbb{E}(X_i \mid Y_i), i = 1, ..., n$. A variância de $X_i|Y_i = y_i$ é dada por $Var(X_i|Y_i = y_i) = \alpha_i^2$.

É possível assumir que há um conjunto de q covariáveis $t_{i1}, t_{i2}, ..., t_{iq}$ para predizer X_i. Assim, o preditor é dado por:

$$\eta_{iX} = \delta_0 + \delta_1 t_{i1} + ... + \delta_q t_{iq} = T_{i1}\Delta_1.$$

Considerando $g_2(\cdot)$ a função de ligação logarítmica, tem-se $\mu_{iX} = exp(\eta_{iX})$. Por fim, os autores propõem algumas estruturas de dependência (distintas funções $h(\cdot)$), sendo a escolha de

$$h(y_i, \mu_{iY}, \mu_{iX}, \gamma) = (1 - y_i + \mu_{iY})\mu_{iX}.$$

a mais adequada para o conjunto de dados analisados. É possível notar que neste caso, tem-se que $\alpha_i = (1 - y_i + \mu_{iY})\mu_{iX}$ e assim, $\mathbb{E}(\alpha_i) = \mathbb{E}[(1 - y_i + \mu_{iY})\mu_{iX}] = \mu_{iX}$ é verificada.

1.3.3.4 Modelo de Stulp (2019)

Stulp (2019) propôs um modelo bivariado de resposta mista, em que a variável resposta discreta segue uma distribuição de Poisson e a variável resposta contínua segue uma distribuição Exponencial, um caso particular da distribuição Weibull quando $a = 1$.

A função conjunta do modelo Poisson-Exponencial foi construída através do método da fatoração, fazendo o produto da função de probabilidade da distribuição Poisson com a função densidade de probabilidade condicional da distribuição Exponencial, que é representada pela seguinte expressão:

$$f(x_i, y_i) = \left(\frac{e^{-\mu_{iy}}\mu_{iy}^{y_i}}{y_i!}\right) * \left(\frac{1}{b_i} exp\left(-\frac{x_i}{b_i}\right)\right), \tag{1.22}$$

na qual adota-se $b_i = \frac{\gamma^{y_i}\mu_{ix}}{1+\mu_{iy}(\gamma-1)}$, uma função duas vezes continuamente diferenciável em relação ao vetor de observações y_i, às médias marginais μ_{iy} e μ_{ix} e ao parâmetro γ que é um parâmetro de associação entre a variável resposta discreta e a variável resposta contínua. Além disso, as médias marginais são relacionadas a covariáveis através de funções de ligação.

Para a estimação dos parâmetros, Stulp (2019) utilizou a estimação por máxima verossimilhança, que tem por definição o objetivo de maximizar a função de verossimilhança ou a função de log-verossimilhança. Muitas vezes, o sistema não é linear e, por isso, são utilizados métodos iterativos, tal como o Newton-Raphson ou escore de Fisher.

É possível observar que, no geral, os trabalhos descritos que utilizam o método da fatoração consideram a estimação dos parâmetros a partir de uma abordagem clássica, e são poucos os trabalhos de modelos bivariados que utilizam o método da fatoração dentro do contexto bayesiano. Dessa forma, com a intenção de corroborar com trabalhos realizados a partir da perspectiva Bayesiana, este trabalho objetiva abordar o modelo apresentado por Stulp (2019), que já foi estudado na abordagem clássica, agora dentro do contexto bayesiano.

Capítulo 2

Modelo Poisson-Exponencial

Resumo

Aqui é proposta uma metodologia Bayesiana para a estimação dos parâmetros no modelo bivariado Poisson-Exponencial. Este é um modelo bivariado de resposta mista, em que a resposta discreta segue uma distribuição Poisson e a resposta contínua, condicionada à discreta, segue uma distribuição Exponencial. Para o método Bayesiano, são adotadas distribuições *a priori* não informativas e para o cálculo das estimativas *a posteriori* é utilizado um algoritmo MCMC. Foi realizada uma análise de resíduos, baseadas na densidade Preditiva Condicional Ordinária (CPO) e os resíduos baseados na distribuição *a posteriori*. É apresentado,também, um estudo de simulação, a fim de verificar a qualidade das estimativas e a performance dos resíduos.

Palavras-chave: Modelos Bivariados, Respostas Discretas e Contínuas, Método de Estimação Bayesiano, Análise de Resíduos Bayesianos.

Abstract

A Bayesian methodology is proposed for the Poisson-Exponential bivariate model. This is a mixed response bivariate model, where the discrete response follows a Poisson distribution and the continuous response, conditioned to the discrete, follows an Exponential distribution. For the Bayesian method, non-informative *a priori* distributions are adopted and for calculating the estimates *a posteriori* an MCMC algorithm is used. An analysis of residues was carried out, based on the Ordinary Conditional Predictive density (CPO) and the residues based on the *a posteriori* distribution. A simulation study is also presented in order to verify the quality of the estimates and the performance of the residues.

Keywords: Bivariate models. Discrete and Continuous Responses. Dependence Between Responses. Analysis of Bayesian Residues.

2.1 Introdução

O modelo de regressão Poisson-Exponencial é um modelo bivariado proposto por Stulp (2019), no qual a variável resposta discreta segue uma distribuição Poisson e a variável resposta contínua, condicionada à variável resposta discreta, segue uma distribuição Exponencial.

Dessa forma, seja Y_i uma variável resposta discreta que segue uma distribuição Poisson tem função de probabilidade é dada por:

$$\mathbb{P}(Y_i) = \frac{e^{-\mu_{iy}}\mu_{iy}^{y_i}}{y_i!}, \tag{2.1}$$

com $\mathbb{E}(Y_i) = \mathbb{V}ar(Y_i) = \mu_{iy}$.

Seja também X_i uma variável resposta contínua, condicionada à variável resposta discreta Y_i, que segue uma distribuição Exponencial de parâmetro b_i, com $b_i > 0$ sua função densidade de probabilidade condicional é dada por:

$$f(x_i \mid y_i) = \frac{1}{b_i} exp\Big(-\frac{x_i}{b_i}\Big), \qquad x_i > 0 \tag{2.2}$$

com $\mathbb{E}(X_i \mid Y_i = y_i) = b_i$ e $\mathbb{V}ar(X_i|Y_i = y_i) = b_i^2$.

Nesse modelo, foram considerados p e q conjuntos de covariáveis que se relacionam às médias marginais através das relações $\eta_{iy} = g_1(\mu_{i_y}) = Z_{ip}\beta_p$ e $\eta_{ix} = g_2(\mu_{i_x}) = T_{iq}\Delta_q$, em que η_{iy} e η_{ix} são preditores lineares e $g_1(\cdot)$ e $g_2(\cdot)$ são funções de ligação canônicas para Y_i e $X_i|Y_i = y_i$, respectivamente. Assim, segue que $\mu_{i_y} = exp(Z_{ip}\beta_p)$ e $\mu_{i_x} = exp(T_{iq}\Delta_q)$.

Para a dependência entre as duas variáveis respostas, foi estabelecida uma estrutura através da média condicional da variável $X_i \mid Y_i = y_i$, ou seja, $b_i = \mathbb{E}(X_i \mid Y_i = y_i) = h(y_i, \mu_{iy}, \mu_{ix}, \gamma)$, cuja função $h(\cdot)$ é conhecida e duas vezes continuamente diferenciável em relação ao vetor de observações y_i, às médias marginais μ_{iy} e μ_{ix} e ao parâmetro γ incluído no modelo.

A função $h(\cdot)$ determina o tipo de dependência que existe entre as respostas e foi determinada de tal forma que tal função tomasse valores no espaço paramétrico da média b_i e, ainda, satisfizesse a propriedade de $\mathbb{E}(\mathbb{E}(X_i \mid Y_i = y_i)) = \mathbb{E}(X_i) = \mu_{ix}$. Oliveira, Diniz e Durbán (2019) propuseram uma possível função para estabelecer uma estrutura de dependência entre as variáveis repostas, representada pela seguinte equação:

$$h(y_i, \mu_{iy}, \mu_{ix}, \gamma) = \frac{\gamma^{y_i}\mu_{ix}}{1 + \mu_{iy}(\gamma - 1)}, \tag{2.3}$$

na qual γ é um parâmetro de associação entre a variável resposta discreta e a variável resposta contínua. Haja vista que o espaço paramétrico do modelo Exponencial é o mesmo

do parâmetro de forma da distribuição Weibull, uma vez que a distribuição Exponencial é um caso particular da distribuição Weibull, foi adotada a mesma estrutura utilizada (OLIVEIRA; DINIZ; DURBÁN, 2019).

Através de 2.3, foi possível verificar a independência entre Y_i e X_i. Para isso, basta testar se o parâmetro γ de $h(\cdot)$ é igual a 1. Assim, um possível teste de independência entre as respostas seria $H_0 : \gamma = 1$ versus $H_1 : \gamma \neq 1$. A rejeição da hipótese nula indica que as variáveis respostas são dependentes.

Utilizando as funções de ligação para relacionar as médias marginais às covariáveis e fazendo o produto de (2.1) por (2.2), é possível escrever a função conjunta do modelo bivariado Poisson-Exponencial, a qual é dada por:

$$f(x_i, y_i) = \left(\frac{e^{-\mu_{iy}}\mu_{iy}^{y_i}}{y_i!}\right)\left(\frac{1}{b_i}exp\left(-\frac{x_i}{b_i}\right)\right), \tag{2.4}$$

com $b_i = \frac{\gamma^{y_i}\mu_{ix}}{1+\mu_{iy}(\gamma-1)}$.

Em seu trabalho, Stulp (2019) apresenta a proposta do modelo bivariado Poisson-Exponencial dentro do contexto de modelagem clássica (frequentista), na qual os parâmetros já são estabelecidos e não se tratam de variáveis aleatórias. Dessa forma, o presente trabalho tem por objetivo apresentar a estimação do modelo bivariado Poisson-Exponencial dentro do contexto Bayesiano, tratando os parâmetros como variáveis aleatórias.

2.2 Estimação

Sejam $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ um conjunto de valores observados de (X_1, Y_1), $(X_2, Y_2), \cdots, (X_n, Y_n)$, com (X_i, Y_i) e (X_j, Y_j) independentes, para $i, j = 1, 2, \cdots, n, i \neq j$, e a função conjunta dada por (2.4). Dada uma amostra de tamanho n, escrevemos a função de verossimilhança de $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$, $\boldsymbol{\Delta} = (\delta_0, \delta_1, \cdots, \delta_q)^T$ e γ da seguinte forma:

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\Delta}, \gamma \mid \textbf{x, y, Z, T}) &= \prod_{i=1}^{n} \mathbb{P}(y_i) * f(x_i) \\ &= \prod_{i=1}^{n} \left(\frac{e^{-\mu_{iy}}\mu_{iy}^{y_i}}{y_i!}\right) * \left(\frac{1}{b_i}\, exp\left(-\frac{x_i}{b_i}\right)\right) \end{aligned} \tag{2.5}$$

com $b_i = \frac{\gamma^{y_i}\mu_{ix}}{1+\mu_{iy}(\gamma-1)}$.

Na análise Bayesiana, é necessário determinar uma distribuição *a priori* específica, pois, caso contrário, não é possível calcular a distribuição *a posteriori* e, portanto, a análise Bayesiana fica comprometida (MAIOLI, 2014). Com o objetivo de comparar com os resultados clássicos vistos em Stulp (2019), consideramos, neste trabalho, distribuições vagas, com grande variabilidade, que não trazem muita informação a respeito dos parâmetros de interesse.

Para encontrarmos a distribuição *a posteriori* dos parâmetros, combinamos as distribuições *à priori* com a função de verossimilhança. Essa operação é possível através do Teorema de Bayes.

Na construção da distribuição *à posteriori* para o modelo Poisson-Exponencial, serão utilizados, como distribuições *à priori*, as distribuições Normal Multivariada para os parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\Delta}$ e Normal para o parâmetro γ. Assim, temos $\boldsymbol{\beta} \sim N_p(\boldsymbol{\theta}_1, \boldsymbol{\Sigma}_1)$, $\boldsymbol{\Delta} \sim N_p(\boldsymbol{\theta}_2, \boldsymbol{\Sigma}_2)$ e $\gamma \sim N(\theta_3, \sigma_3^2)$, independentes, com os hiperparâmetros $\boldsymbol{\theta}_1$, $\boldsymbol{\Sigma}_1$, $\boldsymbol{\theta}_2$, $\boldsymbol{\Sigma}_2$, θ_3 e σ_3^2 conhecidos.

Dessa maneira, a distribuição *a posteriori* para o modelo proposto é:

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\gamma} \mid \textbf{x, y, Z, T}) \;\propto\;& exp\Bigg\{ -\frac{1}{2}\Bigg[\Big((\boldsymbol{\beta}-\boldsymbol{\theta}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\beta}-\boldsymbol{\theta}_1)\Big) \\
&+ \Big((\boldsymbol{\Delta}-\boldsymbol{\theta}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\Delta}-\boldsymbol{\theta}_2)\Big) + \left(\frac{\gamma-\theta_3}{\sigma_1}\right)^2\Bigg]\Bigg\} \\
&\times \prod_{i=1}^{n}\left(\frac{e^{-\mu_{iy}}\mu_{iy}^{y_i}}{y_i!}\right) \times \left(\frac{1}{b_i}exp\left(-\frac{x_i}{b_i}\right)\right)
\end{aligned} \tag{2.6}$$

com $b_i = \frac{\gamma^{y_i}\mu_{ix}}{1+\mu_{iy}(\gamma-1)}$.

A distribuição *à posteriori* que obtivemos não possui uma forma fechada, ou seja, é intratável analiticamente. Para realizarmos a estimação dos parâmetros $\boldsymbol{\beta}, \boldsymbol{\Delta}$ e $\boldsymbol{\gamma}$, aplicamos o procedimento de Monte Carlo via Cadeias de Markov (MCMC).

O algoritmo de Metropolis-Hastings gera uma cadeia de Markov e usa a ideia dos métodos de rejeição, isto é, um valor é gerado de uma distribuição auxiliar e aceito com uma dada probabilidade. Esse mecanismo de correção garante a convergência da cadeia para a distribuição de equilíbrio que, neste caso, é a distribuição *a posteriori* (EHLERS, 2011).

2.3 Análise de Resíduos

Nesta subseção, propomos uma análise de diagnósticos, introduzindo os principais resíduos Bayesianos, com o objetivo de verificar a qualidade do ajuste do modelo aos dados, identificando a existência de pontos *outliers* que possam influenciar no ajuste do modelo.

Consideramos, neste trabalho, um modelo bivariado, construído pela técnica da fatoração, em que a variável aleatória Y_i segue uma distribuição Poisson, a variável aleatória X_i, condicionada a Y_i, segue uma distribuição Exponencial e as observações são independentes. Dessa forma, a análise de resíduos é realizada separadamente para Y_i e para

$X_i|Y_i = y_i$. Neste caso, utilizamos dois tipos de resíduos Bayesianos: o resíduo baseado na densidade preditiva condicional ordinária (CPO) e o resíduo baseado na distribuição *a posteriori* dos parâmetros do modelo.

2.3.1 Resíduos baseados na densidade Preditiva Condicional Ordinária

Cho et al. (2009) apresentam uma forma de avaliar a qualidade do resíduo, através da densidade Preditiva Condicional Ordinária, que compara os valores das respostas preditas pelo modelo ajustado com os valores observados. A densidade preditiva condicional ordinária para a $i-$ésima observação, condicionada a D, considerando o modelo associado à Y_i, é definida como:

$$CPO_i = \left[\int_\Theta \frac{1}{\pi(y_i|D,\boldsymbol{\theta})}\pi(\boldsymbol{\theta}|D)d\boldsymbol{\theta}\right]^{-1} \tag{2.7}$$

em que $D = (\mathbf{x}, \mathbf{y}, \mathbf{Z}, \mathbf{T})$, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Delta}, \gamma)$, $\pi(y_i|D,\boldsymbol{\theta})$ é a função de probabilidade de Y_i e $\pi(\boldsymbol{\theta}|D)$ é a distribuição *à posteriori* de $\boldsymbol{\theta}$

Para encontrar o valor predito para y_i, através da densidade Preditiva Condicional Ordinária, utilizamos uma estimativa de Monte Carlo da função de probabilidade $\pi(y_i|D,\boldsymbol{\theta})$ calculada através de amostras da distribuição *a posteriori* que foram simuladas utilizando técnicas de MCMC $\pi(\boldsymbol{\theta}|D)$ (CHEN; SHAO; IBRAHIM, 2000). O seguinte método numérico fornece o valor predito para y_i:

i. Geramos um conjunto de amostras $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, ..., $\boldsymbol{\theta}_Q$, de tamanho Q, da distribuição *a posteriori* $\pi(\boldsymbol{\theta}|D)$ e, a partir da qual, determinamos $\hat{\mu}_{iy_q}$;

ii. Para cada y_i em $\{0, 1, 2, 3, \ldots\}$, obtenha a estimativa de Monte Carlo para a CPO_i, dada por

$$\widehat{CPO}_i = \left\{\frac{1}{Q}\sum_{q=1}^{Q}\left(\frac{e^{-\hat{\mu}_{iy_q}}\hat{\mu}_{iy_q}^{y_i}}{y_i!}\right)\right\}$$

iii. O valor de $\tilde{y}_i$ em $\{0, 1, 2, 3, \ldots\}$ que maximiza a $\widehat{CPO}_i$ é o valor predito para a variável resposta y_i.

Calculamos o resíduo baseado na CPO como $r_{ppy_i} = y_i - \tilde{y}_i$, em que y_i é a i-ésima resposta observada e $\tilde{y}_i$ é a moda da distribuição preditiva condicional ordinária condicionada ao valor das covariáveis. Padronizamos este resíduo da seguinte maneira:

$$r_{sppy_i} = \frac{r_{ppy_i}}{\sqrt{\widehat{Var}(Y_i)}}, i = 1, \ldots, n$$

em que $\widehat{Var}(Y_i) = \hat{\mu}_{iy}$.

Também, podemos encontrar o valor predito para x_i através da equação (2.7), substituindo a função de probabilidade de Y_i pela função densidade de probabilidade de $X_i|Y_i = y_i$. Assim, utilizamos uma estimativa de Monte Carlo da densidade $\pi(x_i|D, \boldsymbol{\theta})$ com amostras MCMC da distibuição *a posteriori* $\pi(\boldsymbol{\theta}|D)$ (CHEN; SHAO; IBRAHIM, 2000). O seguinte algoritmo fornece o valor predito para x_i:

i. Geramos um conjunto de amostras $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, $\ldots$, $\boldsymbol{\theta}_Q$, de tamanho Q, da distribuição *a posteriori* $\pi(\boldsymbol{\theta}|D)$ e, a partir da qual, determinamos $\hat{\mu}_{iy_q}$, $\hat{\mu}_{ix_q}$, $\hat{\gamma}_q$ e consequentemente, $\hat{b}_{i_q}$;

ii. A integral em (2.7) não possui solução analítica. Para resolvê-la, geramos uma rede de valores na vizinhança de x_i e, para cada observação, retiramos uma amostra dessa rede. Para cada $\hat{x}_i$ na amostra, obtemos a estimativa de Monte Carlo para a CPO, dada por:

$$\widehat{CPO}_i = \left\{ \frac{1}{Q} \sum_{q=1}^{Q} \left(\frac{1}{\hat{b}_{i_q}} exp\Big(- \frac{\hat{x}_i}{\hat{b}_{i_q}} \Big) \right) \right\}$$

com $\hat{b}_{i_q} = \frac{\hat{\gamma}_q^{y_i} \hat{\mu}_{ix_q}}{1+\hat{\mu}_{iy_q}(\hat{\gamma}_q - 1)}$.

iii. O valor de $\tilde{x}_i$ na amostra que maximiza a $\widehat{CPO}_i$ é o valor predito para a variável resposta x_i.

Calculamos o resíduo baseado na CPO como $r_{ppx_i} = x_i - \tilde{x}_i$, em que x_i é a i-ésima resposta observada da variável contínua X e $\tilde{x}_i$ é o i-ésimo valor predito. Padronizamos este resíduo da seguinte maneira:

$$r_{sppx_i} = \frac{r_{ppx_i}}{\sqrt{\widehat{Var}(X_i|Y_i)}}, i = 1, \ldots, n$$

em que $\widehat{Var}(X_i|Y_i) = \hat{b}_i^2$.

É esperado que o conjunto de resíduos, calculados para cada modelo separadamente (Y_i e $X_i|Y_i = y_i$), esteja distribuído de forma aleatória e homogênea, em torno do zero. Caso isso não aconteça, temos indícios de que o modelo está mal ajustado aos dados.

2.3.2 Resíduos baseados na distribuição *a posteriori* dos parâmetros do modelo

Calculamos o resíduo padronizado, para o modelo de Y_i, baseado na distribuição *a posteriori* $\pi(\boldsymbol{\theta}|D)$ como:

$$r_{spdy_i} = \frac{y_i - \hat{E}(Y_i)}{\sqrt{\widehat{Var}(Y_i)}} = \frac{y_i - \hat{\mu}_i y}{\sqrt{\hat{\mu}_{iy}}}$$

em que $\hat{\mu}_{iy} = \Sigma_q^Q \hat{\mu}_{iy_q}$.

De maneira análoga, temos que o resíduo padronizado para o modelo $X_i|Y_i = y_i$, baseado na distribuição *a posteriori* é dado por:

$$r_{spd_i} = \frac{x_1 - \hat{E}(X_i|Y_i)}{\sqrt{\widehat{Var}(X_i|Y_i)}} = \frac{x_i - \hat{b}_i}{\hat{b}_i}$$

em que $\widehat{Var}(X_i|Y_i) = \left(\sum_{q=1}^Q \hat{b}_{i_q}\right)^2$ e $\hat{E}(X_i|Y_i) = \sum_{q=1}^Q \hat{b}_{i_q}$, com $\hat{b}_{i_q} = \frac{\gamma^{y_i}\hat{\mu}_{ix_q}}{1+\hat{\mu}_{iy_q}(\gamma-1)}$.

É esperado que a média dos resíduos esteja centrada em zero e que os valores estejam distribuídos de forma homogênea. A observação pode ser considerada um *outlier* caso a média não esteja centrada em zero ou a amostra de resíduos apresentar uma alta variabilidade.

2.3.3 Interpretação dos Resíduos

Para a análise da adequação do modelo ajustado, utilizamos os gráficos dos resíduos *versus* valores esperados, observando se os pontos são dispersos de maneira aleatória e em uma faixa horizontal centrada em zero. São chamados de *outliers* os pontos que estão distante dessa faixa horizontal centrada no zero, com alguns desvios que devem ser considerados de diferentes tamanhos para cada caso.

Ao observar os gráficos dos resíduos, podemos encontrar algumas características importantes que podem influenciar no modelo em estudo. Se o gráfico tem uma aparência afunilada, isso indica que a variância não é constante, se apresentar uma faixa crescente de pontos pode significar que um termo linear deveria ser incluído no modelo, da mesma forma, se apresentar uma faixa em forma de parábola, nesse caso termos lineares e quadráticos deveriam ser incluídos no modelo. Observações distantes da tendência geral podem afetar o ajuste do modelo.

O gráfico Boxplot também é muito útil para observar os pontos discrepantes, que devem ser analisados conforme a variação dos seus intervalos que representa a variação dos dados e pela presença de pontos extremos, os *outliers*.

2.3.4 Critério de Geweke

A ideia do Critério de Geweke é baseada em um teste de igualdade das médias da primeira e última parte da cadeia de Markov, ou seja, se retirarmos duas amostras da distribuição estacionária da cadeia, as duas médias tendem a ser iguais. Ao fazer o teste

de hipótese, a estatística do teste possuí uma distribuição assintótica normal (BORGES, 2008). A estatística do teste é dada por:

$$\frac{\overline{\theta}_1 - \overline{\theta}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1)$$

em que $\overline{\theta_1}$ é a média das n_1 primeiras observações da cadeia, $\overline{\theta_2}$ é a média das n_2 últimas observações da cadeia, S_1^2, S_2^2 são os estimadores das variâncias de $\hat{\theta}$ no início e no final da cadeia, respectivamente e $\frac{n_1}{N}$ e $\frac{n_2}{N}$ são fixos.

2.4 Critérios de Comparação de Modelos

Uma das etapas mais importantes no processo de modelagem é o estudo de metodologias para a seleção e comparação de modelos, com o objetivo de escolher o que melhor harmoniza com os dados.

São expostas, na literatura, diversas metodologias para esses estudos de adequabilidade. Louzada, Suzuki e Cancho (2013) explanam três critérios Bayesianos de seleção de modelos: o DIC (*Deviance Information Criterion*) que foi proposto por Spiegelhalter et al. (2002), o EAIC (*Expected Akaike Information Criterion*) proposto por Brooks et al. (2002) e o EBIC (*Expected Bayesian (ou Schwarz) Information Criterion*) proposto por Carlin e Louis (2000).

Esses três critérios são baseados na média *a posteriori* da *deviance*, $E[D(\theta)]$, que é uma medida de ajuste e que pode ser aproximada por:

$$\bar{D} = \frac{1}{M}\sum_{m=1}^{M} D(\theta_m), \tag{2.8}$$

em que $D(\theta) = -2\sum_{i=1}^{n} ln(f(t_{1i}, t_{2i}|\theta))$, em que $f(.)$ é a função densidade de probabilidade correspondente ao modelo e m indica a m-ésima realização de um total de M realizações.

Assim, os critérios EAIC, EBIC e DIC podem ser calculados, respectivamente, por $\widehat{EAIC} = \bar{D} + 2q$, $\widehat{EBIC} = \bar{D} + qln(n)$ e $\widehat{DIC} = 2\bar{D} - \hat{D}$, em que q é o número de parâmetros no modelo e $\hat{D} = D\left(\frac{1}{M}\sum_{q=1}^{M}\theta_q\right)$, que é um estimador para $D\{E(\theta)\}$.

O modelo preferível é aquele que, na comparação dos modelos alternativos, apresenta o menor desses critérios.

2.5 Estudos de Simulação

Nesta seção, consideramos um amplo estudo com dados simulados, levando em conta o modelo (2.6), apresentado na Seção 2.2, para ilustrar a metodologia desenvolvida.

O estudo envolve simulação de dados para diferentes tamanhos amostrais, diferentes valores de parâmetros e diferentes escolhas de covariáveis, a fim de constatar o comportamento das estimativas Bayesianas e dos resíduos na análise, considerando os métodos baseados na distribuição *a posteriori* e via CPO, como apresentadas nas seções anteriores.

Foram considerados três cenários, com objetivos diferentes. Nos dois primeiros, apenas uma amostra foi gerada e, então, foram apresentadas medidas associadas às estimação Bayesiana, a performance dos resíduos propostos, além de analisar as propriedades frequentistas dos estimadores Bayesianos. No terceiro cenário, amostras foram perturbadas para verificar a eficiência dos resíduos na detecção de observações influentes.

Para todos os cenários considerados neste estudo de simulação, geramos amostras de tamanhos $50, 100, 500$ e 1000. Ainda, consideramos três covariáveis, z_1, z_2 e z_3, para predizer a variável resposta Y_i e três covariáveis, t_1, t_2 e t_3, para predizer a variável resposta X_i. No caso das distribuições *à priori*, consideramos as distribuições Normal Multivariada para os parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\Delta}$ e Normal para o parâmetro γ.

O primeiro cenário considerou os seguintes valores para os coeficientes de regressão: $\beta_0 = 0,8$, $\beta_1 = 0,1$, $\beta_2 = -0,4$, $\beta_3 = 0,8$, $\delta_0 = 0,1$, $\delta_1 = -0,9$, $\delta_2 = 0,26$ e $\delta_3 = 0,4$. Para o parâmetro que representa uma medida de associação entre as variáveis respostas adotamos três valores $\gamma = 1.6,\ 3,$ e 5.

Para o segundo cenário, adotamos novos valores para os coeficientes de regressão, que segue: $\beta_0 = 0,4$, $\beta_1 = 0,5$, $\beta_2 = -0,6$, $\beta_3 = 0,6$, $\delta_0 = 0,2$, $\delta_1 = -0,5$, $\delta_2 = 0,87$ e $\delta_3 = 0,6$. Para o parâmetro, γ permanecemos com os três valores adotados no cenário 1, que são: $\gamma = 1.6,\ 3,$ e 5.

Em relação às covariáveis, para todos os cenários e configurações, geramos inicialmente apenas covariáveis contínuas a partir da distribuição Uniforme no intervalo [0,1]. Posteriormente, repetimos toda a análise gerando apenas covariáveis discretas a partir da distribuição Uniforme discreta, assumindo os valores $1,\ 2,\ 3,\ 4$ e 5.

Para os dois primeiros cenários, expomos uma tabela resumo *a posteriori* contendo valores de média, mediana, variância, intervalos de credibilidade, intervalos HPD, vício e erro quadrático médio, compreendendo os valores dos parâmetros já estabelecidos, além dos gráficos dos resíduos considerados. Ademais, com o objetivo de verificar as propriedades frequentistas dos estimadores Bayesiano, dispomos as tabelas com as médias dos vícios e dos erros quadráticos médios das médias a posteriori, além das probabilidade de cobertura estimadas para os intervalos de credibilidade inter-quartil e HPD, baseadas em 1000 simulações com $n = 50,\ 100,\ 500$ e 1000.

No cenário 3, o valor da variável resposta discreta Y foi perturbado, na observação 37, e o valor da variável contínua X, na observação 26, em um conjunto de dados de

tamanho $n = 100$, considerando a configuração do cenário 2 com covariáveis contínuas. Assim, apresentamos neste cenário os gráficos dos resíduos com o objetivo de identificar possíveis *outliers* na amostra.

Para a estimação dos parâmetros bayesianos, utilizamos o pacote MCMCpack a função MCMCmetro1R, para o tamanho de cadeia 10000, com saltos de 1000 e intervalos de 1. O desenvolvimento do processo de simulação foi realizado no *software* R, versão 4.0.2 (R Core Team, 2020).

2.6 Resultados

Nesta seção, abordamos os resultados do estudo de simulação para os cenários descritos na seção anterior.

2.6.1 Resultado para o Cenário 1

Na tabela 1, organizamos os resumos das densidades *a posteriori* dos parâmetros do modelo para cada tamanho amostral, quando $\gamma = 1,6$, e considerando as covariáveis contínuas. São apresentadas os valores da média, mediana, variância, intervalo de credibilidade inter-quartil e HPD, vício e erro quadrático médio *a posteriori*, calculados a partir das cadeias MCMC.

Tabela 1 – Medidas resumo para os parâmetros, em que $n = (50, 100, 500, 1000)$, $\gamma = 1.6$, $\boldsymbol{\beta} = (0,8;\ 0,1;\ -0,4;\ 0,8)$ e $\boldsymbol{\Delta} = (0,1;\ -0,9;\ 0,26;\ 0,4)$.

n = 50	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,8	0,8793	0.8836	0,0753	0,3152	1,4084	0,2927	1,3718	0,0793	0,0816
β_1	0,1	0,4078	0,4059	0,0745	-0.1217	0.9477	-0.0679	0.9707	0,3078	0,1692
β_2	-0,4	-0,4258	-0,4262	0,0809	-0.9951	0.1155	-0.9493	0.1205	-0,0258	0,0816
β_3	0,8	0,4683	0,4514	0,0953	-0.1214	1.1034	-0.0656	1.1516	-0,3317	0,2053
δ_0	0,1	0,0418	0,0508	0,2511	-0.9371	1.0202	-0.8521	1.0611	-0,0582	0,2544
δ_1	-0,9	-0,9153	-0,9456	0,2515	-1.8857	0.0504	-1.9527	-0.0233	-0,0153	0,2517
δ_2	0,26	-0,4735	-0,4890	0,2126	-1.3723	0.4619	-1.3048	0.4792	-0,7335	0,7506
δ_3	0,4	1,5935	1,6004	0,3000	0.5318	2.7194	0.4947	2.6045	1,1935	1,7245
γ	1,6	1,4682	1,4672	0,0151	1,2398	1,7223	1,2513	1,7263	-0,1318	0,0324

n = 100	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,8	0,6924	0.6938	0,0284	0,3547	1,0194	0,3762	1,0236	-0,1076	0,0400
β_1	0,1	0,3508	0,3479	0,0368	-0,0030	0,7213	0,0033	0,7213	0,2508	0,0997
β_2	-0,4	-0,1417	-0,1372	0,0416	-0,5419	0,2471	-0,5502	0,2308	0,2583	0,1083
β_3	0,8	0,6470	0,6557	0,0402	0,2434	1,0437	0,2346	1,0257	-0.1530	0,0636
δ_0	0,1	0,4270	0,4040	0,1286	-0,2635	1,1650	-0,3105	1,0975	0,3270	0,2535
δ_1	-0,9	-0,6898	-0,6660	0,1064	-1,3403	-0,0652	-1,3679	-0,1181	0,2102	0,1505
δ_2	0,26	0,2237	-0,2289	0,1015	-0,8812	0,4137	-0,9162	0,3484	-0,4837	0,3355
δ_3	0,4	0,2627	0,2638	0,1294	-0,4669	0,9993	-0,4418	1,0087	-0,1373	0,1483
γ	1,6	1,4919	1,4869	0,0112	1,2866	1,7113	1,2793	1,6892	-0,1081	0,0229

n = 500	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,8	0,8317	0.8331	0,0065	0,6781	0,9915	0,6839	0,9938	0,0317	0,0075
β_1	0,1	0,1399	0,1408	0,0080	-0,0348	0,3160	-0,0394	0,3027	0,0399	0,0096
β_2	-0,4	-0,4699	-0,4700	0,0066	-0,6368	-0,3166	-0,6313	-0,3131	-0,0699	0,0115
β_3	0,8	0,8355	0,8355	0,0078	0,6595	1,0070	0,6629	1,0070	0.0355	0,0091
δ_0	0,1	0,0595	0,0573	0,0213	-0,2143	0,3540	-0,2042	0,3612	-0,0405	0,0229
δ_1	-0,9	-0,6640	-0,6644	0,0212	-0,9422	-0,3991	-0,9391	-0,3991	0,2360	0,0769
δ_2	0,26	0,2077	0,2094	0,0223	-0,0707	0,4871	-0,0744	0,4806	-0,0523	0,0250
δ_3	0,4	0,4339	0,4379	0,0232	0,1321	0,7474	0,1012	0,6996	0,0339	0,0244
γ	1,6	1,5149	1,5145	0,0016	1,4399	1,5979	1,4372	1,5937	-0,0851	0,0089

n = 1000	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,8	0,7728	0.7736	0,0032	0,6617	0,8800	0,6610	0,8778	-0,0109	0,0033
β_1	0,1	0,1198	0,1187	0,0038	-0,0019	0,2331	-0,0007	0,2337	0,0249	0,0044
β_2	-0,4	-0,3765	-0,3771	0,0037	-0,5003	-0,2611	-0,4905	-0,2559	0,0255	0,0040
β_3	0,8	0,8059	0,8052	0,0038	0,6894	0,9314	0,6917	0,9329	-0,0228	0,0044
δ_0	0,1	0,0956	0,0913	0,0119	-0,1103	0,3154	-0,0991	0,3159	0,0927	0,0194
δ_1	-0,9	-0,9875	-0,9882	0,0121	-1,2039	-0,7763	-1,1981	-0,7727	-0,1024	0,0228
δ_2	0,26	0,2262	0,2277	0,0126	0,0022	0,4522	-0,0034	0,4418	0,0290	0,0122
δ_3	0,4	0,4544	0,4611	0,0127	0,2356	0,6682	0,2260	0,6499	-0,1291	0,0270
γ	1,6	1,5751	1,5744	0,0009	1,5180	1,6346	1,5209	1,6364	0,0405	0,0027

Observamos que, inicialmente, alguns dos valores estimados não estão tão próximos aos valores reais, mas a medida que aumenta o valor de n, os valores estimados vão ficando mais próximos dos valores reais e os intervalos de credibilidade e HPD vão ficando menores. O valor real das estimativas estão contidas no intervalo de credibilidade e que existe uma pequena variação delas.

A convergência das cadeias foi verificada por meio do critério de Geweke, que consiste em testar a hipótese nula de que a diferença padronizada entre a média das primeiras $n_A = 0,1n$ interações e a média das $n_B = 0,5n$ últimas segue uma distribuição normal padrão. Ao observarmos a Tabela 2, verificamos que há convergência das cadeias.

Tabela 2 – Teste de Convergência de Geweke, para $\gamma = 1,6$.

	β_0	β_1	β_2	β_3	δ_0	δ_1	δ_2	δ_3	γ
n = 50	0,5746	-0,2236	0,6152	-0,2402	2,2076	-1,2074	0,4012	-2,8376	-1,0595
n = 100	0,5282	1,7946	-1,9909	-0,1115	0,3771	-0,1039	-0,1558	-1,2148	0,0456
n = 500	0,9846	-0,1789	-1,3121	-1,1997	-0,0020	-0,7303	-0,8318	0,6915	-0,0057
n = 1000	0,8724	-0,5278	-0,4548	-0,2732	1,9425	-1,3913	-1,3806	-1,6917	1,0144

Na Tabela 3, apresentamos os resumos das densidades *a posteriori* dos parâmetros do modelo, para cada tamanho amostral, quando $\gamma = 3$ e covariáveis contínuas.

Tabela 3 – Medidas resumo para os parâmetros, em que $n = (50, 100, 500, 1000)$, $\gamma = 3$, $\boldsymbol{\beta} = (0,8;\ 0,1;\ -0,4;\ 0,8)$ e $\boldsymbol{\Delta} = (0,1;\ -0,9;\ 0,26;\ 0,4)$.

n = 50	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,8	0,8789	0.8797	0,0599	0,4157	1,3658	0,4616	1,3923	0,0789	0,0662
β_1	0,1	0,3986	0,3950	0,0606	-0,1080	0,8978	-0,1284	0,8586	0,2986	0,1498
β_2	-0,4	-0,3882	-0,3864	0,0741	-0,9265	0,1354	-0,9339	0,1157	0,0118	0,0742
β_3	0,8	0,4464	0,4602	0,0890	-0,1469	1,0355	-0,0702	1,0922	-0.3536	0,2140
δ_0	0,1	0,1093	0,1011	0,2878	-0,9459	1,1204	-0,9533	1,0711	0,0093	0,2879
δ_1	-0,9	-0,8524	-0,8645	0,2452	-1,8530	0,1390	-1,7552	0,1805	0,0476	0,2474
δ_2	0,26	0,5565	-0,5424	0,2149	-1,4732	0,3513	-1,4287	0,3659	-0,8165	0,8815
δ_3	0,4	1,5819	1,5850	0,3204	0,4890	2,6921	0,4675	2,6586	1,1819	1,7171
γ	3,0	2,7904	2,7793	0,0534	2,3798	3,2892	2,3440	3,1945	-0,2096	0,0973

n = 100	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,8	0,6759	0.6779	0,0315	0,3160	0,9972	0,3525	1,0208	-0,1241	0,0469
β_1	0,1	0,3402	0,3435	0,0368	-0,0430	0,6995	-0,0430	0,6993	0,2402	0,0945
β_2	-0,4	-0,1670	-0,1642	0,0431	-0,5678	0,2477	-0,5572	0,2510	0,2330	0,0974
β_3	0,8	0,6976	0,6960	0,0388	0,3182	1,0879	0,3356	1,1040	-0.1024	0,0492
δ_0	0,1	0,4646	0,4613	0,1447	-0,2661	1,2240	-0,2952	1,1580	0,3646	0,2776
δ_1	-0,9	-0,6417	-0,6402	0,1035	-1,2677	-0,0134	-1,2475	-0,0019	0,2583	0,1702
δ_2	0,26	-0,1981	-0,2039	0,1130	-0,8700	0,4358	-0,8402	0,4513	-0,4581	0,3229
δ_3	0,4	0,2279	0,2238	0,1256	-0,4646	0,8927	-0,4492	0,9000	-0,1721	0,1552
γ	3,0	2,7877	2,7722	0,0390	2,4423	3,2011	2,4391	3,1843	-0,2123	0,0841

n = 500	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,8	0,8310	0.8265	0,0060	0,6821	0,9936	0,6875	0,9960	0,0310	0,0070
β_1	0,1	0,1265	0,1273	0,0071	-0,0341	0,2919	-0,0449	0,2772	0,0265	0,0078
β_2	-0,4	-0,4720	-0,4742	0,0067	-0,6302	-0,3164	-0,6260	-0,3147	-0,0720	0,0119
β_3	0,8	0,8478	0,8506	0,0075	0,6694	1,0174	0,6767	1,0180	0.0478	0,0098
δ_0	0,1	0,0881	0,0881	0,0198	-0,1802	0,3524	-0,1600	0,3680	-0,0119	0,0200
δ_1	-0,9	-0,6699	-0,6635	0,0236	-0,9695	-0,3651	-0,9485	-0,3557	0,2301	0,0766
δ_2	0,26	0,2146	0,2207	0,0207	-0,0733	0,4784	-0,0606	0,4839	-0,0454	0,0228
δ_3	0,4	0,4258	0,4235	0,0218	0,1360	0,7150	0,1241	0,6909	0,0258	0,0224
γ	3,0	2,8527	2,8559	0,0051	2,7107	2,9923	2,7194	2,9934	-0,1473	0,0268

n = 1000	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,8	0,7729	0.7720	0,0030	0,6684	0,8864	0,6651	0,8787	-0,0271	0,0038
β_1	0,1	0,1186	0,1154	0,0034	0,0029	0,2320	-0,0015	0,2244	0,0186	0,0037
β_2	-0,4	-0,3784	-0,3777	0,0030	-0,4917	-0,2786	-0,4849	-0,2732	0,0216	0,0035
β_3	0,8	0,8059	0,8063	0,0038	0,6897	0,9304	0,6892	0,9281	0.0059	0,0038
δ_0	0,1	0,0985	0,0943	0,0115	-0,1214	0,3074	-0,1134	0,3093	-0,0015	0,0115
δ_1	-0,9	-0,9961	-0,9959	0,0129	-1,2061	-0,7579	-1,2297	-0,7858	-0,0961	0,0222
δ_2	0,26	0,2324	0,2366	0,0118	0,0158	0,4435	0,0135	0,4358	-0,0276	0,0125
δ_3	0,4	0,4595	0,4612	0,0128	0,2359	0,6700	0,2332	0,6586	0,0595	0,0163
γ	3,0	2,9528	2,9515	0,0031	2,8477	3,0669	2,8442	3,0618	-0,0472	0,0053

Notamos um comportamento semelhante ao observado anteriormente, ou seja, valores estimados não estão próximos dos valores reais, todavia conforme aumenta o tamanho da amostra, mais próximos os valores reais e os intervalos de credibilidade e HPD vão ficando menores. Existe uma pequena variação nas estimativas e o valor real está contido nos intervalos de credibilidade. Observamos, também, que há convergência nas cadeias, através do teste de Geweke, conforme apresentamos na tabela 4.

Tabela 4 – Teste de Convergência de Geweke, para $\gamma = 3$.

	β_0	β_1	β_2	β_3	δ_0	δ_1	δ_2	δ_3	γ
n = 50	1,6608	-0,0089	0,2454	-2,4317	-0,1196	0,1766	-0,5309	-0,8724	-0,3360
n = 100	-0,0390	1,9583	-1,1413	-0,1621	-0,0006	0,3463	-0,3214	-0,7140	0,3339
n = 500	2,5887	-0,7707	-1,3782	-1,4589	0,5309	-0,6888	0,3049	-1,2263	0.4695
n = 1000	1,5436	0,6615	-2,3047	-0,9820	0,6928	-0,8740	0,1121	-0,8352	0,7673

Na tabela 5, apresentamos os resumos *a posteriori* para cada tamanho amostral, quando $\gamma = 5$.

Tabela 5 – Medidas resumo para os parâmetros, em que $n = (50, 100, 500, 1000)$, $\gamma = 5$, $\boldsymbol{\beta} = (0,8;\ 0,1;\ -0,4;\ 0,8)$ e $\boldsymbol{\Delta} = (0,1;\ -0,9;\ 0,26;\ 0,4)$.

n = 50	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,8	0,9143	0.9298	0,0599	0,4506	1,3838	0,4719	1,3904	0,1143	0,0730
β_1	0,1	0,3829	0,3922	0,0666	-0,1793	0,8659	-0,1294	0,8912	0,2829	0,1465
β_2	-0,4	-0,4297	-0,4341	0,0770	-0,9988	0,1010	-0,9753	0,1016	-0,0297	0,0779
β_3	0,8	0,4290	0,4219	0,0796	-0,0888	1,0067	-0,0798	1,0070	-0.3710	0,2172
δ_0	0,1	0,1567	0,1712	0,2590	-0,8094	1,1338	-0,7983	1,1349	0,0567	0,2621
δ_1	-0,9	-0,8473	-0,8531	0,2497	-1,8177	0,1217	-1,8364	0,1002	0,0527	0,2525
δ_2	0,26	-0,5760	-0,5949	0,2259	-1,4780	0,4196	-1,4705	0,4196	-0,8360	0,9247
δ_3	0,4	1,5706	1,5432	0,3070	0,6068	2,7573	0,5638	2,6482	1,1706	1,6773
γ	5,0	4,6213	4,5864	0,1392	3,9818	5,4389	3,9385	5,3839	-0,3787	0,2826

n = 100	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,8	0,6634	0.6581	0,0265	0,3540	0,9748	0,3645	0,9800	-0,1366	0,0452
β_1	0,1	0,3462	0,3536	0,0314	-0,0204	0,6897	-0,0204	0,6850	0,2462	0,0920
β_2	-0,4	-0,1730	-0,1622	0,0380	-0,5775	0,1967	-0,5775	0,1966	0,2270	0,0896
β_3	0,8	0,7185	0,7143	0,0346	0,3655	1,1088	0,3719	1,1113	-0.0815	0,0412
δ_0	0,1	0,4829	0,4664	0,1548	-0,2703	1,2834	-0,2979	1,2468	0,3829	0,3014
δ_1	-0,9	-0,6559	-0,6509	0,1101	-1,2543	-0,0207	-1,2748	-0,0108	0,2441	0,1697
δ_2	0,26	-0,2184	-0,2184	0,1129	-0,8872	0,4675	-0,7890	0,5270	-0,4784	0,3418
δ_3	0,4	0,2380	0,2326	0,1237	-0,4910	0,9482	-0,4376	0,9708	-0,1620	0,1499
γ	5,0	4,6604	4,6372	0,1084	4,0590	5,3459	4,0324	5,3022	-0,3396	0,2237

n = 500	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,8	0,7784	0.7813	0,0073	0,6099	0,9524	0,6050	0,9353	-0,0216	0,0077
β_1	0,1	0,0552	0,0578	0,0080	-0,1389	0,2283	-0,1478	0,2135	-0,0448	0,0100
β_2	-0,4	-0,5000	-0,5007	0,0069	-0,6684	-0,3446	-0,6747	-0,3534	-0,1000	0,0169
β_3	0,8	0,8803	0,8807	0,0093	0,6919	1,0656	0,6818	1,0477	0.0803	0,0157
δ_0	0,1	0,0750	0,0728	0,0250	-0,2372	0,3929	-0,2418	0,3716	-0,0250	0,0256
δ_1	-0,9	-1,1106	-1,1125	0,0272	-1,4320	-0,7899	-1,4324	-0,7969	-0,2106	0,0716
δ_2	0,26	0,3551	0,3600	0,0250	0,0466	0,6523	0,0480	0,6523	0,0951	0,0340
δ_3	0,4	0,4844	0,4856	0,0247	0,1809	0,7910	0,1675	0,7706	0,0844	0,0318
γ	5,0	4,8531	4,8518	0,0175	4,5946	5,1342	4,5726	5,0945	-0,1469	0,0390

n = 1000	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,8	0,7758	0.7790	0,0033	0,6502	0,8844	0,6442	0,8765	-0,0242	0,0039
β_1	0,1	0,1168	0,1134	0,0037	-0,0014	0,2405	0,0140	0,2498	0,0168	0,0039
β_2	-0,4	-0,3844	-0,3839	0,0034	-0,4996	-0,2680	-0,4935	-0,2623	0,0156	0,0036
β_3	0,8	0,8067	0,8073	0,0037	0,6877	0,9272	0,6850	0,9227	0.0067	0,0038
δ_0	0,1	0,0940	0,0969	0,0124	-0,1366	0,3066	-0,1361	0,3067	-0,0060	0,0125
δ_1	-0,9	-0,9906	-0,9940	0,0144	-1,2272	-0,7578	-1,2311	-0,7660	-0,0906	0,0226
δ_2	0,26	0,2313	0,2378	0,0117	0,0182	0,4429	0,0124	0,4283	-0,0287	0,0125
δ_3	0,4	0,4597	0,4623	0,0132	0,2408	0,6867	0,2373	0,6783	0,0597	0,0168
γ	5,0	4,9251	4,9194	0,0080	4,7441	5,1021	4,7623	5,1136	-0,0749	0,0137

Ao analisarmos as Tabelas 1, 3 e 5 percebemos que existe um comportamento semelhante entre elas. Inicialmente, todas elas apresentam valores estimados diferentes dos valores reais, porém a medida que aumentamos os tamanhos das amostras, os valores vão se aproximando do valor real. O mesmo acontece com os intervalos de credibilidade e HPD, que quanto mais próximo dos valores reais as estimativas vão se aproximando e os intervalos vão diminuindo. Houve convergência nas cadeias, como podemos observar na Tabela 6.

Tabela 6 – Teste de Convergência de Geweke, para $\gamma = 5$.

	β_0	β_1	β_2	β_3	δ_0	δ_1	δ_2	δ_3	γ
n = 50	0,8334	-0,8389	0,9817	-1,4299	0,3313	-0,8387	-0,4424	0,1158	-0.6799
n = 100	0,1875	0,9731	-0,8837	-0,4785	0,9622	-0,3210	-0,5830	-1,4045	0,5842
n = 500	0,6676	0,5351	-1,6799	-0,5508	1,1575	-0,0784	-1,2061	-0,3227	-1,3008
n = 1000	0,4838	0,4125	-1,799	-0,3919	-0,3371	0,0568	-0,3889	-0,2606	0,8403

Os resultados, considerando as covariáveis discretas, são semelhantes aos encontrados na tabela anterior, que considera covariáveis contínuas e, portanto, não serão apresentados.

Em relação à análise de resíduos, ao observar o gráfico dos resíduos via CPO para $n = 50$ e $\gamma = 1,6$, na Figura 1a para a variável resposta discreta Y, que segue uma distribuição Poisson, notamos que os pontos não revelam um padrão isto é, está oscilando em uma faixa horizontal em torno do zero, o que era esperado no gráfico. O mesmo acontece no gráfico da Figura 2a, que corresponde a uma amostra de tamanho $n = 100$, ou seja é um comportamento que observamos nas demais configurações $(n = 500$ e $n = 1000)$.

Ainda observando a Figura 1b, no gráfico dos resíduos via CPO *versus* o valor predito, verificamos que ocorre comportamento idêntico no gráfico da Figura 2b, ambos são valores inteiros, pois o valor predito de uma Poisson é um inteiro, e esta oscilando em torno de zero.

Em relação aos resíduos baseados na distribuição *a posteriori* dos parâmetros, podemos identificar o comportamento esperado, visto que os pontos estão dispersos e em torno de zero, com alguns poucos valores acima do intervalo $(-2, 2)$.

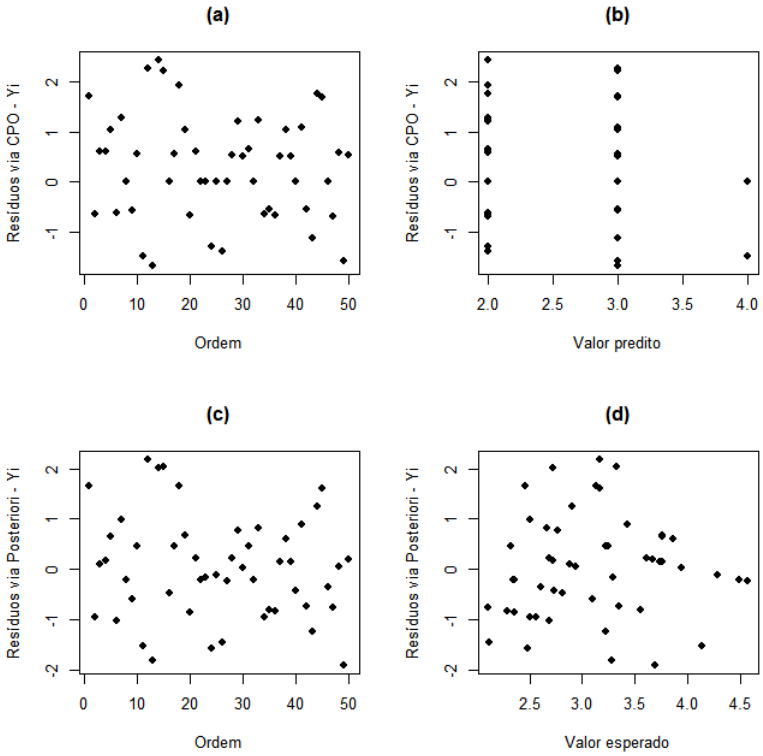

Figura 1 – Gráfico dos Resíduos para Y, em que $n = 50$, $\gamma = 1,6$, $\boldsymbol{\beta} = (0,8;\ 0,1;\ -0,4;\ 0,8)$ e $\boldsymbol{\Delta} = (0,1;\ -0,9;\ 0,26;\ 0,4)$.

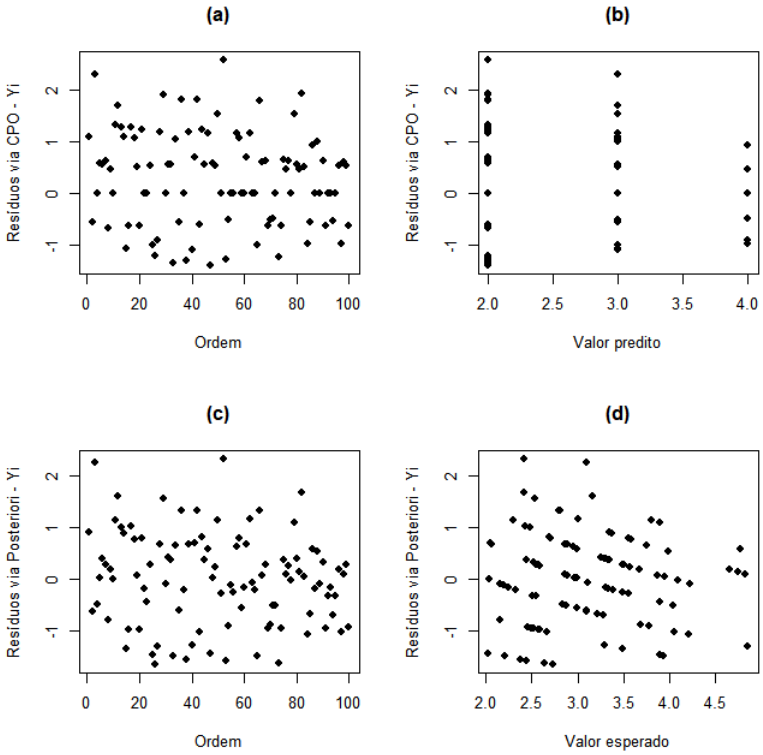

Figura 2 – Gráfico dos Resíduos para Y, em que $n = 100$, $\gamma = 1,6$, $\boldsymbol{\beta} = (0,8;\ 0,1;\ -0,4;\ 0,8)$ e $\boldsymbol{\Delta} = (0,1;\ -0,9;\ 0,26;\ 0,4)$.

Averiguamos que quando $\gamma = 3$ e $\gamma = 5$ os resíduos tem um comportamento muito

semelhante quando $\gamma = 1,6$, como podemos observar na Figura 3 e na Figura 4. Ou seja, os pontos estão oscilando entre $(-2,2)$ e não apresentam um padrão mas estão em torno do zero, o que era esperado no gráfico. Nos gráficos dos resíduos $\times$ valor predito, observamos que os pontos estão em torno de zero, distribuídos de forma aleatória.

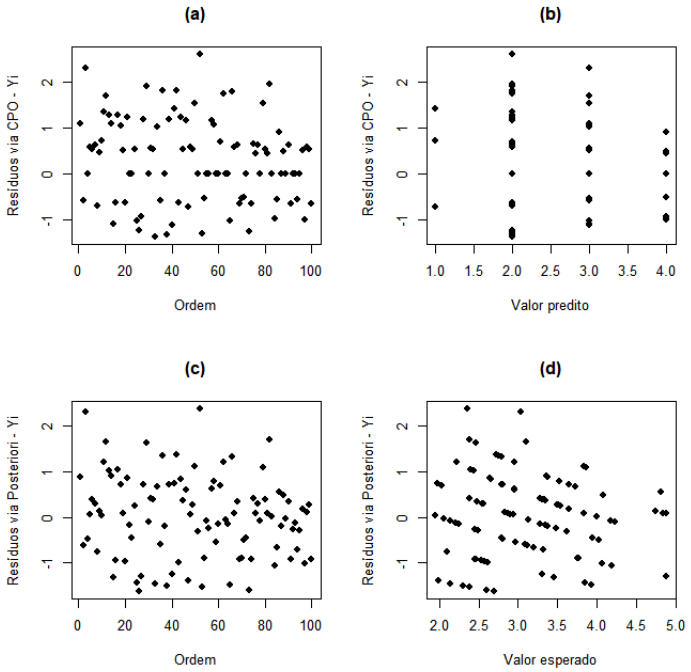

Figura 3 – Gráfico dos Resíduos para Y, em que $n = 100$, $\gamma = 3$, $\boldsymbol{\beta} = (0,8;\ 0,1;\ -0,4;\ 0,8)$ e $\boldsymbol{\Delta} = (0,1;\ -0,9;\ 0,26;\ 0,4)$.

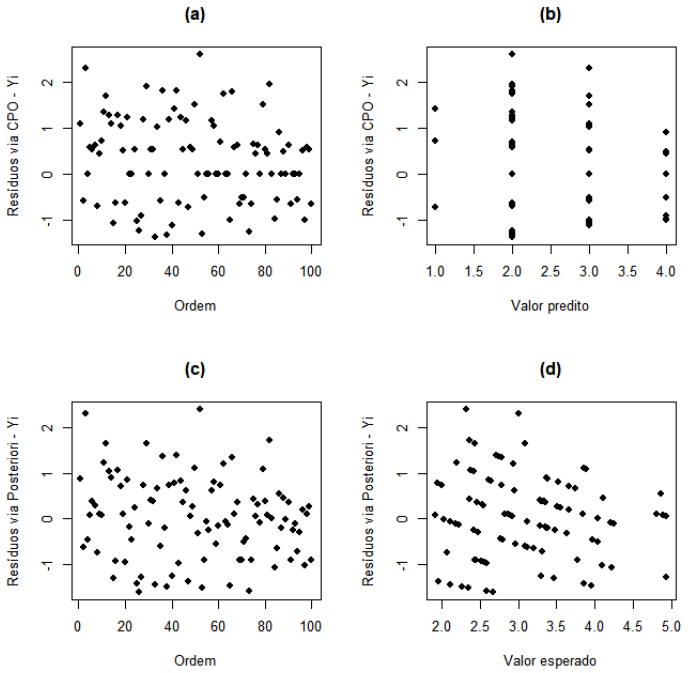

Figura 4 – Gráfico dos Resíduos para Y, em que $n = 100$, $\gamma = 5$, $\boldsymbol{\beta} = (0,8;\ 0,1;\ -0,4;\ 0,8)$ e $\boldsymbol{\Delta} = (0,1;\ -0,9;\ 0,26;\ 0,4)$.

A Figura 5 apresenta os boxplot das amostras MCMC da distribuição *a posteriori* para os resíduos baseados na distribuição *a posteriori* dos parâmetros. O estudo foi realizado para os quatro tamanhos amostrais e para os três valores de γ considerados no estudo de simulação. Porém apresentamos aqui apenas os resultados para o tamanho amostral $n = 100$ e covariáveis contínuas, devido aos demais tamanhos e configurações revelaram conclusões similares. Foram observados pequenos intervalos no boxplot que podem indicar uma pequena variação nos dados e os pontos dispersos dos demais podem indicar pontos *outliers*.

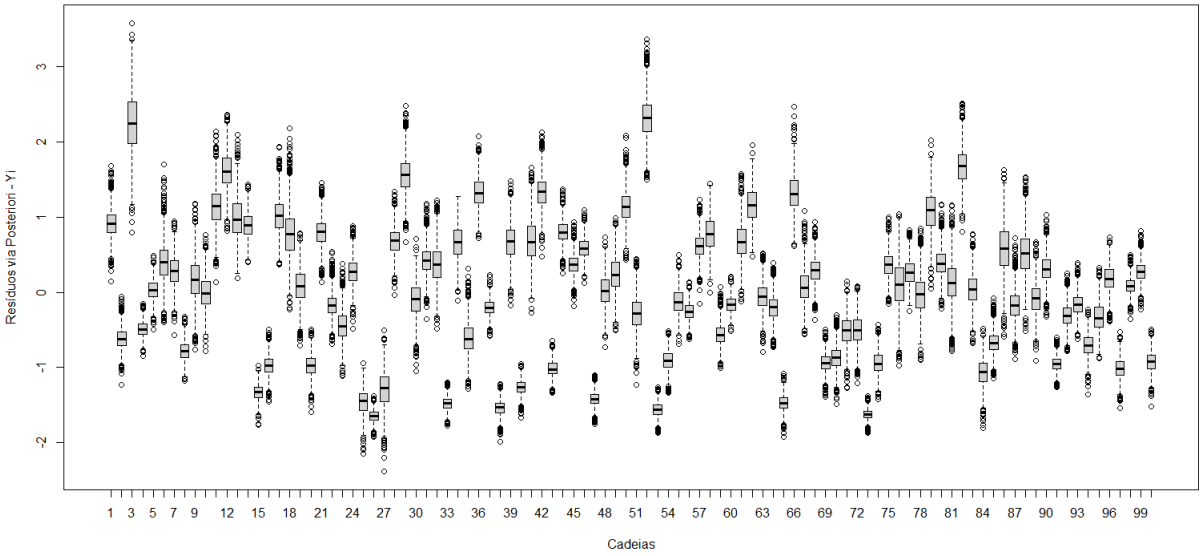

Figura 5 – Boxplot das amostras MCMC da distribuição *a posteriori* dos resíduos para cada observação Y, em que $n = 100$, $\gamma = 1,6$, $\boldsymbol{\beta} = (0,8;\ 0,1;\ -0,4;\ 0,8)$ e $\boldsymbol{\Delta} = (0,1;\ -0,9;\ 0,26;\ 0,4)$.

Na Figura 6, é apresentado o gráfico dos resíduos via CPO e os resíduos baseados na distribuição *a posteriori* dos parâmetros para a variável resposta contínua X, condicionada a resposta discreta Y, que segue uma distribuição Exponencial. Observamos, ainda, que há muita semelhança nos resultados para os quatro tamanhos amostrais $(n = 50, n = 100, n = 500, n = 1000)$ e para os três valores adotados para $\gamma = 1.6,\ 3,$ e 5, considerados no processo. Dessa forma, será apresentado os resultados apenas para o tamanho amostral $n = 100$ e para o parâmetro $\gamma = 1,6$.

Podemos observar que, na Figura 6, todos pontos estão oscilando em torno de zero numa faixa horizontal e não revelam um padrão exato. Além disso, podemos observar que há alguns pontos mais distante do zero. É possível notar uma assimetria na distribuição de ambos os resíduos com valores variando, em geral, entre -2 e 6, com valores concentrados em torno de 0.

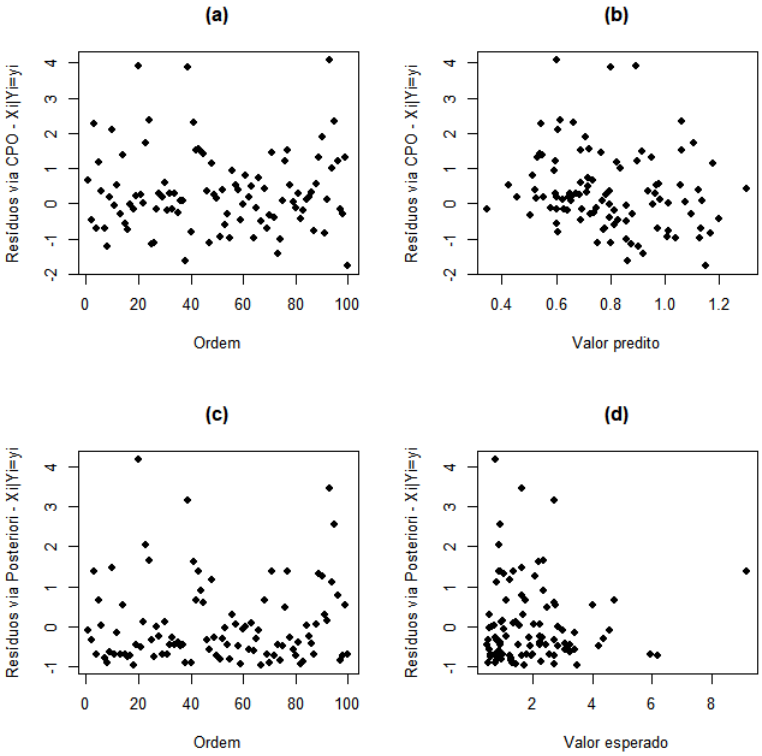

Figura 6 – Gráfico dos Resíduos para X, em que $n = 100$, $\gamma = 1.6$, $\boldsymbol{\beta} = (0,8;\ 0,1;\ -0,4;\ 0,8)$ e $\boldsymbol{\Delta} = (0,1;\ -0,9;\ 0,26;\ 0,4)$.

Ao observar o boxplot dos resíduos, apresentados na Figura 7, verificamos que há observações com pequenos intervalos, o que pode indicar uma pequena variação dos dados e os pontos dispersos que podem indicar pontos *outliers*.

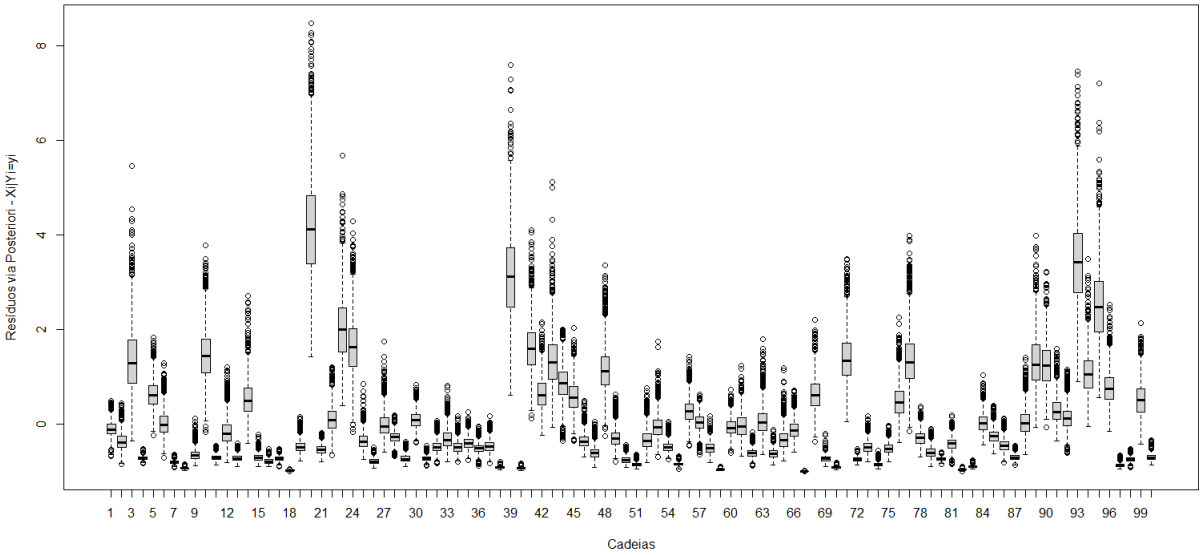

Figura 7 – Boxplot das amostras MCMC da distribuição *a posteriori* dos resíduos para cada observação X, em que $n = 100$, $\gamma = 1,6$, $\boldsymbol{\beta} = (0,8;\ 0,1;\ -0,4;\ 0,8)$ e $\boldsymbol{\Delta} = (0,1;\ -0,9;\ 0,26;\ 0,4)$.

Com o objetivo de verificar as propriedades frequentistas dos estimadores Bayesiano. Na Tabela 7, são apresentados as médias dos vícios e dos erros quadráticos médios da esperança, *a posteriori*, dos parâmetros de interesse, além das probabilidade de cobertura estimadas para os intervalos de credibilidade inter-quartil e HPD, baseadas em 1000 simulações com $n = 50,\ 100,\ 500$ e 1000. Realizamos essa análise para os três valores de γ considerados, tanto para covariáveis contínuas quanto para covariáveis discretas. Como os resultados em todas as configurações do cenário 1 foram similares, dispomos apenas os resultados obtidos para covariáveis contínuas com $\gamma = 1.6$. A probabilidade de cobertura nominal considerada foi de 95%.

Para determinar os valores dos erros quadráticos médios e vícios dos estimadores pontuais para cada amostra de dados gerada, considerando os tamanhos de amostra $(n = 50,\ 100,\ 500,\ 1000)$ são gerados 1000 amostras MCMC das distribuições *a posteriori* para cada parâmetro, de acordo com o cenário 1.

Tabela 7 – Estimativa dos erro quadrático médio (EQM) e dos vícios da média *a posteriori* e probabilidade de cobertura estimadas para os intervalos de credibilidade inter-quartil e HPD, para diferentes tamanhos de amostrais e $\gamma = 1.6$, de acordo com o cenário 1

Parâmetro	Tamanho da amostra	EQM	Vício	Inter-quartil	HPD
β_0	50	0,1425	-0,0353	95,2	94,8
	100	0,0712	-0,0041	94,0	93,8
	500	0,0139	-0,0017	93,4	93,2
	1000	0,0065	-0,0018	94,6	94,2
β_1	50	0,1593	0,0026	95,8	96,0
	100	0,0789	0,0006	96,8	96,8
	500	0,0152	-0,0065	95,0	94,0
	1000	0,0074	-0,0024	94,8	95,4
β_2	50	0,1628	-0,0210	94,2	94,0
	100	0,0768	-0,0213	94,2	94,2
	500	0,0153	-0,0022	93,0	92,6
	1000	0,0072	0,0028	94,2	93,2
β_3	50	0,1680	0,0361	95,8	95,4
	100	0,0800	-0,0008	96,4	96,0
	500	0,0159	0,0040	96,2	94,8
	1000	0,0077	0,0033	94,8	95,4
δ_0	50	0,5203	-0,0350	96,8	95,8
	100	0,2419	-0,0137	96,4	96,2
	500	0,0452	-0,0028	96,2	95,4
	1000	0,0236	-0,0052	94,8	94,4
δ_1	50	0,5956	-0,0095	95,0	94,6
	100	0,2658	-0,0226	96,8	96,4
	500	0,0546	0,0035	96,2	95,8
	1000	0,0248	0,0015	95,6	95,4
δ_2	50	0,5746	-0,0224	95,4	95,4
	100	0,2666	0,0276	93,8	93,6
	500	0,0545	0,0087	94,8	94,6
	1000	0,0233	0,0139	93,4	92,8
δ_3	50	0,6260	0,0325	94,8	94,4
	100	0,2754	0,0038	96,4	95,6
	500	0,0500	0,0054	96,0	95,8
	1000	0,0260	0,0007	93,6	93,4
γ	50	0,0452	0,0388	94,8	94,0
	100	0,0212	0,0116	96,0	96,4
	500	0,0048	-0,0012	95,6	95,0
	1000	0,0037	-0,0030	94,0	93,4

Ao observarmos a Tabela 7, é possível observar que conforme o tamanho amostral cresce, o viés e o EQM se aproximam de zero. Por outro lado, as probabilidades de cobertura para os intervalos de credibilidade inter-quartil e HPD estão próximas da probabilidade de cobertura nominal que é de 95%.

2.6.2 Resultado para o Cenário 2

Na Tabela 8, apresentamos os resumos *a posteriori* para cada tamanho amostral, quando os coeficientes de regressão são $\beta_0 = 0,4$, $\beta_1 = 0,5$, $\beta_2 = -0,6$, $\beta_3 = 0,6$, $\delta_0 = 0,2$, $\delta_1 = -0,5$, $\delta_2 = 0,87$ e $\delta_3 = 0,6$ e $\gamma = 1,6$.

Tabela 8 – Medidas resumo para os parâmetros, em que $n = (50, 100, 500, 1000)$, $\gamma = 1.6$, $\boldsymbol{\beta} = (0,4;\ 0,5;\ -0,6;\ 0,6)$ e $\boldsymbol{\Delta} = (0,2;\ -0,5;\ 0,87;\ 0,6)$.

n = 50	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,4	0,8916	0.8840	0,1091	0,1229	1,6507	0,1834	1,6948	0,4916	0,3931
β_1	0,5	-0,0678	-0,0824	0,1515	-1,0408	0,8796	-0,9979	0,9085	-0,5678	0,5675
β_2	-0,6	-0,6404	-0,6583	0,2451	-1,2142	-0,0364	-1,2350	-0,0721	-0,0404	0,0905
β_3	0,6	0,2562	0,2620	0,0889	-0,4938	0,9561	-0,5067	0,9213	-0,3438	0,2562
δ_0	0,2	-0,1120	-0,1102	0,1381	-1,2263	1,0197	-1,2263	1,0013	-0,3120	0,4191
δ_1	-0,5	-0,3909	-0,3765	0,3006	-1,4593	0,6677	-1,4119	0,6967	0,1091	0,3125
δ_2	0,87	0,6939	0,6840	0,2703	-0,3085	1,6797	-0,3085	1,6767	-0,1761	0,3013
δ_3	0,6	0,9220	0,8939	0,3580	-0,2843	2,1365	-0,2007	2,1465	0,3220	0,4616
γ	1,6	1,3931	1,3855	0,0218	1,1243	1,6963	1,1148	1,6848	-0,2069	0,0646

n = 100	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,4	0,6700	0.6614	0,0406	0,2621	1,6818	0,2361	1,0353	0,2700	0,1135
β_1	0,5	0,3836	0,3804	0,0465	-0,0697	0,8006	-0,0725	0,7883	-0,1164	0,0600
β_2	-0,6	-0,7937	-0,7976	0,0523	-1,2340	-0,3459	-1,2301	-0,3446	-0,1937	0,0898
β_3	0,6	0,4813	0,4813	0,0505	0,0260	0,9401	0,0483	0,9534	-0.1187	0,0646
δ_0	0,2	0,4444	0,4317	0,1155	-0,2230	1,1262	-0,2010	1,1276	0,2444	0,1753
δ_1	-0,5	-0,8551	-0,8849	0,1608	-1,5844	-0,0526	-1,5631	-0,0457	-0,3551	0,2869
δ_2	0,87	0,7995	0,8027	0,1284	0,0846	1,5024	0,1407	1,5390	-0,0705	0,1334
δ_3	0,6	0,7927	0,7886	0,1175	0,1558	1,4661	0,1604	1,4681	0,1927	0,1546
γ	1,6	1,6531	1,6540	0,0148	1,4277	1,9006	1,4425	1,9112	0,0531	0,0176

n = 500	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,4	0,4717	0.4767	0,0096	0,2631	0,6520	0,2810	0,6649	0,0717	0,0147
β_1	0,5	0,2635	0,2617	0,0118	0,0458	0,4682	0,0606	0,4756	-0,2365	0,0678
β_2	-0,6	-0,5358	-0,5308	0,0107	-0,7475	-0,3424	-0,7483	-0,3527	0,0642	0,0149
β_3	0,6	0,7076	0,7073	0,0111	0,5057	0,9213	0,4860	0,8971	0.1076	0,0227
δ_0	0,2	0,1265	0,1255	0,0209	-0,1579	0,4163	-0,1279	0,4334	-0,0735	0,0263
δ_1	-0,5	-0,3220	-0,3225	0,0223	-0,6080	-0,0465	-0,6101	-0,0503	0,1780	0,0540
δ_2	0,87	0,6364	0,6361	0,0227	0,3391	0,9279	0,3371	0,9237	-0,2336	0,0773
δ_3	0,6	0,7805	0,7712	0,0239	0,4713	1,1003	0,4557	1,0725	0,1805	0,0565
γ	1,6	1,6130	1,6137	0,0022	1,5205	1,7054	1,5190	1,7030	0,0130	0,0024

n = 1000	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,4	0,3949	0.3953	0,0048	0,2675	0,5260	0,2684	0,5265	0,0051	0,0048
β_1	0,5	0,4303	0,4302	0,0050	0,2890	0,5616	0,2902	0,5616	-0,0697	0,0098
β_2	-0,6	-0,5386	-0,5424	0,0052	-0,6762	-0,3925	-0,6755	-0,3919	0,0614	0,0089
β_3	0,6	0,6578	0,6563	0,0052	0,5167	0,7994	0,4990	0,7783	0.0578	0,0085
δ_0	0,2	0,3278	0,3242	0,0108	0,1311	0,5379	0,1170	0,5149	0,1278	0,0272
δ_1	-0,5	-0,6293	-0,6279	0,0133	-0,8647	-0,4083	-0,8620	-0,4083	-0,1293	0,0300
δ_2	0,87	0,6908	0,6919	0,0117	0,4860	0,9061	0,4855	0,8982	-0,1792	0,0438
δ_3	0,6	0,7266	0,7240	0,0145	0,4753	0,9547	0,5089	0,9779	0,1266	0,0305
γ	1,6	1,5988	1,5988	0,0011	1,5311	1,6677	1,5307	1,6657	-0,0012	0,0011

Nota-se que, inicialmente, os valores estimados não estão tão próximos aos valores reais, mas, à medida que aumenta o valor de n, os valores estimados vão ficando mais próximos dos valores reais e os intervalos de credibilidade e HPD vão ficando menores. O valor real das estimativas estão contidas no intervalo de credibilidade e existe uma pequena

variação delas. A convergência das cadeias foi verificada por meio do critério de Geweke e, ao analisar a Tabela 9, verificamos que há convergência entre as cadeias.

Tabela 9 – Teste de Convergência de Geweke, para $\gamma = 1,6$.

	β_0	β_1	β_2	β_3	δ_0	δ_1	δ_2	δ_3	γ
n = 50	0,6936	0,4479	-2,4186	-0,8497	1,8351	-1,6612	-0,3593	-1,4465	-0,6308
n = 100	1,1967	-1,0453	-0,5270	0,1395	0,7365	-0,8639	-0,1554	-0,3119	-0.2957
n = 500	-0,5312	1,1752	-1,1555	0,6987	1,1205	-1,5262	0,4707	-1,3544	-0,2807
n = 1000	0,5815	0,6386	-0,8310	-1,2031	0,4431	-0,3756	-1,2206	0,3537	0,0574

Na Tabela 10, apresentamos os resumos *a posteriori* para cada tamanho amostral, quando $\gamma = 3$.

Tabela 10 – Medidas descritivas para os parâmetros, em que $n = (50, 100, 500, 1000)$, $\gamma = 3$, $\boldsymbol{\beta} = (0,4;\ 0,5;\ -0,6;\ 0,6)$ e $\boldsymbol{\Delta} = (0,2;\ -0,5;\ 0,87;\ 0,6)$.

n = 50	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,4	0,3592	0.3622	0,1075	-0,2786	1,0030	-0,3255	0,9326	-0,0438	0,1094
β_1	0,5	0,4616	0,4647	0,1190	-0,2279	1,1778	-0,2416	1,1086	-0,0384	0,1205
β_2	-0,6	-0,7963	-0,7920	0,1079	-1,4394	-0,1593	-1,4467	0,1776	-0,1963	0,1464
β_3	0,6	0,6379	0,6401	0,1269	-0,0773	1,3429	-0,0773	1,3322	0.0379	0,1283
δ_0	0,2	0,5242	0,5344	0,3168	-0,4755	1,6863	-0,4859	1,6481	0,3242	0,4219
δ_1	-0,5	-0,5053	-0,4948	0,3952	-1,7082	0,7550	-1,7350	0,6755	-0,0053	0,3953
δ_2	0,87	1,2789	1,3014	0,1958	0,3808	2,1494	0,3605	2,0721	0,4089	0,3629
δ_3	0,6	-0,3334	-0,3509	0,2808	-1,3956	0,7661	-0,4175	0,7207	-0,9334	1,1522
γ	3,0	3,0486	3,0197	0,0859	2,5271	3,6689	2,5246	3,6513	0,0486	0,0883

n = 100	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,4	0, 067	0.5037	0,0371	0,1201	0,8722	0,1562	0,8876	0,1067	0,0485
β_1	0,5	0,2199	0,2188	0,0517	-0,2272	0,6626	-0,2622	0,6126	-0,2801	0,1302
β_2	-0,6	-0,2921	-0,2894	0,0461	-0,7263	0,1237	-0,7438	0,0855	0,3079	0,1409
β_3	0,6	0,2888	0,2874	0,0492	-0,1715	0,7166	-0,1357	0,7461	-0.3112	0,1461
δ_0	0,2	0,5836	0,5776	0,0980	-0,0294	1,2107	-0,0561	1,1599	0,3836	0,2451
δ_1	-0,5	-0,8238	-0,8223	0,1132	-1,4971	-0,1550	-1,5299	-0,2038	-0,3238	0,2181
δ_2	0,87	0,6755	0,6766	0,1270	-0,0303	1,3725	-0,0300	1,3725	-0,1945	0,1648
δ_3	0,6	0,4859	0,4726	0,1298	-0,1890	1,2088	-0,1923	1,1854	0,1141	0,1428
γ	3,0	2,9122	2,9033	0,0398	2,5396	3,3152	2,5548	3,3211	-0,0878	0,0475

n = 500	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,4	0,3717	0.3735	0,0099	0,1784	0,5696	0,1867	0,5711	-0,0283	0,0107
β_1	0,5	0,5535	0,5589	0,0107	0,3420	0,7512	0,3420	0,7510	0,0535	0,0136
β_2	-0,6	-0,6541	-0,6509	0,0092	-0,8555	-0,4819	-0,8525	-0,4819	-0,0541	0,0121
β_3	0,6	0,6548	0,6585	0,0097	0,4599	0,8502	0,4772	0,8619	0.0548	0,0127
δ_0	0,2	-0,0674	-0,0711	0,0228	-0,3776	0,2438	-0,3823	0,2240	-0,2674	0,0942
δ_1	-0,5	-0,3433	-0,3445	0,0242	-0,6586	-0,0365	-0,6758	-0,0658	0,1567	0,0488
δ_2	0,87	0,7661	0,7683	0,0203	0,4756	1,0491	0,4697	1,0352	-0,1039	0,0311
δ_3	0,6	1,0353	1,0388	0,0249	0,7435	1,3157	0,7457	1,3157	0,4353	0,2145
γ	3,0	3,0739	3,0689	0,0090	2,9019	3,2657	2,8886	3,2403	0,0739	0,0145

n = 1000	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,4	0,4754	0.4782	0,0042	0,1784	0,5696	0,3374	0,5905	0,0754	0,0099
β_1	0,5	0,4622	0,4599	0,0048	0,3420	0,7512	0,3315	0,5982	-0,0378	0,0063
β_2	-0,6	-0,7575	-0,7563	0,0051	-0,8555	-0,4819	-0,8975	0,6171	-0,1575	0,0299
β_3	0,6	0,5894	0,5893	0,0052	0,4599	0,8502	0,4302	0,7224	-0.0106	0,0054
δ_0	0,2	0,1438	0,1455	0,0108	-0,3776	0,2438	-0,0570	0,3445	-0,0562	0,0140
δ_1	-0,5	-0,3769	-0,3794	0,0134	-0,6586	-0,0365	-0,5981	-0,1464	0,1231	0,0286
δ_2	0,87	-0,7918	0,7895	0,0117	0,4756	1,0491	0,5940	0,9938	-0,0782	0,0178
δ_3	0,6	0,5915	0,5868	0,0123	0,7435	1,3157	0,3767	0,8089	-0,0085	0,0123
γ	3,0	2,9290	2,9245	0,0045	2,9019	3,2657	2,7995	3,0633	-0,0710	0,0095

Observamos um comportamento semelhante quando o $\gamma = 1,6$, ou seja, valores estimados não estão próximos dos valores reais, contudo conforme aumenta o tamanho da amostra, mais próximos os valores reais e os intervalos de credibilidade e HPD vão ficando menores. Existe uma pequena variação nas estimativas e o valor real está contido nos intervalos de credibilidade. Verificamos, também, que há convergência nas cadeias, através do teste de Geweke, conforme ilustramos na tabela 11.

Tabela 11 – Teste de Convergência de Geweke, para $\gamma = 3$.

	β_0	β_1	β_2	β_3	δ_0	δ_1	δ_2	δ_3	γ
n = 50	-0,5212	0,9764	-0,3898	-0,0536	0,8484	-1,1528	0,0497	-1,0503	0,5793
n = 100	0,8631	0,8203	-2,1023	-1,0651	0,3974	-1,3557	-1,2024	0,2269	1,0567
n = 500	1,3588	-0,6397	-1,7665	-0,1910	0,4818	-0,3002	-0,3165	-0,2307	-0,9971
n = 1000	1,8206	-1,0871	-1,0578	-1,1966	1,1034	-1,3400	-0,2844	-0,3467	-0,3482

Na Tabela 12, apresentamos os resumos *a posteriori* para cada tamanho amostral, quando $\gamma = 5$.

Tabela 12 – Medidas descritivas para os parâmetros, em que $n = (50, 100, 500, 1000)$, $\gamma = 5$, $\boldsymbol{\beta} = (0,4;\ 0,5;\ -0,6;\ 0,6)$ e $\boldsymbol{\Delta} = (0,2;\ -0,5;\ 0,87;\ 0,6)$.

n = 50	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,4	0,5124	0.5169	0,0481	0,0986	0,9673	0,0986	0,9673	0,1124	0,0607
β_1	0,5	0,9416	1,0223	0,1140	0,3505	1,6238	0,3744	1,6238	0,4416	0,3090
β_2	-0,6	-1,2265	-1,3334	0,1705	-1,8761	-0,4296	-1,8761	-0,4296	-0,6265	0,5630
β_3	0,6	0,5663	0,6325	0,1200	0,0945	1,2200	0,0945	1,2200	-0,0337	0,1211
δ_0	0,2	2,1750	2,2331	0,1390	1,4577	3,1443	1,4577	2,8896	1,9750	4.0397
δ_1	-0,5	-0,6734	-0,6317	0,0959	-1,4224	-0,0108	-1,4224	-0,2059	-0,1734	0,1259
δ_2	0,87	-4,7232	-4,7983	0,1416	-5,7403	-4,0284	-5,2098	-4,0150	-5,5932	31,4250
δ_3	0,6	3,1100	-3,2212	0,0866	2,6171	3,5433	2,6171	3,5411	2,5100	6,3867
γ	5,0	1,1484	1,1283	0,0098	0,9724	1,2808	0,9724	1,2808	-3,8516	14,8448

n = 100	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,4	-0,0522	-0,0602	0,0528	-0,4890	0,4061	-0,5147	0,3646	-0,4522	0,2574
β_1	0,5	0,5366	0,5349	0,0421	0,1221	0,9240	0,1560	0,9410	0,0366	0,0435
β_2	-0,6	-0,4452	-0,4436	0,0494	-0,9136	-0,0088	-0,8546	0,0270	0,1548	0,0733
β_3	0,6	1,1116	1,1060	0,0541	0,6530	1,5850	0,6155	1,5139	0.5116	0,3158
δ_0	0,2	0,5978	0,5919	0,1033	-0,0063	1,2934	-0,0984	1,1732	0,3978	0,2615
δ_1	-0,5	-0,5629	-0,5699	0,1460	-1,3184	-0,2035	-1,2780	0,2176	-0,0629	0,1500
δ_2	0,87	0,3846	0,3729	0,1293	-0,3243	1,0810	-0,2776	1,0953	-0,4854	0,3649
δ_3	0,6	0,1751	0,1890	0,1050	-0,4806	0,8034	-0,5277	0,7341	-0,4249	0,2856
γ	5,0	4,9080	4,8955	0,1405	4,2247	5,6647	4,2391	5,6753	-0,0920	0,1489

n = 500	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,4	0,4714	0.4711	0,0066	0,3193	0,6325	0,3120	0,6240	0,0714	0,0118
β_1	0,5	0,4565	0,4543	0,0090	0,2674	0,6445	0,2824	0,6504	-0,0435	0,0109
β_2	-0,6	-0,5912	-0,5921	0,0075	-0,7550	-0,4255	-0,7550	-0,4255	0,0088	0,0075
β_3	0,6	0,6072	0,6026	0,0080	0,4286	0,7866	0,4391	0,7904	0.0072	0,0080
δ_0	0,2	0,2329	0,2299	0,0213	-0,0420	0,5246	-0,0367	0,5249	0,0329	0,0224
δ_1	-0,5	-0,5553	-0,5527	0,0214	-0,8410	-0,2762	-0,8308	-0,2732	-0,0553	0,0245
δ_2	0,87	0,8519	0,8504	0,0235	0,5567	1,1533	0,5525	1,1389	-0,0181	0,0238
δ_3	0,6	0,6667	0,6733	0,0262	0,3519	0,9898	0,3638	0,9933	0,0667	0,0306
γ	5,0	5,1327	5,1313	0,0265	4,8101	5,4648	4,8129	5,4648	0,1327	0,0441

n = 1000	Real	Média	Mediana	Variância	IC		HPD		Vício	EQM
β_0	0,4	0,4378	0.4364	0,0044	0,3070	0,5679	0,3097	0,5692	0,0378	0,0059
β_1	0,5	0,4206	0,4196	0,0050	0,2811	0,5577	0,2912	0,5653	-0,0794	0,0113
β_2	-0,6	-0,6383	-0,6390	0,0045	-0,7708	-0,5073	-0,7753	-0,5212	-0,0383	0,0059
β_3	0,6	0,6611	0,6618	0,0050	0,5249	0,7960	0,5249	0,7952	0.0611	0,0087
δ_0	0,2	0,2649	0,2620	0,0118	0,0572	0,4861	0,0649	0,4915	0,0649	0,0160
δ_1	-0,5	-0,5311	-0,5304	0,0118	-0,7351	-0,3110	-0,7351	-0,3111	0,0311	0,0127
δ_2	0,87	-0,8504	0,8560	0,0115	0,6352	1,0530	0,6237	1,0302	-0,0196	0,0119
δ_3	0,6	0,5476	0,5510	0,0130	0,3183	0,7697	0,3152	0,7624	-0,0524	0,0157
γ	5,0	5,1082	5,1015	0,0117	4,9025	5,3458	4,8960	5,3327	0,1082	0,0234

Ao observarmos as Tabelas 8, 10 e 12, percebemos que existe um comportamento semelhante entre elas. Inicialmente, todas elas revelam valores estimados diferentes dos valores reais, porém à medida que aumentamos os valores da amostra, os valores vão se aproximando do valor real. O mesmo acontece com os intervalos de credibilidade e HPD, pois quanto mais próximo dos valores reais, as estimativas vão se aproximando, os intervalos vão diminuindo. Houve convergência nas cadeias, como podemos observar na Tabela 13.

Tabela 13 – Teste de Convergência de Geweke, para $\gamma = 5$.

	β_0	β_1	β_2	β_3	δ_0	δ_1	δ_2	δ_3	γ
n = 50	1,3649	-1,6935	-1,0567	0,0290	-0,0439	0,5906	-1,0138	0,4240	-1,4255
n = 100	1,5105	0,1986	-1,5803	-1,0921	1,9101	-0,4447	-2,2009	-1,1879	-0,5218
n = 500	-0,1216	0,3790	-0,4903	0,2233	1,4189	-1,5817	-1,1617	-1,1221	0,2007
n = 1000	0,1605	1,2781	0,4056	-2,1141	0,9822	0,1963	-0,7979	-1,1136	-1,8148

No Cenário 2, fixamos os valores de $\gamma = (1.6,\ 3,\ 5)$ e alteramos os valores dos parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\Delta}$ do cenário 1. Ao analisar os resultados dos resíduos para os quatro tamanhos amostrais $(n = 50, n = 100, n = 200, n = 500, n = 1000)$ e para os três parâmetros $(\gamma = 1.6, \gamma = 3, \gamma = 5)$ considerados no processo, percebemos que há muita similaridade e, portanto, serão apresentado os gráficos dos resíduos para o tamanho amostral $n = 100$ e para o parâmetro $\gamma = 3$.

Ao observar o gráfico dos resíduos via CPO, na Figura 8, para a variável resposta discreta Y, percebemos que há uma oscilação dos pontos, em uma faixa horizontal, em torno de zero e não pussui um padrão, o que era esperado no gráfico. Além disso, no gráfico dos resíduos via CPO $\times$ valor predito exibe apresentam um bom comportamento.

Em relação aos resíduos baseados na distribuição *a posteriori* dos parâmetros, podemos notar um comportamento esperado, visto que os pontos estão bastante dispersos e em torno de zero, com alguns poucos valores fora do intervalo $(-2, 2)$.

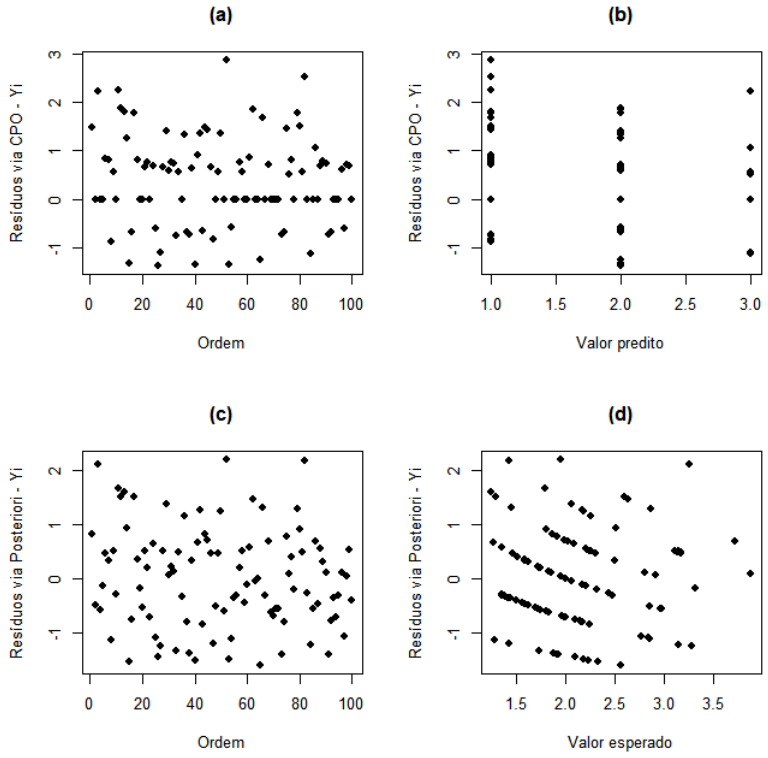

Figura 8 – Gráfico dos Resíduos para Y, em que $n = 100$, $\gamma = 1.6$ e $\boldsymbol{\beta} = (0,4; 0,5; -0,6; 0,6)$ e $\boldsymbol{\Delta} = (0,2; -0,5; 0,87; 0,6)$.

A Figura 9, apresenta os boxplot das amostras MCMC da distribuição *a posteriori* para os resíduos baseados na distribuição a posteriori dos parâmetros. Observamos inúmeros pontos dispersos, o que pode indicar possíveis *outliers*. Ademais, o boxplot apresenta uma pequena variação nos dados.

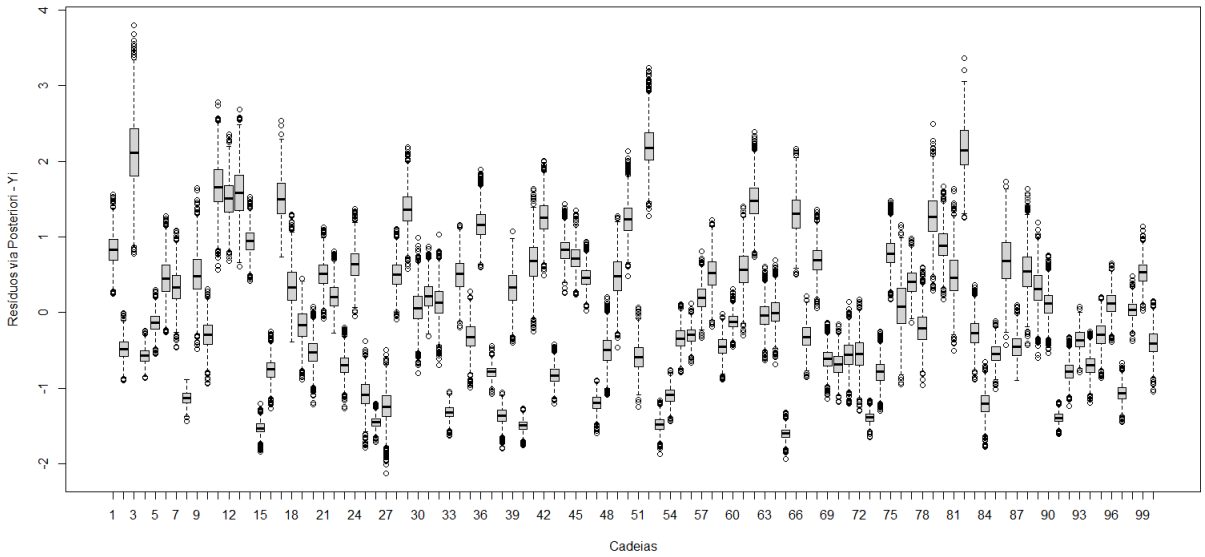

Figura 9 – Boxplot das amostras MCMC da distribuição *a posteriori* dos resíduos para cada observação Y, em que $n = 100$, $\gamma = 1.6$ e $\boldsymbol{\beta} = (0,4; 0,5; -0,6; 0,6)$ e $\boldsymbol{\Delta} = (0,2; -0,5; 0,87; 0,6)$.

Na Figura 10, consta o gráfico dos resíduos via CPO e os resíduos baseados na distribuição *a posteriori* dos parâmetros para a variável resposta contínua X, condicionada a resposta discreta Y, que segue uma distribuição Exponencial.

Há uma oscilação em uma faixa horizontal torno do zero e não apresenta um padrão exato, como observamos na Figura 10. Assim, como no cenário 1, verificamos uma assimetria na distribuição de ambos os resíduos com valores variando, em geral, entre -2 e 6, com valores concentrados em torno de 0.

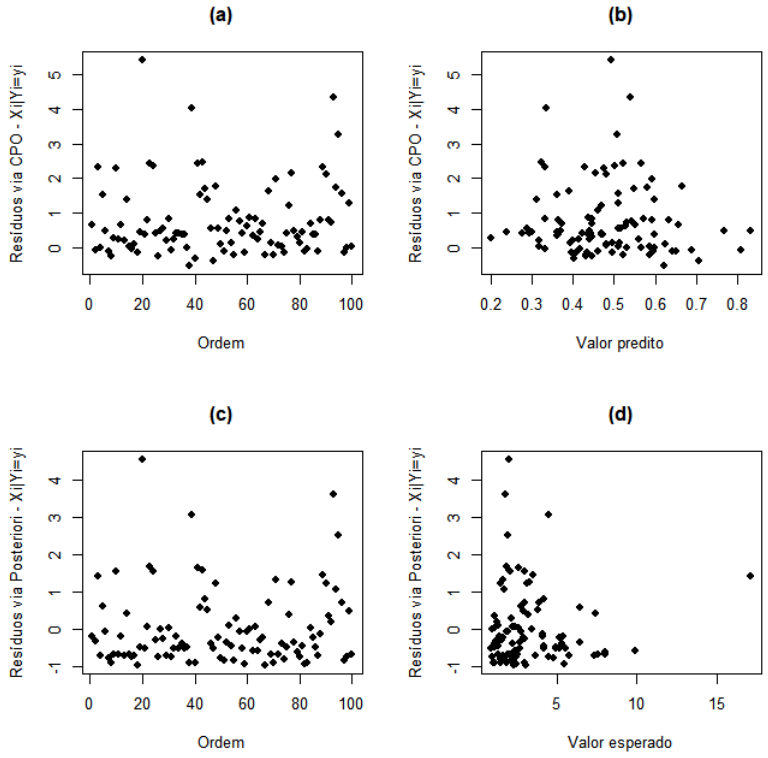

Figura 10 – Gráfico dos Resíduos para X, em que $n = 100$, $\gamma = 1.6$ e $\boldsymbol{\beta} = (0,4; 0,5; -0,6; 0,6)$ e $\boldsymbol{\Delta} = (0,2; -0,5; 0,87; 0,6)$.

Como esperado, o boxplot dos resíduos apresentou observações com intervalos pequenos, o que pode significar que há pouca variabilidade dos dados e há alguns pontos *outliers*, como podemos contemplar na Figura 11.

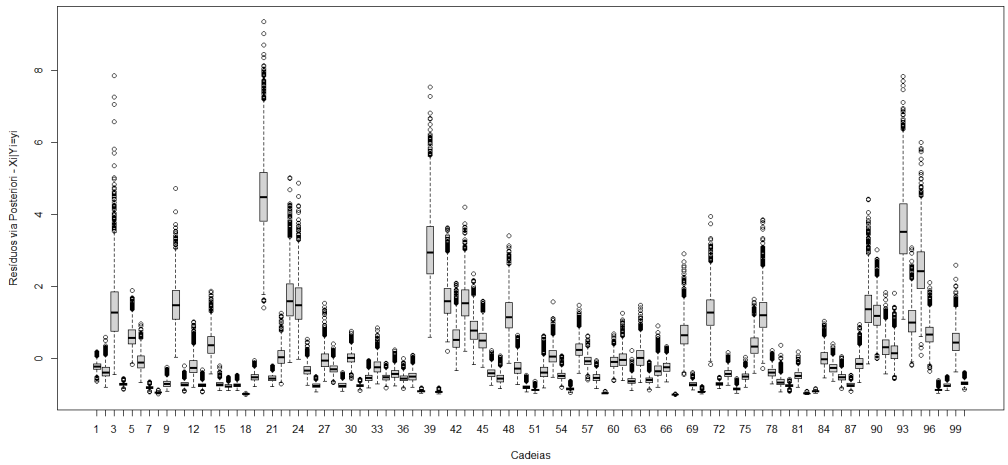

Figura 11 – Boxplot das amostras MCMC da distribuição *a posteriori* dos resíduos para cada observação X, em que $n = 100$, $\gamma = 1.6$ e $\boldsymbol{\beta} = (0,4; 0,5; -0,6; 0,6)$ e $\boldsymbol{\Delta} = (0,2; -0,5; 0,87; 0,6)$

Com o objetivo de verificar as propriedades frequentistas dos estimadores Bayesiano, na Tabela 14 são apresentadas as médias dos vícios e dos erros quadráticos médios das médias *a posteriori*. Além das probabilidade de cobertura estimadas para os intervalos de credibilidade inter-quartil e HPD, baseadas em 1000 simulações com $n = 50,\ 100,\ 500$ e 1000.

Operamos essa análise para os três valores de γ considerados, tanto para covariáveis contínuas quanto para covariáveis discretas. Visto que os resultados em todas as configurações do cenário 1 foram similares, dispomos apenas os resultados obtidos para covariáveis contínuas com $\gamma = 1.6$. A probabilidade de cobertura nominal considerada foi de 95%.

Para determinar os valores dos erros quadráticos médios e vícios das médias *a posteriori* para cada amostra de dados gerada, considerando os tamanhos de amostra $(n = 50,\ 100,\ 500,\ 1000)$ são gerados 1000 amostras MCMC das distribuições *a posteriori* para cada parâmetro, de acordo com o cenário 2.

Tabela 14 – Médias dos erro quadrático médio (EQM) e dos vícios da média *a posteriori* e probabilidade de cobertura estimadas para os intervalos de credibilidade inter-quartil e HPD, para diferentes tamanhos de amostrais e $\gamma = 1.6$, de acordo com o cenário 2

Parâmetro	Tamanho da amostra	EQM	Vício	Inter-quartil	HPD
β_0	50	0,2228	-0,0406	94,2	94,0
	100	0,1003	-0,0257	94,2	94,4
	500	0,0191	0,0070	94,8	95,0
	1000	0,0096	-0,0019	95,4	95,8
β_1	50	0,2700	0,0110	93,4	93,4
	100	0,1202	0,0141	94,4	94,4
	500	0,0221	-0,0083	95,6	94,2
	1000	0,0108	0,0007	95,8	95,2
β_2	50	0,2602	-0,0430	92,4	92,0
	100	0,1088	-0,0002	94,6	94,4
	500	0,0208	-0,0096	95,4	95,2
	1000	0,0112	-0,0040	92,8	92,6
β_3	50	0,2420	0,0183	95,4	95,4
	100	0,1100	-0,0037	96,8	96,6
	500	0,0227	0,0001	95,2	94,0
	1000	0,0120	0,0036	94,4	93,2
δ_0	50	0,5094	-0,0103	96,4	95,0
	100	0,2391	-0,0166	95,4	93,6
	500	0,0445	0,0008	95,0	95,0
	1000	0,0232	-0,0025	93,6	93,0
δ_1	50	0,5933	-0,0148	95,2	94,4
	100	0,2668	-0,0262	95,2	94,8
	500	0,0500	0,0004	95,6	95,0
	1000	0,0325	0,0075	94,6	94,0
δ_2	50	0,5923	-0,0183	94,2	93,6
	100	0,2694	0,0197	93,6	94,2
	500	0,0486	-0,0032	93,2	93,4
	1000	0,0245	0,0008	94,4	94,4
δ_3	50	0,6262	-0,0130	94,0	93,8
	100	0,2833	0,0025	93,8	93,6
	500	0,0474	-0,0050	96,4	95,8
	1000	0,0269	0,0031	95,6	94,8
γ	50	0,0750	0,0483	94,6	94,4
	100	0,0317	0,0219	94,4	94,0
	500	0,0050	0,0063	95,0	94,6
	1000	0,0045	-0,0012	94,0	93,2

Ao observarmos a Tabela 14, é possível notar uma semelhança ao cenário 4, ou seja, quando aumenta o tamanho da amostra, os valores das médias dos EQM e dos vícios vão se aproximando de zero. Já em relação às probabilidades de cobertura para os intervalos de credibilidade inter-quartil e HPD, estão próximas da probabilidade de cobertura nominal que é de 95%.

2.6.3 Resultado para o Cenário 3

Os resíduos podem apresentar pontos *outliers*, indicando que o ajuste não é adequado. Diante disso, realizamos um estudo de perturbação para verificar a performance dos resíduos na detecção de observações atípicas, *outliers*. O propósito é identificar observações ou conjuntos de observações que exercem uma influência desproporcional nos resultados do modelo. Utilizamos o cenário 2 com $\gamma = 1,6$ para realizar o estudo.

Assim, foi perturbado o valor da variável resposta discreta Y, na observação 37, e o valor da variável contínua X, na observação 26, em um conjunto de dados de tamanho $n = 100$, considerando a configuração do cenário 2 com covariáveis contínuas.

Na figura 12, são dispostos os resíduos para Y, considerando a perturbação na observação 37 para respostas discretas. Na figura 13, constam os resíduos para X, considerando a perturbação na observação 26 para respostas contínuas. Podemos expor que, por meio da análise, foi possível verificar que as observações estavam em discrepância com as demais, já que o comportamento dos resíduos, associados às observações perturbadas, ficaram diferentes do que se observou no estudo nos cenários 1 e 2, o que significa que, através, da análise de resíduos é possível identificar possíveis observações *outliers*.

É importante destacar que perturbações em outras observações, em dados com outros tamanhos de amostra e nas configurações do cenário 1, também foram verificados, sendo os resultados semelhantes e, portanto, não apresentados neste trabalho.

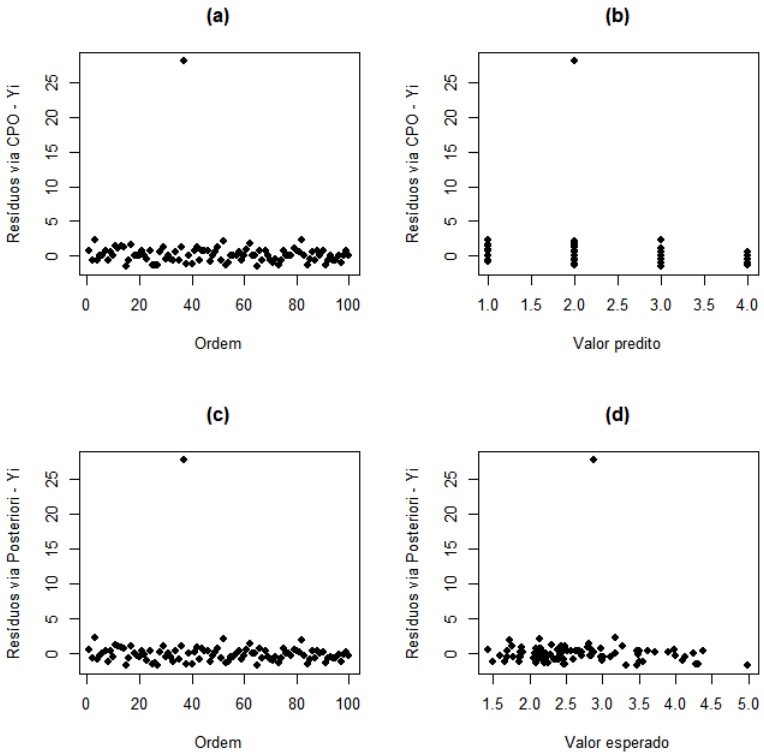

Figura 12 – Resíduos para Y, considerando o Cenário 3 (dados perturbados), em que $n = 100$, $\gamma = 1.6$.

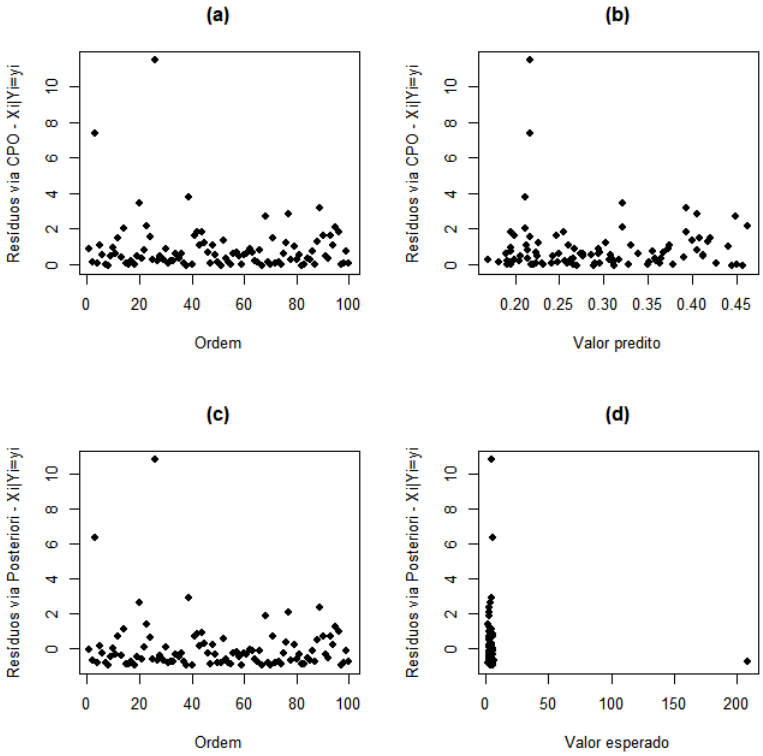

Figura 13 – Resíduos para X, considerando o Cenário 3 (dados perturbados), em que $n = 100$, $\gamma = 1.6$.

2.7 Conclusão

Utilizamos o modelo bivariado Poisson-Exponencial, proposto por Stulp (2019), dentro de um contexto Bayesiano para a estimação dos parâmetros. Tal modelo apresenta uma estrutura de dependência entre as variáveis respostas a partir da média da distribuição condicional e covariáveis que se associam às médias marginais através de funções de ligação.

Outrossim, realizamos uma abordagem Bayesiana para estimar os parâmetros do modelo. Apresentamos uma análise de resíduos baseados na densidade preditiva condicional ordinária (CPO) e o resíduo baseado na distribuição *a posteriori* para o modelo proposto.

Desenvolvemos um estudo de simulação para ilustrar a metodologia considerando os quatro tamanhos amostrais diferentes, a fim de verificar o bom funcionamento da técnica, isto é, a qualidade do ajuste Bayesiano e a performance dos resíduos propostos. Com base nas análises de resíduos apresentadas, observamos a adequabilidade dos modelos aos dados, ademais desenvolvemos um estudo de perturbação, no qual detectamos a presença de alguns pontos extremos, os *outliers*.

A estimação Bayesiana mostrou-se bastante interessante, visto que os resultados não foram muito destoantes quando comparado com os resultado do método frequentista.

Isso indica que a abordagem Bayesiana é uma boa alternativa.

Capítulo 3

Considerações Finais e Propostas Futuras

Considerações Finais

O estudo de modelos bivariados têm se tornado muito importante, pois inúmeras situações permitem dois resultados possíveis como resposta. O trabalho de (STULP, 2019) vem para colaborar com os modelos já propostos na literatura diante de uma perspectiva frequentista. O objetivo deste trabalho foi expandir o modelo Poisson-Exponencial para a metodologia Bayesiana. Assim, foram encontrados suas distribuições *a posteriori* e, então, foram realizados estudos de simulação, para diferentes tamanhos de amostras bem como análise de resíduos, para verificar a robustez desse modelo, a fim de avaliar seus pressupostos. Os resultados apresentados no decorrer deste estudo, mostram que a estimação Bayesiana é uma boa abordagem para este modelo. Além disso, a partir desse modelo, foram realizados alguns esquemas de perturbação para verificar se o modelo é bem ajustado. Percebemos que, de fato, a metodologia reconhece as observações perturbadas e, portanto, o modelo é bem ajustado aos dados.

Propostas Futuras

Para os próximos trabalhos, propomos trabalhar na análise de resíduos *deviance*, como apresentada nos trabalhos de Prado (2013) e Pires (2012), bem como ajustar o modelo para um conjunto de dados reais. Propomos, também, outras distribuições *à priori* para os parâmetros do modelo. Assim, o diagnóstico de influência é outro tópico a ser estudado futuramente.

Referências

BASTOS, M. M. *Modelagem probabilística da dinâmica da Zika usando modelos hierárquicos bayesianos*. Tese (Doutorado), 2018.

BOLFARINE, H.; SANDOVAL, M. C. *Introdução à inferência estatística*. [S.l.]: SBM, 2001. v. 2.

BORGES, L. C. *Análise bayesiana do modelo fatorial dinâmico para um vetor de séries temporais usando distribuições elípticas.* Tese (Doutorado) — Universidade de São Paulo, 2008.

BROOKS, S.; SMITH, J.; VEHTARI, A.; PLUMMER, M.; STONE, M.; ROBERT, C. P.; TITTERINGTON, D.; NELDER, J.; ATKINSON, A.; DAWID, A. et al. Discussion on the paper by spiegelhalter, best, carlin and van der linde. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, Wiley-Blackwell, v. 64, n. 4, p. 616–639, 2002.

CARLIN, B. P.; LOUIS, T. A. *Bayes and empirical Bayes methods for data analysis*. [S.l.]: Chapman & Hall/CRC„ 2000.

CATALANO, P. J.; RYAN, L. M. Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 87, n. 419, p. 651–658, 1992.

CHEN, M.-H.; SHAO, Q.-M.; IBRAHIM, J. G. *Monte Carlo Methods in Bayesian Computation*. New York, NY: Springer, 2000. 19–66 p.

CHO, H.; IBRAHIM, J. G.; SINHA, D.; ZHU, H. Bayesian case influence diagnostics for survival models. *Biometrics*, Wiley Online Library, v. 65, n. 1, p. 116–124, 2009.

COLES, S.; JR, T. P. J. R. Inferência estatística. 2016.

CUNHA, D. R. d. et al. Modelos de regressão bivariada: uma aplicação em equações mincerianas de rendimento. Universidade Federal de Goiás, 2018.

EHLERS, R. S. Inferência bayesiana. *Departamento de Matemática Aplicada e Estatıstica, ICMC-USP*, v. 64, 2011.

FARIA, C. U. de; MAGNABOSCO, C. de U.; REYES, A. de los; LÔBO, R. B.; BEZERRA, L. A. F. Inferência bayesiana e sua aplicação na avaliação genética de bovinos da raça nelore: revisão bibliográfica. *Ciência Animal Brasileira*, v. 8, n. 1, p. 75–86, 2007.

FITZMAURICE, G. M.; LAIRD, N. M. Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American statistical Association*, Taylor & Francis, v. 90, n. 431, p. 845–852, 1995.

JUNG, R. C.; WINKELMANN, R. Two aspects of labor mobility: a bivariate poisson regression approach. *Empirical economics*, Springer, v. 18, n. 3, p. 543–556, 1993.

KHAFRI, S.; KAZEMNEJAD, A.; ESKANDARI, F. Hierarchical bayesian analysis of bivariate poisson regression model 1. Citeseer, 2008.

LOUZADA, F.; SUZUKI, A.; CANCHO, V. The fgm long-term bivariate survival copula model: modeling, bayesian estimation, and case influence diagnostics. *Communications in Statistics-Theory and Methods*, Taylor & Francis, v. 42, n. 4, p. 673–691, 2013.

MAIOLI, M. C. *Inferência Bayesiana como um procedimento de decisão.* 2014. Monografia (PIBIC/CNPq), UNICAMP (Universidade Estadual de Campinas), Campinas, Brazil.

MELLO, C. R. d.; SILVA, A. M. d. Modelagem estatística da precipitação mensal e anual e no período seco para o estado de minas gerais. *Revista Brasileira de Engenharia Agrícola e Ambiental*, SciELO Brasil, v. 13, n. 1, p. 68–74, 2009.

OLIVEIRA, M. B. d. Bayes te bayes tn: modelos bayesianos robustos para seleção genômica ampla. Universidade Federal de Piauí, p. 71, 2019.

OLIVEIRA, W. L. de; DINIZ, C. A. R.; DURBÁN, M. A class of bivariate regression models for discrete and/or continuous responses. *Communications in Statistics-Simulation and Computation*, Taylor & Francis, v. 48, n. 8, p. 2359–2383, 2019.

OLKIN, I.; TATE, R. F. et al. Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 32, n. 2, p. 448–465, 1961.

PIRES, R. M. Modelos de regressão binomial correlacionada. Universidade Federal de São Carlos, 2012.

PRADO, F. B. d. Modelos de regressão bivariados bernoulli: exponencial. Universidade Federal de São Carlos, 2013.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. Disponível em: <http://www.R-project.org/>.

RIBEIRO, T. R. Modelagens estatística para dados de sobrevivência bivariados: uma abordagem bayesiana. Universidade Federal de São Carlos, 2017.

SCOLLNIK, D. P. Regression models for bivariate loss data. *North American Actuarial Journal*, Taylor & Francis, v. 6, n. 4, p. 67–80, 2002.

SONG, J.; BARNHART, H. X.; LYLES, R. H. A gee approach for estimating correlation coefficients involving left-censored variables. *Journal of Data Science*, v. 2, n. 3, p. 245–257, 2004.

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)*, Wiley Online Library, v. 64, n. 4, p. 583–639, 2002.

STULP, P. *Diagnóstico de Influência Local no modelo Poisson-Exponencial*. Dissertação (Dissertação de Mestrado) — Universidade Estadual de Maringá, 2019.

WIKLE, C. K.; MILLIFF, R. F.; HERBEI, R.; LEEDS, W. B. Modern statistical methods in oceanography: A hierarchical perspective. *Statistical Science*, JSTOR, p. 466–486, 2013.