UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelos de regressão para dados de contagem *k*-Modificados

Milene Alves Garcia

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

ICMC USP SÃO CARLOS

Milene Alves Garcia

Modelos de regressão para dados de contagem *k*-Modificados

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

Área de Concentração: Estatística

Orientadora: Profa. Dra. Katiane Silva Conceição

USP – São Carlos
Fevereiro de 2020

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

Garcia, Milene Alves
G634m Modelos de Regressão para Dados de Contagem k-Modificados / Milene Alves Garcia; orientadora Katiane Silva Conceição. - São Carlos - SP, 2020.
130 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-Graduação em Estatística) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2020.

1. Dados de Contagem; Dados $k-$ Deflacionados; Dados $k-$Inflacionados; Distribuição *Hurdle*; Distribuições Discretas; Modelos de Regressão. I. Conceição, Katiane Silva, orient. II. Título.

Milene Alves Garcia

k-Modified regression models to count data

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

Concentration Area: Statistics

Advisor: Profa. Dra. Katiane Silva Conceição

USP – São Carlos
February 2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

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

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado da candidata Milene Alves Garcia, realizada em 21/01/2020:

Katiane S. Conceição

Profa. Dra. Katiane Silva Conceição
USP

Prof. Dr. Caio Lucidius Naberezny Azevedo
UNICAMP

Prof. Dr. Anderson Luiz Ara Souza
UFBA

Este trabalho é dedicado a todas as pessoas que acreditam no poder da educação.

AGRADECIMENTOS

Agradeço primeiramente a Deus pois sei que Ele está sempre comigo.

Aos meus pais Fernando e Ilsa por todo amor e cuidado, e por sempre incentivarem meus estudos e meu desenvolvimento pessoal.

Ao meu namorado Jhonata pelo amor, companheirismo e toda a ajuda ao longo dos anos. Você torna meus dias mais leves.

À minha orientadora e amiga Katiane, por compartilhar comigo o conhecimento e pela amizade construída.

Ao professor e amigo Marinho de Andrade, por todo apoio, ajuda e amizade.

Aos meus colegas de mestrado, e em especial aos amigos Gabriela, Fabiana, Juliana e Victor.

À minha família e amigos.

Aos professores da USP e da UFSCar pelo empenho, e a todos os professores que contribuíram para minha formação.

Pesquisa desenvolvida com o auxílio dos recursos de HPC (*High Performance Computing*) disponibilizados pela Superintendência de Tecnologia da Informação da Universidade de São Paulo.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

“Aprenda como se você fosse viver para sempre.
Viva como se você fosse morrer amanhã.”
(Santo Isidoro de Sevilha)

RESUMO

GARCIA, M. A.. **Modelos de Regressão para Dados de Contagem k-Modificados**. 2020. 130 f. Dissertação (Mestrado em em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Neste trabalho é proposto o modelo de regressão para a família de distribuições k-Modificadas. Entende-se como modificação, a inclusão de um parâmetro na função massa de probabilidade das distribuições discretas tradicionais, capaz de modelar a inflação ou deflação da observação k no conjunto de dados. A modificação em relação à distribuição original torna-se imprescindível quando, em muitas situações práticas, uma determinada observação k ocorre no conjunto de dados com uma frequência maior ou menor do que a esperada ao considerar uma determinada distribuição discreta. Para o contexto de modelos de regressão, distribuições discretas cuja função massa de probabilidade pode ser escrita em função de sua média serão consideradas.

Palavras-chave: Dados de Contagem; Dados $k-$ Deflacionados; Dados $k-$Inflacionados; Distribuição *Hurdle*; Distribuições Discretas; Modelos de Regressão.

ABSTRACT

GARCIA, M. A.. **Modelos de Regressão para Dados de Contagem *k*-Modificados**. 2020. 130 f. Dissertação (Mestrado em em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

In this work the regression model to *k*-Modified family distributions is proposed. It is understood as modification, the inclusion of a parameter in the probability mass function of the traditional discrete distributions, which is able to model inflation or deflation of the observation k in the dataset. The modification in relation to the original distribution becomes necessary when, in many practical situations, a specific observation k occurrs in the dataset with higher or lower frequency than what is expected by considering a specific discrete distribution. In the context of regression models, discrete distributions with probability mass function that can be written through their means will be considered.

Key-words: Count Data; Discrete Distributions; Hurdle Distribution; *k*-Deflated Data; *k*-Inflated Data; Regression Models.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

χ^2 ..... Qui-Quadrado
k-DB ..... Binomial k-Deflacionada
k-DG ..... Geométrica k-Deflacionada
k-DP ..... Poisson k-Deflacionada
k-IB ..... Binomial k-Inflacionada
k-IG ..... Geométrica k-Inflacionada
k-IP ..... Poisson k-Inflacionada
k-MB .... Binomial k-Modificada
k-MG .... Geométrica k-Modificada
k-MP ..... Poisson k-Modificada
k-MPS ... Série de Potência k-Modificada
$k-$DPS ... Série de Potência $k-$Deflacionada
$k-$IPS .... Série de Potência $k-$Inflacionada
$k-$MD ... Discretas $k-$Modificadas
$k-$SPS ... Série de Potência k-Subtraída
B ..... Equivalente ao critério *Logarithm of the Pseudo Marginal Likelihood (LPML)*
C. Log-log Complemento Log-log
DIC ..... Critério de Informação Deviance
EAIC ..... Critério de Informação Akaike Esperado
EBIC ..... Critério de Informação Bayesiano Esperado
IBGE ..... Instituto Brasileiro de Geografia e Estatística
IDH ..... Índice de Desenvolvimento Humano
Jags ..... Just Another Gibs Sample
KS ..... Kolmogorov-Smirnov
LPML .... Logaritmo da Verossimilhança Pseudo Marginal
MCMC ... Monte Carlo Cadeia de Markov
MV ..... Máxima Verossimilhança
PS ..... Série de Potência
ZMPS .... Série de Potência Zero-Modificada

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Na Estatística, muitas situações reais estão relacionadas a problemas de contagem. Nesse contexto, distribuições da família Série de Potência são geralmente empregadas, como pode ser visto em Khatri (1959), Patil (1962), Gupta (1974), Consul (1990) e Johnson, Kotz e Kemp (1992). Alguns exemplos de distribuições dessa família são: distribuição Binomial, Geométrica, Poisson, Poisson Generalizada e Binomial Negativa, e essas distribuições podem ser utilizadas para descrever, por exemplo, o número de novos casos de câncer de mama por dia, o número de células contadas usando um hemacitômetro, o número de nematóides encontrados em amostras de solo, entre outros. Porém, em certos conjuntos de dados, as seguintes situações podem ocorrer:

a) A frequência observada de um determinado valor k pode apresentar discrepância ao ser comparada com a frequência esperada de uma distribuição discreta tradicional, podendo esta discrepância ser alta, baixa ou nula;

b) Há sobredispersão nos dados, ou seja, a variância amostral é maior que a sua média;

c) Há subdispersão nos dados, ou seja, a variância amostral é menor que a sua média.

Desta forma, considerar distribuições tradicionais na sua forma original pode ser inadequado para estas situações.

Em algumas situações práticas, conjuntos de dados podem apresentar bimodalidade (ou até mais que duas modas), e então torna-se necessário encontrar um modelo probabilístico que explique adequadamente o comportamento destes dados. No contexto mais simples, com ilustração na Figura 1, ao considerar como exemplo um conjunto de dados que apresenta bimodalidade (Mo_1 e Mo_2), pode-se pensar que este comportamento se deve a uma baixa frequência da observação k, sendo $Mo_1 < k < Mo_2$ ou ainda, a uma alta frequência da observação k, sendo $k = Mo_1$ ou $k = Mo_2$. Neste sentido, identificar essa discrepância (alta ou baixa) na

frequência da observação k é fundamental para realizar boas inferências sobre o(s) parâmetro(s) de interesse.

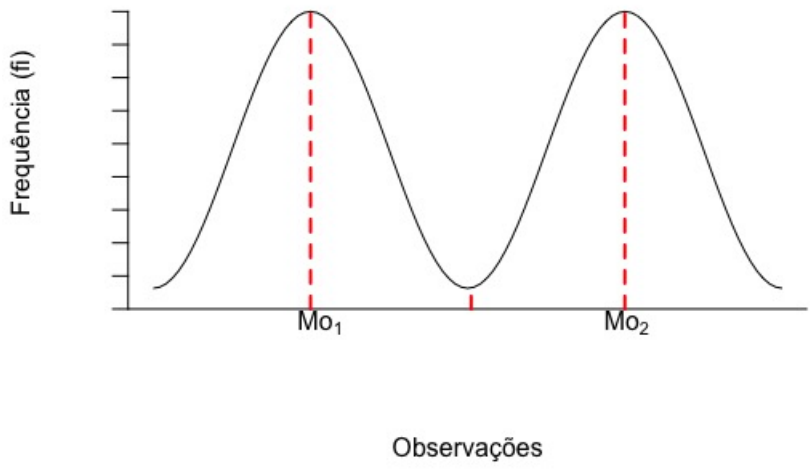

Figura 1 – Conjunto de dados artificiais com bimodalidade.

Fonte: Elaborada pelo autor.

Outros exemplos de dados de contagem com modificação na frequência da observação k podem ser vistos nas seguintes situações reais:

a) Pandey (1965) realizou um estudo com plantas *Primula veris* e observou que muitas plantas produziram 8 flores. Ao conjunto de dados que representa o número de flores por cada planta, ele ajustou a distribuição de Poisson 8-inflacionada ($k = 8$);

b) Lambert (1992) discutiu que equipamentos manufaturados podem estar em dois diferentes estados: no primeiro estado a máquina está perfeita e não produz nenhum defeito, e no segundo estado a máquina está imperfeita e produz um número de erros que segue uma distribuição de Poisson. Desta forma, ao conjunto de dados que representa o número de erros em cada máquina, é possível ajustar a distribuição de Poisson 0-Inflacionada ($k = 0$);

c) Considerando o número de visitas ao dentista por indivíduo, muitas vezes, observa-se uma grande quantidade de zero visitas na amostra (ROSENQVIST; ARIEN; SINTONEN, 1995), e esse comportamento pode ocorrer por diversos motivos: medo de ir ao dentista, pelos indivíduos terem dentes saudáveis, ou simplesmente por falta de tempo ou dinheiro, por exemplo. Desta forma, ao conjunto de dados que representa o número de visitas ao dentista por cada indivíduo, podemos também ajustar a distribuição de Poisson 0-Inflacionada;

d) Considere famílias carentes que participam de um programa do governo, em que benefícios são dados por número de filhos. Suponha um caso hipotético em que muitas famílias optam por ter mais filhos para ter mais benefícios. Neste sentido, o conjunto de dados formado pelo número de filhos nas famílias sugere uma deflação na observação 1, no caso a distribuição Poisson 1-Deflacionada ($k = 1$).

Seguindo este contexto, Carvalho (2017) propôs a família de distribuições Série de Potência k-Modificadas (k-MPS[1]), uma extensão da família Série de Potência comumente usada em problemas de dados de contagem, em que um novo parâmetro é adicionado na função massa de probabilidade das distribuições Série de Potência, o qual é responsável por explicar adequadamente a ocorrência da observação discrepante no conjunto de dados, bem como a frequência de todas as outras observações. Assim, a distribuição k-MPS comporta dados com observação k em excesso (k-inflação), dados com excassez de observação k (k-deflação), dados com ausência da observação k (k-subtração), ou até mesmo dados com a frequência usual nas suas observações (ou seja, a própria distribuição Série de Potência tradicional).

No contexto mais amplo, neste trabalho estendemos a ideia da $k-$modificação para qualquer distribuição discreta, sendo a família k-MPS o seu caso particular. Para a estimação pontual dos parâmetros utilizaremos o método de máxima verossimilhança e também uma abordagem Bayesiana, em que esta última será a principal abordagem considerada neste trabalho por permitir incoporar informações *a priori* sobre os parâmetros. Finalmente, apresentaremos a distribuição discretas $k-$Modificada no contexto de modelos de regressão.

1.1 Motivação

Este trabalho é motivado pelo seguinte exemplo: considere o conjunto de dados referente às notificações de óbitos fetais (segundo a fonte dos dados, IBGE (2016), óbito fetal é a morte "ocorrida antes da expulsão ou de sua extração completa do corpo materno, independentemente da duração da gestação") ocorridos e registrados no local de residência da mãe do feto, em cidades do Estado da Bahia (Brasil) em 2014.

O Estado da Bahia tem 417 cidades e a distribuição de frequência de notificações de mortes fetais nas cidades baianas está apresentada na Tabela 1:

Tabela 1 – Notificações de óbitos fetais em cidades do estado da Bahia em 2014.

y_i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
f_i	96	84	57	45	40	28	12	9	8	5	4	6	2	3	1
y_i	15	16	18	20	23	24	30	43	50	53	56	71	116	601	Total
f_i	1	3	1	1	1	1	1	1	1	2	1	1	1	1	417

[1] Usaremos a sigla k-MPS, do inglês *k-Modified Power Series*, referindo-se às distribuições Série de Potência k-Modificadas.

É importante enfatizar que, embora 96 cidades no conjunto de dados notificaram observações zero, isto não significa necessariamente que não existam casos de óbitos fetais nesses lugares, mas essa situação pode ocorrer por deficiência do sistema em incluir casos notificados da variável investigada. Desta forma, é fundamental avaliar a probabilidade de não-notificações de óbitos fetais.

Para explicar as ocorrências de notificações de óbitos fetais, consideramos como variável explicativa o Índice de Desenvolvimento Humano (IDH), que é uma medida comparativa baseada em três pilares: saúde, educação e renda (PNDU, 2016), isto é, uma forma padronizada de medir o bem estar da população. As 417 cidades do estado do Bahia tem IDH entre 0.486 e 0.759, que corresponde ao IDH das cidades de Itapicuru e Salvador, respectivamente (ATLAS, 2011)[2].

Embora seja possível ajustar o modelo Poisson para os dados de notificações de óbitos fetais, ao analisar as frequências dos casos notificados na Tabela 1, notamos que há uma possível discrepância na frequência da observação zero, fornecendo evidências de um excesso (inflação) desta observação, e assim, através da teoria proposta nesse trabalho será possível ajustar uma regressão de Poisson k-Modificada, considerando $k = 0$, para obter melhores ajustes a este conjunto de dados reais.

1.2 Objetivo

O principal objetivo deste trabalho é ampliar a classe de distribuições k-modificadas apresentadas em Carvalho (2017) e introduzí-las no contexto de modelos de regressão. Esta nova classe será denominada Modelos Discretos k-Modificados. De forma mais específica queremos:

i) ampliar a classe de distribuições k-modificadas introduzindo mais distribuições discretas;

ii) apresentar as distribuições discretas k-modificadas na versão *Hurdle*;

iii) descrever o procedimento de estimação e inferência dos parâmetros das distribuições discretas k-modificadas no contexto geral;

iv) incluir esta família de distribuições no contexto de modelos de regressão;

v) introduzir variáveis explicativas tanto para o parâmetro de média da distribuição discreta associada quanto para o parâmetro responsável pela modificação das probabilidades, principalmente da observação k.

A dissertação está organizada como descrito a seguir: no Capítulo 2 é apresentada a família de distribuições discretas e, em particular, a família Série de Potência uniparamétrica e a Série de Potência k-Subtraída. Apresentamos também as definições da distribuição Discreta k-Modificada e da distribuição Série de Potência k-Modificada (usual e versão *Hurdle*), suas

[2] Informações baseadas no Censo 2010.

propriedades relevantes e particularidades. Demonstramos a estimação de seus parâmetros através de abordagem Clássica e Bayesiana; no Capítulo 3 apresentamos algumas aplicações reais para a distribuição Série de Potência k-Modificada; no Capítulo 4 estendemos a ideia de k-Modificação para o contexto de modelos de regressão e apresentamos o modelo de regressão Série de Potência k-Modificado, suas propriedades e estimação dos parâmetros; no Capítulo 5 apresentamos um estudo de simulação considerando os modelos de regressão Binomial, Geométrica e Poisson k-Modificadas; no Capítulo 6 são apresentadas aplicações dos modelos de regressão Série de Potência k-Modificados, considerando dados artificiais e reais. Por fim, as conclusões e algumas propostas futuras para dar continuidade a este trabalho podem ser vistas no Capítulo 7.

CAPÍTULO

2

CONCEITOS E NOÇÕES PRELIMINARES

Seja Y uma variável aleatória discreta definida sobre os inteiros $\mathscr{A}_s = \{s, s+1, s+2, \ldots\}$, $s \in \mathbb{Z}$, o suporte da variável. Considere $\pi_D(y;\boldsymbol{\theta})$ a função massa de probabilidade associada a variável aleatória Y, com $y \in \mathscr{A}_s$, e $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_r)$, $r \geq 1$, o vetor de parâmetros da distribuição. A média e variância de Y são $\mu = E(Y) = \mu(\boldsymbol{\theta})$ e $\sigma^2 = Var(Y) = \sigma^2(\boldsymbol{\theta})$.

Considerando os parâmetros $\mu = \mu(\boldsymbol{\theta})$ e $\boldsymbol{\phi} = (\phi_1(\boldsymbol{\theta})\ \phi_2(\boldsymbol{\theta})\ \ldots\ \phi_{r-1}(\boldsymbol{\theta}))$ podemos reescrever a função $\pi_D(y;\boldsymbol{\theta})$ como $\pi_D(y;\mu,\boldsymbol{\phi})$. Esta parametrização é conveniente para o contexto de modelos de regressão. Vale ressaltar que para as distribuições discretas mais utilizadas temos, em geral, $r \leq 2$. Um caso particular, temos as distribuições da família Série de Potência (PS), que constituem uma classe ampla de distribuições para dados de contagem e já foram estudadas por Khatri (1959), Patil (1962), Gupta (1974), Consul (1990) e Johnson, Kotz e Kemp (1992), dentre outros.

Neste Capítulo, apresentamos a família Série de Potência Uniparamétrica e sua extensão, que comporta distribuições para dados de contagem com modificação principalmente na probabilidade da observação k, e podem ser construídas a partir das distribuições discretas tradicionais, chamadas de distribuições discretas k-modificadas.

2.1 Família Série de Potência Uniparamétrica

É apresentada a seguir, a família Série de Potência (PS[1]) Uniparamétrica:

Definição 1 (Família Série de Potência Uniparamétrica)**.** Seja Y uma variável aleatória inteira e não-negativa, pertencente à família PS com função massa de probabilidade dada por:

$$\pi_{PS}(y;\mu) = \frac{a(y)g(\mu)^y}{f(\mu)},\ y \in \mathscr{A}_s = \{s, s+1, s+2, \ldots\},\ s \in \mathbb{N}$$

[1] Usaremos a sigla PS, do inglês *Power Series*, referindo-se à família Série de Potência.

em que μ é o parâmetro de média, com $\mu \in \mathscr{M} \subseteq \mathbb{R}^+$; $a(y)$ é uma função positiva; $f(\mu)$ e $g(\mu)$ são funções positivas e duas vezes diferenciáveis, com $f(\mu) = \sum_{y \in A_s} a(y) g(\mu)^y$. Para mais detalhes referente à família PS, ver Gupta (1974) e Cordeiro, Andrade e Castro (2009).

A Tabela 2 nos fornece as funções $a(y)$, $f(\mu)$ e $g(\mu)$ para algumas distribuições da família PS, cujo suporte se inicia em s.

Tabela 2 – Algumas distribuições uniparamétricas da família PS.

PS	Distribuição	$f(\mu)$	$g(\mu)$	$a(y)$	$\mathscr{A}_s$	$\mathscr{M}$
B	Binomial	$\left(\frac{m}{m-\mu}\right)^m$	$\frac{\mu}{m-\mu}$	$\binom{m}{y}$	$\{0,1,\ldots,m\}$	$0<\mu<m$
BO	Borel	$1-\frac{1}{\mu}$	$\left(1-\frac{1}{\mu}\right)e^{-1+\frac{1}{\mu}}$	$\frac{y^{y-2}}{(y-1)!}$	$\{1,2,\ldots\}$	$\mu>0$
BT	Borel-Tanner	$\left(1-\frac{1}{\mu}\right)^m$	$\left(1-\frac{m}{\mu}\right)e^{-1+\frac{m}{\mu}}$	$\frac{my^{y-m-1}}{(y-m)!}$	$\{m,m+1,\ldots\}$	$\mu>0$
G	Geométrica	$1+\mu$	$\frac{\mu}{1+\mu}$	1	$\{0,1,\ldots\}$	$\mu>0$
H	Haight	$\frac{\mu-1}{2\mu-1}$	$\frac{\mu(\mu-1)}{(2\mu-1)^2}$	$\frac{(2y-2)!}{y!(y-1)!}$	$\{1,2,\ldots\}$	$\mu>0$
P	Poisson	e^{μ}	μ	$\frac{1}{y!}$	$\{0,1,\ldots\}$	$\mu>0$

Fonte: Adaptada de Carvalho (2017).

Visando modificar as distribuições pertencentes à família Série de Potência, tornando nula a probabilidade de ocorrência de uma determinada observação, a definição da distribuição Série de Potência k-Subtraída é necessária.

2.1.1 Distribuição Série de Potência $k-$Subtraída

Ao remover o ponto k do suporte da variável Y, $\mathscr{A}_s = \{s, s+1, s+2, \ldots\}$, Carvalho (2017) modificou a distribuição da família Série de Potência, dando origem a distribuição Série de Potência k-Subtraída ($k-$SPS[2]), cuja função massa de probabilidade é dada por:

$$\begin{aligned} \pi_{k-SPS}(y;\mu) &= \frac{\pi_{PS}(y;\mu)}{1-\pi_{PS}(k;\mu)} \\ &= \frac{a(y)g(\mu)^y}{f(\mu)-a(k)g(\mu)^k},\ y \in \mathscr{A}_{\{-k\}} = \{s, s+1, \ldots, k-1, k+1, \ldots\}, \quad (2.1.1) \end{aligned}$$

com $k \geq s$.

Para verificar se $\pi_{k-SPS}(y;\mu)$ é uma função massa de probabilidade, os seguintes itens devem ser satisfeitos:

[2] Usaremos a sigla k-SPS, do inglês *k-Subtracted Power Series*, referindo-se às distribuições Série de Potência k-Subtraídas.

$$\begin{cases} i)\ \ \pi_{k-SP}(y;\mu) > 0,\ \forall\, y \in \mathscr{A}_{\{-k\}} \\ ii)\ \sum\limits_{y\in\mathscr{A}_{-k}} \pi_{k-SP}(y;\mu) = 1 \end{cases}.$$

No item i), já temos conhecimento que $\pi_{PS}(y;\mu) > 0$ e ainda $(1-\pi_{PS}(k;\mu)) > 0$. Portanto, temos que a razão entre estes dois termos é sempre positiva:

$$\frac{\pi_{PS}(y;\mu)}{1-\pi_{PS}(k;\mu)} > 0,\ \forall\, y \in \mathscr{A}_{\{-k\}}.$$

Em ii) temos que:

$$\begin{aligned} \sum_{y\in A_{\{-k\}}} \pi_{k-SP}(y;\mu) &= \sum_{y\in A_{\{-k\}}} \frac{a(y)g(\mu)^y}{f(\mu)-a(k)g(\mu)^k} \\ &= \frac{1}{f(\mu)-a(k)g(\mu)^k} \sum_{y\in A_{\{-k\}}} a(y)g(\mu)^y \\ &= \frac{f(\mu)}{f(\mu)-a(k)g(\mu)^k} \sum_{y\in A_{\{-k\}}} \frac{a(y)g(\mu)^y}{f(\mu)} \\ &= \frac{f(\mu)}{f(\mu)\left(1-\dfrac{a(k)g(\mu)^k}{f(\mu)}\right)} \left(1-\frac{a(k)g(\mu)^k}{f(\mu)}\right) \\ &= \frac{1}{(1-\pi_{PS}(k;\mu))}(1-\pi_{PS}(k;\mu)) = 1. \end{aligned}$$

A média da variável aleatória Y com distribuição $k-$SPS é:

$$\begin{aligned} \mu_{k-SPS} = E[Y] &= \sum_{y\in A_{\{-k\}}} y\frac{a(y)g(\mu)^y}{f(\mu)-a(k)g(\mu)^k} \\ &= \frac{f(\mu)}{f(\mu)-a(k)g(\mu)^k} \sum_{y\in A_{\{-k\}}} y\frac{a(y)g(\mu)^y}{f(\mu)} \\ &= \frac{f(\mu)}{f(\mu)\left(1-\dfrac{a(k)g(\mu)^k}{f(\mu)}\right)} \left(\mu - k\frac{a(k)g(\mu)^k}{f(\mu)}\right) \\ &= \frac{\mu - k\pi_{PS}(k;\mu)}{1-\pi_{PS}(k;\mu)} \end{aligned}$$

e a variância é:

$$\sigma^2_{k-SPS} = Var[Y] \ = \ E[Y^2] - \big\{E[Y]\big\}^2,$$

tal que

$$\begin{aligned} E[Y^2] &= \sum_{y\in A_{\{-k\}}} y^2 \frac{a(y)g(\mu)^y}{f(\mu)-a(k)g(\mu)^k} \\ &= \frac{f(\mu)}{f(\mu)-a(k)g(\mu)^k} \sum_{y\in A_{\{-k\}}} y^2 \frac{a(y)g(\mu)^y}{f(\mu)} \\ &= \frac{f(\mu)}{f(\mu)\left(1-\frac{a(k)g(\mu)^k}{f(\mu)}\right)} \left(\mu(1+\mu)-k^2\frac{a(k)g(\mu)^k}{f(\mu)}\right) \\ &= \frac{\mu^2}{(1-\pi_{PS}(k;\mu))}+\frac{\sigma^2}{(1-\pi_{PS}(k;\mu)}-\frac{k^2\pi_{PS}(k;\mu)}{(1-\pi_{PS}(k;\mu))}, \end{aligned}$$

e sendo $\mu_{k-SPS}=\dfrac{\mu-k\pi_{PS}(k;\mu)}{1-\pi_{PS}(k;\mu)}$, obtemos:

$$\begin{aligned} Var[Y] &= \frac{\mu^2}{(1-\pi_{PS}(k;\mu))}+\frac{\sigma^2}{(1-\pi_{PS}(k;\mu))}-\frac{k^2\pi_{PS}(k;\mu)}{(1-\pi_{PS}(k;\mu))}-\left(\frac{\mu-k\pi_{PS}(k;\mu)}{1-\pi_{PS}(k;\mu)}\right) \\ &= \frac{\sigma^2(1-\pi_{PS}(k;\mu))-\mu^2\pi_{PS}(k;\mu)-k^2\pi_{PS}(k;\mu)+2\mu\pi_{PS}(k;\mu)}{(1-\pi_{PS}(k;\mu))^2}. \end{aligned}$$

Portanto,

$$\sigma^2_{k-SPS}=Var[Y] = \frac{\sigma^2(1-\pi_{PS}(k;\mu))-(\mu-k)^2\pi_{PS}(k;\mu)}{(1-\pi_{PS}(k;\mu))^2}.$$

Estendendo a ideia de modificação nas distribuições da família Série de Potência presente em Carvalho (2017), introduziremos um parâmetro adicional na função massa de probabilidade da família de distribuições discretas de tal forma que esta nova família de distribuições comporte as distribuições discretas tradicionais, incluindo também as distribuições k-subtraídas.

2.2 Distribuição Discreta k-Modificada

Definição 2 (Distribuições $k-$MD[3])**.** Uma variável aleatória discreta Y tem distribuição $k-$modificada, para algum $k\in\mathscr{A}_s$ tal que $k\geq s$, se sua função massa de probabilidade é dada por:

$$\pi_{k-MD}(y;\boldsymbol{\theta},p)=(1-p)I_{\{k\}}(y)+p\pi_D(y;\boldsymbol{\theta}),\ y\in\mathscr{A}_s, \tag{2.2.1}$$

em que p é o parâmetro responsável pela modificação das probabilidades em relação às distribuições tradicionais, satisfazendo a restrição

$$0\leq p\leq \frac{1}{1-\pi_D(k;\boldsymbol{\theta})}, \tag{2.2.2}$$

[3] Usaremos a sigla k-MD, do inglês *k-Modified Discrete*, referindo-se às distribuições Discretas k-Modificadas.

ou seja, $p \in \mathscr{P} = \left[0; \frac{1}{1-\pi_D(k;\boldsymbol{\theta})}\right] \subset \mathbb{R}$; e $I_{\{k\}}(y)$ é uma função indicadora, tal que

$$I_{\{k\}}(y) = \begin{cases} 1, & \text{se } y = k \\ 0, & \text{se } y \neq k \end{cases}.$$

Notação: $Y \sim k-\text{MD}(\boldsymbol{\theta}, p)$.

Similarmente, podemos escrever a distribuição k-MD em função da média da variável aleatória com distribuição discreta tradicional, e neste trabalho iremos considerar as distribuições pertencentes a família Série de Potência.

2.3 Distribuição Série de Potência k-Modificada

Definição 3. Seja Y uma variável aleatória com distribuição pertencente a família Série de Potência $k-$Modificada ($k-$MPS) com parâmetros μ e p. Assim, a função massa de probabilidade de Y é dada por:

$$\pi_{k-MPS}(y;\mu,p) = (1-p)I_{\{k\}}(y) + p\pi_{PS}(y;\mu),\ \forall\, y \in \mathscr{A}_s = \{s, s+1, s+2, \ldots\}, \tag{2.3.1}$$

em que p é o parâmetro responsável pela modificação das probabilidades em relação à distribuição Série de Potência usual, satisfazendo a restrição

$$0 \leq p \leq \frac{1}{1-\pi_{PS}(k;\mu)}. \tag{2.3.2}$$

Se uma variável aleatória Y tem distribuição $k-$MPS, sua média é dada por:

$$\begin{aligned}
\mu_{k-MPS} &= E(Y) = \sum_{y \in PS} y\pi_{k-MPS}(y;\mu,p) \\
&= \sum_{y \in PS} y\left\{(1-p)I_{\{k\}}(y) + p\pi_{PS}(y;\mu)\right\} \\
&= \sum_{y \in PS} y((1-p)I_{\{k\}}(y)) + \sum_{y \in PS} yp\pi_{PS}(y;\mu) \\
&= k(1-p) + p\mu.
\end{aligned}$$

Pela definição de variância, podemos calcular inicialmente $E(Y^2)$:

$$\begin{aligned}
E(Y^2) &= \sum_{y \in PS} y^2\left\{(1-p)I_{\{k\}}(y) + p\pi_{PS}(y;\mu)\right\} \\
&= \sum_{y \in PS} y^2(1-p)I_{\{k\}}(y) + \sum_{y \in PS} y^2 p\pi_{PS}(y;\mu) \\
&= k^2(1-p) + pE_{PS}(Y^2) \\
&= k^2(1-p) + p\left\{Var_{PS}(Y) + \left(E_{PS}(Y)\right)^2\right\} \\
&= k^2(1-p) + p(\sigma^2 + \mu^2),
\end{aligned}$$

uma vez que $Var_{PS}()$ e $E_{PS}()$ correspondem, respectivamente, a média e a variância de uma variável aleatória com distribuição PS. Sendo $E(Y) = \mu_{k-MPS} = k(1-p) + p\mu$, temos:

$$\begin{aligned}
\sigma^2_{k-MPS} &= k^2(1-p) + p(\sigma^2+\mu^2) - \{k(1-p)+p\mu\}^2 \\
&= k^2 - k^2p + p\sigma^2 + p\mu^2 - k^2(1-p)^2 - 2p\mu k(1-p) - p^2\mu^2 \\
&= p\{k^2 + \sigma^2 + \mu^2 - pk^2 - 2\mu k + 2p\mu k - p\mu^2\} \\
&= p\{(k-\mu)^2 - p(k^2 - 2\mu k + \mu^2) + \sigma^2\} \\
&= p\{(k-\mu)^2 - p(k-\mu)^2 + \sigma^2\} \\
&= p\{(k-\mu)^2(1-p) + \sigma^2\}.
\end{aligned}$$

As variâncias das distribuições da família k-MPS e das distribuições PS associadas são apresentadas na Tabela 3. Podemos notar que a variância σ^2_{k-MPS} tem ordem quadrática para valores de k e μ e, além disso, nos casos em que $k = \mu$, a variância σ^2_{k-MPS} atinge seu menor valor (para mais detalhes, ver Carvalho (2017)).

Tabela 3 – Variância das distribuições PS e k-MPS.

PS	Distribuição	σ^2	σ^2_{k-MPS}
B	Binomial	$\mu\left(1-\frac{\mu}{m}\right)$	$p\left\{(1-p)(k-\mu)^2 + \mu\left(1-\frac{\mu}{m}\right)\right\}$
BO	Borel	$\mu^2(\mu-1)$	$p\left\{(1-p)(k-\mu)^2 + \mu^2(\mu-1)\right\}$
BT	Borel-Tanner	$\left(\frac{\mu}{m}\right)^2(\mu-m)$	$p\left\{(1-p)(k-\mu)^2 + \left(\frac{\mu}{m}\right)^2(\mu-m)\right\}$
G	Geométrica	$\mu(1+\mu)$	$p\left\{(1-p)(k-\mu)^2 + \mu(1+\mu)\right\}$
H	Haight	$\frac{2\mu}{(2\mu-1)^3}$	$p\left\{(1-p)(k-\mu)^2 + \frac{2\mu}{(2\mu-1)^3}\right\}$
P	Poisson	μ	$p\left\{(1-p)(k-\mu)^2 + \mu\right\}$

Fonte: Adaptada de Carvalho (2017).

2.3.1 Casos Particulares da Distribuição $k-MPS(\mu, p)$

Diferentes valores de p levam a diferentes particularidades da distribuição $k-$MPS, como pode ser visto ao comparar a diferença entre as probabilidades da observação k obtidas com a família de distribuições $k-$MPS e as distribuições Série de Potência usuais (ver Conceição, Andrade e Louzada (2013)):

$$\begin{aligned}
\pi_{k-MPS}(k;\mu,p) - \pi_{PS}(k;\mu) &= (1-p) + p\pi_{PS}(k;\mu) - \pi_{PS}(k;\mu) \\
&= (1-p)\{1-\pi_{PS}(k;\mu)\}. \qquad (2.3.3)
\end{aligned}$$

Dessa forma, temos:

(i) Quando $p = 0$ em (2.3.3), temos que $\pi_{k-MPS}(k;\mu,p) = 1$. Portanto, $\pi_{k-MPS}(y;\mu,p)$ é uma distribuição degenerada com toda massa no ponto k.

(ii) Para todo $0 < p < 1$ em (2.3.3), temos que $\pi_{k-MPS}(k;\mu,p) > \pi_{PS}(k;\mu)$. Logo, $\pi_{k-MPS}(y;\mu,p)$ é uma distribuição Série de Potência k-Inflacionada ($k-$IPS[4]), a qual tem uma propoção adicional de observações k.

(iii) Quando $p = 1$ em (2.3.3), temos que $\pi_{k-MPS}(k;\mu,p) = \pi_{PS}(k;\mu)$. Portanto, $\pi_{k-MPS}(y;\mu,p)$ é uma distribuição PS usual.

(iv) Para todo $1 < p < \frac{1}{1-\pi_{PS}(k;\mu)}$ em (2.3.3), temos que $\pi_{k-MPS}(k;\mu,p) < \pi_{PS}(k;\mu)$. Logo, $\pi_{k-MPS}(y;\mu,p)$ é uma distribuição Série de Potência k-Deflacionada ($k-$DPS[5]).

(v) Quando $p = \frac{1}{1-\pi_{PS}(k;\mu)}$ em (2.3.3), temos que $\pi_{k-MPS}(k;\mu,p) = 0$. Portanto, $\pi_{k-MPS}(y;\mu,p)$ é uma distribuição $k-$SPS, com função massa de probabilidade dada em (2.1.1).

Para melhor compreensão dos casos particulares (i) - (v) da distribuição $k-$MPS, a Figura 2 ilustra um diagrama com as informações sintetizadas.

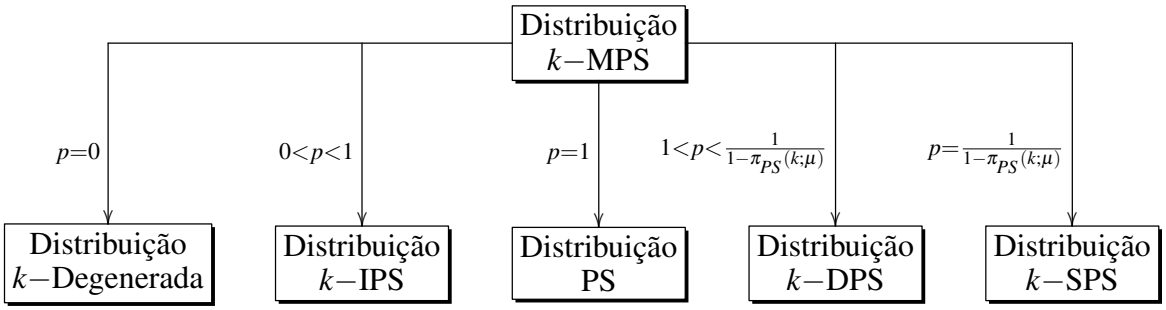

Figura 2 – Diagrama dos casos particulares da distribuição $k-$MPS.

Fonte: Carvalho (2017).

2.3.2 Coeficiente de Dispersão

Conceição (2013) definiu em seu trabalho o coeficiente de dispersão d para as distribuições ZMPS. Estendemos aqui esse conceito para as distribuições k-MPS.

O coeficiente de dispersão é uma medida de agregação ou desagregação (CONCEIÇÃO; ANDRADE; LOUZADA, 2013), responsavél por fornecer o grau de dispersão existente na distribuição. Ele é dado pela razão entre a variância da variável aleatória e sua média, isto é,

[4] Usaremos a sigla k-IPS, do inglês *k-Inflated Power Series*, referindo-se às distribuições da família Série de Potência k-Inflacionada.

[5] Usaremos a sigla k-DPS, do inglês *k-Deflated Power Series*, referindo-se às distribuições da família Série de Potência k-Deflacionadas.

$d(Y) = Var(Y)/E(Y)$. Desta forma, o coeficiente de dispersão da família de distribuições k-MPS é dado por

$$\begin{aligned} d_{k-MPS} &= \frac{\sigma^2_{k-MPS}}{\mu_{k-MPS}} \\ &= \frac{(k-\mu)^2(1-p)+\sigma^2}{\frac{k}{p}-k+\mu}. \end{aligned} \tag{2.3.4}$$

Da Equação (2.3.4) temos que:

i) Se $d_{k-MPS} = 1$, tem-se que $\sigma^2_{k-MPS} = \mu_{k-MPS}$, e portanto a distribuição k-MPS é equidispersa;

ii) Se $d_{k-MPS} > 1$, tem-se que $\sigma^2_{k-MPS} > \mu_{k-MPS}$, e portanto a distribuição k-MPS é sobredispersa;

iii) Se $d_{k-MPS} < 1$ a distribuição k-MPS é subdispersa pois $\sigma^2_{k-MPS} < \mu_{k-MPS}$.

É importante enfatizar que, embora estejamos definindo o coeficiente de dispersão para a família k-MPS, esse conceito pode ser estendido para qualquer distribuição discreta k-modificada.

2.3.3 Distribuição $k-$MPS e sua versão Hurdle

A função $\pi_{k-MPS}(y;\mu,p)$, dada pela Equação (2.3.1), pode ser reescrita da seguinte forma:

$$\begin{aligned} \pi_{k-MPS}(y;\mu,p) &= \pi_{k-MPS}(k;\mu,p)I_{\{k\}}(y) + \pi_{k-MPS}(y;\mu,p)(1-I_{\{k\}}(y)) \\ &= \left\{1-p(1-\pi_{PS}(k;\mu))\right\}I_{\{k\}}(y) \\ &+ p(1-\pi_{PS}(k;\mu))\frac{\pi_{PS}(y;\mu)}{1-\pi_{PS}(k;\mu)}\left\{1-I_{\{k\}}(y)\right\}. \end{aligned} \tag{2.3.5}$$

Sabendo que $\dfrac{\pi_{PS}(y;\mu)}{1-\pi_{PS}(k;\mu)}\left\{1-I_{\{k\}}(y)\right\}$ é a distribuição $k-$SPS e definindo ω por $\omega = p(1-\pi_{PS}(k;\mu))$, $\omega \in \Omega \subseteq [0,1]$, temos a versão "barreira" da distribuição $k-$MPS, que neste trabalho será chamada de versão *Hurdle* (ver Dalrymple, Hudson e Ford (2003)) da distribuição $k-$MPS, cuja função massa de probabilidade é:

$$\pi_{k-MPS}(y;\mu,\omega) = (1-\omega)I_{\{k\}}(y) + \omega\pi_{k-SPS}(y;\mu),\ \forall\, y \in \mathscr{A}_s, \tag{2.3.6}$$

tal que $0 \leq \omega \leq 1$.

A distribuição $k-$MPS parametrizada por ω será denotada por $k-MPS(\mu,\omega)$. Vale mencionar que, apesar do parâmetro $\omega \in [0,1]$, a distribuição $k-$modificada continua representando $k-$inflação e $k-$deflação.

É fácil verificar na versão *Hurdle* da distribuição que a probabilidade de ocorrência do evento $Y = k$ é dado por $1-\omega$ e a probabilidade do evento $Y \neq k$, é dada por $\omega\pi_{k-SPS}(y;\mu)$.

Consequentemente, podemos interpretar a distribuição $k-MPS(\mu,\omega)$ como a superposição de dois processos, sendo que um explicará apenas as observações k e o outro processo explicará as observações diferentes de k, baseando-se na distribuição $k-$SPS. Ou seja, a distribuição $k-MPS(\mu,\omega)$ pode ser vista como uma distribuição de mistura, o que é frequentemente proposto ao analisar conjuntos de dados com inflação em algum ponto $k \in \mathscr{A}_s$, como pode ser visto no artigo de Murat e Szynal (1998).

A média e variância da variável aleatória Y com distribuição $k-MPS(\mu,\omega)$ são, respectivamente:

$$\begin{aligned}\mu_{k-MPS} &= k\left(1-\frac{\omega}{(1-\pi_{PS}(k;\mu))}\right)+\frac{\mu\omega}{(1-\pi_{PS}(k;\mu))}\\ &= (1-\omega)k+\omega\mu_{k-SPS}\end{aligned}$$

e

$$\sigma^2_{k-MPS} = k^2+\frac{\omega(\sigma^2-\mu^2)}{1-\pi_{PS}(k;\mu)}-((1-\omega)k+\omega\mu_{K-SPS})^2,$$

em que μ e σ^2 são a média e variância da distribuição Série de Potência usual.

2.3.3.1 Casos Particulares da Distribuição $k-MPS(\mu,\omega)$

Assim como na distribuição $k-MPS(\mu,p)$, a versão *Hurdle* da distribuição também comporta os casos particulares apresentados na Subseção 2.3.1, ao considerar diferentes valores de ω. De maneira similar, avaliaremos a diferença entre as probabilidades da observação k obtidas com a distribuição $k-MPS(\mu,\omega)$ e a distribuição PS usual:

$$\pi_{k-MPS}(k;\mu,\omega)-\pi_{PS}(k;\mu) = 1-\omega-\pi_{PS}(k;\mu). \tag{2.3.7}$$

Ao avaliar a Equação (2.3.7) considerando diferentes valores de ω, temos os seguintes casos particulares:

(i) Quando $\omega=0$ em (2.3.7), temos que $\pi_{k-MPS}(k;\mu,\omega)=1$. Portanto, $\pi_{k-MPS}(y;\mu,\omega)$ é uma distribuição degenerada com toda massa no ponto k.

(ii) Para todo $0<\omega<1-\pi_{PS}(k;\mu)$ em (2.3.7), temos que $\pi_{k-MPS}(k;\mu,\omega)>\pi_{PS}(k;\mu)$. Logo, $\pi_{k-MPS}(y;\mu,\omega)$ é uma distribuição $k-$IPS.

(iii) Quando $\omega=1-\pi_{PS}(k;\mu)$ em (2.3.7), temos que $\pi_{k-MPS}(k;\mu,\omega)=\pi_{PS}(k;\mu)$. Portanto, $\pi_{k-MPS}(y;\mu,\omega)$ é uma distribuição PS usual.

(iv) Para todo $1-\pi_{PS}(k;\mu)<\omega<1$ em (2.3.7), temos que $\pi_{k-MPS}(k;\mu,\omega)<\pi_{PS}(k;\mu)$. Logo, $\pi_{k-MPS}(y;\mu,\omega)$ é uma distribuição $k-$DPS.

(v) Quando $\omega=1$ em (2.3.7), temos que $\pi_{k-MPS}(k;\mu,\omega)=0$. Portanto, $\pi_{k-MPS}(y;\mu,\omega)$ é uma distribuição $k-$SPS.

2.3.4 Estimação dos Parâmetros

Apresentamos a seguir os procedimentos de estimação dos parâmetros da distribuição $k-$MPS. Para obtermos as estimativas pontuais dos parâmetros da distribuição, consideramos o método de máxima verossimilhança e uma abordagem Bayesiana.

2.3.4.1 Método de Máxima Verossimilhança

Para a estimação dos parâmetros da distribuição $k-$MPS via máxima verossimilhança, vamos considerar as duas parametrizações estudadas nas Seções anteriores: a parametrizada por p e a parametrizada por ω. Sendo Y uma variável aleatória com distribuição $k-$MPS, considere:

- $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ uma amostra aleatória de Y. Isto é, n realizações independentes e identicamente distribuídas a Y.
- $\boldsymbol{y} = (y_1\ y_2\ \ldots\ y_n)$ o vetor de observações associado a amostra aleatória $\boldsymbol{Y}$.
- n_j, $j = 0,1,2,\ldots$, que corresponde o número de observações j no vetor $\boldsymbol{y}$, tal que $n = \sum_{j=0}^{\infty} n_j$, o número total de observações.

a) Método de Máxima Verossimilhança para $k - MPS(\mu, p)$

A função de verossimilhança associada ao vetor de observações $\boldsymbol{y}$ de n realizações independente da variável aleatória Y com distribuição $k - MPS(\mu, p)$ é dada por:

$$\begin{aligned} L(\mu,p) &= \prod_{i=1}^{n} \left\{(1-p)I_{\{k\}}(y_i) + p\pi_{PS}(y_i;\mu)\right\} \\ &= \prod_{i=1}^{n} \left\{(1-p+p\pi_{PS}(k;\mu))^{I_{\{k\}}(y_i)} (p\pi_{PS}(y_i;\mu))^{1-I_{\{k\}}(y_i)}\right\} \\ &= \left(1-p(1-\pi_{PS}(k;\mu))\right)^{n_k} \prod_{j:j\in A_0 | j\neq k}^{\infty} \left\{(p\pi_P(j;\mu))^{n_j}\right\}. \end{aligned}$$

O logaritmo natural da função de verossimilhança é:

$$\ell(\mu,p) = n_k \log(1-p(1-\pi_{PS}(k;\mu))) + \sum_{j:j\in A_0 | j\neq k}^{\infty} n_j \log(p\pi_{PS}(j;\mu)).$$

Derivando $\ell(\mu,p)$ em relação a cada parâmetro, obtemos o vetor escore $U = (U_p, U_\mu)$, cujos elementos são:

$$U_p = \frac{\partial \ell(\mu,p)}{\partial p} = \frac{-n_k(1-\pi_{PS}(k;\mu))}{1-p(1-\pi_{PS}(k;\mu))} + \frac{n-n_k}{p}$$

e

$$U_\mu = \frac{\partial \ell(\mu,p)}{\partial \mu} = \frac{n_k p\left(\frac{\partial}{\partial \mu}\pi_{PS}(k;\mu)\right)}{1-p(1-\pi_{PS}(k;\mu))} + \sum_{j:j\in A_0|j\neq k}^{\infty} \frac{n_j\left(\frac{\partial}{\partial \mu}\pi_{PS}(j;\mu)\right)}{\pi_{PS}(j;\mu)}. \quad (2.3.8)$$

Para encontrar o estimador de máxima verossimilhança (MV) de p e de μ basta igualar U_p e U_μ a zero. Apenas para o parâmetro p conseguimos obter o estimador explicitamente:

$$\hat{p} = \frac{n-n_k}{n(1-\pi_{PS}(k;\hat{\mu}))}.$$

Para obter o estimador de MV de μ, podemos substituir p por $\hat{p}$ na Equação (2.3.8) para simplificações. Dessa forma, a função U_μ é reduzida a:

$$\begin{aligned}\frac{\partial \ell(\mu,p)}{\partial \mu} &= \frac{(n-n_k)\left(\frac{\partial}{\partial \mu}\pi_{PS}(k;\mu)\right)}{1-\pi_{PS}(k;\mu)} + \sum_{j:j\in A_0|j\neq k}^{\infty} \frac{n_j\left(\frac{\partial}{\partial \mu}\pi_{PS}(j;\mu)\right)}{\pi_{PS}(j;\mu)} \\ &= -\left(\frac{\partial}{\partial \mu}\log(f(\mu))\right)\left[\frac{(n-n_k)f(\mu)}{f(\mu)-1}\right] + \left(\frac{\partial}{\partial \mu}\log(g(\mu))\right)\sum_{j:j\in A_0|j\neq k}^{\infty} jn_j,\end{aligned}$$

que ao igualar a zero

$$-\left(\frac{\partial}{\partial \mu}\log(f(\mu))\right)\left[\frac{(n-n_k)f(\mu)}{f(\mu)-1}\right] + \left(\frac{\partial}{\partial \mu}\log(g(\mu))\right)\sum_{j:j\in A_0|j\neq k}^{\infty} jn_j = 0$$

e realizar algumas manipulações algébricas, chegamos na equação simplificada

$$\bar{y}_{\{-k\}} = \frac{\mu f(\mu)}{f(\mu)-1}, \quad (2.3.9)$$

em que $\bar{y}_{\{-k\}} = \frac{1}{(n-n_k)}\sum_{j:j\in A_0|j\neq k}^{\infty} jn_j$ corresponde à média aritmética das observações de $\boldsymbol{y}$ diferentes de k.

b) Método de Máxima Verossimilhança para $k-MPS(\mu,\omega)$

Para a variável aleatória Y com distribuição $k-MPS(\mu,\omega)$, a função de verossimilhança associada ao vetor de observações $\boldsymbol{y}$ é:

$$\begin{aligned}L(\mu,\omega) &= \prod_{i=1}^{n}\left\{(1-\omega)I_{\{k\}}(y_i) + \omega\pi_{k-SPS}(y_i;\mu)\right\} \\ &= \prod_{i=1}^{n}\left\{(1-\omega)^{I_{\{k\}}(y_i)}(\omega\pi_{k-SPS}(y_i;\mu))^{1-I_{\{k\}}(y_i)}\right\} \\ &= (1-\omega)^{n_k}\omega^{(n-n_k)}\prod_{j:j\in A_0|j\neq k}^{\infty}\left\{\left(\frac{\pi_{PS}(j;\mu)}{1-\pi_{PS}(k;\mu)}\right)^{n_j}\right\}.\end{aligned}$$

Sendo assim, o logaritmo natural da função de verossimilhança é dado por:

$$\begin{aligned}\ell(\mu,\omega) &= n_k\log(1-\omega)+(n-n_k)\log(\omega)+\sum_{j:j\in A_0|j\neq k}^{\infty} n_j\log\left(\frac{\pi_{PS}(j;\mu)}{1-\pi_{PS}(k;\mu)}\right)\\ &= \sum_{j:j\in A_0|j\neq k}^{\infty} n_j\log(\pi_{PS}(j;\mu))-(n-n_k)\log(1-\pi_{PS}(k;\mu))+\\ &\quad n_k\log(1-\omega)+(n-n_k)\log(\omega)\\ &= \ell_1(\mu)+\ell_2(\omega).\end{aligned}$$

Com a Equação acima, podemos notar que ω e μ são ortogonais, pois podemos fatorar o termo $\ell(\mu,\omega)$ em dois termos:

$$\ell_1(\mu)=\sum_{j:j\in A_0|j\neq k}^{\infty} n_j\log(\pi_{PS}(j;\mu))-(n-n_k)\log(1-\pi_{PS}(k;\mu))$$

e

$$\ell_2(\omega)=n_k\log(1-\omega)+(n-n_k)\log(\omega),$$

tais que $\ell_1(\mu)$ não depende de ω e $\ell_2(\omega)$ não depende de μ. Os elementos do vetor escore, $\boldsymbol{U}=(U_\omega,U_\mu)$, são dados por:

$$U_\omega=\frac{\partial\ell_2(\omega)}{\partial\omega} = \frac{-n_k}{1-\omega}+\frac{n-n_k}{\omega}$$

e

$$\begin{aligned}U_\mu=\frac{\partial\ell_1(\mu)}{\partial\mu} &= \sum_{j:j\in A_0|j\neq k}^{\infty} n_j\left(\frac{\partial}{\partial\mu}\log(\pi_{PS}(j;\mu))\right)-(n-n_k)\left(\frac{\partial}{\partial\mu}\log(1-\pi_{PS}(k;\mu))\right)\\ &= \sum_{j:j\in A_0|j\neq k}^{\infty}\frac{n_j\left(\frac{\partial}{\partial\mu}\pi_{PS}(j;\mu)\right)}{\pi_{PS}(j;\mu)}+\frac{(n-n_k)\left(\frac{\partial}{\partial\mu}\pi_{PS}(k;\mu)\right)}{(1-\pi_{PS}(k;\mu))}.\end{aligned}$$

Então podemos obter o estimador de MV de ω e de μ igualando, respectivamente, U_ω e U_μ a zero. Assim, também como no caso anterior, apenas o parâmetro de modificação, ω, encontramos explicitamente o estimador:

$$\hat{\omega}=\frac{n-n_k}{n}.$$

Na estimação do parâmetro μ, obtemos diretamente a mesma equação com a parametrização em p:

$$\bar{y}_{\{-k\}}=\frac{\mu f(\mu)}{f(\mu)-1}.$$

Um ponto que deve ser ressaltado nos procedimentos de estimação para os parâmetros p e ω é: a estimativa de p depende da estimativa de μ, enquanto a estimativa do parâmetro ω depende somente de informações sobre as observações da amostra. Desta forma, é mais conveniente usarmos a parametrização da distribuição na versão *Hurdle* de forma a obter o estimador de MV de p dado por:

$$\hat{p} = \frac{\hat{\omega}}{(1-\pi_{PS}(k;\hat{\mu}))}, \tag{2.3.10}$$

o qual fornece diretamente o caso particular da distribuição.

2.3.4.2 Abordagem Bayesiana

Para a estimação dos parâmetros da distribuição k-MPS via abordagem Bayesiana, consideramos apenas a parametrização na versão *Hurdle*, já que obtendo a estimativa para ω é possível estimar p, a partir da relação $p = \dfrac{\omega}{(1-\pi_{PS}(k;\mu))}$.

a) Abordagem Bayesiana para $k-MPS(\mu,\omega)$

Consideremos novamente Y uma variável aleatória com distribuição $k-$MPS, de forma que $\boldsymbol{Y}$ seja uma amostra aleatória de Y (n realizações independentes e identicamente distribuídas de Y) e $\boldsymbol{y}$ seja um vetor de observações associado a $\boldsymbol{Y}$.

Sejam μ e ω, independentes, os parâmetros da distribuição k-MPS. Iremos considerar para μ uma distribuição *a priori* Gama com hiperparâmetros $a>0$ e $b>0$ ($\mu \sim \text{Gama}(a,b)$) e para ω iremos considerar uma distribuição *a priori* Beta com hiperparâmetros $c>0$ e $d>0$ ($\omega \sim \text{Beta}(c,d)$), uma vez que estas distribuições satisfazem os suportes de μ e ω. A *priori* conjunta é dada por $\pi(\mu,\omega) = \pi(\mu)\pi(\omega)$. Os valores considerados para os hiperparâmetros foram escolhidos de tal forma que resultassem em *priori* vagas para μ e ω. Desta forma, a densidade a *posteriori* conjunta é

$$\pi(\mu,\omega|\boldsymbol{y}) \propto L(\mu,\omega)\pi(\mu,\omega).$$

Nota: Para a distribuição Binomial, $0<\mu<m$. Portanto, ao considerar a distribuição Binomial k-Modificada, vamos supor como *priori* vaga para μ a distribuição Uniforme, isto é, $\mu \sim U(0,m)$.

Do ponto de vista Bayesiano, inferências sobre os parâmetros podem ser baseadas nas suas densidades marginais a *posteriori*, o que pode ser obtido ao integrar a densidade conjunta a *posteriori*. Neste caso, para ω temos por hipótese que

$$\pi(\omega) \propto \omega^{c-1}(1-\omega)^{d-1},$$

e portanto,

$$\begin{aligned} \pi(\omega|\boldsymbol{y}) &\propto \exp\{\ell_2(\omega)\}\pi(\omega) \\ &\propto (1-\omega)^{n_k}\omega^{n-n_k}\omega^{c-1}(1-\omega)^{d-1} \\ &\propto \omega^{c+n-n_k-1}(1-\omega)^{b+n_k-1}, \end{aligned} \tag{2.3.11}$$

isto é, de (2.3.11) concluimos que a *posteriori* de ω tem distribuição Beta, logo $\omega|\boldsymbol{y} \sim \text{Beta}(c+n-n_k, b+n_k)$. Já para o parâmetro μ, temos, por hipótese

$$\pi(\mu) \propto \mu^{a-1}e^{-b\mu}, \quad \mu > 0,$$

e assim,

$$\begin{aligned} \pi(\mu|\boldsymbol{y}) &\propto \exp\{\ell_1(\mu)\}\pi(\mu) \\ &\propto \prod_{j:j\in A_0|j\neq k}^{\infty} \left\{ \left(\frac{\pi_{PS}(j;\mu)}{1-\pi_{PS}(k;\mu)} \right)^{n_j} \right\} \mu^{a-1}e^{-b\mu}. \end{aligned} \tag{2.3.12}$$

Uma vez que a distribuição a *posteriori* (2.3.12) não tem uma forma padrão, uma possibilidade para a estimação do parâmetro é utilizar o algoritmo de Metropolis-Hastings (CHIB; GREENBERG, 1995), (HASTINGS, 1970), que é um procedimento iterativo de uma larga classe de métodos Monte Carlo em Cadeia de Markov (MCMC[6]).

[6] Usaremos a sigla MCMC, do inglês *Monte Carlo Markov Chain*, referindo-se a Monte Carlo em Cadeia de Markov.

CAPÍTULO

3

APLICAÇÕES: DISTRIBUIÇÃO k-MPS

A seguir apresentamos aplicações das distribuição da família k-MPS. As aplicações consideradas foram retiradas de Carvalho (2017), o qual analisou cada conjunto de dados considerando a abordagem de máxima verossimilhança. Elas são baseadas em conjuntos de dados reais e estão relacionadas aos temas: futebol, variação climática e linguística. Para todos os exemplos, foi considerada uma abordagem Bayesiana para a estimação dos parâmetros, dado que as estimativas clássicas estão disponíveis no trabalho de Carvalho (2017).

3.1 Números de Gols do Barcelona em Confrontos com o Real Madrid

Esta aplicação consiste em analisar o conjunto de dados formado pelo número de gols marcados pelo time do Barcelona em todos os jogos contra o Real Madrid entre 1955 e 2015. Neste período foram realizadas 131 partidas, das quais foram 68 vitórias do Barcelona, 62 vitórias do Real Madrid e 1 empate. A distribuição de frequência do número de gols marcados pelo Barcelona em cada partida, a média amostral ($\bar{y}$) e o desvio-padrão (dp) estão apresentados na Tabela 4.

Tabela 4 – Distribuição de frequência e estatísticas descritivas do número de gols marcados pelo Barcelona em confrontos com o Real Madrid entre 1955 e 2015.

y_i	0	1	2	3	4	5	6	$\bar{y}$	dp	Total
f_i	27	39	34	21	5	3	2	1.656	1.335	131

Como podemos notar, a média amostral para o conjunto de dados descrito acima é baixa e esperamos altas frequências de valores baixos, sob a suposição de uma distribuição de Poisson. Diante do desconhecimento da distribuição mais adequada para explicar as frequências observadas, consideramos a distribuição Poisson k-Modificada (k-MP) nos pontos $k = 0$, 1 e

2 por ser uma distribuição mais geral, a qual tem como caso particular a distribuição Poisson. As estimativas Bayesianas dos parâmetros da distribuição e seus respectivos intervalos de credibilidade com 95% estão apresentados na Tabela 5.

Tabela 5 – Estimativas Bayesianas dos parâmetros da distribuição Poisson k-Modificada e seus respectivos intervalos de credibilidade de 95%, considerando os pontos de modificação k = 0, 1 e 2.

k	$\hat{\mu}$	$\hat{p}$	$\hat{\omega}$	$\hat{n}_k$	
				k-M	P
0	1.700 (1.429; 1.996)	0.966 (0.870; 1.045)	0.789 (0.715; 0.854)	3	24
1	1.636 (1.420; 1.879)	1.026 (0.895; 1.124)	0.699 (0.621; 0.773)	-3	42
2	1.652 (1.440; 1.878)	0.998 (0.898; 1.087)	0.737 (0.658; 0.810)	0	34

Analisando os resultados apresentados na Tabela 5, considerando o ponto de modificação $k = 0$, temos que a estimativa do parâmetro p foi $\hat{p} = 0.966$ e seu respectivo intervalo de credibilidade contém o valor 1 e, portanto, podemos concluir que os dados podem ser explicados por uma distribuição Poisson. Vale ressaltar que, considerando a distribuição Poisson com $\hat{\mu} = 1.700$, a quantidade esperada de zero é de 24 observações, isto é, $n\pi_P(0;\hat{\mu}) = 131 * 0.183 \simeq 24$. Desta forma, teríamos 3 zeros provenientes do processo de modificação no zero (0-inflação, já que $p < 1$), resultando nos 27 zeros observados no conjunto de dados, os quais não foram suficientes para ocasionar uma modificação significativa na frequência desta observação (zero-inflação).

De maneira análoga, concluímos também que, para $k = 1$ e $k = 2$ as estimativas de p são bem próximas a 1, como pode ser visto na Tabela 5, e os respectivos intervalos de credibilidade também contém o valor 1, e portanto os dados referentes ao número de gols marcados pelo Barcelona em confrontos com o Real Madrid entre 1955 e 2015 podem ser explicados por uma distribuição Poisson tradicional. Ainda nesta Tabela é possível verificar que, considerando a distribuição Poisson com $\hat{\mu} = 1.636$, a quantidade esperada de um é igual a 42 e o processo de modificação no um (1-deflação, pois $p > 1$) é responsável por remover 3 observações um; já a distribuição de Poisson com com $\hat{\mu} = 1.652$ é responsável por todos as 34 valores dois observados no conjunto de dados.

A partir destas análises, podemos notar uma total concordância dos resultados das distribuições k-MP ($k = 0$, 1 e 2) ajustadas, os quais apontaram a inexistência de qualquer modificação nas frequências das observações, indicando a distribuição Poisson para explicar o comportamento do conjunto de dados em todos os pontos de k-Modificação considerados.

A Figura 3 apresenta os gráficos comparativos das frequências observadas com as frequências esperadas segundo as distribuições Poisson e k-MP com $k = 0$, 1 e 2. A adequação das distribuições ajustadas foi obtida através dos testes de aderência Qui-Quadrado (χ^2) e Kolmogorov-Smirnov (KS) (para mais detalhes, ver Conover (1999)), cujos resultados estão

apresentados na Tabela 6. É possível observar nesta Tabela que os valores obtidos das estatísticas de teste para as distribuições k-MP ($k = 0$, 1 e 2) ajustadas são menores do que os respectivos valores críticos ao utilizar um nível de significância de 5%. Consequentemente, para cada distribuição, a hipótese nula H_0 não é rejeitada, concluindo que a distribuição Poisson é adequada para explicar o comportamento dos dados. Vale ressaltar que podemos chegar nesta conclusão ao observar que os intervalos de credibilidade de p apresentados na Tabela 5 contém o valor 1.

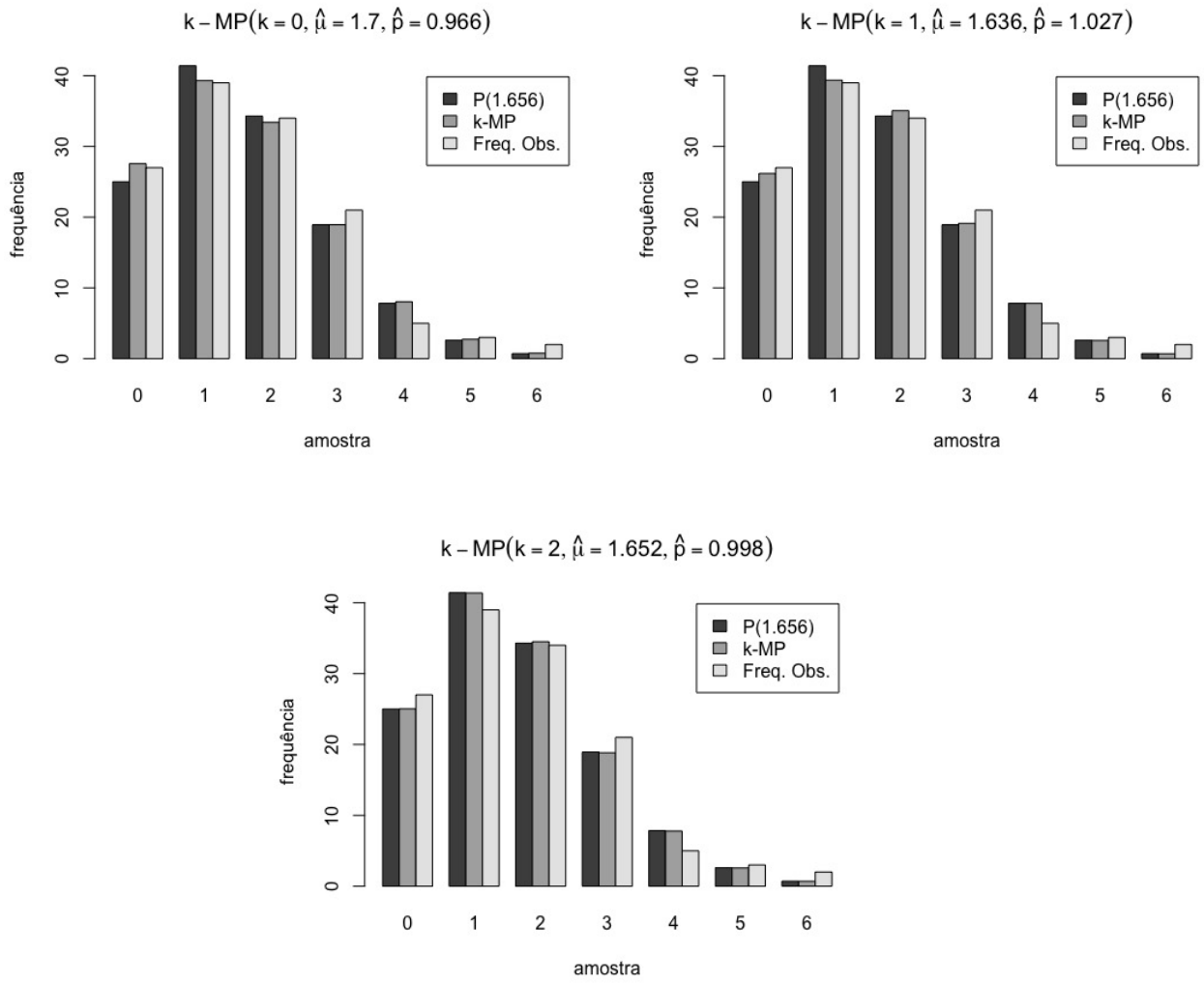

Figura 3 – Gráficos comparativos entre as frequências observadas dos dados e as frequências esperadas segundo as distribuições Poisson e k-MP, considerando $k = 0$, 1 e 2.

Fonte: Elaborada pelo autor.

Tabela 6 – Resultados dos testes de aderência Qui-Quadrado e Kolmogorov-Smirnov considerando a distribuição k-MP ajustada, com $k = 0$, 1 e 2.

Teste	k	Estatística de Teste	Valor Crítico
χ^2	0	0.473	9.488
	1	0.358	
	2	0.656	
KS	0	0.014	0.119
	1	0.012	
	2	0.015	

3.2 Variação da Temperatura Global

O aquecimento global tem sido um tema amplamente debatido e de interesse de estudo por vários pesquisadores devido ao seu grande impacto ambiental. Por esse motivo, a análise de um conjunto de dados referente a variação anual da temperatura global foi considerado neste trabalho. O conjunto de dados correspondente às variações anuais da temperatura média global entre 1958 e 2008 está associado a um estudo prévio realizado em 1978 pelos pesquisadores Angell e Korshover, que publicaram uma análise da variação da temperatura média global (em graus Celsius) entre os anos de 1958 e 1977 (ANGELL; KORSHOVER, 1978).

Para este estudo, as variações anuais da temperatura média global foram obtidas considerando o desvio entre as temperaturas médias anuais (de 1958 a 2008) e a média das temperaturas anuais dos primeiros 20 anos considerados no estudo (1958 a 1977). A Figura 4 apresenta as variações anuais da temperatura média global entre 1958 e 2008, onde é possível notar oscilações entre os anos que em sua maioria são variações positivas.

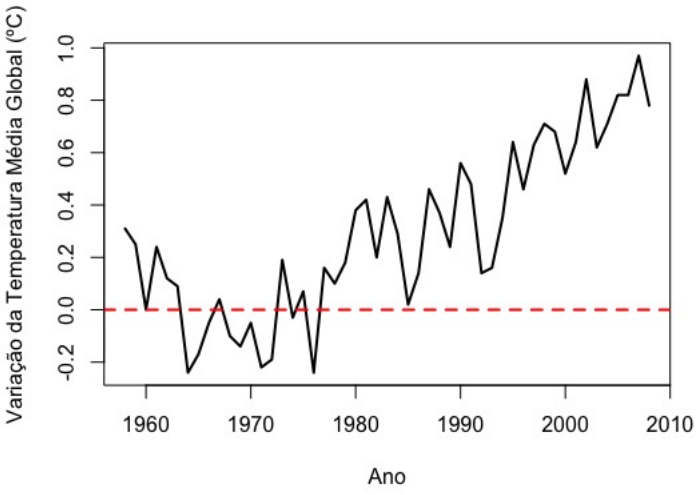

Figura 4 – Variações anuais (em ° C) da temperatura média global entre os anos de 1958 e 2008.

A princípio, consideramos então o experimento que consiste em verificar a ocorrência de

uma variação positiva (sucesso - aumento relativo da temperatura média global) ou uma variação negativa (fracasso - diminuição relativa da temperatura média global) em um determinado ano. Denotando por λ $(0 < \lambda < 1)$ o parâmetro associado ao processo de Bernoulli, podemos interpretá-lo como sendo a probabilidade de observar uma variação positiva da temperatura média global em um determinado ano (probabilidade de sucesso).

A partir do processo de Bernoulli, definimos uma variável aleatória Y que representa a contagem de anos consecutivos com variações negativas da temperatura média global (fracassos) até a ocorrência de uma variação positiva da temperatura (sucesso), o que define um processo geométrico. Vale ressaltar que ao considerar a distribuição geométrica parametrizada na média μ, temos que $\lambda = 1/(1+\mu)$. A Tabela 7 apresenta a distribuição de frequência dos dados observados das realizações independentes de Y e as estatísticas resumo (média e desvio-padrão). Podemos notar uma alta frequência da observação zero, dando indícios de zero-inflação.

Tabela 7 – Distribuição de frequência e estatísticas descritivas do número de anos consecutivos de variação negativa da temperatura média global até a ocorrência de uma variação positiva entre 1958 e 2008.

y_i	0	1	3	5	$\bar{y}$	dp	Total
f_i	34	3	1	1	0.282	0.944	39

Ajustamos a distribuição Geométrica k-Modificada (k-MG) aos dados considerando apenas $k = 0$. Na Tabela 8 apresentamos as estimativas Bayesianas dos parâmetros e seus respectivos intervalos com 95% de credibilidade. Analisando os resultados da Tabela, temos que a estimativa do parâmetro de modificação p foi $\hat{p} = 0.263$ e o intervalo de credibilidade não contém o valor 1. Portanto, podemos concluir que os dados são explicados por uma distribuição Geométrica 0-Inflacionada. Vale ressaltar que, considerando a distribuição Geométrica com $\hat{\mu} = 1.253$, o número esperado de zero é de 17 observações, isto é, $n\pi_G(0;\hat{\mu}) = 39 * 0.444 \simeq 17$. Desta forma, temos adicionalmente outros 17 zeros provenientes do processo de modificação no zero (0-inflação), totalizando em 34 zeros observados no conjunto de dados. A adição de zeros ocasinou uma mudança significativa na frequência desta observação, caracterizando o conjunto de dados como zero-inflacionado.

Tabela 8 – Estimativas Bayesianas dos parâmetros da distribuição Geométrica k-Modificada e os respectivos intervalos de credibilidade de 95%, considerando k = 0.

k	$\hat{\mu}$	$\hat{p}$	$\hat{\omega}$	$\hat{n}_k$	
				k-M	G
0	1.253 (0.001; 2.553)	0.262 (0.121; 0.421)	0.146 (0.040; 0.251)	17	17

A Figura 5 apresenta o gráfico comparativo das frequências observadas e as frequências esperadas segundo as distribuições Geométrica e 0-MG (ou, equivalentemente, 0-IG).

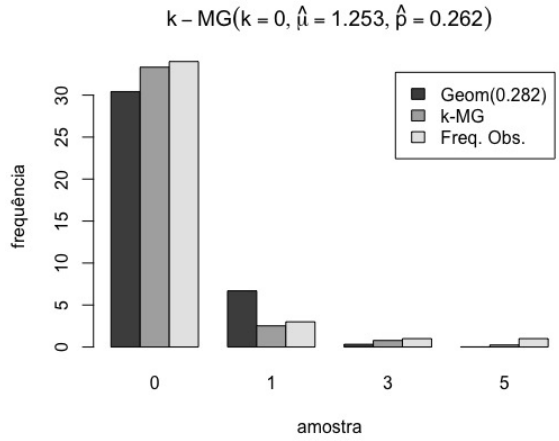

Figura 5 – Gráfico comparativo entre as frequências observadas dos dados e as frequências esperadas segundo as distribuições Geométricas e 0-MG (ou 0-IG).

Fonte: Elaborada pelo autor.

A Tabela 9 apresenta os resultados dos testes de aderência Qui-Quadrado e KS considerados para avaliar a adequabilidade da distribuição 0-MG (ou 0-IG) ajustada. Como os valores obtidos das estatísticas de teste considerando são inferiores aos valores críticos, ao utilizar um nível de significância de 5%, não temos evidências para rejeitar a hipótese nula H_0, isto é, há evidências de que os dados podem ser explicados adequadamente pela distribuição 0-MG (no caso, 0-IG).

Tabela 9 – Resultados dos testes de aderência Qui-Quadrado e KS considerando a distribuição 0-MG ajustada.

Teste	Estatística de Teste	Valor Crítico
χ^2	0.066	3.841
KS	0.030	0.218

3.3 Significado de Palavras Técnicas da Estatística

Esta aplicação consiste na análise do conjunto de dados referente ao experimento no qual foram selecionados 59 alunos de Graduação em Estatística em uma determinada Universidade com ao menos um ano de curso. Para verificar o conhecimento sobre o inglês técnico relacionado à Estatística, solicitaram-lhes a tradução em português das palavras *Average* e *Standard Deviation*.

Considerando que a tradução correta de cada palavra é um ensaio de Bernoulli com a mesma probabilidade de sucesso λ ($0 < \lambda < 1$), definimos Y a variável aleatória que representa o número de acertos obtidos por cada aluno. A Tabela 10 apresenta a distribuição de frequência e as estatísticas resumo (média e desvio-padrão) do número de acertos dos 59 alunos.

Tabela 10 – Distribuição de frequência e estatísticas descritivas do número de acertos na tradução de palavras técnicas de Estatística.

y_i	0	1	2	$\bar{y}$	sd	Total
f_i	18	15	26	1.135	0.860	59

Para este conjunto de dados, ajustamos a distribuição Binomial k-Modificada (k-MB) com $m = 2$ (m ensaios de Bernoulli independentes) e considerando os pontos de modificação $k = 0, 1$ e 2. As estimativas Bayesianas dos parâmetros e seus respectivos intervalos com 95% de credibilidade podem ser vistos na Tabela 11 a seguir.

Tabela 11 – Estimativas Bayesianas dos parâmetros da distribuição Binomial k-Modificada e os respectivos intervalos de credibilidade de 95%, com $m = 2$ e considerando $k = 0, 1$ e 2.

k	$\hat{\mu}$	$\hat{p}$	$\hat{\omega}$	$\hat{n}_k$	
				k-M	B
0	1.527 (1.303; 1.750)	0.731 (0.597; 0.851)	0.690 (0.575; 0.804)	15	3
1	1.092 (0.944; 1.241)	1.463 (1.281; 1.689)	0.738 (0.629; 0.846)	-14	29
2	0.617 (0.336; 0.897)	0.616 (0.474; 0.748)	0.557 (0.434; 0.681)	20	6

Com os resultados apresentados na Tabela acima, temos que, considerando o ponto de modificação $k = 0$, a estimativa de p é $\hat{p} = 0.731$ com intervalo de credibilidade contendo valores abaixo de 1, evidenciando que os dados podem ser explicados por uma distribuição Binomial 0-Inflacionada (0-IB), isto é, o conjunto de dados é caracterizado como inflacionado de zero. Similarmente, considerando a distribuição Binomial com $\hat{\mu} = 1.527$, o número esperado de zero é de apenas 3 observações ($n\pi_B(0;\hat{\mu}) = 59 * 0.055 \simeq 3$). Assim, temos adicionalmente 15 zeros provenientes do processo de modificação desta observação (0-inflação), totalizando em 18 zeros observados no conjunto de dados. Esta adição de observações zeros ocasionou uma mudança significativa na frequência desta observação, acarretando na caracterização do conjunto de dados como zero-inflacionado.

Por outro lado, quando o ponto de modificação $k = 1$ foi considerado, notamos que $\hat{p} = 1.463$ e o intervalo de credibilidade contém valores acima de 1. Portanto, neste caso, há evidências de que os dados são explicados por uma distribuição Binomial 1-Deflacionada (1-DB), isto é, o conjunto de dados é caracterizado como deflacionado de um. Neste contexto, o processo binomial com $\hat{\mu} = 1.092$ contribuiria para a ocorrência de 29 observações um ($n\pi_B(1;\hat{\mu}) = 59 * 0.496 \simeq 29$), porém o processo de modificação no um (1-deflação) removeu 14 observações, resultando em 15 observações um no conjunto de dados.

Finalmente, considerando o ponto de modificação $k = 2$ obtivemos $\hat{p} = 0.616$ e o intervalo de credibilidade contém valores abaixo de 1. Logo, também podemos considerar estes dados como sendo explicados por uma distribuição Binomial 2-Inflacionada (2-IB), sendo que o pro-

cesso Binomial com $\hat{\mu} = 0.617$ contribuiu para 6 observações dois ($n\pi_B(2;\hat{\mu}) = 59 * 0.095 \simeq 6$), enquanto que o processo de modificação no dois (2-inflação) acrescentou 20 destas observações no conjunto de dados.

A Figura 6 apresenta os gráficos comparativos das frequências observadas e as frequências esperadas segundo as distribuições Binomial e k-MB, considerando $k = 0, 1$ e 2. Novamente utilizamos os testes de aderência Qui-Quadrado e KS para verificar a adequabilidade das distribuições ajustadas e os resultados estão apresentados na Tabela 12. Podemos observar que os valores obtidos das estatísticas de teste para as distribuições k-MB ajustadas considerando os três pontos de modificação ($k = 0$, 1 e 2) são menores do que os respectivos valores críticos (nível de significância de 5%). Consequentemente, para cada distribuição, a hipótese nula H_0 não é rejeitada, concluindo que as três distribuições $k-$MB ajustadas são adequadas para explicar o comportamento dos dados. Este fato é reforçado principalmente ao verificar graficamente as proximidades das frequências observadas e esperadas para cada distribuição ajustada (Figura 6).

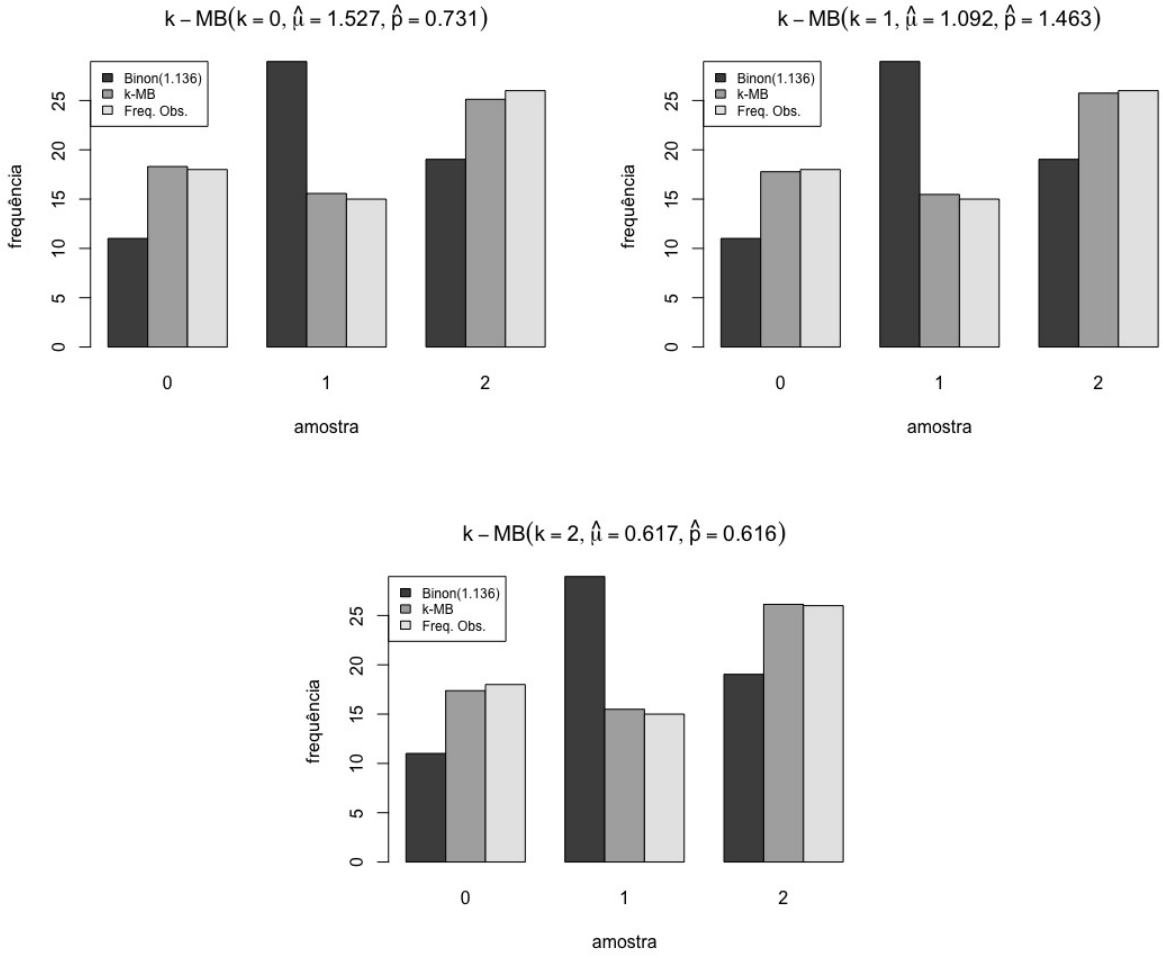

Figura 6 – Gráficos comparativos entre as frequências observadas dos dados e as frequências esperadas segundo as distribuições Binomial e k-MB, considerando $k = 0$, 1 e 2.

Fonte: Elaborada pelo autor.

Tabela 12 – Resultados dos testes de aderência Qui-Quadrado e KS considerando a distribuição k-MB, com $k = 0, 1$ e 2.

Teste	k	Estatística	Valor Crítico
χ^2	0	0.057	5.991
	1	0.020	
	2	0.039	
KS	0	0.015	0.177
	1	0.004	
	2	0.011	

Portanto, para os dados referentes ao significado de palavras técnicas da estatística, podemos concluir que os dados podem ser caracterizados de diversas formas, podendo ser inflacionados ou deflacionados, dependendo do ponto k que consideramos como ponto de modificação.

CAPÍTULO

4

MODELOS DE REGRESSÃO PARA A FAMÍLIA k-MPS

Apresentamos neste Capítulo a família de distribuições k-MD no contexto de modelos de regressão, mais especificamente o modelo $k-$MPS, permitindo assim modelar conjuntos de dados de contagem com modificação no ponto k em função de covariáveis.

4.1 O Modelo $k-$MPS

Considerando as distribuições discretas parametrizadas na média μ pertencentes à família k-MPS, introduzimos então os modelos de regressão k-MPS, uma extensão dos modelos ZMPS[1] propostos por Conceição (2013).

Definição 4. Considere os vetores $\boldsymbol{x}$ e $\boldsymbol{z}$ contendo respectivamente q_1 e q_2 covariáveis, $\boldsymbol{x} = (1\ x_1\ x_2\ \ldots\ x_{q_1})$ e $\boldsymbol{z} = (1\ z_1\ z_2\ \ldots\ z_{q_2})$. O modelo de regressão para dados discretos com distribuição pertencente à família $k-MPS(\mu, p)$ pode ser obtido escrevendo a função (2.3.1) da seguinte forma:

$$\pi_{k-MPS}(y;\mu(\boldsymbol{x}),p(\boldsymbol{z})) = (1-p(\boldsymbol{z}))I_{\{k\}}(y) + p(\boldsymbol{z})\pi_{PS}(y;\mu(\boldsymbol{x})), \tag{4.1.1}$$

em que a restrição de p apresentada em (2.3.2) é reescrita como:

$$0 \leq p(\boldsymbol{z}) \leq \frac{1}{1-\pi_{PS}(k;\mu(\boldsymbol{x}))}. \tag{4.1.2}$$

Definimos $h_\mu(\mu)$ e $h_p(p)$ as funções de ligação para μ e p, respectivamente, de tal forma que $\mu(\boldsymbol{x}) = h_\mu^{-1}(\boldsymbol{x}\boldsymbol{\beta}_1)$ e $p(\boldsymbol{z}) = h_p^{-1}(\boldsymbol{z}\boldsymbol{\beta}_2)$, sendo $\boldsymbol{\beta}_1^\top = (\beta_{10}\ \beta_{11}\ \ldots\ \beta_{1q_1})$ e $\boldsymbol{\beta}_2^\top = (\beta_{20}\ \beta_{21}\ \ldots\ \beta_{2q_2})$ os vetores de parâmetros dos respectivos preditores linear. A função de ligação considerada para

[1] Sigla definida por Conceição (2013), do inglês *Zero-Modified Power Series*, referindo-se aos distribuições Série de Potência Zero-Modificadas.

μ (que relaciona μ ao preditor linear) é $h_\mu(\mu) = \log(\mu) = \boldsymbol{x}\boldsymbol{\beta}_1$. Para a distribuição Binomial, em que $0 < \mu < m$, iremos considerar $h_\mu(\mu) = \log\left(\frac{\mu}{m-\mu}\right)$.

Um exemplo de função de ligação que relaciona p ao preditor linear é

$$h_p(p) = \log\left(\frac{p}{\frac{1}{1-\pi_{PS}(k;\mu(\boldsymbol{x}))} - p}\right).$$

Essa função de ligação para p garante que a restrição apresentada em (4.1.2) seja satisfeita.

De fato, considerando $h_p(p) = \boldsymbol{z}\boldsymbol{\beta}_2$, temos:

$$\log\left(\frac{p}{\frac{1}{1-\pi_{PS}(k;\mu(\boldsymbol{x}))} - p}\right) = \boldsymbol{z}\boldsymbol{\beta}_2.$$

Explicitando p, obtemos:

$$p(\boldsymbol{z}) = \left(\frac{e^{\boldsymbol{z}\boldsymbol{\beta}_2}}{1+e^{\boldsymbol{z}\boldsymbol{\beta}_2}}\right)\left(\frac{1}{1-\pi_{PS}(k;\mu(\boldsymbol{x}))}\right).$$

Denotando

$$\omega(\boldsymbol{z}) = \frac{e^{\boldsymbol{z}\boldsymbol{\beta}_2}}{1+e^{\boldsymbol{z}\boldsymbol{\beta}_2}},$$

temos que $0 < \omega(\boldsymbol{z}) < 1$. Sendo assim,

$$p(\boldsymbol{z}) = \frac{\omega(\boldsymbol{z})}{1-\pi_{PS}(k;\mu(\boldsymbol{x}))}.$$

Portanto, $0 < p(\boldsymbol{z}) < \frac{1}{1-\pi_{PS}(k;\mu(\boldsymbol{x}))}$.

Observação: A função de ligação $h_p(p)$ exclui dois casos específicos da variável aleatória Y: quando a variável tem distribuição degenerada em k e quando a variável tem distribuição $k-$SPS.

4.1.1 O Modelo k-MPS e sua versão Hurdle

Podemos escrever o modelo de regressão na versão *Hurdle* e, uma vez que $\omega(\boldsymbol{z}) = p(\boldsymbol{z})(1-\pi_{PS}(k;\mu(\boldsymbol{x}))$ de (2.3.5), temos:

$$\pi_{k-MPS}(y;\mu(\boldsymbol{x}),\omega(\boldsymbol{z})) = (1-\omega(\boldsymbol{z}))I_{\{k\}}(y) + \omega(\boldsymbol{z})\pi_{k-SPS}(y;\mu(\boldsymbol{x})), \tag{4.1.3}$$

em que $0 \leq \omega(\boldsymbol{z}) \leq 1$.

A partir da função de ligação $h_p(p)$ para p, obtivemos que a relação entre uma possível função de ligação $h_\omega(\omega)$ e o preditor linear $\boldsymbol{z}\boldsymbol{\beta}_2$ retorna

$$\omega(\boldsymbol{z}) = \frac{e^{\boldsymbol{z}\boldsymbol{\beta}_2}}{1+e^{\boldsymbol{z}\boldsymbol{\beta}_2}}, \tag{4.1.4}$$

garantindo que $0 < \omega(\boldsymbol{z}) < 1$. Ou seja, para este parâmetro a função de ligação considerada foi a Logito:

$$h_\omega(\omega) = \log\left(\frac{\omega}{1-\omega}\right).$$

Considerando a distribuição k-MPS parametrizada em ω, há uma grande variedade de funções de ligação para serem escolhidas, por exemplo, a Logito, Complemento Log-log ou Gumbel. Porém, ao considerar diferentes funções de ligação para ω resulta em diferentes funções de ligação para p, como pode ser visto na Tabela 13.

Tabela 13 – Algumas funções de ligação para o parâmetro de modificação.

$h_\omega(\omega)$	$h_p(p)$
$\log\left(\frac{\omega(\boldsymbol{z})}{1-\omega(\boldsymbol{z})}\right)$: Logito	$\log\left(\frac{p(\boldsymbol{z})}{\frac{1}{1-\pi_{PS}(k;\mu(\mathbf{x}))}} - p(\boldsymbol{z})\right)$
$\log(-\log(1-\omega(\boldsymbol{z})))$: C. Log-log	$\log\left(-\log(1-p(\boldsymbol{z})(1-\pi_{PS}(k;\mu(\mathbf{x}))))\right)$
$-\log(-\log(\omega(\boldsymbol{z})))$: Gumbel	$-\log(-\log(p(\boldsymbol{z})(1-\pi_{PS}k;\mu(\mathbf{x}))))$

Considerar as diferentes funções de ligação apresentadas na Tabela 13, resulta em diferentes funções para os parâmetros $p(\mathbf{z})$ e $\omega(\mathbf{z})$, quando é considerada a versão usual do modelo de regressão ou sua versão *Hurdle*, o que pode ser visto na Tabela 14:

Tabela 14 – Parâmetros de modificação para as funções de ligação apresentadas na Tabela 13.

$p(\mathbf{z})$	$\omega(\mathbf{z})$
$\frac{e^{\mathbf{z}\boldsymbol{\beta}_2}}{1+e^{\mathbf{z}\boldsymbol{\beta}_2}} \cdot \frac{1}{(1-\pi_{PS}(k;\mu(\mathbf{x})))}$	$\frac{e^{\mathbf{z}\boldsymbol{\beta}_2}}{1+e^{\mathbf{z}\boldsymbol{\beta}_2}}$
$(1-e^{-e^{\mathbf{z}\boldsymbol{\beta}_2}}) \cdot \frac{1}{(1-\pi_{PS}(k;\mu(\mathbf{x})))}$	$1-e^{-e^{\mathbf{z}\boldsymbol{\beta}_2}}$
$e^{-e^{-\mathbf{z}\boldsymbol{\beta}_2}} \cdot \frac{1}{(1-\pi_{PS}(k;\mu(\mathbf{x})))}$	$e^{-e^{-\mathbf{z}\boldsymbol{\beta}_2}}$

A Figura 7 apresenta uma comparação do comportamento de cada uma destas funções de ligação.

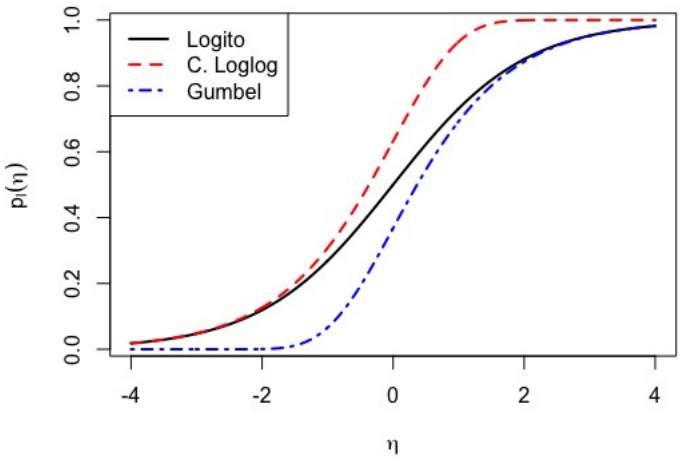

Figura 7 – Ilustração do comportamento de cada função de ligação: Logito, Complemento Loglog e Gumbel.

Fonte: Elaborada pelo autor.

Vale ressaltar que, para ambas parametrizações do modelo $k-$MPS, consideramos a mesma função de ligação para μ, $h_\mu(\mu) = \log(\mu)$, que ao relacionar essa função com o preditor linear $\boldsymbol{x}\boldsymbol{\beta}_1$, temos:

$$\mu(\boldsymbol{x}) = e^{\boldsymbol{x}\boldsymbol{\beta}_1}. \tag{4.1.5}$$

Para o modelo Binomial,

$$\mu(\boldsymbol{x}) = \frac{me^{\boldsymbol{x}\boldsymbol{\beta}_1}}{1+e^{\boldsymbol{x}\boldsymbol{\beta}_1}}, \quad m \in \mathbb{N}^*.$$

O modelo de regressão $k-$MPS tem (q_1+q_2+2) parâmetros, que correspondem aos vetores $\boldsymbol{\beta}_1^\intercal$ e $\boldsymbol{\beta}_2^\intercal$. Para inferir sobre estes parâmetros vamos apresentar os procedimentos com as abordagens clássica e Bayesiana.

4.2 Função de Verossimilhança e seu Logaritmo Natural

Seja $\boldsymbol{y} = (y_1,\ y_2,\ \ldots,\ y_n)$ um vetor de observações. Considere que y_i, $i = 1,2,\ldots,n$, foram provenientes de n realizações independentes de variável aleatória Y_i com distribuição k-MPS(μ_i, p_i). Sejam $\boldsymbol{x}_i = (1,\ x_{i1},\ x_{i2},\ \ldots,\ x_{iq_1})$ e $\boldsymbol{z}_i = (1,\ z_{i1},\ z_{i2},\ \ldots,\ z_{iq_2})$ dois vetores de covariáveis associados a y_i. Assim, a função de verossimilhança associada ao vetor de observações $\boldsymbol{y}$ é dada por

$$L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \prod_{i=1}^{n} \left\{ \Big(1 - p(\boldsymbol{z}_i) + p(\boldsymbol{z}_i)\pi_{PS}(y_i, \mu(\boldsymbol{x}_i))\Big)^{I_{\{k\}}(y_i)} \cdot \Big(p(\boldsymbol{z}_i)\pi_{PS}(y_i, \mu(\boldsymbol{x}_i))\Big)^{1-I_{\{k\}}(y_i)} \right\}.$$

Então, o logaritmo natural da função de verossimilhança é

$$\begin{aligned}\ell(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) &= \sum_{i=1}^{n}\left\{I_{\{k\}}(y_i)\left(\log\left(1-p(\boldsymbol{z}_i)+p(\boldsymbol{z}_i)\pi_{PS}(y_i,\mu(\boldsymbol{x}_i))\right)\right)\right.\\ &+ \left.(1-I_{\{k\}}(y_i))\left(\log\left(p(\boldsymbol{z}_i)\right)+\log\left(\pi_{PS}(y_i,\mu(\boldsymbol{x}_i))\right)\right)\right\}.\end{aligned}$$

Considerando o modelo versão *Hurdle* apresentado na Equação (4.1.3), temos que a função de verossimilhança associada ao vetor de observações $\boldsymbol{y}$ é dada por

$$L_H(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)=\prod_{i=1}^{n}\left\{(1-\omega(\boldsymbol{z}_i))^{I_{\{k\}}(y_i)}\left(\frac{\omega(\boldsymbol{z}_i)\pi_{PS}(y_i;\mu(\boldsymbol{x}_i))}{1-\pi_{PS}(k;\mu(\boldsymbol{x}_i))}\right)^{1-I_{\{k\}}(y_i)}\right\}. \tag{4.2.1}$$

O logaritmo natural da função verossimilhança é

$$\begin{aligned}\ell_H(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) &= \sum_{i=1}^{n}\left\{I_{\{k\}}(y_i)\log(1-\omega(\boldsymbol{z}_i))+(1-I_{\{k\}}(y_i))\log\left(\frac{\omega(\boldsymbol{z}_i)\pi_{PS}(y_i;\mu(\boldsymbol{x}_i))}{1-\pi_{PS}(k;\mu(\boldsymbol{x}_i))}\right)\right\}\\ &= \sum_{i=1}^{n}\left(1-I_{\{k\}}(y_i)\right)\log\left(\frac{\pi_{PS}(y_i;\mu(\boldsymbol{x}_i))}{1-\pi_{PS}(k;\mu(\boldsymbol{x}_i))}\right)+\\ &\quad \sum_{i=1}^{n}\left\{I_{\{k\}}(y_i)\log(1-\omega(\boldsymbol{z}i))+(1-I_{\{k\}}(y_i))\log(\omega(\boldsymbol{z}_i))\right\}\\ &= \ell_1(\boldsymbol{\beta}_1)+\ell_2(\boldsymbol{\beta}_2).\end{aligned} \tag{4.2.2}$$

Com a equação (4.2.2), notamos que $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ são ortogonais, pois podemos fatorar a função $\ell_H(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)$ em dois termos,

$$\ell_1(\boldsymbol{\beta}_1)=\sum_{i=1}^{n}\left(1-I_{\{k\}}(y_i)\right)\log\left(\frac{\pi_{PS}(y_i;\mu(\boldsymbol{x}_i))}{1-\pi_{PS}(k;\mu(\boldsymbol{x}_i))}\right)$$

e

$$\ell_2(\boldsymbol{\beta}_2)=\sum_{i=1}^{n}\{I_{\{k\}}(y_i)\log(1-\omega(\boldsymbol{z}i))+(1-I_{\{k\}}(y_i))\log(\omega(\boldsymbol{z}_i))\},$$

tais que $\ell_1(\boldsymbol{\beta}_1)$ não depende de $\boldsymbol{\beta}_2$ e $\ell_2(\boldsymbol{\beta}_2)$ não depende de $\boldsymbol{\beta}_1$. Em outras palavras, podemos estimar $\boldsymbol{\beta}_1$ independentemente de $\boldsymbol{\beta}_2$ e vice-versa, e por causa dessa simplicidade na estimação dos parâmetros, vamos trabalhar com a versão *Hurdle*.

Desta forma, ao considerarmos as funções de ligação Logito, Complemento Log-log e Gumbel para ω, teremos diferentes funções para $\ell_2(\boldsymbol{\beta}_2)$, como pode ser visto na Tabela 15.

4.3 Estimação dos Parâmetros do Modelo $k-$MPS

4.3.1 Abordagem Clássica

Considerando o método de máxima verossimilhança, o procedimento para encontrar os estimadores de $\boldsymbol{\beta}_{ij_i}$, $i=1,2$ e $j_i=0,1,\ldots,q_i$, consiste em encontrar as soluções das equações

Tabela 15 – Expressões de $\ell_2(\boldsymbol{\beta}_2)$, considerando as funções de ligação Logito, Complemento Log-log e Gumbel.

Função de Ligação	$\ell_2(\boldsymbol{\beta}_2)$
Logito	$\sum_{i=1}^{n}\left\{(1-I_{\{k\}}(y_i))\boldsymbol{z}\boldsymbol{\beta}_2-\log(1+e^{\boldsymbol{z}\boldsymbol{\beta}_2})\right\}$
C. Log-log	$\sum_{i=1}^{n}\left\{-I_{\{k\}}(y_i)e^{\boldsymbol{z}\boldsymbol{\beta}_2}+(1-I_{\{k\}}(y_i))\log(1-e^{-e^{\boldsymbol{z}\boldsymbol{\beta}_2}})\right\}$
Gumbel	$\sum_{i=1}^{n}\left\{I_{\{k\}}(y_i)\log(1-e^{-e^{-\boldsymbol{z}\boldsymbol{\beta}_2}})+(1-I_{\{k\}}(y_i))(-e^{-\boldsymbol{z}\boldsymbol{\beta}_2})\right\}$

de verossimilhança:

$$\frac{\partial \ell_1(\boldsymbol{\beta}_1)}{\partial \beta_{1j}}=0, \quad \forall\ j=0,1,\ldots,q_1$$

e

$$\frac{\partial \ell_2(\boldsymbol{\beta}_2)}{\partial \beta_{2j}}=0, \quad \forall\ j=0,1,\ldots,q_2.$$

Escrevendo $\ell_1(\boldsymbol{\beta}_1)$ em termos das funções a, f e g, temos

$$\begin{aligned}
\ell_1(\beta_1) &= \sum_{i=1}^{n}(1-I_{\{k\}}(y_i))\log(\pi_{k-MPS}(y_i,\mu(\boldsymbol{x}_i))) \\
&= \sum_{i=1}^{n}\big(1-I_{\{k\}}(y_i)\big)\log\left(\frac{\pi_{PS}(y_i;\mu(\boldsymbol{x}_i))}{1-\pi_{PS}(k;\mu(\boldsymbol{x}_i))}\right) \\
&= \sum_{i=1}^{n}(1-I_{\{k\}}(y_i))\left[\log(a(y_i))+y_i\log(g(\mu(\boldsymbol{x}_i)))-\log(f(\mu(\boldsymbol{x}_i))-a(k)g(\mu(\boldsymbol{x}_i))^k)\right].
\end{aligned}$$

Os elementos do vetor score relacionados a $\boldsymbol{\beta}_1$ podem ser obtidos da seguinte forma:

$$\frac{\partial \ell_1(\boldsymbol{\beta}_1)}{\partial \beta_{1j}} = \sum_{i=1}^{n}(1-I_{\{k\}}(y_i))\left[y_i\frac{1}{g(\mu(\boldsymbol{x}_i))}\frac{\partial g(\mu(\boldsymbol{x}_i))}{\partial \beta_{1j}}-\frac{\frac{\partial f(\mu(\boldsymbol{x}_i))}{\partial \beta_{1j}}-a(k)\frac{\partial g(\mu(\boldsymbol{x}_i))^k}{\partial \beta_{1j}}}{f(\mu(\boldsymbol{x}_i))-a(k)g(\mu(\boldsymbol{x}_i))^k}\right],$$

$j=0,1,\ldots,q_1$. Sendo $\mu(\boldsymbol{x}_i)=e^{\boldsymbol{x}_i\boldsymbol{\beta}_i}$, temos

$$\frac{\partial \ell_1(\boldsymbol{\beta}_1)}{\partial \beta_{1j}} = \sum_{i=1}^{n}(1-I_{\{k\}}(y_i))\left[y_i\frac{1}{g(e^{\boldsymbol{x}_i\boldsymbol{\beta}_i})}\frac{\partial g(e^{\boldsymbol{x}_i\boldsymbol{\beta}_i})}{\partial \beta_{1j}}-\frac{\frac{\partial f(e^{\boldsymbol{x}_i\boldsymbol{\beta}_i})}{\partial \beta_{1j}}-a(k)\frac{\partial g(e^{\boldsymbol{x}_i\boldsymbol{\beta}_i})^k}{\partial \beta_{1j}}}{f(e^{\boldsymbol{x}_i\boldsymbol{\beta}_i})-a(k)g(e^{\boldsymbol{x}_i\boldsymbol{\beta}_i})^k}\right].$$

O bloco da matriz de informação observada de Fisher referente a $\boldsymbol{\beta}_1$ é

$$\begin{aligned}
\frac{\partial^2 \ell_1(\boldsymbol{\beta}_1)}{\partial \beta_{1r}\partial \beta_{1j}} &= \sum_{i=1}^{n}(1-I_{\{k\}}(y_i))\left[y_i\left(\frac{g(\mu(\boldsymbol{x}_i)\frac{\partial^2 g(\mu(\boldsymbol{x}_i))}{\partial \beta_{1r}\partial \beta_{1j}}\frac{\partial g(\mu(\boldsymbol{x}_i))}{\partial \beta_{1}j}\frac{\partial g(\mu(\boldsymbol{x}_i))}{\partial \beta_{1r}}}{g(\mu(\boldsymbol{x}_i))^2}\right)\right] \\
&- \left(\frac{\frac{\partial^2 f(\mu(\boldsymbol{x}_i))}{\partial \beta_{1r}\partial \beta_{1j}}-a(k)\frac{\partial^2 g(\mu(\boldsymbol{x}_i))^k}{\partial \beta_{1r}\partial \beta_{1j}}}{(f(\mu(\mu(\boldsymbol{x}_i)))-a(k)g(\mu(\boldsymbol{x}_i))^k)^2}\right) \\
&+ \frac{(\frac{\partial f(\mu(\boldsymbol{x}_i))}{\partial \beta_1 j}-a(k)\frac{\partial g(\mu(\boldsymbol{x}_i))^k}{\partial \beta_1 j})(\frac{\partial g(\mu(\boldsymbol{x}_i))}{\partial \beta_1 r}-a(k)\frac{\partial g(\mu(\boldsymbol{x}_i))^k}{\partial \beta_1 r})}{(f(\mu(\boldsymbol{x}_i))-a(k)g(\mu(\boldsymbol{x}_i))^k)^2},
\end{aligned}$$

com $r, j = 0, 1, \ldots, q_1$, o qual tem dimensão $(q_1+1)\times(q_1+1)$.

Novamente, considerando $\mu(\boldsymbol{x}_i) = e^{\boldsymbol{x}_i\boldsymbol{\beta}_1}$, temos:

$$\begin{aligned}\frac{\partial^2\ell_1(\boldsymbol{\beta}_1)}{\partial\beta_{1r}\partial\beta_{1j}} &= \sum_{i=1}^{n}(1-I_{\{k\}}(y_i))\left[y_i\left(\frac{g(e^{\boldsymbol{x}_i\boldsymbol{\beta}_1})\frac{\partial^2 g(e^{\boldsymbol{x}_i\boldsymbol{\beta}_1})}{\partial\beta_{1r}\partial\beta_{1j}}\frac{\partial g(e^{\boldsymbol{x}_i\boldsymbol{\beta}_1})}{\partial\beta_{1j}}\frac{\partial g(e^{\boldsymbol{x}_i\boldsymbol{\beta}_1})}{\partial\beta_{1r}}}{g(e^{\boldsymbol{x}_i\boldsymbol{\beta}_1})^2}\right)\right]\\ &- \left(\frac{\frac{\partial^2 f(e^{\boldsymbol{x}_i\boldsymbol{\beta}_1})}{\partial\beta_{1r}\partial\beta_{1j}} - a(k)\frac{\partial^2 g(e^{\boldsymbol{x}_i\boldsymbol{\beta}_1})^k}{\partial\beta_{1r}\partial\beta_{1j}}}{(f(e^{\boldsymbol{x}_i\boldsymbol{\beta}_1}) - a(k)g(e^{\boldsymbol{x}_i\boldsymbol{\beta}_1})^k)^2}\right)\\ &+ \frac{(\frac{\partial f(e^{\boldsymbol{x}_i\boldsymbol{\beta}_1})}{\partial\beta_{1j}} - a(k)\frac{\partial g(e^{\boldsymbol{x}_i\boldsymbol{\beta}_1})^k}{\partial\beta_{1j}})(\frac{\partial g(e^{\boldsymbol{x}_i\boldsymbol{\beta}_1})}{\partial\beta_{1r}} - a(k)\frac{\partial g(e^{\boldsymbol{x}_i\boldsymbol{\beta}_1})^k}{\partial\beta_{1r}})}{(f(e^{\boldsymbol{x}_i\boldsymbol{\beta}_1}) - a(k)g(e^{\boldsymbol{x}_i\boldsymbol{\beta}_1})^k)^2}.\end{aligned}$$

Para a função $\ell_2(\boldsymbol{\beta}_2)$,

$$\ell_2(\boldsymbol{\beta}_2) = \sum_{i=1}^{n}\left\{I_{\{k\}}(y_i)\log(1-\omega(\boldsymbol{z}_i)) + (1-I_{\{k\}}(y_i))\log(\omega(\boldsymbol{z}_i))\right\}, \qquad (4.3.1)$$

os elementos do vetor score e o bloco da matriz de informação observada de Fisher relacionados a $\boldsymbol{\beta}_2$ dependerá da função de ligação considerada para ω.

a) Função de ligação Logito:

Ao considerar a função de ligação Logito, temos que $\omega(\boldsymbol{z}) = \frac{e^{\boldsymbol{z}\boldsymbol{\beta}_2}}{1+e^{\boldsymbol{z}\boldsymbol{\beta}_2}}$, e então

$$\begin{aligned}\ell_2(\boldsymbol{\beta}_2) &= \sum_{i=1}^{n}\left\{I_{\{k\}}(y_i)\log\left(\frac{1}{1+e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}\right) + (1-I_{\{k\}}(y_i))\log\left(\frac{e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}{1+e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}\right)\right\}\\ &= \sum_{i=1}^{n}\{(1-I_{\{k\}}(y_i))\log(e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}) - \log(1-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2})\}.\end{aligned}$$

Assim, os elementos do vetor score relacionados a $\boldsymbol{\beta}_2$ podem ser obtidos das seguintes formas:

$$\begin{aligned}\frac{\partial\ell_2(\boldsymbol{\beta}_2)}{\partial\beta_{20}} &= \sum_{i=1}^{n}\left\{(1-I_{\{k\}}(y_i))\left(\frac{\partial\boldsymbol{z}_i\boldsymbol{\beta}_2}{\partial\boldsymbol{\beta}_{20}}\right) - \frac{e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}{1+e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}\frac{\partial\boldsymbol{z}_i\boldsymbol{\beta}_2}{\partial\boldsymbol{\beta}_{20}}\right\}\\ &= \sum_{i=1}^{n}\left\{(1-I_{\{k\}}(y_i)) - \frac{e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}{1+e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}\right\}\end{aligned}$$

e

$$\begin{aligned}\frac{\partial\ell_2(\boldsymbol{\beta}_2)}{\partial\beta_{2r}} &= \sum_{i=1}^{n}\left\{(1-I_{\{k\}}(y_i))\left(\frac{\partial\boldsymbol{z}_i\boldsymbol{\beta}_2}{\partial\boldsymbol{\beta}_{2r}}\right) - \frac{e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}{1+e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}\frac{\partial\boldsymbol{z}_i\boldsymbol{\beta}_2}{\partial\boldsymbol{\beta}_{2r}}\right\}\\ &= \sum_{i=1}^{n}\left\{(1-I_{\{k\}}(y_i))z_{ir} - \frac{e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}z_{ir}}{1+e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}\right\}, \quad \forall\, r = 1,2,\ldots,q_2.\end{aligned}$$

Os elementos do bloco da matriz de informação observada de Fisher referente ao vetor de parâmetros $\boldsymbol{\beta}_2$ são

$$\begin{aligned}\frac{\partial^2 \ell_2(\boldsymbol{\beta_2})}{\partial \beta_{20}^2} &= \sum_{i=1}^{n}\left\{-\frac{e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}{1+e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}+\left(\frac{e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}{1+e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}\right)^2\right\}\\ &= -\sum_{i=1}^{n}\frac{e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}{(1+e^{\boldsymbol{z}_i\boldsymbol{\beta}_2})^2},\end{aligned}$$

$$\frac{\partial^2 \ell_2(\boldsymbol{\beta_2})}{\partial \beta_{2j}\partial \beta_{20}} = \sum_{i=1}^{n}\left\{-\left(\frac{e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}{1+e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}\right)z_{ij}+\left(\frac{e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}{1+e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}\right)^2 z_{ij}\right\}, \quad \forall\, j=1,2,\ldots,q_2$$

e

$$\frac{\partial^2 \ell_2(\boldsymbol{\beta_2})}{\partial \beta_{2j}\partial \beta_{2r}} = \sum_{i=1}^{n}\left\{-\left(\frac{e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}{1+e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}\right)z_{ir}z_{ij}+\left(\frac{e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}{1+e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}\right)^2 z_{ir}z_{ij}\right\}, \quad \forall\, r,j=1,2,\ldots,q_2.$$

b) Função de ligação Complemento Log-log:

Ao considerar a função de ligação Complemento Log-log, tem-se que $\omega(\boldsymbol{z}) = 1-e^{-e^{\boldsymbol{z}\boldsymbol{\beta}_2}}$, e então

$$\begin{aligned}\ell_2(\boldsymbol{\beta}_2) &= \sum_{i=1}^{n}\{I_{\{k\}}(y_i)\log(e^{-e^{\boldsymbol{z}\boldsymbol{\beta}_2}})+(1-I_{\{k\}}(y_i))\log(1-e^{-e^{\boldsymbol{z}\boldsymbol{\beta}_2}})\}\\ &= \sum_{i=1}^{n}\{I_{\{k\}}(y_i)(-e^{\boldsymbol{z}\boldsymbol{\beta}_2})+(1-I_{\{k\}}(y_i))\log(1-e^{-e^{\boldsymbol{z}\boldsymbol{\beta}_2}})\}.\end{aligned}$$

Assim, os elementos do vetor score relacionados a $\boldsymbol{\beta}_2$ podem ser obtidos das seguintes formas:

$$\begin{aligned}\frac{\partial \ell_2(\boldsymbol{\beta}_2)}{\partial \beta_{20}} &= \sum_{i=1}^{n}\left\{I_{\{k\}}(y_i)e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}\frac{\partial \boldsymbol{z}_i\boldsymbol{\beta}_2}{\boldsymbol{\beta}_{20}}+(1-I_{\{k\}}(y_i))\frac{(-e^{-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}})(-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2})}{1-e^{-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}}\frac{\partial \boldsymbol{z}_i\boldsymbol{\beta}_2}{\partial \boldsymbol{\beta}_{20}}\right\}\\ &= \sum_{i=1}^{n}\left\{I_{\{k\}}(y_i)(e^{\boldsymbol{z}_i\boldsymbol{\beta}_2})+(1-I_{\{k\}}(y_i))\frac{(-e^{-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}})(-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2})}{1-e^{-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}}\right\},\end{aligned}$$

e

$$\begin{aligned}\frac{\partial \ell_2(\boldsymbol{\beta}_2)}{\partial \beta_{2r}} &= \sum_{i=1}^{n}\left\{I_{\{k\}}(y_i)(e^{\boldsymbol{z}_i\boldsymbol{\beta}_2})\frac{\partial \boldsymbol{z}_i\boldsymbol{\beta}_2}{\boldsymbol{\beta}_{2r}}+(1-I_{\{k\}}(y_i))\frac{(-e^{-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}})(-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2})}{1-e^{-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}}\frac{\partial \boldsymbol{z}_i\boldsymbol{\beta}_2}{\partial \boldsymbol{\beta}_{2r}}\right\}\\ &= \sum_{i=1}^{n}\left\{I_{\{k\}}(y_i)(e^{\boldsymbol{z}_i\boldsymbol{\beta}_2})z_{ir}+(1-I_{\{k\}}(y_i))\frac{(-e^{-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}})(-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2})z_{ir}}{1-e^{-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}}\right\}, \quad \forall\, r=1,2,\ldots,q_2.\end{aligned}$$

Os elementos do bloco da matriz de informação observada de Fisher referente ao vetor de parâmetros $\boldsymbol{\beta}_2$ são:

$$\frac{\partial^2 \ell_1(\boldsymbol{\beta_2})}{\partial \beta_{20}^2} = \sum_{i=1}^{n}\left\{(I_{\{k\}}(y_i))[-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}]+(1-I_{\{k\}}(y_i))\left[\frac{(-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}+1)(e^{-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}e^{\boldsymbol{z}_i\boldsymbol{\beta}_2})}{1-e^{-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}}-\left(\frac{e^{-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}{1-e^{-e^{\boldsymbol{z}_i\boldsymbol{\beta}_2}}}\right)^2\right]\right\},$$

$$\frac{\partial^2 \ell_1(\boldsymbol{\beta_2})}{\partial \beta_{2j} \partial \beta_{20}} = \sum_{i=1}^{n} \Bigg\{ (I_{\{k\}}(y_i))[-e^{\boldsymbol{z_i\beta}_2} z_{ij}] + (1-I_{\{k\}}(y_i)) \Bigg[\frac{(-e^{\boldsymbol{z_i\beta}_2}+1)(e^{-e^{\boldsymbol{z_i\beta}_2}} e^{\boldsymbol{z_i\beta}_2} z_{ij})}{1-e^{-e^{\boldsymbol{z_i\beta}_2}}} - \left(\frac{e^{-e^{\boldsymbol{z_i\beta}_2}} e^{\boldsymbol{z_i\beta}_2}}{1-e^{-e^{\boldsymbol{z_i\beta}_2}}} \right)^2 z_{ij} \Bigg] \Bigg\}, \quad \forall\, j = 1,2,\ldots,q_2,$$

e

$$\frac{\partial^2 \ell_1(\boldsymbol{\beta_2})}{\partial \beta_{2j} \partial \beta_{2r}} = \sum_{i=1}^{n} \Bigg\{ (I_{\{k\}}(y_i))[-e^{\boldsymbol{z_i\beta}_2} z_{ir} z_{ij}] + (1-I_{\{k\}}(y_i)) \Bigg[\frac{(-e^{\boldsymbol{z_i\beta}_2}+1)(e^{-e^{\boldsymbol{z_i\beta}_2}} e^{\boldsymbol{z_i\beta}_2} z_{ir} z_{ij})}{1-e^{-e^{\boldsymbol{z_i\beta}_2}}} - \left(\frac{e^{-e^{\boldsymbol{z_i\beta}_2}} e^{\boldsymbol{z_i\beta}_2}}{1-e^{-e^{\boldsymbol{z_i\beta}_2}}} \right)^2 z_{ir} z_{ij} \Bigg] \Bigg\}, \quad \forall\, r,j = 1,2,\ldots,q_2.$$

c) Função de ligação Gumbel:

Considerando a função de ligação Gumbel, temos que $\omega(\boldsymbol{z}) = e^{-e^{-\boldsymbol{z\beta}_2}}$, e então

$$\begin{aligned} \ell_2(\boldsymbol{\beta}_2) &= \sum_{i=1}^{n} \Big\{ I_{\{k\}}(y_i) \log(1-e^{-e^{-\boldsymbol{z_i\beta}_2}}) + (1-I_{\{k\}}(y_i)) \log(e^{-e^{-\boldsymbol{z_i\beta}_2}}) \Big\} \\ &= \sum_{i=1}^{n} \Big\{ I_{\{k\}}(y_i) \log(1-e^{-e^{-\boldsymbol{z_i\beta}_2}}) + (1-I_{\{k\}}(y_i))(-e^{-\boldsymbol{z_i\beta}_2}) \Big\}. \end{aligned}$$

Assim, os elementos do vetor score relacionados a $\boldsymbol{\beta}_2$ podem ser obtidos das seguintes formas:

$$\begin{aligned} \frac{\partial \ell_2(\boldsymbol{\beta}_2)}{\partial \beta_{20}} &= \sum_{i=1}^{n} \Bigg\{ I_{\{k\}}(y_i) \frac{(-e^{-e^{-\boldsymbol{z_i\beta}_2}})(-e^{-\boldsymbol{z_i\beta}_2})(-1)\frac{\partial \boldsymbol{z_i\beta}_2}{\boldsymbol{\beta}_{20}}}{1-e^{-e^{-\boldsymbol{z_i\beta}_2}}} + (1-I_{\{k\}}(y_i))(-e^{-\boldsymbol{z_i\beta}_2})(-1)\frac{\partial \boldsymbol{z_i\beta}_2}{\partial \boldsymbol{\beta}_{20}} \Bigg\} \\ &= \sum_{i=1}^{n} \Bigg\{ I_{\{k\}}(y_i) \frac{(-e^{-e^{-\boldsymbol{z_i\beta}_2}})(-e^{-\boldsymbol{z_i\beta}_2})(-1)}{1-e^{-e^{-\boldsymbol{z_i\beta}_2}}} + (1-I_{\{k\}}(y_i))(-e^{-\boldsymbol{z_i\beta}_2})(-1) \Bigg\}, \end{aligned}$$

e

$$\begin{aligned} \frac{\partial \ell_2(\boldsymbol{\beta}_2)}{\partial \beta_{2q_j}} &= \sum_{i=1}^{n} \Bigg\{ I_{\{k\}}(y_i) \frac{(-e^{-e^{-\boldsymbol{z_i\beta}_2}})(-e^{-\boldsymbol{z_i\beta}_2})(-1)\frac{\partial \boldsymbol{z_i\beta}_2}{\boldsymbol{\beta}_{2q_j}}}{1-e^{-e^{-\boldsymbol{z_i\beta}_2}}} + (1-I_{\{k\}}(y_i))(-e^{-\boldsymbol{z_i\beta}_2})(-1)\frac{\partial \boldsymbol{z_i\beta}_2}{\partial \boldsymbol{\beta}_{2j}} \Bigg\} \\ &= \sum_{i=1}^{n} \Bigg\{ I_{\{k\}}(y_i) \frac{(-e^{-e^{-\boldsymbol{z_i\beta}_2}})(-e^{-\boldsymbol{z_i\beta}_2})(-z_{iq_j})}{1-e-e^{-\boldsymbol{z_i\beta}_2}} + (1-I_{\{k\}}(y_i))(-e^{-\boldsymbol{z_i\beta}_2})(-z_{ij}) \Bigg\}, \quad \forall\, q_j = 1,2,\ldots,q_2. \end{aligned}$$

Os elementos do bloco da matriz de informação observada de Fisher referente ao vetor de parâmetros $\boldsymbol{\beta}_2$ são:

$$\frac{\partial^2 \ell_1(\boldsymbol{\beta_2})}{\partial \beta_{20}^2} = \sum_{i=1}^{n} \Bigg\{ (I_{\{k\}}(y_i)) \Bigg[\frac{(-e^{-e^{-\boldsymbol{z_i\beta}_2}} e^{-\boldsymbol{z_i\beta}_2})(e^{\boldsymbol{z_i\beta}_2}-1)}{1-e^{-e^{-\boldsymbol{z_i\beta}_2}}} - \left(\frac{e^{-e^{-\boldsymbol{z_i\beta}_2}} e^{-\boldsymbol{z_i\beta}_2}}{1-e^{-e^{-\boldsymbol{z_i\beta}_2}}} \right)^2 \Bigg] + (1-I_{\{k\}}(y_i))[-e^{-\boldsymbol{z_i\beta}_2}] \Bigg\},$$

$$\frac{\partial^2 \ell_1(\boldsymbol{\beta_2})}{\partial \beta_{2j} \partial \beta_{20}} = \sum_{i=1}^{n} \left\{ (I_{\{k\}}(y_i)) \left[\frac{(-e^{-e^{-\mathbf{z}_i\boldsymbol{\beta}_2}} e^{-\mathbf{z}_i\boldsymbol{\beta}_2} z_{ij})(e^{-\mathbf{z}_i\boldsymbol{\beta}_2} - 1)}{1 - e^{-e^{-\mathbf{z}_i\boldsymbol{\beta}_2}}} - \left(\frac{e^{-e^{-\mathbf{z}_i\boldsymbol{\beta}_2}} e^{-\mathbf{z}_i\boldsymbol{\beta}_2}}{1 - e^{-e^{-\mathbf{z}_i\boldsymbol{\beta}_2}}} \right)^2 z_{ij} \right] + (1 - I_{\{k\}}(y_i))[-e^{-\mathbf{z}_i\boldsymbol{\beta}_2} z_{ij}] \right\}, \quad \forall\, j = 1, 2, \ldots, q_2,$$

e

$$\frac{\partial^2 \ell_1(\boldsymbol{\beta_2})}{\partial \beta_{2j} \partial \beta_{2r}} = \sum_{i=1}^{n} \left\{ (I_{\{k\}}(y_i)) \left[\frac{(-e^{-e^{-\mathbf{z}_i\boldsymbol{\beta}_2}} e^{-\mathbf{z}_i\boldsymbol{\beta}_2} z_{ir} z_{ij})(e^{-\mathbf{z}_i\boldsymbol{\beta}_2} - 1)}{1 - e^{-e^{-\mathbf{z}_i\boldsymbol{\beta}_2}}} - \left(\frac{e^{-e^{-\mathbf{z}_i\boldsymbol{\beta}_2}} e^{-\mathbf{z}_i\boldsymbol{\beta}_2}}{1 - e^{-e^{-\mathbf{z}_i\boldsymbol{\beta}_2}}} \right)^2 z_{ir} z_{ij} \right] + (1 - I_{\{k\}}(y_i))[-e^{-\mathbf{z}_i\boldsymbol{\beta}_2} z_{ir} z_{ij}] \right\}, \quad \forall\, r, j = 1, 2, \ldots, q_2.$$

De maneira geral, uma solução analítica em forma fechada para o estimador de máxima verossimilhança de β_{ij_i}, $i = 1,2$ e $j_i = 0, 1, \ldots, q_i$, não pode ser encontrada. Sendo assim, torna-se necessário recorrer a algum procedimento de otimização numérica. A função `optim` do *Software* R (R Core Team, 2015) pode ser utilizada para a obtenção das estimativas de máxima verossimilhança dos vetores de parâmetros $\boldsymbol{\beta}_1^\top = (\beta_{10}\ \beta_{11}\ \ldots\ \beta_{1q_1})$ e $\boldsymbol{\beta}_2^\top = (\beta_{20}\ \beta_{21}\ \ldots\ \beta_{2q_2})$.

4.3.2 Abordagem Bayesiana

Para o procedimento Bayesiano, considere novamente y_i, $i = 1, 2, \ldots, n$, n observações independentes da variável aleatória Y_i com distribuição k-MPS(μ_i, ω_i). Sejam $\boldsymbol{\beta}_1 = (\beta_{11}, ..., \beta_{1q_1})$ e $\boldsymbol{\beta}_2 = (\beta_{21}, ..., \beta_{2q_2})$ os vetores de parâmetros dos preditores lineares e, então, para cada $\boldsymbol{\beta}_{ij_i}$, $i = 1,2$ e $j = 1, .., q_i$, adotamos uma *priori* vaga para que as informações contidas nos dados sejam as mais relevantes. Supondo independência dos parâmetros, a *priori* conjunta é dada por

$$\pi(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \prod_{i=1}^{2} \prod_{j=1}^{q_i} \pi(\beta_{ij_i}).$$

Consequentemente, a densidade a *posteriori* conjunta é dada por

$$\pi(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \boldsymbol{y}) \propto L_H(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \pi(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2),$$

em que $L_H(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ a função de verossimilhança associada ao vetor de observações $\boldsymbol{y}$ dada em (4.2.1).

Do ponto de vista Bayesiano, inferências sobre os parâmetros podem ser baseados nas densidades a *posteriori* marginais, o que pode ser obtido ao integrar a densidade a *posteriori* conjunta (CONCEIÇÃO; ANDRADE; LOUZADA, 2013). Neste caso, entretanto, não é possível obter soluções analíticas para as integrais. Assim, para resolver este problema, utilizamos o algoritmo *Metropolis-Hastings* (CHIB; GREENBERG, 1995), (HASTINGS, 1970), que é um procedimento iterativo de uma grande classe de métodos MCMC. Para implementar este algoritmo, consideramos as densidades condicionais dos parâmetros definidos nos vetores $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$, dadas por

$$\pi(\beta_{1j_1} | \boldsymbol{\beta}_1^*, \boldsymbol{y}) \propto \exp\{\ell_1(\boldsymbol{\beta}_1)\} \pi(\beta_{1j_1}) \qquad \text{e} \qquad \pi(\beta_{2j_2} | \boldsymbol{\beta}_2^*, \boldsymbol{y}) \propto \exp\{\ell_2(\boldsymbol{\beta}_2)\} \pi(\beta_{2j_2}),$$

em que $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_1 - \{\beta_{1j_1}\}$ e $\boldsymbol{\beta}_2^* = \boldsymbol{\beta}_2 - \{\beta_{2j_2}\}$, com $j_1 = 1,\ldots,q_1$ e $j_2 = 1,\ldots,q_2$.

Todas as implementações computacionais foram feitas utilizando os pacotes `R2jags` (versão 0.5-7) e `rjags` (versão 4-6) no sistema Jags (PLUMMER, 2017) por meio do *Software* R (R Core Team, 2015). Para verificar a convergência das cadeias foi considerado o procedimento de diagnóstico de Gelman-Rubin (GELMAN; RUBIN, 2009).

4.4 Estudo Bayesiano de Pontos Influentes

Considere $\boldsymbol{y}_{(-i)}$ como sendo um vetor de observações após a remoção da i-ésima observação de $\boldsymbol{y}$, isto é, $\boldsymbol{y}_{(-i)} = (y_1, y_2, ..., y_{i-1}, y_{i+1}, ..., y_n)$. Seja $\pi(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\boldsymbol{y})$ a densidade conjunta a *posteriori* do vetor de parâmetros $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ considerando o conjunto de dados original, e seja $\pi(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\boldsymbol{y}_{(-i)})$ a densidade conjunta a *posteriori* considerando o conjunto de dados após a remoção da i-ésima observação. Desta forma, a influência da observação y_i pode ser avaliada através da distância de Kullback-Leibler (KL), medida que calcula a distância entre duas densidades a *posteriori* (CHO *et al.*, 2009). Isto é,

$$KL(\pi, \pi_{(-i)}) = \int \pi(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\boldsymbol{y}) \log\left(\frac{\pi(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\boldsymbol{y})}{\pi(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\boldsymbol{y}_{(-i)})}\right) d\boldsymbol{\beta}. \tag{4.4.1}$$

A equação (4.4.1) pode ser expressa como esperança a *posteriori*:

$$KL(\pi, \pi_{(-i)}) = -\log(CPO_i) + E_{\pi(\boldsymbol{\beta}|\boldsymbol{y})}\{\log(\pi_{k-MPS}(y_i; \boldsymbol{\beta}_1, \boldsymbol{\beta}_2))\},$$

em que CPO_i é a densidade condicional preditiva ordenada (CPO) da observação y_i.

Desta forma, gerando uma amostra $\{(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^{(1)}, \ldots, (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^{(Q)}\}$ a partir da densidade a *posteriori* $\pi(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\boldsymbol{y})$ é possível estimar o efeito da observação y_i através da seguinte equação

$$\widehat{KL}(\pi, \pi_{(-i)}) = -\log(\widehat{CPO}_i) + \frac{1}{Q}\sum_{q=1}^{Q} \log(\pi_{k-MPS}(y_i; (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^{(q)})), \tag{4.4.2}$$

em que $\widehat{CPO}_i$ em (4.4.2) é dado por $\widehat{CPO}_i = \left[\frac{1}{Q}\sum_{q=1}^{Q} \frac{1}{\log(\pi_{k-MPS}(y_i; (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^{(q)}))}\right]^{-1}$.

McCulloch (1989) propôs a medida de calibração ρ_i para a distância $KL(\pi, \pi_{(-i)})$, em que esta é derivada da solução da equação $KL(\pi, \pi_{(-i)}) = KL(B(0.5), B(\rho_i)) = -\log\left(4\rho_i(1-\rho_i)\right)/2$, de forma que $B(\rho_i)$ denota a distribuição de Bernoulli com uma probabilidade de sucesso ρ. Isto implica que, descrever os resultados usando a densidade a *posteriori* completa $\pi(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\boldsymbol{y})$ ao invés da densidade a *posteriori* removida a i-ésima observação, $\pi(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\boldsymbol{y}_{(-i)})$, é equivalente a descrever um evento não observado como tendo probabilidade ρ_i, quando a probabilidade correta é 0.5 (CONCEIÇÃO; ANDRADE; LOUZADA, 2013). Desta forma, ao resolver a equação para ρ_i teremos $\rho_i = \frac{1}{2}\{1 + \sqrt{1 - \exp(-2KL(\pi, \pi_{(-i)}))}\}$, e portanto $0.5 \leq \rho_i \leq 1$. Assim, para $\rho \gg 0.5$ a i-ésima observação pode ser considerada como um ponto influente.

Nos Capítulos 6 e 7 a seguir, iremos apresentar um estudo de simulação e aplicações com conjuntos de dados artificiais e reais dos modelos de regressão *k*-MPS. Embora tenhamos apresentado duas abordagens para a estimação dos parâmetros, para estes estudos consideramos apenas a abordagem Bayesiana por incorporar informações prévias sobre os parâmetros.

CAPÍTULO

5

ESTUDO DE SIMULAÇÃO

Para avaliar as propriedades dos estimadores Bayesianos dos parâmetros do modelo, consideramos alguns estudos de simulação com conjuntos de dados k-modificados. Cada estudo consistiu em gerar $N = 500$ conjuntos de dados k-inflacionados de tamanho n ($n = 100$ e 500) e $N = 500$ conjuntos de dados k-deflacionados de tamanho n ($n = 200$[1] e 500), a partir de uma variável aleatória Y_i com distribuição k-MPS(μ_i,ω_i), $i = 1,2,\ldots,n$, considerando diferentes valores de k.

Consideramos a matriz $\boldsymbol{X}$ de covariáveis com dimensão $n \times 2$, a qual na primeira coluna é constituída de 1's (intercepto) e a segunda coluna é constituída da variável $\boldsymbol{X}_1 = (X_{11}, X_{12}, \ldots, X_{1n})$, sendo cada elemento X_{1i} gerado de uma distribuição Uniforme $U(0,1)$. Consideramos também a matriz $\boldsymbol{Z}$ de covariáveis $n \times 2$, de forma que $\boldsymbol{Z} \equiv \boldsymbol{X}$. Aos vetores de parâmetros $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ atribuímos diferentes valores e para os casos k-inflacionados, consideramos as funções de ligação Logito, Complemento Log-log e Gumbel. Já nos casos k-deflacionados, apenas as funções de ligação Logito e Gumbel foram consideradas, pois a Complemento Log-log não apresentou um comportamento adequado ao problema nesta situação. Ressaltamos que, para cada função de ligação utilizada para gerar as amostras, consideramos a mesma para o ajuste dos modelos.

As *priori* consideradas para os vetores de parâmetros $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ foram distribuições Normais Multivariadas tal que $\boldsymbol{\beta}_1 \sim NMult(\hat{\boldsymbol{\beta}}_1, (\tau_1 * \boldsymbol{H}_1)^{-1})$ e $\boldsymbol{\beta}_2 \sim NMult(\hat{\boldsymbol{\beta}}_2, (\tau_2 * \boldsymbol{H}_2)^{-1})$, em que $\hat{\boldsymbol{\beta}}_i$ é o vetor de médias (escolhido com base em informações prévias) para $\boldsymbol{\beta}_i$, $i = 1,2$, e $\boldsymbol{H}_i$ é uma matriz diagonal 2×2 cuja diagonal principal é 10^{-1}. Também consideramos as constantes de precisão, $\tau_1 > 0$ e $\tau_2 > 0$, conhecidas, que são usadas para controlar a taxa de rejeição do algoritmo de geração de cada cadeia.

Como estimador Bayesiano, consideramos a média a *posteriori* para cada parâmetro

[1] Consideramos o tamanho amostral de $n = 200$ para amostras k-deflacionadas para garantir que elas fossem, de fato, caracterizadas como deflacionadas.

(perda quadrática) e avaliamos seu desempenho usando a raiz quadrada relativa do erro quadrático médio $\left(\sqrt{EQM(\hat{\beta})/\beta}\right)$, o desvio-padrão $\left(\sqrt{Var(\hat{\beta})}\right)$ e o vício médio $\left(\mathscr{B}(\hat{\beta})\right)$, em que $EQM(\hat{\beta}) = E(\hat{\beta}-\beta)^2$ e $\mathscr{B}(\hat{\beta}) = E(\hat{\beta}-\beta)$. Os resultados deste estudo são apresentados a seguir para conjuntos de dados caracterizados como k-inflacionados e k-deflacionados.

Para cada caso do estudo de simulação, utilizamos do *software Jags*[2] (PLUMMER, 2017) para analisar os conjuntos, gerando duas cadeias de tamanho 50.000 das condicionais a *posteriori* de cada parâmetro e considerando um período de aquecimento (*burn in*) de 10.000 iterações em cada cadeia. A convergência da cadeia foi verificada através do critério de Gelman-Rubin (GELMAN; RUBIN, 2009). Para obter amostras pseudo-independentes, consideramos saltos de tamanho 10, resultando em cadeias de tamanho 8.000 para cada parâmetro.

5.1 Modelo Binomial k-Modificado (k-MB)

Apresentamos os estudos de simulação com conjunto de dados inflacionados e deflacionadados gerados a partir de modelos k-MB.

5.1.1 Modelo Binomial k-Inflacionado (k-IB)

Para este estudo de simulação, conjuntos de dados k-inflacionados ($k = 0, 1$ e 2) foram gerados de um modelo k-MB[3] com $0 < p(\boldsymbol{z}_i) < 1, \forall\, i = 1, \ldots, n$ (k-IB[4]), atribuindo os seguintes valores a cada vetor de parâmetros: $\boldsymbol{\beta}_1 = (3, -2)$, $\boldsymbol{\beta}_2 = (-2, 3)$ e $m = 10$. Consideramos os parâmetros de precisão $\tau_1 = 1$ e $\tau_2 = 0.5$. Sobre as funções de ligação, as mesmas utilizadas para gerar os conjuntos de dados foram utilizadas para o procedimento de estimação e, assim, analisar os resultados.

As estimativas pontuais dos parâmetros $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ (aqui consideradas como a média a *posteriori* para cada parâmetro), em sua maioria, estão sempre próximas dos verdadeiros valores, como podem ser vistos a partir dos baixos valores apresentados pelas medidas de eficiências (ver Tabelas 16, 17 e 18). Apresentamos, nas Tabelas 16, 17 e 18, os resultados do estudo de simulação para dados 0-MB, 1-MB e 2-MB respectivamente, na qual podemos observar que todas as medidas de eficiência dos estimadores aproximam-se de zero quando o tamanho do conjunto de dados aumenta.

[2] Usaremos a sigla Jags, do inglês *Just Another Gibs Sample*, referindo-se ao software Jags para análise Bayesiana de modelos hierárquicos usando simulações MCMC.
[3] Usaremos a sigla k-MB, do inglês *k-Modified Binomial*, referindo-se à distribuição Binomial k-Modificada.
[4] Usaremos a sigla k-IB, do inglês *k-Inflated Binomial*, referindo-se à distribuição Binomial k-Inflacionada.

Tabela 16 – Medidas de eficiência do estimador Bayesiano para cada parâmetro do modelo k-IB, com $k = 0$ e $m = 10$, considerando diferentes funções de ligação para ω.

Função de Ligação	n	(τ_1, τ_2)	Parâmetros	$\sqrt{EQM(\hat{\beta}_{jk})/\beta_{jk}}$	$\sqrt{Var(\hat{\beta}_{jk})}$	$\mathscr{B}(\hat{\beta}_{jk})$
Logito	100	(1, 0.5)	β_{10}	0.169	0.497	0.388
			β_{11}	-0.335	0.662	0.524
			β_{20}	-0.265	0.515	0.406
			β_{21}	0.294	0.858	0.682
	500	(1, 0.5)	β_{10}	0.154	0.458	0.167
			β_{11}	-0.295	0.586	0.226
			β_{20}	-0.226	0.440	0.201
			β_{21}	0.225	0.661	0.322
CLogLog	100	(1, 0.5)	β_{10}	0.148	0.437	0.339
			β_{11}	-0.288	0.573	0.449
			β_{20}	-0.241	0.464	0.378
			β_{21}	0.255	0.744	0.599
	500	(1, 0.5)	β_{10}	0.064	0.193	0.154
			β_{11}	-0.127	0.254	0.204
			β_{20}	-0.091	0.182	0.144
			β_{21}	0.094	0.281	0.219
Gumbel	100	(1, 0.5)	β_{10}	0.295	0.874	0.681
			β_{11}	-0.542	1.076	0.841
			β_{20}	-0.276	0.501	0.418
			β_{21}	0.288	0.780	0.660
	500	(1, 0.5)	β_{10}	0.122	0.367	0.298
			β_{11}	-0.223	0.446	0.361
			β_{20}	-0.093	0.184	0.148
			β_{21}	0.101	0.297	0.240

Tabela 17 – Medidas de eficiência do estimador Bayesiano para cada parâmetro do modelo k-IB, com $k = 1$ e $m = 10$, considerando diferentes funções de ligação para ω.

Função de Ligação	n	(τ_1, τ_2)	Parâmetros	$\sqrt{EQM(\hat{\beta}_{jk})/\beta_{jk}}$	$\sqrt{Var(\hat{\beta}_{jk})}$	$\mathscr{B}(\hat{\beta}_{jk})$
Logito	100	(1, 0.5)	β_{10}	0.295	0.874	0.395
			β_{11}	-0.542	1.076	0.508
			β_{20}	-0.276	0.501	0.484
			β_{21}	0.288	0.780	0.749
	500	(1, 0.5)	β_{10}	0.068	0.203	0.162
			β_{11}	-0.133	0.267	0.213
			β_{20}	-0.122	0.240	0.190
			β_{21}	0.134	0.397	0.319
CLogLog	100	(1, 0.5)	β_{10}	0.164	0.485	0.380
			β_{11}	-0.336	0.668	0.514
			β_{20}	-0.228	0.446	0.350
			β_{21}	0.242	0.715	0.560
	500	(1, 0.5)	β_{10}	0.065	0.195	0.155
			β_{11}	-0.131	0.261	0.208
			β_{20}	-0.092	0.181	0.148
			β_{21}	0.092	0.274	0.221
Gumbel	100	(1, 0.5)	β_{10}	0.285	0.843	0.652
			β_{11}	-0.501	0.991	0.776
			β_{20}	-0.286	0.529	0.421
			β_{21}	0.292	0.813	0.656
	500	(1, 0.5)	β_{10}	0.129	0.385	0.307
			β_{11}	-0.236	0.471	0.375
			β_{20}	-0.098	0.188	0.152
			β_{21}	0.101	0.293	0.236

Tabela 18 – Medidas de eficiência do estimador Bayesiano para cada parâmetro do modelo k-IB, com $k = 2$ e $m = 10$, considerando diferentes funções de ligação para ω.

Função de Ligação	n	(τ_1, τ_2)	Parâmetros	$\sqrt{EQM(\hat{\beta}_{jk})/\beta_{jk}}$	$\sqrt{Var(\hat{\beta}_{jk})}$	$\mathscr{B}(\hat{\beta}_{jk})$
Logito	100	(1, 0.5)	β_{10}	0.165	0.490	0.378
			β_{11}	-0.323	0.644	0.503
			β_{20}	-0.287	0.551	0.438
			β_{21}	0.305	0.880	0.700
	500	(1, 0.5)	β_{10}	0.076	0.227	0.183
			β_{11}	-0.152	0.302	0.244
			β_{20}	-0.121	0.242	0.190
			β_{21}	0.138	0.411	0.325
CLogLog	100	(1, 0.5)	β_{10}	0.154	0.458	0.361
			β_{11}	-0.295	0.586	0.467
			β_{20}	-0.226	0.440	0.349
			β_{21}	0.225	0.661	0.517
	500	(1, 0.5)	β_{10}	0.063	0.186	0.150
			β_{11}	-0.123	0.243	0.195
			β_{20}	-0.094	0.188	0.153
			β_{21}	0.093	0.278	0.225
Gumbel	100	(1, 0.5)	β_{10}	0.154	0.458	0.642
			β_{11}	-0.295	0.586	0.789
			β_{20}	-0.226	0.440	0.409
			β_{21}	0.225	0.661	0.678
	500	(1, 0.5)	β_{10}	0.122	0.367	0.292
			β_{11}	-0.226	0.451	0.359
			β_{20}	-0.092	0.183	0.145
			β_{21}	0.103	0.305	0.243

5.1.2 Modelo Binomial k-Deflacionado (k-DB)

Para este estudo de simulação, conjuntos de dados k-deflacionados ($k = 0$ e 1) foram gerados de um modelo k-MB com $1 < p(\boldsymbol{z}_i) < 1/(1 - \pi_B(y_i, \mu(\boldsymbol{x}_i)))$, $\forall\ i = 1, \ldots, n$ (k-DB[5]), atribuindo os seguintes valores a cada vetor de parâmetros: $\boldsymbol{\beta}_1 = (-2, -2)$ e $\boldsymbol{\beta}_2 = (2, 3)$. Consideramos os parâmetros de precisão $\tau_1 = 1$ e $\tau_2 = 0.5$ e as funções de ligação Logito e Gumbel. Como resultado, temos que as estimativas Bayesianas dos vetores de parâmetros $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ (aqui consideradas como a média a *posteriori* para cada parâmetro) estão sempre próximas dos verdadeiros valores, como podem ser vistos a partir dos baixos valores apresentados pelas medidas de eficiências (ver Tabelas 19 e 20).

Ainda sobre as Tabelas 19 e 20, temos os resultados do estudo de simulação para dados 0-MB e 1-MB respectivamente, na qual podemos observar que todas as medidas de eficiência consideradas apresentam valores relativamente baixos, aproximando-se de zero quando o tamanho do conjunto de dados aumenta. Portanto, o procedimento Bayesiano de estimação

[5] Usaremos a sigla k-DB, do inglês *k-Deflated Binomial*, referindo-se à distribuição Binomial k-Deflacionada.

torna-se mais preciso quanto maior o tamanho das amostras.

Tabela 19 – Medidas de eficiência do estimador Bayesiano para cada parâmetro do modelo k-DB, com $k = 0$ e $m = 10$, considerando diferentes funções de ligação para ω.

Função de Ligação	n	(τ_1, τ_2)	Parâmetros	$\sqrt{EQM(\hat{\beta}_{jk})/\beta_{jk}}$	$\sqrt{Var(\hat{\beta}_{jk})}$	$\mathscr{B}(\hat{\beta}_{jk})$
Logito	200	(1, 0.5)	β_{10}	-0.126	0.250	0.198
			β_{11}	-0.279	0.555	0.438
			β_{20}	0.507	0.993	0.609
			β_{21}	0.785	2.300	1.618
	500	(1, 0.5)	β_{10}	-0.069	0.138	0.107
			β_{11}	-0.163	0.326	0.254
			β_{20}	0.193	0.386	0.295
			β_{21}	0.397	1.164	0.856
Gumbel	200	(1, 0.5)	β_{10}	-0.118	0.234	0.184
			β_{11}	-0.259	0.517	0.411
			β_{20}	0.354	0.689	0.522
			β_{21}	0.811	2.375	1.562
	500	(1, 0.5)	β_{10}	-0.071	0.143	0.115
			β_{11}	-0.178	0.353	0.282
			β_{20}	0.163	0.325	0.260
			β_{21}	0.353	1.041	0.810

Tabela 20 – Medidas de eficiência do estimador Bayesiano para cada parâmetro do modelo k-DB, com $k = 1$ e $m = 10$, considerando diferentes funções de ligação para ω.

Função de Ligação	n	(τ_1, τ_2)	Parâmetros	$\sqrt{EQM(\hat{\beta}_{jk})/\beta_{jk}}$	$\sqrt{Var(\hat{\beta}_{jk})}$	$\mathscr{B}(\hat{\beta}_{jk})$
Logito	200	(1, 0.5)	β_{10}	-0.073	0.146	0.115
			β_{11}	-0.223	0.430	0.343
			β_{20}	0.311	0.615	0.478
			β_{21}	0.805	2.329	1.559
	500	(1, 0.5)	β_{10}	-0.050	0.099	0.081
			β_{11}	-0.135	0.269	0.211
			β_{20}	0.203	0.406	0.324
			β_{21}	0.387	1.132	0.867
Gumbel	200	(1, 0.5)	β_{10}	-0.080	0.160	0.125
			β_{11}	-0.252	0.488	0.390
			β_{20}	0.339	0.654	0.500
			β_{21}	0.724	2.136	1.458
	500	(1, 0.5)	β_{10}	-0.050	0.099	0.079
			β_{11}	-0.143	0.284	0.227
			β_{20}	0.187	0.374	0.292
			β_{21}	0.346	1.020	0.804

5.2 Modelo Geométrico k-Modificado (k-MG)

Apresentamos a seguir os estudos de simulação com conjunto de dados inflacionados e deflacionadados gerados a partir de modelos k-MG.

5.2.1 Modelo Geométrico k-Inflacionado (k-IG)

Para este estudo de simulação, conjuntos de dados k-inflacionados ($k = 0, 1$ e 2) foram gerados de um modelo k-MG[6] com $0 < p(\boldsymbol{z}_i) < 1, \forall\, i = 1, \ldots, n$ (k-IG[7]), atribuindo os seguintes valores aos vetores de parâmetros: $\boldsymbol{\beta}_1 = (4, 2)$ e $\boldsymbol{\beta}_2 = (-2, 3)$. Consideramos os parâmetros de precisão τ_1 e τ_2 entre 0.4 e 1. As estimativas Bayesianas dos vetores de parâmetros $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ (a média a *posteriori* para cada parâmetro) estão, em sua maioria, próximas dos verdadeiros valores quando observamos os baixos valores apresentados pelas medidas de eficiências (ver Tabelas 21, 22 e 23).

Ainda nas Tabelas 21, 22 e 23, apresentamos os resultados do estudo de simulação para dados 0-MG, 1-MG e 2-MG respectivamente, na qual podemos observar que todas as medidas de eficiência aproximam-se de zero quando o tamanho do conjunto de dados aumenta. Portanto, o procedimento Bayesiano para a estimação dos parãmetros torna-se mais preciso quanto maior o tamanho das amostras.

[6] Usaremos a sigla k-MG, do inglês *k-Modified Geometric*, referindo-se à distribuição Geométrica k-Modificada.
[7] Usaremos a sigla k-IG, do inglês *k-Inflated Geometric*, referindo-se à distribuição Geométrica k-Inflacionada.

Tabela 21 – Medidas de eficiência do estimador Bayesiano para cada parâmetro do modelo k-IG, com $k = 0$, considerando diferentes funções de ligação para ω.

Função de Ligação	n	(τ_1, τ_2)	Parâmetros	$\sqrt{EQM(\hat{\beta}_{jk})/\beta_{jk}}$	$\sqrt{Var(\hat{\beta}_{jk})}$	$\mathscr{B}(\hat{\beta}_{jk})$
Logito	100	(1, 1)	β_{10}	0.109	0.434	0.346
			β_{11}	0.298	0.594	0.478
			β_{20}	-0.281	0.552	0.437
			β_{21}	0.299	0.877	0.708
	500	(1, 1)	β_{10}	0.049	0.197	0.158
			β_{11}	0.146	0.292	0.235
			β_{20}	-0.114	0.224	0.184
			β_{21}	0.129	0.384	0.314
CLogLog	100	(0.5, 0.5)	β_{10}	0.112	0.447	0.350
			β_{11}	0.337	0.675	0.528
			β_{20}	-0.232	0.448	0.365
			β_{21}	0.237	0.688	0.557
	500	(0.5, 0.5)	β_{10}	0.046	0.183	0.145
			β_{11}	0.135	0.270	0.214
			β_{20}	-0.095	0.190	0.150
			β_{21}	0.097	0.290	0.226
Gumbel	100	(0.5, 0.5)	β_{10}	0.225	0.903	0.688
			β_{11}	0.574	1.150	0.876
			β_{20}	-0.288	0.542	0.423
			β_{21}	0.290	0.821	0.650
	500	(0.5, 0.5)	β_{10}	0.098	0.391	0.310
			β_{11}	0.255	0.510	0.408
			β_{20}	-0.091	0.177	0.144
			β_{21}	0.100	0.293	0.239

Tabela 22 – Medidas de eficiência do estimador Bayesiano para cada parâmetro do modelo k-IG, com $k = 1$, considerando diferentes funções de ligação para ω.

Função de Ligação	n	(τ_1, τ_2)	Parâmetros	$\sqrt{EQM(\hat{\beta}_{jk})/\beta_{jk}}$	$\sqrt{Var(\hat{\beta}_{jk})}$	$\mathscr{B}(\hat{\beta}_{jk})$
Logito	100	(1, 1)	β_{10}	0.101	0.404	0.317
			β_{11}	0.298	0.596	0.473
			β_{20}	-0.253	0.490	0.390
			β_{21}	0.305	0.885	0.703
	500	(1, 1)	β_{10}	0.050	0.198	0.157
			β_{11}	0.142	0.283	0.221
			β_{20}	-0.121	0.241	0.191
			β_{21}	0.132	0.395	0.314
CLogLog	100	(0.5, 0.5)	β_{10}	0.106	0.423	0.338
			β_{11}	0.282	0.563	0.445
			β_{20}	-0.224	0.428	0.343
			β_{21}	0.217	0.628	0.496
	500	(0.5, 0.5)	β_{10}	0.046	0.185	0.148
			β_{11}	0.131	0.262	0.211
			β_{20}	-0.095	0.188	0.148
			β_{21}	0.095	0.285	0.226
Gumbel	100	(0.5, 0.5)	β_{10}	0.244	0.975	0.767
			β_{11}	0.625	1.248	0.975
			β_{20}	-0.309	0.576	0.427
			β_{21}	0.321	0.899	0.696
	500	(0.5, 0.5)	β_{10}	0.102	0.410	0.325
			β_{11}	0.262	0.524	0.417
			β_{20}	-0.097	0.191	0.147
			β_{21}	0.105	0.310	0.245

Tabela 23 – Medidas de eficiência do estimador Bayesiano para cada parâmetro do modelo k-IG, com $k = 2$, considerando diferentes funções de ligação para ω.

Função de Ligação	n	(τ_1, τ_2)	Parâmetros	$\sqrt{EQM(\hat{\beta}_{jk})/\beta_{jk}}$	$\sqrt{Var(\hat{\beta}_{jk})}$	$\mathscr{B}(\hat{\beta}_{jk})$
Logito	100	(1, 1)	β_{10}	0.104	0.416	0.327
			β_{11}	0.315	0.631	0.503
			β_{20}	-0.281	0.541	0.428
			β_{21}	0.305	0.879	0.705
	500	(1, 1)	β_{10}	0.048	0.190	0.155
			β_{11}	0.134	0.268	0.214
			β_{20}	-0.116	0.230	0.184
			β_{21}	0.126	0.377	0.302
CLogLog	100	(0.5, 0.5)	β_{10}	0.110	0.441	0.357
			β_{11}	0.303	0.604	0.484
			β_{20}	-0.252	0.478	0.393
			β_{21}	0.243	0.693	0.574
	500	(0.5, 0.5)	β_{10}	0.045	0.181	0.141
			β_{11}	0.128	0.256	0.205
			β_{20}	-0.096	0.192	0.152
			β_{21}	0.093	0.278	0.219
Gumbel	100	(0.4, 0.4)	β_{10}	0.235	0.939	0.732
			β_{11}	0.597	1.193	0.939
			β_{20}	-0.261	0.488	0.376
			β_{21}	0.276	0.779	0.616
	500	(0.4, 0.4)	β_{10}	0.097	0.388	0.305
			β_{11}	0.256	0.513	0.407
			β_{20}	-0.094	0.186	0.142
			β_{21}	0.106	0.313	0.244

5.2.2 Modelo Geométrico k-Deflacionado (k-DG)

Para este estudo de simulação, conjuntos de dados k-deflacionados foram gerados de um modelo k-MG com $1 < p(\boldsymbol{z}_i) < 1/(1-\pi_G(y_i, \mu(\boldsymbol{x}_i)))$ (k-DG[8]), atribuindo os seguintes valores aos vetores de parâmetros: $\boldsymbol{\beta}_1 = (-1, 1)$ e $\boldsymbol{\beta}_2 = (2, 1)$. Consideramos os parâmetros de precisão τ_1 e τ_2 iguais a 1 e as funções de ligação Logito e Gumbel.

Apresentamos nas Tabelas 24 e 25 os resultados do estudo de simulação para dados 0-MG e 1-MG respectivamente, na qual podemos observar que todas as medidas aproximam-se de zero quando o tamanho do conjunto de dados aumenta. Portanto, o procedimento Bayesiano de estimação torna-se mais preciso quanto maior o tamanho das amostras. Estes resultados são mais uma evidência para afirmarmos que as estimativas Bayesianas de cada parâmetro de $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ (a média a *posteriori* para cada parâmetro) estão próximas dos verdadeiros valores.

[8] Usaremos a sigla k-DG, do inglês *k-Deflated Geometric*, referindo-se à distribuição Geométrica k-Deflacionada.

Tabela 24 – Medidas de eficiência do estimador Bayesiano para cada parâmetro do modelo k-DG, com $k = 0$, considerando diferentes funções de ligação para ω.

Função de Ligação	n	(τ_1, τ_2)	Parâmetros	$\sqrt{EQM(\hat{\beta}_{jk})/\beta_{jk}}$	$\sqrt{Var(\hat{\beta}_{jk})}$	$\mathscr{B}(\hat{\beta}_{jk})$
Logito	200	(1,1)	β_{10}	-0.275	0.272	0.217
			β_{11}	0.435	0.434	0.349
			β_{20}	0.259	0.513	0.402
			β_{21}	0.979	0.973	0.757
	500	(1,1)	β_{10}	-0.173	0.173	0.136
			β_{11}	0.282	0.282	0.224
			β_{20}	0.153	0.305	0.234
			β_{21}	0.618	0.613	0.489
Gumbel	200	(1,1)	β_{10}	-0.252	0.251	0.199
			β_{11}	0.401	0.401	0.319
			β_{20}	0.239	0.476	0.367
			β_{21}	0.969	0.948	0.749
	500	(1,1)	β_{10}	-0.172	0.172	0.139
			β_{11}	0.282	0.282	0.223
			β_{20}	0.149	0.297	0.230
			β_{21}	0.607	0.607	0.477

Tabela 25 – Medidas de eficiência do estimador Bayesiano para cada parâmetro do modelo k-DG, com $k = 1$, considerando diferentes funções de ligação para ω.

Função de Ligação	n	(τ_1, τ_2)	Parâmetros	$\sqrt{EQM(\hat{\beta}_{jk})/\beta_{jk}}$	$\sqrt{Var(\hat{\beta}_{jk})}$	$\mathscr{B}(\hat{\beta}_{jk})$
Logito	200	(1,1)	β_{10}	-0.245	0.242	0.194
			β_{11}	0.387	0.385	0.310
			β_{20}	0.241	0.481	0.382
			β_{21}	1.009	0.992	0.777
	500	(1,1)	β_{10}	-0.160	0.159	0.123
			β_{11}	0.251	0.250	0.198
			β_{20}	0.146	0.291	0.231
			β_{21}	0.586	0.585	0.468
Gumbel	200	(1,1)	β_{10}	-0.244	0.242	0.195
			β_{11}	0.375	0.373	0.298
			β_{20}	0.231	0.457	0.359
			β_{21}	1.008	0.993	0.760
	500	(1,1)	β_{10}	-0.166	0.166	0.132
			β_{11}	0.250	0.250	0.198
			β_{20}	0.145	0.290	0.231
			β_{21}	0.588	0.583	0.467

5.3 Modelo Poisson k-Modificado (k-MP)

Os estudos de simulação com conjunto de dados inflacionados e deflacionadados gerados a partir de modelos k-MP podem ser vistos a seguir.

5.3.1 Modelo Poisson k-Inflacionado (k-IP)

Para este estudo de simulação, conjuntos de dados k-inflacionados ($k = 0, 1$ e 3) foram gerados de um modelo k-MP com $0 < p(\boldsymbol{z}_i) < 1, \forall\, i = 1, \ldots, n$ (k-IP[9]), atribuindo os seguintes valores aos vetores de parâmetros: $\boldsymbol{\beta}_1 = (2, -2)$ e $\boldsymbol{\beta}_2 = (1, -5)$. Consideramos os parâmetros de precisão τ_1 e τ_2 entre 0.4 e 1. Como resultado intuitivo, temos que as estimativas Bayesianas dos vetores de parâmetros $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ (a média a *posteriori* para cada parâmetro) estão, em sua maioria, próximas dos verdadeiros valores quando observamos os baixos valores apresentados pelas medidas de eficiências (ver Tabelas 21, 22 e 23).

Ainda nas Tabelas 26, 27 e 28, apresentamos os resultados do estudo de simulação para dados 0-MP, 1-MP e 3-MP respectivamente, na qual podemos observar que todas as medidas de eficiência aproximam-se de zero quando o tamanho do conjunto de dados aumenta. Portanto, o procedimento Bayesiano para a estimação dos parâmetros torna-se mais preciso quanto maior o tamanho das amostras.

[9] Usaremos a sigla k-IP, do inglês *k-Inflated Poisson*, referindo-se à distribuição Poisson k-Inflacionada.

Tabela 26 – Medidas de eficiência do estimador Bayesiano para cada parâmetro do modelo k-IP, com $k = 0$, considerando diferentes funções de ligação para ω.

Função de Ligação	n	(τ_1, τ_2)	Parâmetros	$\sqrt{EQM(\hat{\beta}_{jk})/\beta_{jk}}$	$\sqrt{Var(\hat{\beta}_{jk})}$	$\mathscr{B}(\hat{\beta}_{jk})$
Logito	100	(1, 0.5)	β_{10}	0.071	0.143	0.110
			β_{11}	-0.315	0.616	0.485
			β_{20}	0.503	0.497	0.403
			β_{21}	-0.255	1.233	1.007
	500	(1, 0.5)	β_{10}	0.033	0.066	0.053
			β_{11}	-0.137	0.272	0.223
			β_{20}	0.213	0.212	0.170
			β_{21}	-0.111	0.544	0.437
CLogLog	100	(1, 0.5)	β_{10}	0.062	0.122	0.096
			β_{11}	-0.268	0.535	0.423
			β_{20}	0.462	0.451	0.319
			β_{21}	-0.272	1.285	0.939
	500	(1, 0.5)	β_{10}	0.026	0.052	0.042
			β_{11}	-0.120	0.239	0.189
			β_{20}	0.150	0.150	0.117
			β_{21}	-0.080	0.399	0.317
Gumbel	100	(1, 0.5)	β_{10}	0.099	0.199	0.151
			β_{11}	-0.688	1.363	1.049
			β_{20}	0.582	0.542	0.427
			β_{21}	-0.451	2.045	1.501
	500	(1, 0.5)	β_{10}	0.048	0.095	0.076
			β_{11}	-0.288	0.576	0.450
			β_{20}	0.216	0.211	0.167
			β_{21}	-0.151	0.723	0.561

Tabela 27 – Medidas de eficiência do estimador Bayesiano para cada parâmetro do modelo k-IP, com $k = 1$, considerando diferentes funções de ligação para ω.

Função de Ligação	n	(τ_1, τ_2)	Parâmetros	$\sqrt{EQM(\hat{\beta}_{jk})/\beta_{jk}}$	$\sqrt{Var(\hat{\beta}_{jk})}$	$\mathscr{B}(\hat{\beta}_{jk})$
Logito	100	(0.5, 0.5)	β_{10}	0.065	0.130	0.105
			β_{11}	-0.277	0.551	0.437
			β_{20}	0.512	0.499	0.394
			β_{21}	-0.281	1.315	1.062
	500	(0.5, 0.5)	β_{10}	0.033	0.064	0.053
			β_{11}	-0.121	0.242	0.197
			β_{20}	0.234	0.234	0.185
			β_{21}	-0.120	0.592	0.470
CLogLog	100	(0.5, 0.5)	β_{10}	0.069	0.138	0.110
			β_{11}	-0.257	0.508	0.412
			β_{20}	0.418	0.408	0.303
			β_{21}	-0.251	1.191	0.889
	500	(0.5, 0.5)	β_{10}	0.029	0.057	0.045
			β_{11}	-0.113	0.225	0.181
			β_{20}	0.152	0.152	0.123
			β_{21}	-0.086	0.430	0.336
Gumbel	100	(0.5, 0.5)	β_{10}	0.097	0.194	0.151
			β_{11}	-0.567	1.136	0.891
			β_{20}	0.522	0.479	0.405
			β_{21}	-0.393	1.766	1.461
	500	(0.5, 0.5)	β_{10}	0.036	0.073	0.058
			β_{11}	-0.246	0.492	0.390
			β_{20}	0.198	0.191	0.159
			β_{21}	-0.132	0.627	0.526

Tabela 28 – Medidas de eficiência do estimador Bayesiano para cada parâmetro do modelo k-IP, com $k = 3$, considerando diferentes funções de ligação para ω.

Função de Ligação	n	(τ_1, τ_2)	Parâmetros	$\sqrt{EQM(\hat{\beta}_{jk})/\beta_{jk}}$	$\sqrt{Var(\hat{\beta}_{jk})}$	$\mathscr{B}(\hat{\beta}_{jk})$
Logito	100	(0.5, 0.5)	β_{10}	0.066	0.133	0.105
			β_{11}	-0.277	0.540	0.432
			β_{20}	0.529	0.521	0.414
			β_{21}	-0.288	1.374	1.054
	500	(0.5, 0.5)	β_{10}	0.033	0.065	0.053
			β_{11}	-0.123	0.246	0.195
			β_{20}	0.209	0.209	0.164
			β_{21}	-0.105	0.517	0.411
CLogLog	100	(0.5, 0.5)	β_{10}	0.065	0.131	0.104
			β_{11}	-0.269	0.530	0.426
			β_{20}	0.416	0.406	0.317
			β_{21}	-0.243	1.134	0.910
	500	(0.5, 0.5)	β_{10}	0.029	0.059	0.046
			β_{11}	-0.113	0.225	0.183
			β_{20}	0.155	0.154	0.127
			β_{21}	-0.093	0.451	0.368
Gumbel	100	(1, 0.4)	β_{10}	0.085	0.169	0.135
			β_{11}	-0.527	1.055	0.828
			β_{20}	0.498	0.468	0.369
			β_{21}	-0.383	1.726	1.370
	500	(1, 0.4)	β_{10}	0.039	0.078	0.062
			β_{11}	-0.231	0.460	0.364
			β_{20}	0.201	0.196	0.160
			β_{21}	-0.139	0.672	0.528

5.3.2 Modelo Poisson k-Deflacionado (k-DP)

Para este estudo de simulação, conjuntos de dados k-deflacionados foram gerados de um modelo k-MP com$1 < p(\boldsymbol{z}_i) < 1/(1-\pi_P(y_i, \mu(\boldsymbol{x}_i)))$ (k-DP[10]), atribuindo os seguintes valores aos vetores de parâmetros: $\boldsymbol{\beta}_1 = (-0.5, 0.8)$ e $\boldsymbol{\beta}_2 = (1.5, 2.5)$. Consideramos os parâmetros de precisão $\tau_1 = \tau_2 = 0.5$ e as funções de ligação Logito e Gumbel. Como resultado, é possível afirmar que as estimativas Bayesianas de cada parâmetro dos vetores $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ (a média a *posteriori*) estão, em sua maioria, próximas dos verdadeiros valores (ver Tabelas 29 e 27).

Apresentamos nas Tabelas 29 e 27 os resultados do estudo de simulação para dados 0-MP e 1-MP respectivamente, na qual podemos observar que todas as medidas aproximam-se de zero quando o tamanho do conjunto de dados aumenta. Portanto, o procedimento Bayesiano para a estimação dos parâmetros torna-se mais preciso quanto maior o tamanho das amostras.

[10] Usaremos a sigla k-DP, do inglês *k-Deflated Poisson*, referindo-se à distribuição Poisson k-Deflacionada.

Tabela 29 – Medidas de eficiência do estimador Bayesiano para cada parâmetro do modelo k-DP, com $k = 0$, considerando diferentes funções de ligação para ω.

Função de Ligação	n	(τ_1, τ_2)	Parâmetros	$\sqrt{EQM(\hat{\beta}_{jk})/\beta_{jk}}$	$\sqrt{Var(\hat{\beta}_{jk})}$	$\mathscr{B}(\hat{\beta}_{jk})$
Logito	200	(0.5, 0.5)	β_{10}	-0.469	0.234	0.186
			β_{11}	0.473	0.378	0.300
			β_{20}	0.348	0.520	0.411
			β_{21}	0.555	1.359	1.052
	500	(0.5, 0.5)	β_{10}	-0.273	0.136	0.107
			β_{11}	0.271	0.217	0.171
			β_{20}	0.196	0.294	0.232
			β_{21}	0.280	0.696	0.549
Gumbel	200	(0.5, 0.5)	β_{10}	-0.522	0.258	0.203
			β_{11}	0.462	0.369	0.288
			β_{20}	0.377	0.563	0.425
			β_{21}	0.491	1.212	0.937
	500	(0.5, 0.5)	β_{10}	-0.264	0.132	0.105
			β_{11}	0.277	0.222	0.177
			β_{20}	0.158	0.237	0.187
			β_{21}	0.270	0.670	0.539

Tabela 30 – Medidas de eficiência do estimador Bayesiano para cada parâmetro do modelo k-DP, com $k = 1$, considerando diferentes funções de ligação para ω.

Função de Ligação	n	(τ_1, τ_2)	Parâmetros	$\sqrt{EQM(\hat{\beta}_{jk})/\beta_{jk}}$	$\sqrt{Var(\hat{\beta}_{jk})}$	$\mathscr{B}(\hat{\beta}_{jk})$
Logito	200	(0.5, 0.5)	β_{10}	-0.293	0.146	0.117
			β_{11}	0.292	0.233	0.188
			β_{20}	0.358	0.536	0.411
			β_{21}	0.521	1.290	0.992
	500	(0.5, 0.5)	β_{10}	-0.179	0.090	0.072
			β_{11}	0.179	0.143	0.115
			β_{20}	0.205	0.305	0.246
			β_{21}	0.289	0.720	0.577
Gumbel	200	(0.5, 0.5)	β_{10}	-0.284	0.141	0.112
			β_{11}	0.263	0.210	0.167
			β_{20}	0.306	0.455	0.361
			β_{21}	0.488	1.192	0.945
	500	(0.5, 0.5)	β_{10}	-0.184	0.092	0.072
			β_{11}	0.168	0.134	0.106
			β_{20}	0.165	0.247	0.194
			β_{21}	0.280	0.690	0.531

CAPÍTULO

6

APLICAÇÕES: MODELO k-MPS

Neste Capítulo apresentamos aplicações para os modelos de regressão k-MPS, considerando conjuntos de dados artificiais e conjuntos de dados reais.

6.1 Dados Artificiais

Para o estudo com dados artificiais, considere as n observações de uma variável explicativa X_1, ditas $x_1, x_2, \ldots, x_n$. Cada observação x_i, $i = 1, 2, \ldots, n$, foi gerada de uma distribuição Uniforme(0,1), $X_1 \sim U(0,1)$. Uma amostra de tamanho n ($n = 100$ e 200) foi gerada a partir da distribuição k-IPS (e da distribuição k-DPS) considerando as funções de ligação Logito, Complemento Log-log e Gumbel (funções de ligação Logito e Gumbel). As densidades a *priori* consideradas para os vetores de parâmetros $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ foram distribuições Normais Multivariadas, isto é, $\boldsymbol{\beta}_1 \sim NMult(\boldsymbol{\beta}_1^*, 1/(\tau_1 \boldsymbol{H}_1))$ e $\boldsymbol{\beta}_2 \sim NMult(\boldsymbol{\beta}_2^*, 1/(\tau_2 \boldsymbol{H}_2))$, em que $\boldsymbol{\beta}_i^*$ e $\boldsymbol{H}_i$ ($i = 1, 2$) são os hiperparâmetros associados, respectivamente, aos vetores de médias e as matrizes covariâncias, de dimensão 2×2 cada (ver Apêndice A). A fim de evitar repetições, consideramos cenários distintos para as distribuições PS aqui consideradas.

Considerando conjuntos de dados artificiais em diferentes cenários, selecionamos duas observações em cada conjunto para "pertubá-las", de maneira similar ao que foi feito em Cancho *et al.* (2011). Desta forma, para tornar as observações influentes no conjunto de dados consideramos três situações distintas: pertubar cada ponto individual e pertubar os dois pontos simultaneamente. A observação y_i foi pertubada da seguinte forma: $\tilde{y}_i = y_i + \delta s_y$, em que $\tilde{y}_i$ é a nova observação (um valor inteiro obtido através da função *round()*), δ é uma constante e s_y é o desvio padrão do conjunto de observações $\boldsymbol{y}$. O objetivo deste estudo é mostrar a necessidade de modelos robustos para lidar com a presença de pontos influentes nos conjuntos de dados.

6.1.1 Modelo k-MB

6.1.1.1 Modelo k-IB

Considerando o modelo k-MB, os valores atribuídos para os parâmetros que garante a $k-$inflação foram $\boldsymbol{\beta}_1^\top = (3,-2)$, $\boldsymbol{\beta}_2^\top = (-2,3)$ e $m = 10$. Foi considerado o ponto de modificação $k = 2$.

O modelo $k-$MB foi ajustado ao conjunto de dados gerados e as estimativas Bayesianas dos parâmetros, juntamente com os respectivos intervalos com 95% de credibilidade, são apresentados na Tabela 31. Analisando esta Tabela, observamos que as estimativas obtidas (média e mediana a *posteriori* ao considerar, respectivamente, a função de perda quadrática e a absoluta) estão próximas dos verdadeiros valores e todos os intervalos contêm o verdadeiro valor do parâmetro. Vale ressaltar que, através do critério Gelman-Rubin asseguramos a convergência das cadeias pois em todas as cadeias os valores aproximaram-se de 1.

Tabela 31 – Sumário a *posteriori* e intervalos com 95% de credibilidade dos parâmetros do modelo 2−IB, com $m = 10$.

Função de Ligação	Parâmetro	Valor Real	Média	Mediana	Desvio Padrão	IC(95%)
Logito	β_{10}	3	3.644	3.642	0.373	(2.937; 4.389)
	β_{11}	-2	-2.832	-2.827	0.484	(-3.791; -1.904)
	β_{20}	-2	-2.003	-1.994	0.415	(-2.830; -1.212)
	β_{21}	3	2.766	2.748	0.675	(1.468; 4.084)
C. Log-log	β_{10}	3	3.107	3.101	0.330	(2.479; 3.761)
	β_{11}	-2	-2.011	-2.011	0.424	(-2.850; -1.197)
	β_{20}	-2	-1.799	-1.793	0.330	(-2.465; -1.173)
	β_{21}	3	2.644	2.636	0.461	(1.744; 3.559)
Gumbel	β_{10}	3	3.516	3.502	0.573	(2.412; 4.677)
	β_{11}	-2	-2.679	-2.669	0.722	(-4.123; -1.287)
	β_{20}	-2	-2.120	-2.113	0.340	(-2.814; -1.498)
	β_{21}	3	3.313	3.293	0.605	(2.190; 4.530)

Considerando a função de ligação Gumbel, a distribuição de frequência dos dados gerados da distribuição 2-IB é apresentada na Tabela 32.

Tabela 32 – Distribuição de frequência da amostra gerada do modelo 2−IB, considerando a função de ligação Gumbel.

y_i	2	5	6	7	8	9	10
f_i	77	1	2	6	3	5	6

A Figura 8 apresenta os gráficos de algumas características dos dados, obtidos a partir das estimativas Bayesianas dos parâmetros (considerando apenas a média a *posteriori*) do modelo 2−IB ajustado. O Gráfico 8 **(A)** apresenta as probabilidades e os intervalos com 95% de credibilidade da ocorrência da observação $k = 2$ e notamos que conforme a covariável X_1 aumenta as probabilidades diminuem. Notamos ainda que as probabilidades estimadas da observação $k = 2$

estão próximas das verdadeiras. No Gráfico 8 **(B)** temos os valores verdadeiros, as estimativas Bayesianas de p e os seus respectivos intervalos com 95% de credibilidade. Observamos que para todo valor da covariável X_1, o conjunto de dados foi caracterizado como $2-$Inflacionado. No Gráfico 8 **(C)** apresentamos as médias verdadeiras, as médias ajustadas e os intervalos com 95% de credibilidade. Podemos observar uma concordância entre as curvas (a verdadeira e a ajustada).

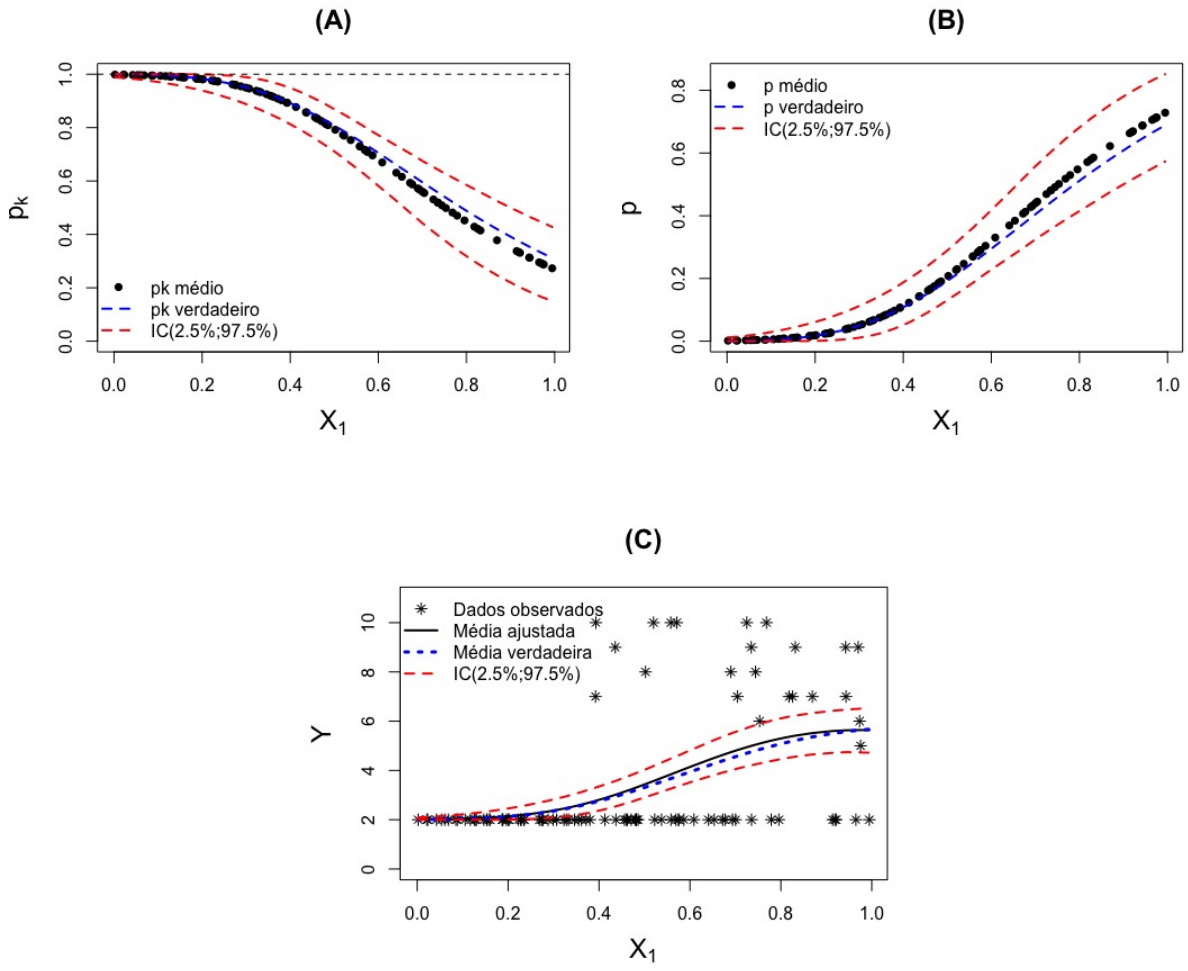

Figura 8 – Gráficos de algumas características do modelo Bayesiano $2-$IB ajustado, considerando a função de ligação Gumbel. **(A)** Estimativas da probabilidade de k. **(B)** Estimativas do parâmetro p. **(C)** Médias reais e ajustadas em função da covariável X_1.

Fonte: Elaborada pelo autor.

Ainda sobre os dados referentes a Tabela 32, temos uma amostra Binomial 2-Inflacionada, cuja média amostral corresponde a 3.420, a mediana é 2.000 e desvio padrão $s_y = 2.709$. Para a realização de um estudo de pontos influentes, consideramos $\delta = -0.8$ e selecionamos as observações das posições 2 e 15 para pertubá-las, ou seja, $y_2 = 2$ e $y_{15} = 2$. As estimativas Bayesianas dos parâmetros foram obtidas considerando os casos já descritos (cada observação pertubada individualmente e as duas observações pertubadas simultaneamente). A Tabela 33

apresenta os valores para cada β_{ij}, $i = 1,2$ e $j = 0,1$. Podemos observar nesta Tabela que o **caso a** é referente à análise Bayesiana original; os **casos b** e **c** referem-se às pertubações individuais das observações y_2 e y_{15}, respectivamente; já o **caso d** é referente à pertubação das duas observações conjuntamente. Observamos que, ao pertubar os pontos nos casos descritos acima, ocorreram impactos que resultaram em aumento ou redução das estimativas (a maioria deles com diferenças significativas). A Figura 9 apresenta as nuvens de pontos para os **casos a–d** considerados neste estudo.

Tabela 33 – Sumário a *posteriori* dos parâmetros do modelo 2−IB, considerando diferentes casos de pertubação.

Caso	Pertubação	β_{10}			β_{11}			β_{20}			β_{21}		
		Média	Mediana	SD	Média	Mediana	SD	Média	Mediana	SD	Média	Mediana	SD
a	[–]	3.516	3.502	0.573	-2.679	-2.669	0.722	-2.120	-2.113	0.340	3.313	3.293	0.605
b	[y_2]	0.474	0.472	0.328	1.201	1.207	0.464	-1.523	-1.523	0.226	2.349	2.348	0.422
c	[y_{15}]	0.681	0.678	0.361	0.882	0.885	0.507	-1.689	-1.684	0.255	2.644	2.628	0.472
d	[y_2,y_{15}]	-0.389	-0.384	0.285	2.327	2.326	0.422	-1.349	-1.345	0.214	2.096	2.091	0.409

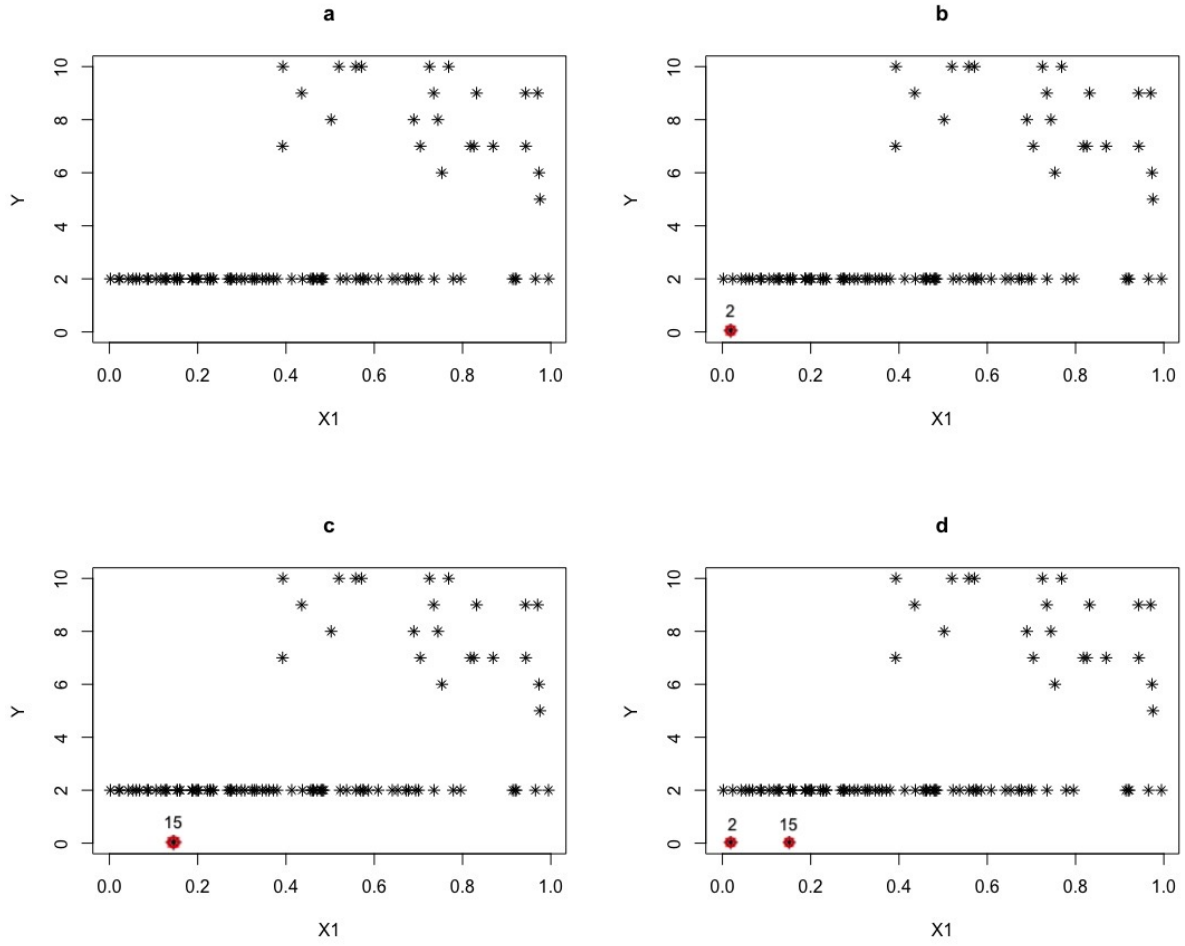

Figura 9 – Plotagem das nunvens de pontos nos diferentes casos de pertubação, considerando o modelo 2−IB ajustado.

Para avaliarmos a existência de ponto(s) influente(s), consideramos o valor da distância KL e a sua calibração ρ_i para cada caso e os resultados são apresentados na Tabela 34. É possível

notar que antes da pertubação (**caso a**) os valores da distância KL das observações y_2 e y_{15} eram baixos, e eles aumentaram de forma expressiva quando pertubados. Além disso, a calibração ρ_i de cada observação era próxima de 0.5 e, após as pertubações nos diferentes casos, os valores desta medida aproximaram-se de 1, levando-nos a concluir que estas observações pertubadas são pontos influentes, já que neste trabalho consideramos $\rho_i > 0.950$ como indicativo de ponto influente. Portanto, o procedimento adotado conseguiu de fato identificar os pontos discrepantes.

Tabela 34 – Distância KL e a sua calibração ρ_i para as observações pertubadas nos diferentes casos, considerando o modelo 2−IB ajustado.

Caso	Pertubação	KL		ρ_i	
		$[y_2]$	$[y_{15}]$	$[y_2]$	$[y_{15}]$
a	[–]	0.000	0.000	0.502	0.506
b	$[y_2]$	3.118	–	0.999	–
c	$[y_{15}]$	–	2.734	–	0.999
d	$[y_2,y_{15}]$	1.286	0.902	0.980	0.957

A Figura 10 ilustra a distância KL considerando os pontos de pertubação identificados nos **casos a – d**.

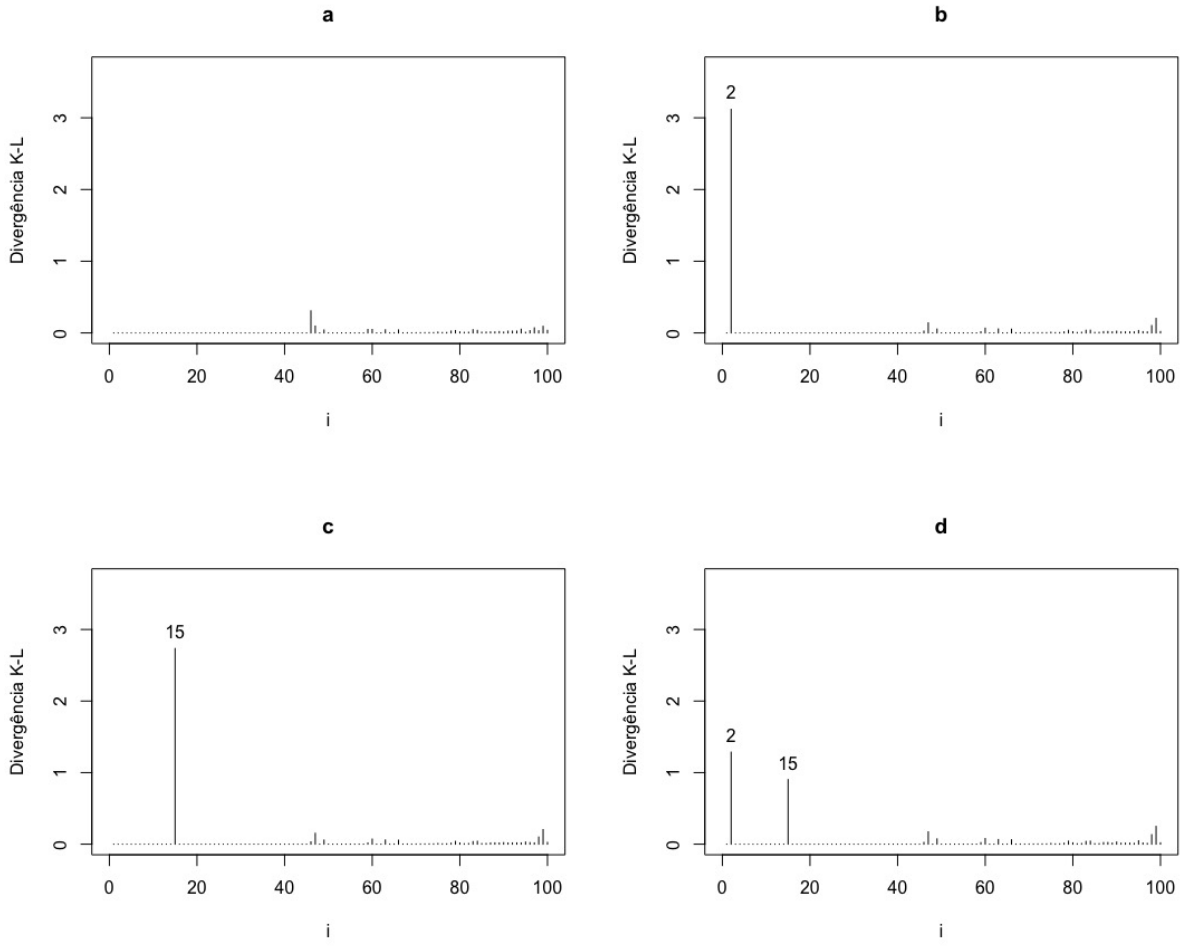

Figura 10 – Distância KL$(\pi, \pi_{(-i)})$ para os diferentes casos de pertubação, considerando o modelo 2−IB ajustado.

6.1.1.2 Modelo k-DB

Ainda considerando o modelo k-MB, os valores atribuídos para os parâmetros que garantem a $k-$deflação foram $\boldsymbol{\beta}_1^\intercal = (-2,-2)$, $\boldsymbol{\beta}_2^\intercal = (2,3)$ e $m = 10$. Aqui foi considerado o ponto de modificação $k = 0$.

Novamente, o modelo $k-$MB foi ajustado ao conjunto de dados gerados e as estimativas Bayesianas e seus respectivos intervalos com 95% de credibilidade, são apresentados na Tabela 35. Nesta Tabela, observamos que as estimativas obtidas (média e mediana a *posteriori*) também estão próximas dos verdadeiros valores e os intervalos contêm o verdadeiro valor do parâmetro. Através do critério Gelman-Rubin asseguramos a convergência das cadeias pois, novamente, em todas as cadeias esses valores aproximaram-se de 1.

Tabela 35 – Sumário a *posteriori* e intervalos com 95% de credibilidade dos parâmetros do modelo $0-$DB, com $m = 10$.

Função de Ligação	Parâmetro	Valor Real	Média	Mediana	Desvio Padrão	IC(95%)
Logito	β_{10}	-2	-1.938	-1.936	0.152	(-2.238; -1.647)
	β_{11}	-2	-2.200	-2.201	0.364	(-2.916; -1.496)
	β_{20}	2	1.826	1.821	0.431	(1.001; 2.694)
	β_{21}	3	3.489	3.413	1.238	(1.219; 6.064)
Gumbel	β_{10}	-2	-2.121	-2.120	0.164	(-2.444; -1.805)
	β_{11}	-2	-1.769	-1.771	0.348	(-2.431; -1.085
	β_{20}	2	2.138	2.113	0.459	(1.283; 3.096)
	β_{21}	3	3.198	3.175	1.212	(0.953; 5.623)

Assim como no caso anterior, também consideramos a função de ligação Gumbel e a distribuição de frequência da amostra 0-DB é apresentada na Tabela 36.

Tabela 36 – Distribuição de frequência da amostra gerada do modelo $0-$DB, considerando a função de ligação Gumbel.

y_i	0	1	2	3	4
f_i	7	148	42	2	1

Os gráficos com algumas características dos dados gerados (Tabela 36), obtidos a partir das estimativas Bayesianas (média a *posteriori*) dos parâmetros do modelo 0-DB são apresentados na Figura 11. O Gráfico 11 **(A)** apresenta as probabilidades e os intervalos com 95% de credibilidade da ocorrência da observação $k = 0$ e notamos que conforme a covariável X_1 aumenta as probabilidades diminuem e que as probabilidades estimadas da observação $k = 0$ também estão próximas das verdadeiras. Já no Gráfico 11 **(B)** temos os valores verdadeiros, as estimativas Bayesianas e os intervalos com 95% de credibilidade do parâmetro p. Ressaltamos que o conjunto de dados foi caracterizado como 0-Deflacionado, uma vez que para todo valor da covariável X_1, as estimativas Bayesianas e respectivos intervalos de credibilidade estão abaixo de 1. O Gráfico 11 **(C)** apresenta as médias verdadeiras, as médias ajustadas e os intervalos

com 95% de credibilidade. Percebemos novamente a concordância entre as curvas verdadeira e ajustada.

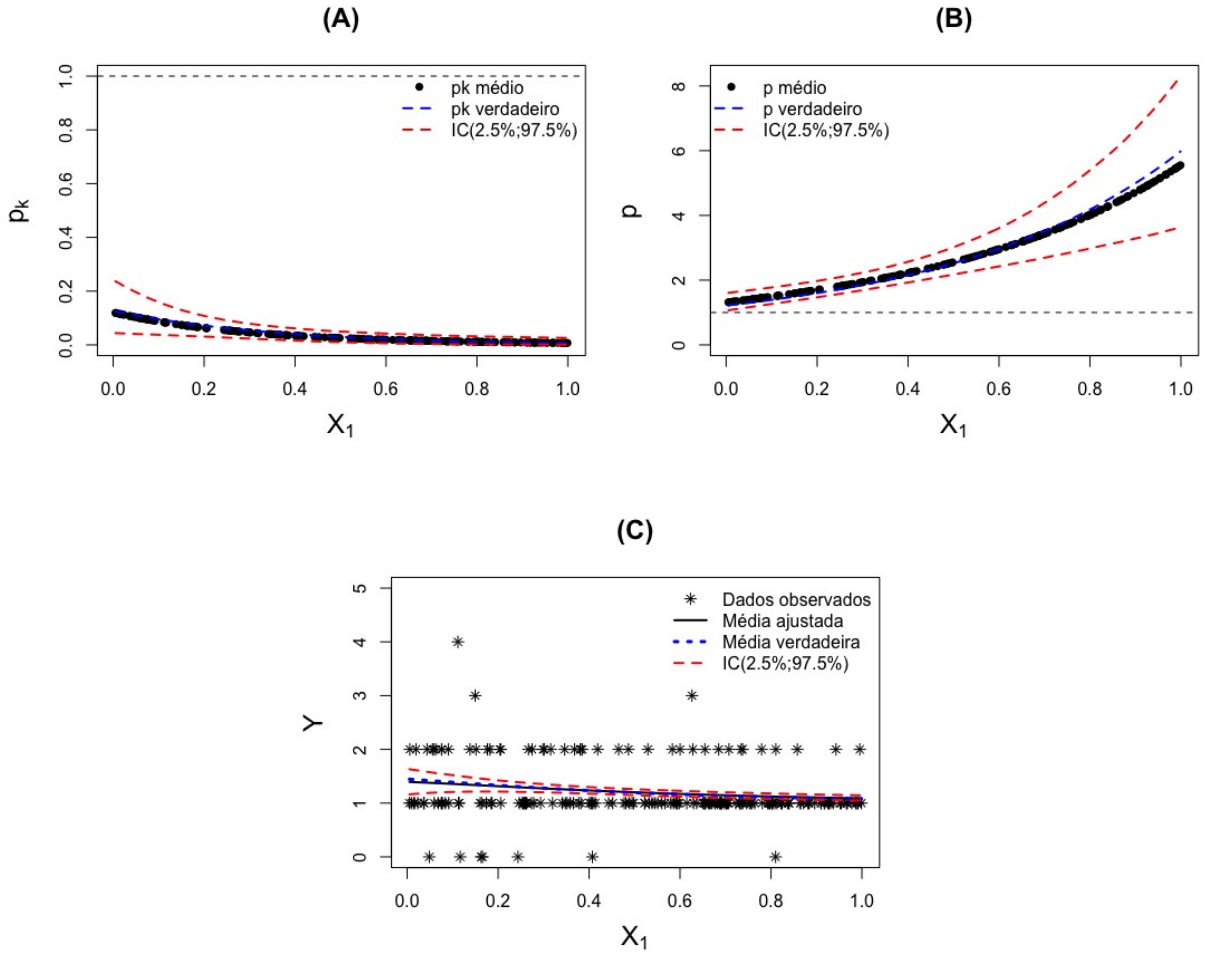

Figura 11 – Gráficos de algumas características do modelo Bayesiano 0−DB ajustado, considerando a função de ligação Gumbel. **(A)** Estimativas da probabilidade de k. **(B)** Estimativas do parâmetro p. **(C)** Médias reais e ajustadas em função da covariável X_1.

Fonte: Elaborada pelo autor.

Nos dados referentes à Tabela 36, temos uma amostra Binomial 0-Deflacionada com média amostral de 1.210, a mediana é 1.000 e desvio padrão $s_y = 0.536$. Para a realização do estudo de pontos influentes, consideramos $\delta = 14$ e selecionamos as observações das posições 199 e 200 para pertubá-las, isto é, $y_{199} = 2$ e $y_{200} = 1$. As estimativas Bayesianas dos parâmetros foram obtidas considerando os casos que já descrevemos anteriormente. Na Tabela 37 são apresentados os valores para os $\beta_{i,j}$, $i = 1,2$ e $j = 0,1$. Novamente, o **caso a** refere-se à análise Bayesiana original; os **casos b** e **c** referem-se às pertubações individuais das observações y_{199} e y_{200}; e o **caso d** é referente à pertubação das duas observações simultaneamente. Ressaltamos que, os pontos de pertubação resultaram em impactos de redução ou aumento nas estimativas

Bayesianas e, em alguns desses casos, essa diferença foi significativa. A Figura 12 apresenta as nuvens de pontos nos **casos a – d** considerados.

Tabela 37 – Sumário a *posteriori* dos parâmetros do modelo 0−DB, considerando diferentes casos de pertubação.

Caso	Pertubação	β_{10}			β_{11}			β_{20}			β_{21}		
		Média	Mediana	SD	Média	Mediana	SD	Média	Mediana	SD	Média	Mediana	SD
a	[–]	-2.121	-2.120	0.164	-1.769	-1.771	0.348	2.138	2.113	0.459	3.198	3.175	1.212
b	$[y_{199}]$	-2.380	-2.374	0.170	-0.786	-0.784	0.305	2.149	2.139	0.458	3.193	3.154	1.239
c	$[y_{200}]$	-2.385	-2.382	0.169	-0.778	-0.777	0.308	2.149	2.139	0.458	3.193	3.154	1.239
d	$[y_{199}, y_{200}]$	-2.608	-2.605	0.176	-0.067	-0.070	0.290	2.149	2.139	0.458	3.193	3.154	1.239

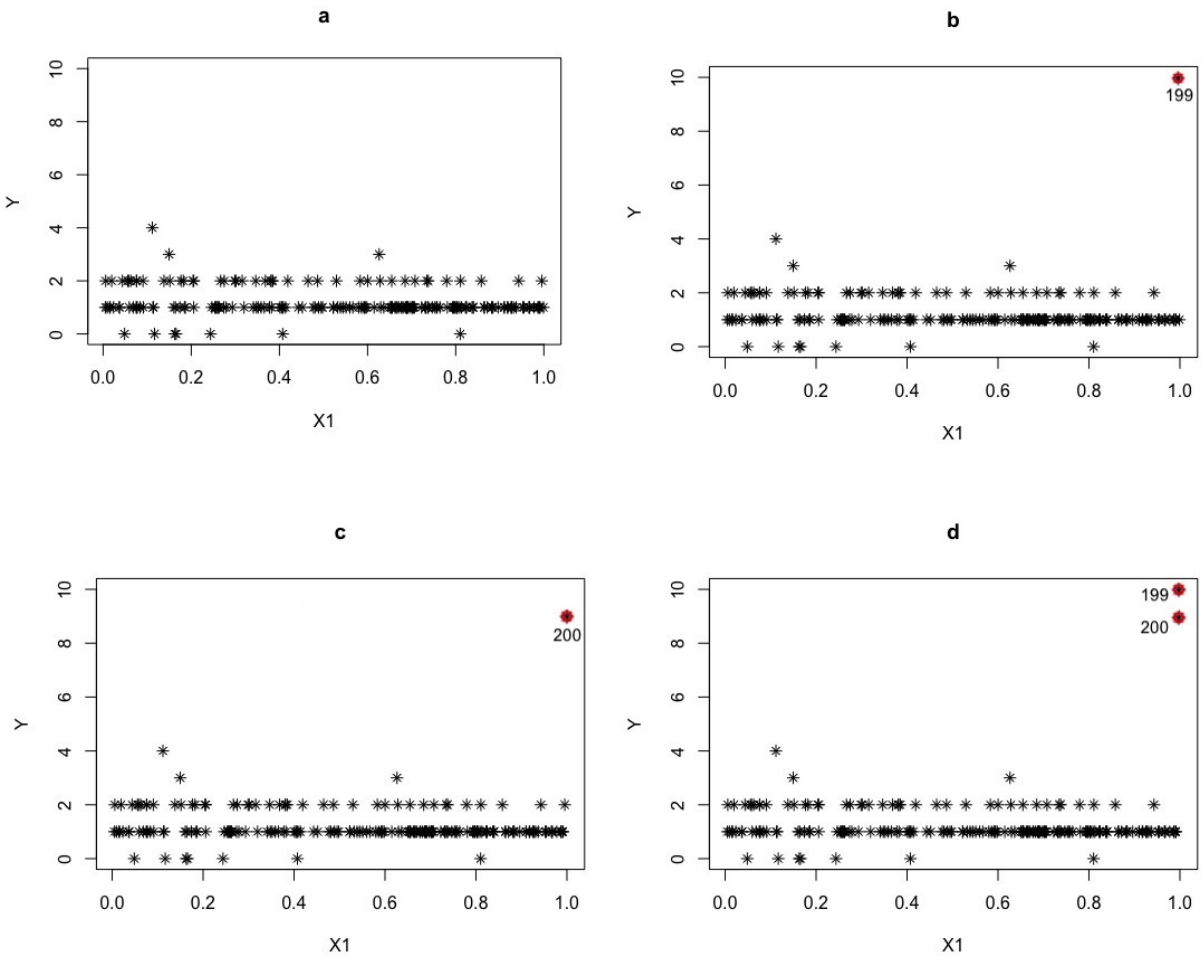

Figura 12 – Plotagem das nuvens de pontos nos diferentes casos de pertubação, considerando o modelo 0−DB ajustado.

Considerando novamente o valor da distância KL e a sua calibração ρ_i para cada um dos casos, podemos avaliar a existência de ponto(s) influente(s) na Tabela 38. Notamos que, antes da pertubação (**caso a**) os valores da distância KL nos pontos y_{199} e y_{200} eram 0.023 e 0.000, respectivamente, e eles aumentaram de forma expressiva quando pertubados. Além disso, a calibração ρ_i para as observações eram próximas de 0.5 e após as pertubações (**casos b–d**) os valores desta medida aproximam-se de 1, o que é um indicativo de que esses pontos pertubados podem ser considerados como influentes.

Tabela 38 – Distância KL e a sua calibração ρ_i para as observações pertubadas nos diferentes casos, considerando o modelo 0−DB ajustado.

Caso	Pertubação	KL		ρ_i	
		$[y_{199}]$	$[y_{200}]$	$[y_{199}]$	$[y_{200}]$
a	[–]	0.023	0.000	0.606	0.512
b	$[y_{199}]$	1.349	–	0.983	–
c	$[y_{200}]$	–	1.105	–	0.972
d	$[y_{199}, y_{200}]$	0.992	0.783	0.964	0.945

A Figura 13 ilustra a distância KL considerando os pontos de pertubação identificados nos **casos a – d**.

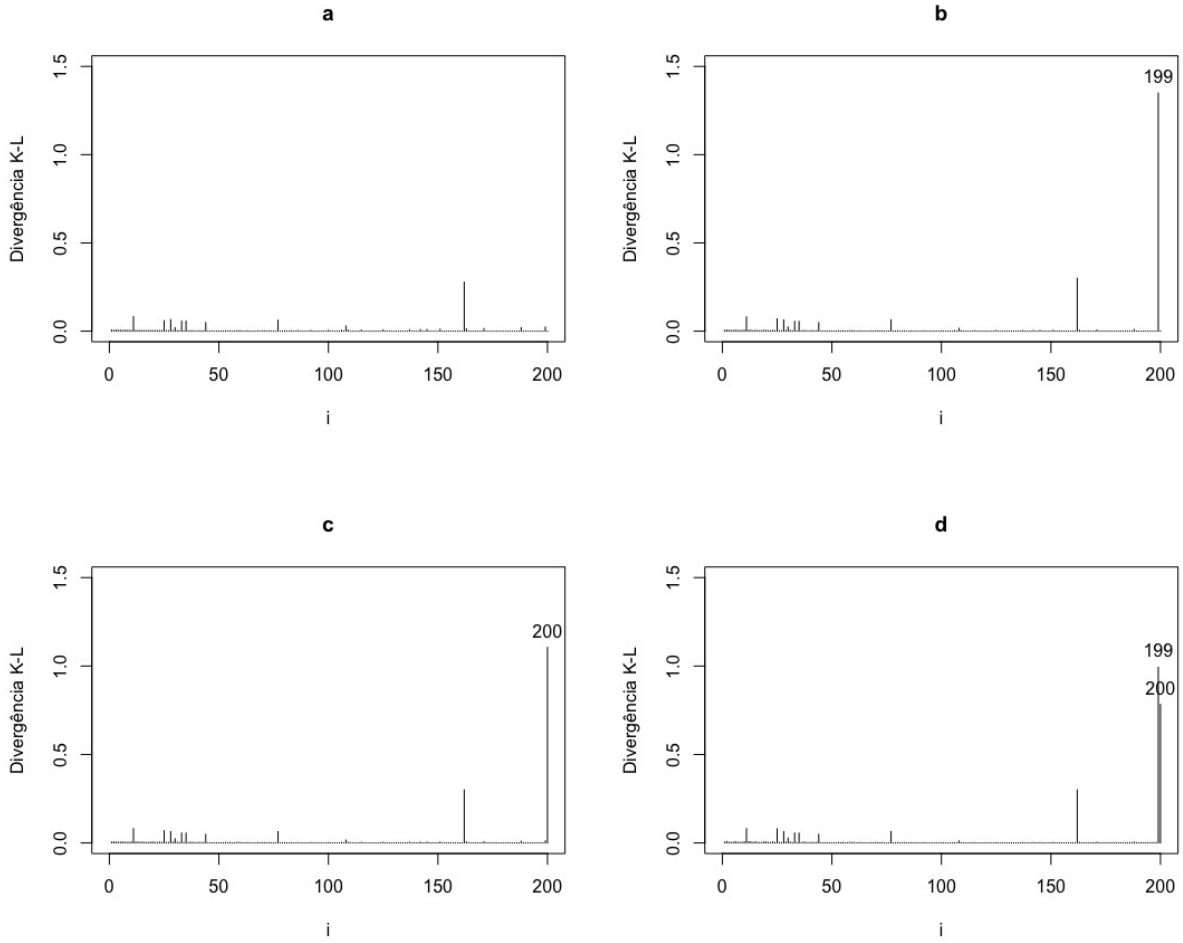

Figura 13 – Distância KL$(\pi, \pi_{(-i)})$ para os diferentes casos de pertubação, considerando o modelo 0−DB ajustado.

6.1.2 Modelo k-MG

6.1.2.1 Modelo k-IG

Considerando o modelo k-MG, os valores atribuídos para os parâmetros que garantem a $k-$inflação foram $\boldsymbol{\beta}_1^\intercal = (4,2)$ e $\boldsymbol{\beta}_2^\intercal = (-2,3)$. Foi considerado o ponto de modificação $k = 0$.

O modelo $k-$MG foi ajustado ao conjunto de dados gerados e as estimativas Bayesianas juntamente com os respectivos intervalos com 95% de credibilidade são apresentados na Tabela 39. Observamos que as estimativas obtidas (média e mediana a *posteriori* quando consideramos, respectivamente a função de perda quadrática e absoluta) estão próximas dos verdadeiros valores e os respectivos intervalos contêm o verdadeiro valor do parâmetro. Novamente, convergência das cadeias foi assegurado através do critério Gelman-Rubin, cujos valores aproximaram-se de 1.

Tabela 39 – Sumário a *posteriori* e intervalos com 95% de credibilidade dos parâmetros do modelo $0-$IG.

Função de Ligação	Parâmetro	Valor Real	Média	Mediana	Desvio Padrão	IC(95%)
Logito	β_{10}	4	3.685	3.680	0.282	(3.153; 4.248)
	β_{11}	2	2.583	2.583	0.397	(1.786; 3.353)
	β_{20}	-2	-1.944	-1.942	0.336	(-2.614; -1.294)
	β_{21}	3	3.102	3.095	0.561	(2.017; 4.238)
C. Log-log	β_{10}	4	4.199	4.193	0.279	(3.664; 4.766)
	β_{11}	2	1.780	1.784	0.415	(0.957; 2.584)
	β_{20}	-2	-2.178	-2.167	0.340	(-2.884; -1.551)
	β_{21}	3	3.638	3.620	0.572	(2.536; 4.814)
Gumbel	β_{10}	4	4.283	4.269	0.634	(3.084; 5.573)
	β_{11}	2	1.869	1.885	0.758	(0.349; 3.304)
	β_{20}	-2	-2.277	-2.267	0.342	(-2.960; -1.635)
	β_{21}	3	3.452	3.438	0.563	(2.402; 4.600)

Ao considerar a função de ligação Logito, a distribuição de frequências da amostra gerada do modelo $0-$IG é apresentada na Tabela 40.

Tabela 40 – Distribuição de frequência da amostra gerada do modelo $0-$IG, considerando a função de ligação Logito.

y_i	0	1	5	6	14	15	16	40	45	50	72	73	80	84	96	≥ 109
f_i	58	2	1	1	1	2	1	1	1	1	1	1	1	1	1	26

A Figura 14 apresenta os gráficos de algumas características dos dados (Tabela 40), obtidos a partir das estimativas Bayesianas dos parâmetros (considerando apenas a média a *posteriori*) do modelo $k-$IG. O Gráfico 14 **(A)** contém as probabilidades e os intervalos com 95% de credibilidade da ocorrência da observação $k = 0$ e é possível ver que, conforme a covariável X_1 aumenta as probabilidades diminuem. Além disso, podemos notar que as probabilidades estimadas da observação $k = 0$ estão próximas das verdadeiras. Já no Gráfico 14 **(B)** temos os valores verdadeiros, as estimativas Bayesianas e os intervalos com 95% de credibilidade do parâmetro p. Notamos que, para todo valor da covariável X_1, o conjunto de dados foi caracterizado como $0-$Inflacionado. O Gráfico 14 **(C)** apresenta as médias verdadeiras, as médias ajustadas e os intervalos com 95% de credibilidade, o qual apresentou um bom ajuste para os dados, dada a concordância entre as curvas.

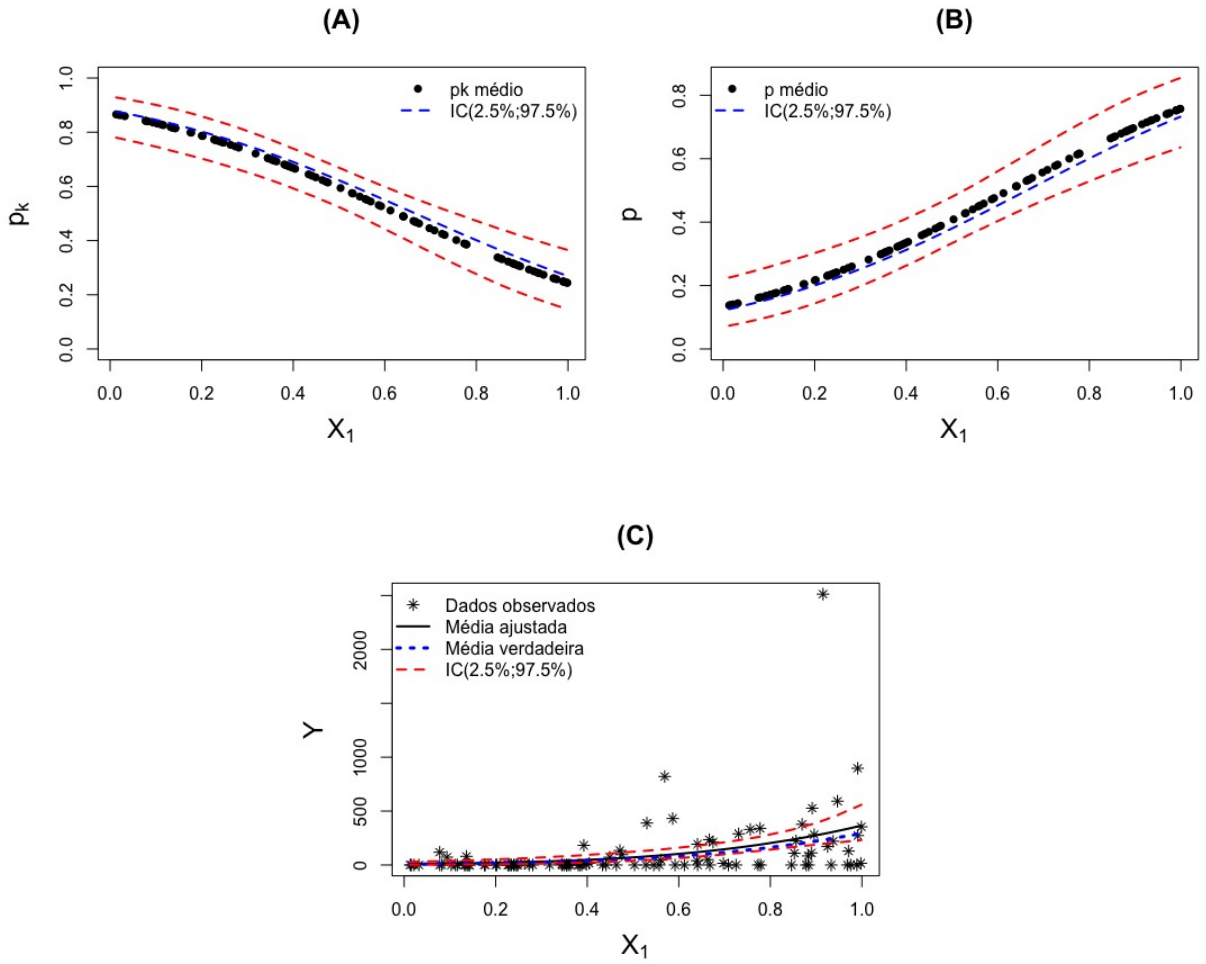

Figura 14 – Gráficos de algumas características do modelo Bayesiano 0−IG ajustado, considerando a função de ligação Logito. **(A)** Estimativas da probabilidade de k. **(B)** Estimativas do parâmetro p. **(C)** Médias reais e ajustadas em função da covariável X_1.

Fonte: Elaborada pelo autor.

Ainda considerando os dados referentes à Tabela 40, temos que média amostral é 110.650, mediana igual a 0.000 e desvio padrão $S_y = 296.015$. Para realizarmos um estudo de pontos influentes, foi considerado $\delta = 7$ e selecionamos as observações nas posições 5 e 88 para pertubá-las, isto é, $y_5 = 118$ e $y_{88} = 2513$. Assim, as estimativas Bayesianas dos parâmetros considerando os diferentes casos de pertubação (pertubando cada observação individualmente e com as duas observações sendo pertubadas de forma simultânea) foram obtidas. As estimativas para cada $\beta_{i,j}$, $i = 1,2$ e $j = 0,1$ são apresentadas na Tabela 41. Podemos observar que o **caso a** refere-se à análise Bayesiana original, enquanto os **casos b** e **c** são referentes às pertubações que ocorreram de forma individual nas observações y_5 e y_{88}; já o **caso d** é referente à pertubação simultânea das observações. Observamos que esses casos de pertubação influenciaram em aumento ou redução nessas estimativas (alguns deles com variações significativas). A Figura 15 mostra as nuvens de pontos nos **casos a – d**.

Tabela 41 – Sumário a *posteriori* dos parâmetros do modelo 0−IG, considerando diferentes casos de pertubação.

Caso	Pertubação	β_{10}			β_{11}			β_{20}			β_{21}		
		Média	Mediana	SD	Média	Mediana	SD	Média	Mediana	SD	Média	Mediana	SD
a	[–]	3.685	3.680	0.282	2.583	2.583	0.397	-1.944	-1.942	0.336	3.102	3.095	0.561
b	$[y_5]$	5.546	5.538	0.239	0.303	0.311	0.328	-1.935	-1.933	0.342	3.087	3.081	0.568
c	$[y_{88}]$	3.555	3.548	0.273	2.949	2.956	0.385	-1.935	-1.933	0.342	3.087	3.081	0.568
d	$[y_5, y_{88}]$	5.478	5.472	0.233	0.611	0.611	0.321	-1.935	-1.933	0.342	3.087	3.081	0.568

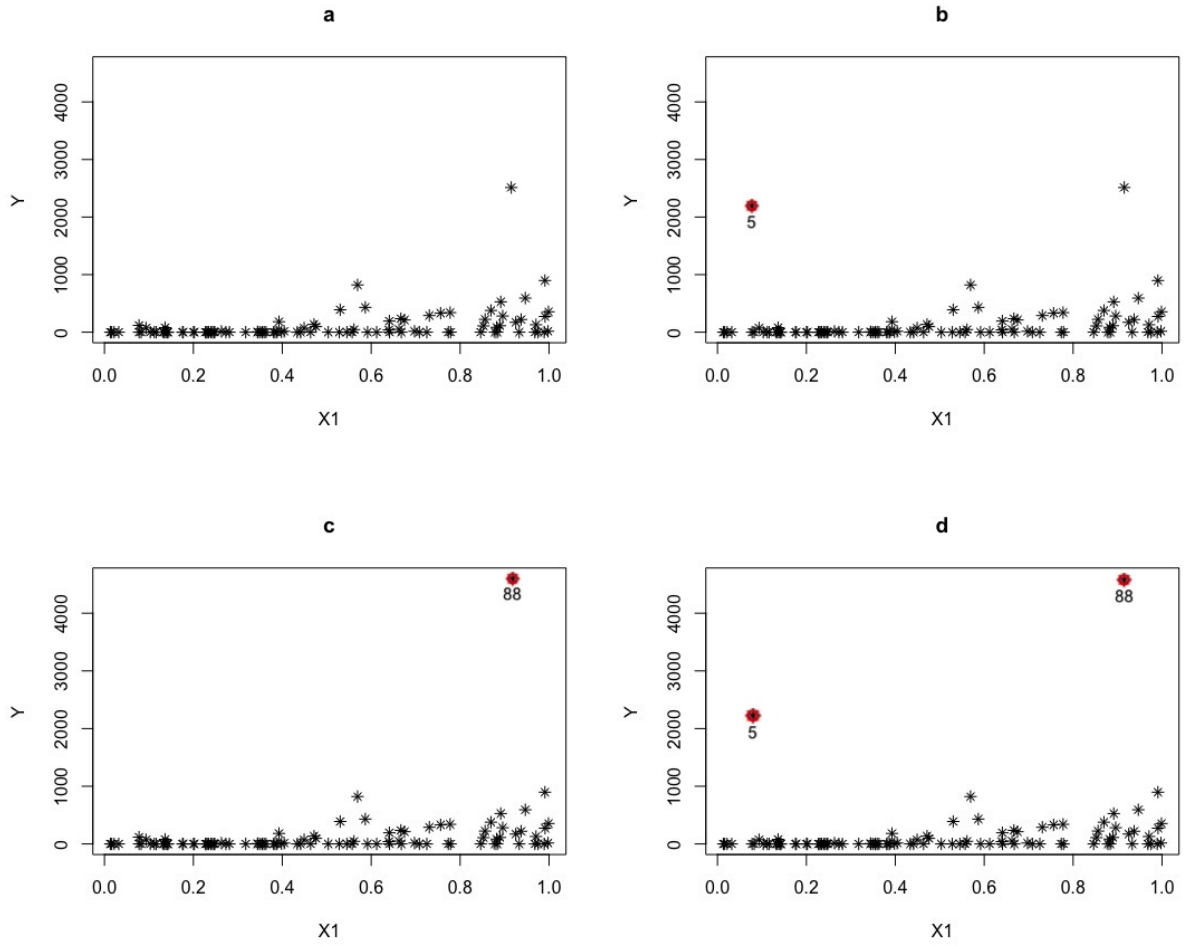

Figura 15 – Plotagem das nunvens de pontos nos diferentes casos de pertubação, considerando o modelo 0−IG ajustado.

Para avaliarmos a existência de ponto(s) influente(s), consideramos o valor da distância KL e a sua calibração ρ_i para os casos descritos acima e os resultados estão na Tabela 42. Podemos notar que no **caso a** (antes da pertubação) os valores da distância KL nas observações y_5 e y_{88} eram baixos e eles aumentaram significativamente quando pertubados. Além disso, a calibração ρ_i que a princípio era no máximo 0.852 nas observações, aproximando-se de 1 após os casos de pertubação (**casos b–d**), indicando que esses pontos pertubados são influentes, dado que consideramos $\rho_i > 0.950$ como indicativo de ponto influente. Logo, podemos concluir que o procedimento conseguiu identificar os pontos discrepantes.

Tabela 42 – Distância KL e a sua calibração ρ_i para as observações pertubadas nos diferentes casos, considerando o modelo $0-$IG ajustado.

Caso	Pertubação	KL		ρ_i	
		$[y_5]$	$[y_{88}]$	$[y_5]$	$[y_{88}]$
a	[–]	0.116	0.341	0.728	0.852
b	$[y_5]$	2.038	–	0.996	–
c	$[y_{88}]$	–	1.036	–	0.967
d	$y_5, y_{88}]$	2.253	1.174	0.997	0.976

A Figura 16 ilustra a distância KL considerando os pontos de pertubação identificados nos **casos a – d**.

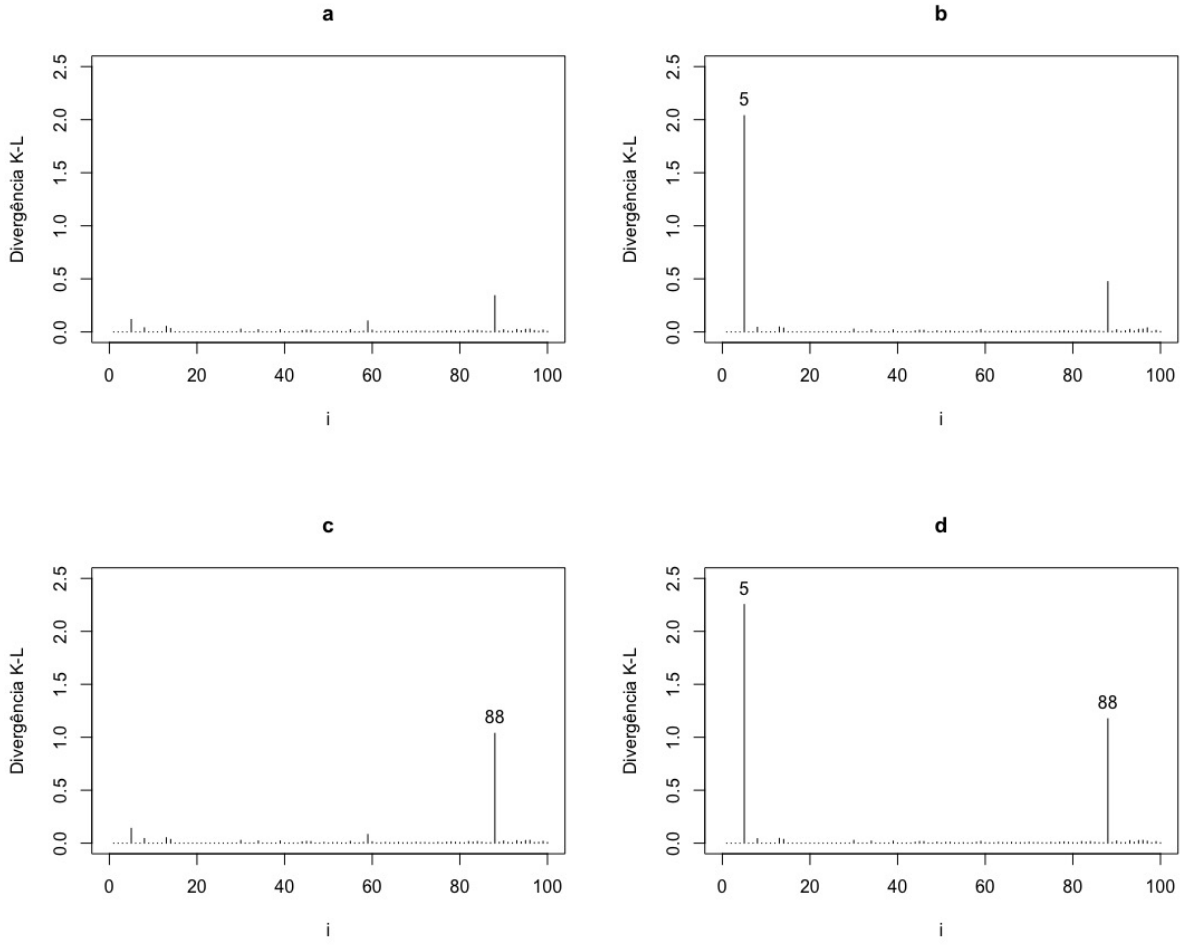

Figura 16 – Distância KL$(\pi, \pi_{(-i)})$ para os diferentes casos de pertubação, considerando o modelo $0-$IG ajustado.

6.1.2.2 Modelo k-DG

Ainda considerando o modelo k-MG, os valores atribuídos para os parâmetros que garante a $k-$deflação foram $\boldsymbol{\beta}_1^\top = (-1, 1)$ e $\boldsymbol{\beta}_2^\top = (2, 1)$. Foi considerado o ponto de modificação $k = 0$.

O modelo $k-$MG Bayesiano foi ajustado ao conjunto de dados gerados e as estimativas Bayesianas e os intervalos com 95% de credibilidade são apresentados na Tabela 43. As estimativas obtidas (média e mediana a *posteriori*) estão próximas dos verdadeiros valores e os intervalos contêm o verdadeiro valor do parâmetro. Através do critério Gelman-Rubin asseguramos a convergência das cadeias com valores aproximando-se de 1.

Tabela 43 – Sumário a *posteriori* e intervalos com 95% de credibilidade dos parâmetros do modelo $0-$DG.

Função de Ligação	Parâmetro	Valor Real	Média	Mediana	Desvio Padrão	IC(95%)
Logito	β_{10}	-1	-0.930	-0.930	0.182	(-1.288; -0.573)
	β_{11}	1	0.964	0.968	0.298	(0.373; 1.534)
	β_{20}	2	1.973	1.968	0.337	(1.309; 2.647)
	β_{21}	1	1.433	1.424	0.705	(0.074; 2.824)
Gumbel	β_{10}	-1	-1.119	-1.117	0.193	(-1.501; -0.752)
	β_{11}	1	1.410	1.406	0.299	(0.830; 2.009)
	β_{20}	2	1.739	1.735	0.271	(1.216; 2.289)
	β_{21}	1	1.165	1.161	0.553	(0.075; 2.254)

Considerando a função de ligação Logito, a distribuição de frequência dos dados gerado do modelo 0-DG é apresentada na Tabela 44:

Tabela 44 – Distribuição de frequência da amostra gerada do modelo $0-$DG, considerando a função de ligação Logito.

y_i	0	1	2	3	4	5	6
f_i	14	112	43	17	11	2	1

Na Figura 17 apresentamos os gráficos com algumas características dos dados, obtidos a partir das estimativas Bayesianas dos parâmetros do modelo $k-$DG (considerando apenas a média a *posteriori*). O Gráfico 17 **(A)** contém as probabilidades e os intervalos com 95% de credibilidade da ocorrência da observação $k = 0$ e vemos que conforme a covariável X_1 aumenta as probabilidades diminuem. Vemos também que as probabilidades estimadas da observação $k = 0$ estão próximas das verdadeiras. No Gráfico 17 **(B)** temos os valores verdadeiros, as estimativas Bayesianas e os intervalos com 95% de credibilidade do parâmetro p. Observe que para todo valor da covariável X_1, o conjunto de dados foi caracterizado como $0-$Inflacionado. Já o Gráfico 17 **(C)** apresenta as médias verdadeiras, as médias ajustadas e os intervalos com 95% de credibilidade, e podemos notar uma boa concordância das curvas.

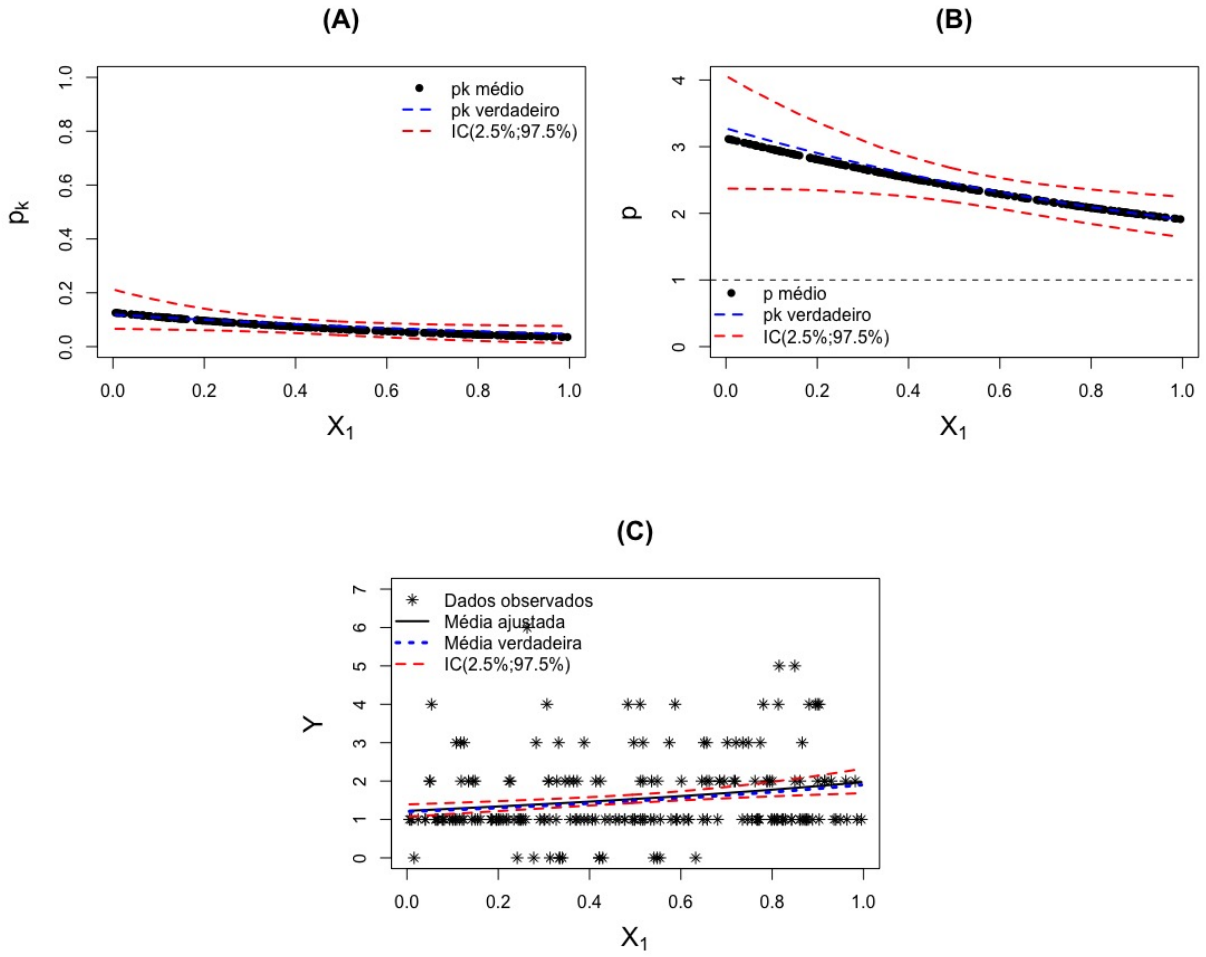

Figura 17 – Gráficos de algumas características do modelo Bayesiano 0−DG ajustado, considerando a função de ligação Logito. **(A)** Estimativas da probabilidade de k. **(B)** Estimativas do parâmetro p. **(C)** Médias reais e ajustadas em função da covariável X_1.

Fonte: Elaborada pelo autor.

Considerando o conjunto de dados da Tabela 44, temos que a média amostral é 1.545, a mediana igual a 1.000 e desvio padrão $S_y = 1.055$. Novamente, para realizar um estudo de pontos influentes, consideramos $\delta = 15$ e selecionamos as observações das posições 9 e 22 para pertubá-las, ou seja, as observações $y_9 = 2$ e $y_{22} = 3$. Dados os casos de pertubação já descritos, a Tabela 45 apresenta as estimativas Bayesianas para cada parâmetro $\beta_{i,j}$, $i = 1,2$ e $j = 0,1$. O **caso a** é referente à análise Bayesiana original, enquanto os **casos b** e **c** referem-se às pertubações individuais das observações y_9 e y_{22}, respectivamente; já o **caso d** é referente à pertubação simultânea das duas observações. Vemos que os **casos b – d** de pertubação resultaram em impactos na redução ou aumento das estimativas, sendo alguns com diferenças significativas. A Figura 18 apresenta as nuvens de pontos nos **casos a – d**.

Tabela 45 – Sumário a *posteriori* dos parâmetros do modelo 0−DG, considerando diferentes casos de pertubação.

Caso	Pertubação	β_{10}			β_{11}			β_{20}			β_{21}		
		Média	Mediana	SD	Média	Mediana	SD	Média	Mediana	SD	Média	Mediana	SD
a	[–]	-0.930	-0.930	0.182	0.964	0.968	0.298	1.973	1.968	0.337	1.433	1.424	0.705
b	$[y_9]$	-0.391	-0.393	0.158	0.197	0.202	0.268	1.985	1.975	0.349	1.411	1.407	0.727
c	$[y_{22}]$	-0.432	-0.430	0.160	0.279	0.279	0.269	1.985	1.975	0.349	1.411	1.407	0.727
d	$[y_9, y_{22}]$	-0.391	-0.393	0.158	0.197	0.202	0.268	1.985	1.975	0.349	1.411	1.407	0.727

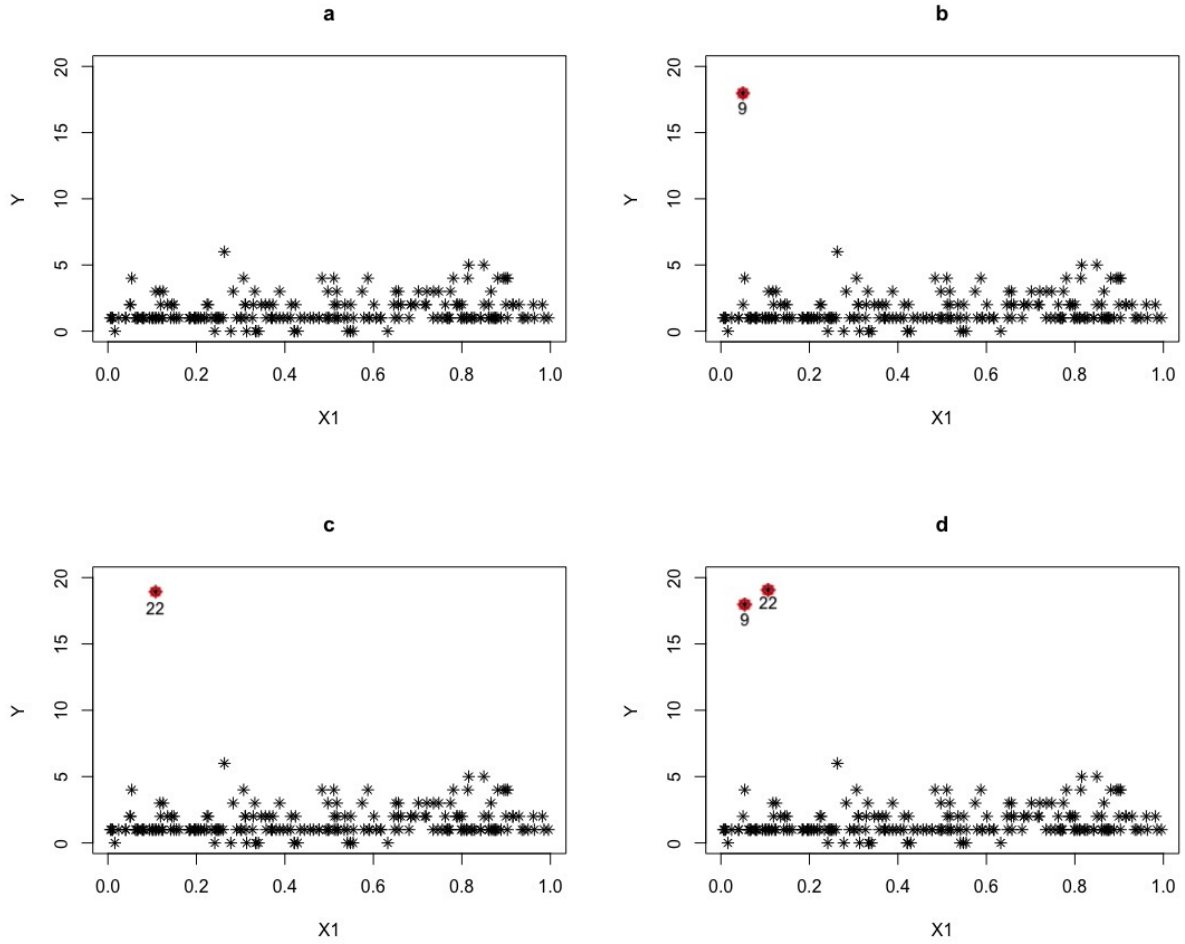

Figura 18 – Plotagem das nunvens de pontos nos diferentes casos de pertubação, considerando o modelo 0−DB ajustado.

Para avaliarmos a existência de ponto(s) influente(s), consideramos novamente o valor da distância KL e a sua calibração ρ_i para os diferentes casos de pertubação. Os resultados deste estudo são apresentados na Tabela 46. Notamos que antes da pertubação (**caso a**) a distância KL era aproximadamente 0 nos pontos y_2 e y_{15}, e os valores dessa medida aumentaram após os casos de pertubação. Além disso, antes da pertubação a calibração ρ_i era próxima a 0.5 em cada um desses pontos e nos casos **casos b – d** os valores dessa medida aproximaram-se de 1 (aqui consideramos pontos influentes aqueles cujo $\rho_i > 0.950$). Dessa forma, concluimos que nestes casos os pontos pertubados foram considerados pontos influentes. Portanto, o método adotado conseguiu mais uma vez identificar os pontos discrepantes.

Tabela 46 – Distância KL e a sua calibração ρ_i para as observações pertubadas nos diferentes casos, considerando o modelo 0−DG ajustado.

Caso	Pertubação	KL		ρ_i	
		$[y_9]$	$[y_{22}]$	$[y_9]$	$[y_{22}]$
a	[–]	0.003	0.015	0.540	0.585
b	$[y_9]$	1.088	–	0.971	–
c	$[y_{22}]$	–	1.072	–	0.970
d	$[y_9, y_{22}]$	1.088	1.005	0.971	0.965

A Figura 19 apresenta a distância KL considerando os **casos a – d** e pontos de pertubação identificados.

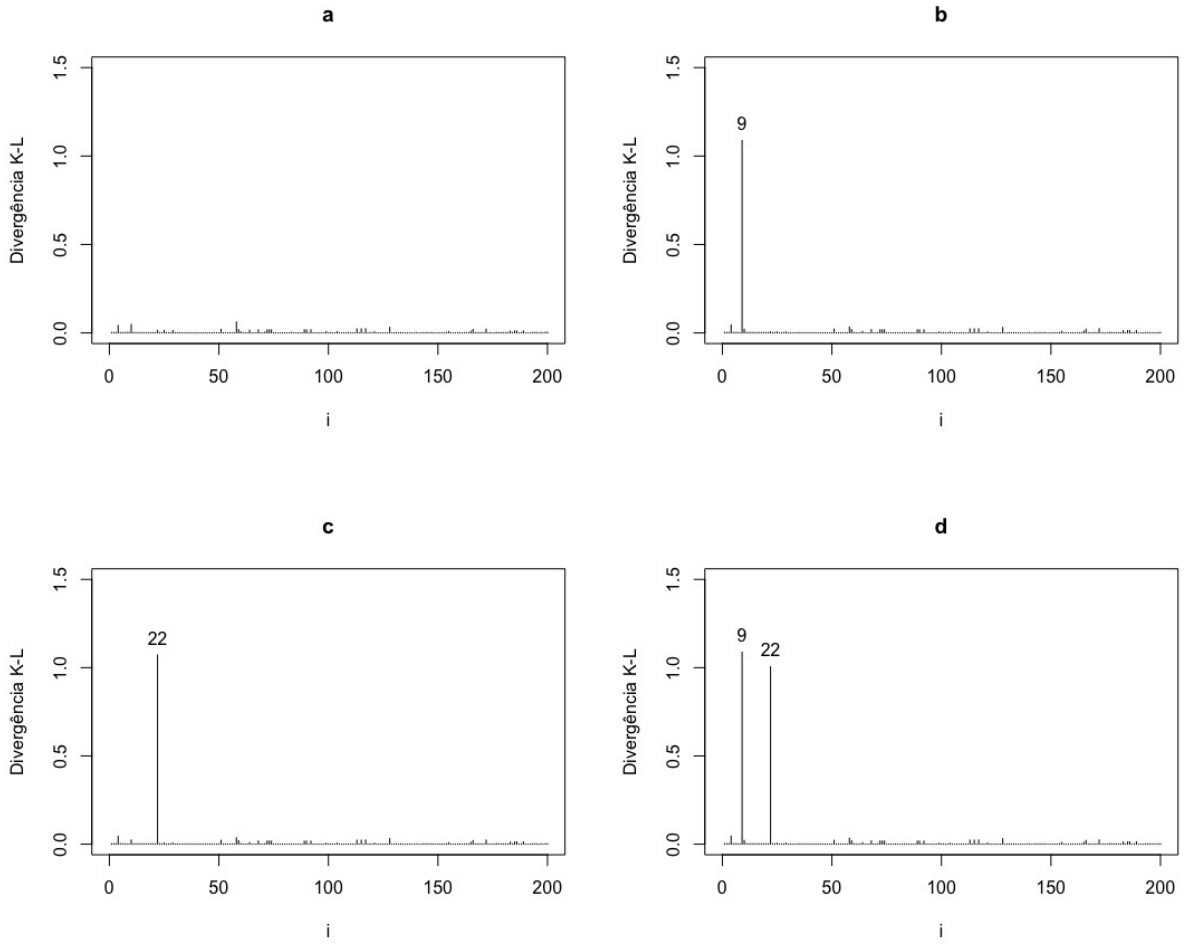

Figura 19 – Distância KL$(\pi, \pi_{(-i)})$ para os diferentes casos de pertubação, considerando o modelo 0−DG ajustado.

6.1.3 Modelo k-MP

6.1.3.1 Modelo k-IP

Para o estudo com o modelo k-MP, os valores atribuídos aos parâmetros que garantem a $k-$inflação foram $\boldsymbol{\beta}_1^\intercal = (2,-2)$ e $\boldsymbol{\beta}_2^\intercal = (1,-5)$. Foi considerado o ponto de modificação $k = 1$.

O modelo $k-$MP foi ajustado ao conjunto de dados gerados e as estimativas Bayesianas dos parâmetros, juntamente com os respectivos intervalos com 95% de credibilidade, são apresentados na Tabela 47. Vale observar que as estimativas Bayesianas obtidas através da média e mediana a *posteriori* (considerando a função de perda quadrática e a absoluta, respectivamente) estão próximas dos verdadeiros valores e os intervalos de credibilidade contêm os verdadeiros valores dos parâmetros. Ressaltamos que, através do critério Gelman-Rubin asseguramos a convergência das cadeias, pois em todas elas os valores aproximaram-se de 1.

Tabela 47 – Sumário a *posteriori* e intervalos com 95% de credibilidade dos parâmetros do modelo $1-$IP.

Função de Ligação	Parâmetro	Valor Real	Média	Mediana	Desvio Padrão	IC(95%)
Logito	β_{10}	2	2.101	2.104	0.154	(1.793; 2.394)
	β_{11}	-2	-2.425	-2.412	0.606	(-3.620; -1.260)
	β_{20}	1	1.055	1.036	0.487	(0.144; 2.027)
	β_{21}	-5	-5.495	-5.414	1.274	(-8.204; -3.198)
C. Log-log	β_{10}	2	2.074	2.076	0.103	(1.870; 2.272)
	β_{11}	-2	-1.936	-1.939	0.455	(-2.830; -1.064)
	β_{20}	1	1.027	1.030	0.288	(0.449; 1.578)
	β_{21}	-5	-5.091	-5.052	0.860	(-6.854; -3.478)
Gumbel	β_{10}	2	1.871	1.875	0.166	(1.540; 2.185)
	β_{11}	-2	-1.492	-1.473	0.991	(-3.461; 0.413)
	β_{20}	1	1.063	1.045	0.387	(0.367; 1.876)
	β_{21}	-5	-5.101	-5.024	1.240	(-7.656; -2.899)

Considerando neste caso a função de ligação Complemento Log-log, a distribuição de frequência dos dados gerados do modelo $1-$IP é apresentada na Tabela 48.

Tabela 48 – Distribuição de frequência da amostra gerada do modelo $1-$IP, considerando a função de ligação Complemento Log-log.

y_i	1	2	3	4	5	6	7	8	9	10	12
f_i	66	5	4	5	3	3	6	3	2	1	2

A Figura 20 apresenta os gráficos com algumas características da amostra, obtidas a partir das estimativas dos parâmetros (considerando apenas a média a *posteriori*) do modelo $k-$IP ajustado. O gráfico 20 **(A)** apresenta as probabilidades e os intervalos com 95% de credibilidade da ocorrência da observação $k = 1$ e notamos que conforme a covariável X_1 aumenta as probabilidades também aumentam. Observe que as probabilidades estimadas da observação $k = 1$ estão próximas das verdadeiras. No gráfico 20 **(B)** temos os valores verdadeiros, as estimativas Bayesianas de p e os respectivos intervalos com 95% de credibilidade. Observe

que para todo valor da covariável X_1, o conjunto de dados foi caracterizado como 1−Inflacionado. Já o gráfico 20 **(C)** apresenta as médias verdadeiras, as médias ajustadas e os intervalos com 95% de credibilidade, o qual apresentou um bom ajuste do modelo aos dados.

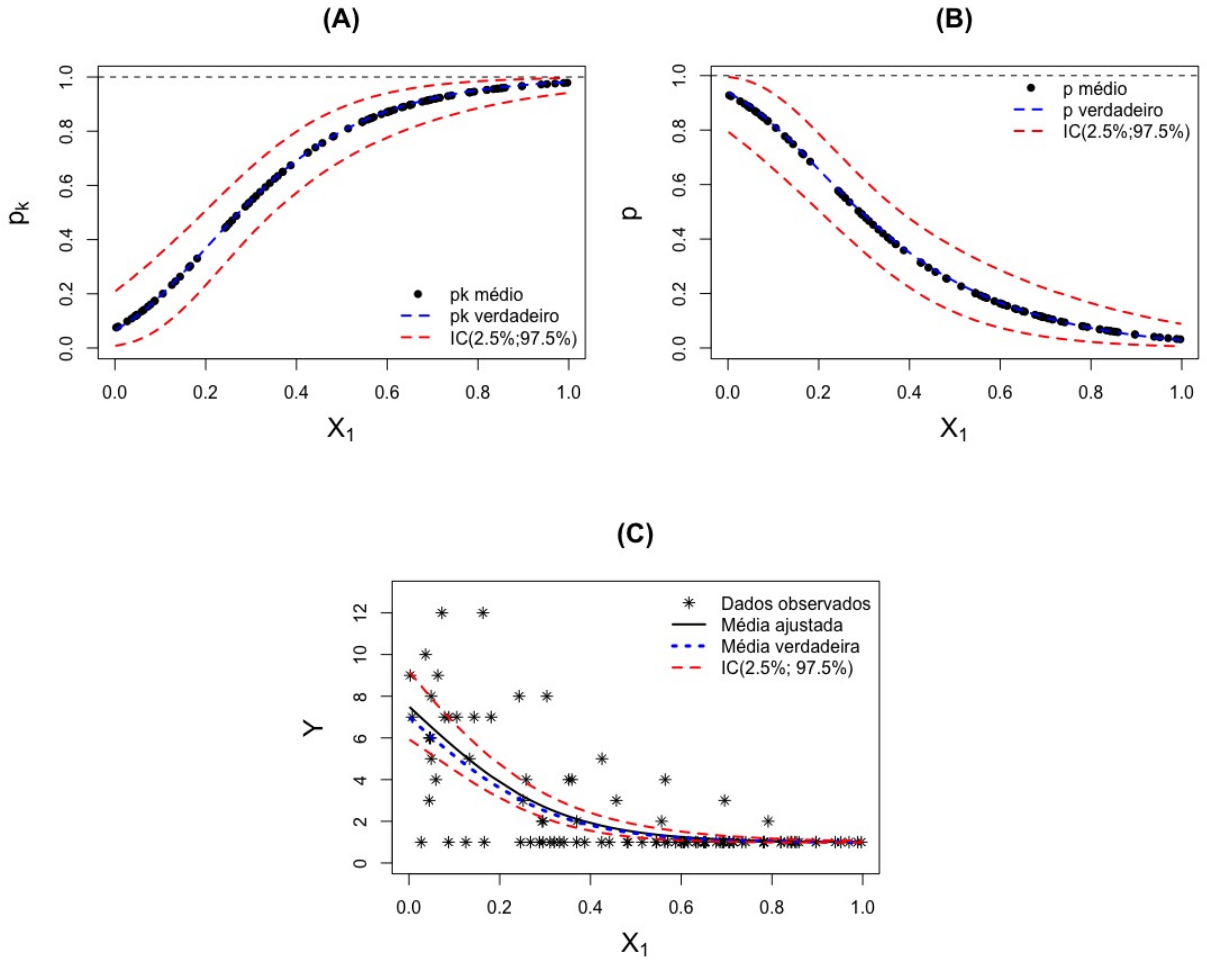

Figura 20 – Gráficos de algumas características do modelo Bayesiano 1−IP ajustado, considerando a função de ligação Complemento Log-log. **(A)** Estimativas da probabilidade de k. **(B)** Estimativas do parâmetro p. **(C)** Médias reais e ajustadas em função da covariável X_1.

Considerando ainda os dados apresentados na Tabela 48, temos uma amostra 1-IP com média amostral de 2.590, a mediana é 1.000 e desvio padrão $s_y = 2.753$. Para o estudo de pontos influentes, selecionamos as observações 44 e 85 para pertubá-las, isto é $y_{44} = 5$ e $y_{85} = 2$, e consideramos $\delta = 5$. A Tabela 49 apresenta as estimativas Bayesianas para cada $\beta_{i,j}$, $i = 1,2$ e $j = 0,1$, considerando os casos já descritos (pertubando cada observação de forma individual e as duas observação sendo pertubadas simultaneamente). Observamos que o **caso a** é referente à análise Bayesiana original; já os **casos b** e **c** referem-se às pertubações individuais das observações y_{44} e y_{85}, respectivamente; e o **caso d** é referente à pertubação nas duas observações simultaneamente. Ressaltamos que, ao pertubar as observações nos casos descritos acima, ocorreram impactos de aumento ou redução das estimativas. A Figura 21 mostra as nuvens de pontos nos **casos a – d**.

Tabela 49 – Sumário a *posteriori* dos parâmetros do modelo 1−IP, considerando diferentes casos de pertubação.

Caso	Pertubação	β_{10}			β_{11}			β_{20}			β_{21}		
		Média	Mediana	SD	Média	Mediana	SD	Média	Mediana	SD	Média	Mediana	SD
a	[–]	2.074	2.076	0.103	-1.936	-1.939	0.455	1.027	1.030	0.288	-5.091	-5.052	0.860
b	$[y_{44}]$	2.037	2.038	0.102	-1.266	-1.261	0.415	1.020	1.025	0.288	-5.084	-5.063	0.863
c	$[y_{85}]$	1.888	1.888	0.105	-0.473	-0.467	0.375	1.020	1.025	0.288	-5.084	-5.063	0.863
d	$[y_{44}, y_{85}]$	1.875	1.875	0.103	-0.083	-0.072	0.346	1.020	1.025	0.288	-5.084	-5.063	0.863

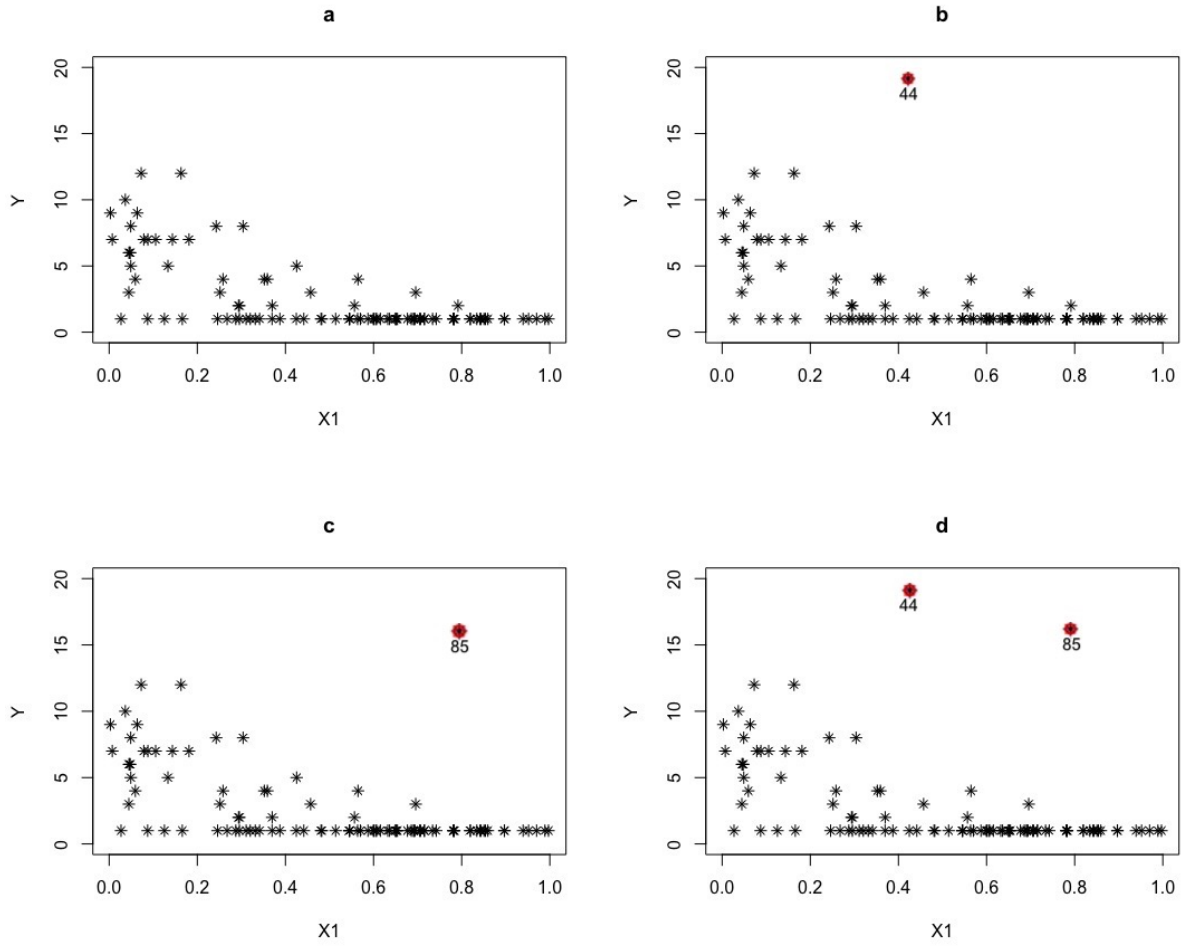

Figura 21 – Plotagem das nunvens de pontos nos diferentes casos de pertubação, considerando o modelo 1−IP ajustado.

Para avaliarmos a existência de ponto(s) influente(s), consideramos novamente o valor da distância KL e a sua calibração ρ_i para cada um dos casos descritos acima e os resultados são apresentados na Tabela 50. Novamente, notamos que antes da pertubação (**caso a**) os valores da distância KL nas observações y_{44} e y_{85} eram baixos e eles aumentaram expressivamente quando pertubados. Além disso, os valores da calibração ρ_i nos **casos b – d** de pertubação aproximam-se de 1, indicando que estes são considerados pontos influentes (consideramos $\rho_i > 0.950$ como indicativo de ponto influente). Portanto, que o procedimento utilizado conseguiu identificar os pontos discrepantes.

Tabela 50 – Distância KL e a sua calibração ρ_i para as observações pertubadas nos diferentes casos, considerando o modelo 1−IP ajustado.

Caso	Pertubação	KL		ρ_i	
		$[y_{44}]$	$[y_{85}]$	$[y_{44}]$	$[y_{85}]$
a	[–]	0.039	0.136	0.637	0.744
b	$[y_{44}]$	1.949	–	0.995	–
c	$[y_{85}]$	–	5.181	–	1.000
d	$[y_{44}, y_{85}]$	0.939	3.366	0.960	1.000

A Figura 22 ilustra a distância KL considerando os pontos de pertubação identificados nos **casos a – d**.

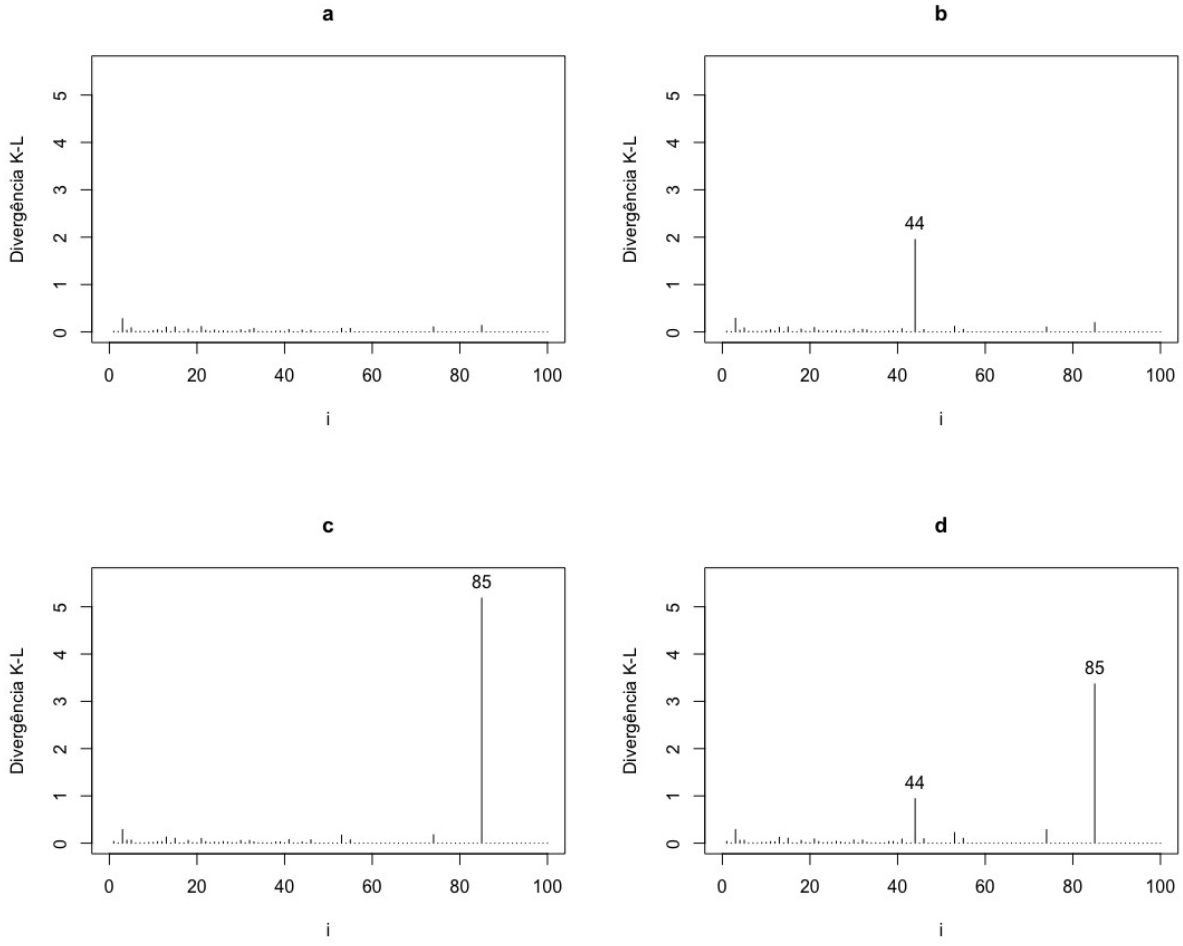

Figura 22 – Distância KL$(\pi, \pi_{(-i)})$ para os diferentes casos de pertubação, considerando o modelo 1−IP ajustado.

6.1.3.2 Modelo k-DP

Novamente, considerando o modelo k-MP, os valores atribuídos para os parâmetros que garantem a $k-$deflação foram $\boldsymbol{\beta}_1^\top = (-0.5, 0.8)$ e $\boldsymbol{\beta}_2^\top = (1.5, 2, 5)$. Foi considerado o ponto de modificação $k = 1$.

O modelo $k-$MP foi ajustado ao conjunto de dados gerados e as estimativas Bayesianas dos parâmetros juntamente com os respectivos intervalos com 95% de credibilidade são apresentados na Tabela 51. Podemos notar que as estimativas Bayesianas obtidas (média e mediana a *posteriori* ao considerar as funções de perda quadrática e Bayesiana, respectivamente) estão próximas dos verdadeiros valores, além dos intervalos de credibilidades conter os respectivos valores verdadeiros dos parâmetros. Assim como em todos os casos anteriores, convergência das cadeias foi assegurada pelo critério Gelman-Rubin dado que em todas as cadeias os valores aproximaram-se de 1.

Tabela 51 – Sumário a *posteriori* e intervalos com 95% de credibilidade dos parâmetros do modelo $1-$DP.

Função de Ligação	Parâmetro	Valor Real	Média	Mediana	Desvio Padrão	IC(95%)
Logito	β_{10}	-0.5	-0.510	-0.507	0.144	(-0.800; -0.233)
	β_{11}	0.8	0.770	0.768	0.227	(0.328; 1.225)
	β_{20}	1.5	1.831	1.819	0.490	(0.918; 2.836)
	β_{21}	2.5	2.466	2.422	1.160	(0.302; 4.841)
Gumbel	β_{10}	-0.5	-0.396	-0.393	0.142	(-0.683; -0.128)
	β_{11}	0.8	0.668	0.667	0.217	(0.249; 1.096)
	β_{20}	1.5	1.405	1.394	0.377	(0.700; 2.184)
	β_{21}	2.5	2.211	2.190	0.864	(0.567; 3.970)

Considerando a função de ligação Logito, a distribuição de frequência dos dados gerados do modelo $1-$DP pode ser vista na Tabela 52.

Tabela 52 – Distribuição de frequência da amostra gerada do modelo $1-$DP, considerando a função de ligação Logito.

y_i	0	1	2	3	4	5
f_i	119	12	51	14	2	2

Na Figura 23 temos os gráficos com algumas características dos dados, obtidos a partir das estimativas Bayesianas dos parâmetros (considerando apenas a média a **posteriori**) do modelo $1-$DP ajustado. O Gráfico 23 **(A)** apresenta as probabilidades e os intervalos com 95% de credibilidade da ocorrência da observação $k = 1$ e notamos que conforme a covariável X_1 aumenta as probabilidades diminuem. Notemos também que probabilidades estimadas da observação $k = 1$ são próximas das verdadeiras. Já o Gráfico 23 **(B)** mostra os valores verdadeiros, as estimativas Bayesianas de p e os intervalos com 95% de credibilidade. Observe que, para todo valor da covariável X_1, o conjunto de dados foi caracterizado como $1-$Deflacionado. Finalmente, o Gráfico 23 **(C)** contém as médias verdadeiras, as médias ajustadas e os intervalos com 95% de credibilidade, o qual apresentou um bom ajuste do modelo aos dados, dada a concordância entre as curvas (a verdadeira e a ajustada).

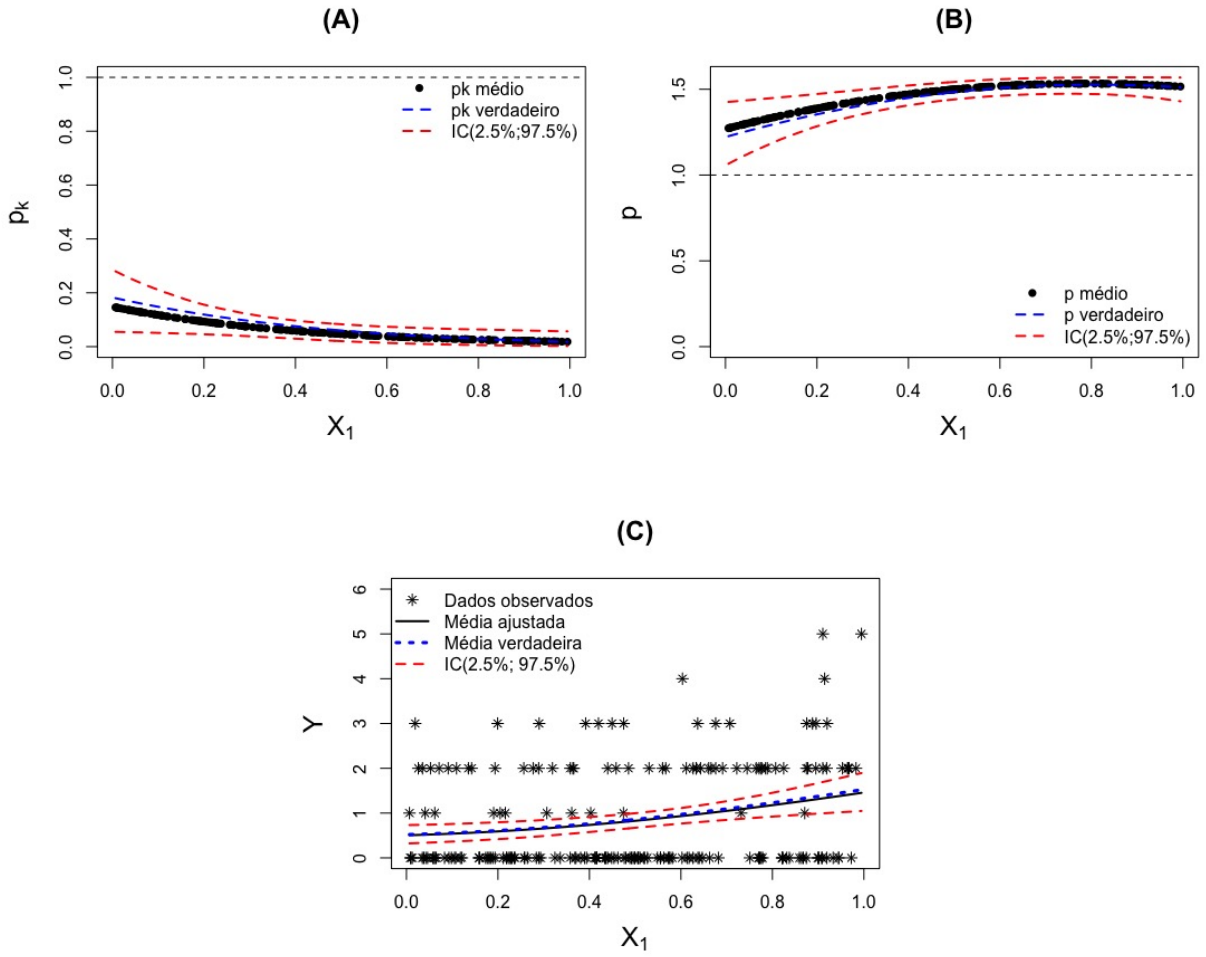

Figura 23 – Gráficos de algumas características do modelo Bayesiano 1−DP ajustado, considerando a função de ligação Logito. **(A)** Estimativas da probabilidade de k. **(B)** Estimativas do parâmetro p. **(C)** Médias reais e ajustadas em função da covariável X_1.

Ainda considerando os dados referentes à Tabela 52, temos uma amostra 1-DP com média de 0.870, mediana igual a 0.000 e desvio padrão $s_y = 1.170$. Selecionamos as observações das posições 19 e 42, isto é, $y_{19} = 2$ e $y_{42} = 3$ para pertubação e consideramos $\delta = 12$. As estimativas Bayesianas de cada parâmetro $\beta_{i,j}$, $i = 1,2$ e $j = 0,1$ foram obtidas considerando os casos anteriormente descritos (pertubando cada observação de forma individual e pertubando as duas observações simultaneamente) e os resultados são apresentados na Tabela 53. Podemos observar que o **caso a** refere-se à análise Bayesiana original; os **casos b** e **c** são referentes às pertubações individuais das observações y_{19} e y_{42} e o **caso d** refere-se às pertubações simultâneas das observações. Desta forma, ao pertubar os pontos nos casos descritos acima, ocorreram impactos de redução ou aumento nas estimativas. A Figura 24 mostra as nuvens de pontos nos **casos a – d** considerados neste estudo.

Tabela 53 – Sumário a *posteriori* dos parâmetros do modelo $1-$DP, considerando diferentes casos de pertubação.

Caso	Pertubação	β_{10}			β_{11}			β_{20}			β_{21}		
		Média	Mediana	SD	Média	Mediana	SD	Média	Mediana	SD	Média	Mediana	SD
a	[–]	-0.510	-0.507	0.144	0.770	0.768	0.227	1.831	1.819	0.490	2.466	2.422	1.160
b	$[y_{19}]$	-0.277	-0.275	0.129	0.447	0.445	0.210	1.829	1.817	0.485	2.471	2.448	1.144
c	$[y_{42}]$	-0.318	-0.315	0.132	0.522	0.522	0.212	1.829	1.817	0.485	2.471	2.448	1.144
d	$[y_{19}, y_{42}]$	-0.121	-0.120	0.121	0.245	0.248	0.203	1.829	1.817	0.485	2.471	2.448	1.144

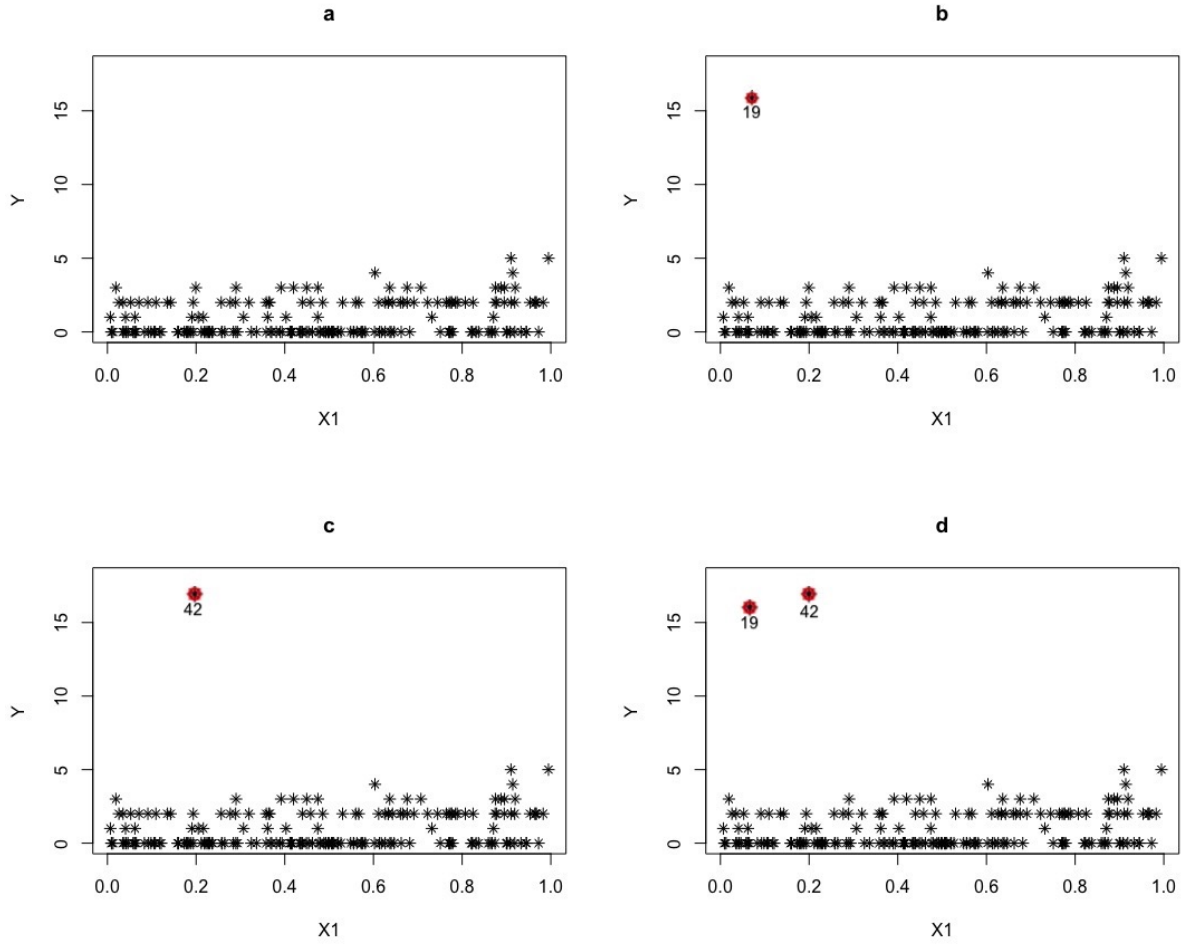

Figura 24 – Plotagem das nunvens de pontos nos diferentes casos de pertubação, considerando o modelo $1-$DP ajustado.

Para avaliarmos a existência de ponto(s) influente(s), consideramos o valor da distância KL e a sua calibração ρ_i. A Tabela 54 apresenta os resultados desse estudo. Notamos que antes da pertubação (**caso a**) os valores da distância KL das observações y_{19} e y_{42} eram próximos de 0 e eles aumentaram de forma expressiva quando pertubados. Além disso, o valor a calibração ρ_i era próxima de 0.6 e, após a pertubação das observações nos **casos b – d**, houve um aumento no valor desta medida, aproximando-se de 1 e reforçando que estas observações são consideradas pontos influentes (já que $\rho_i > 0.950$). Portanto, através deste procedimento, conseguimos mais uma vez identificar os pontos discrepantes.

Tabela 54 – ivergência KL e a sua calibração ρ_i para as observações pertubadas nos diferentes casos, considerando o modelo 1−DP ajustado.

Caso	Pertubação	KL		ρ_i	
		$[y_{19}]$	$[y_{42}]$	$[y_{19}]$	$[y_{42}]$
a	[–]	0.022	0.034	0.603	0.629
b	$[y_{19}]$	1.818	–	0.993	–
c	$[y_{42}]$	–	1.315	–	0.982
d	$[y_{19}, y_{42}]$	1.391	1.020	0.984	0.966

A Figura 25 ilustra a distância KL considerando os pontos de pertubação identificados nos **casos a – d**.

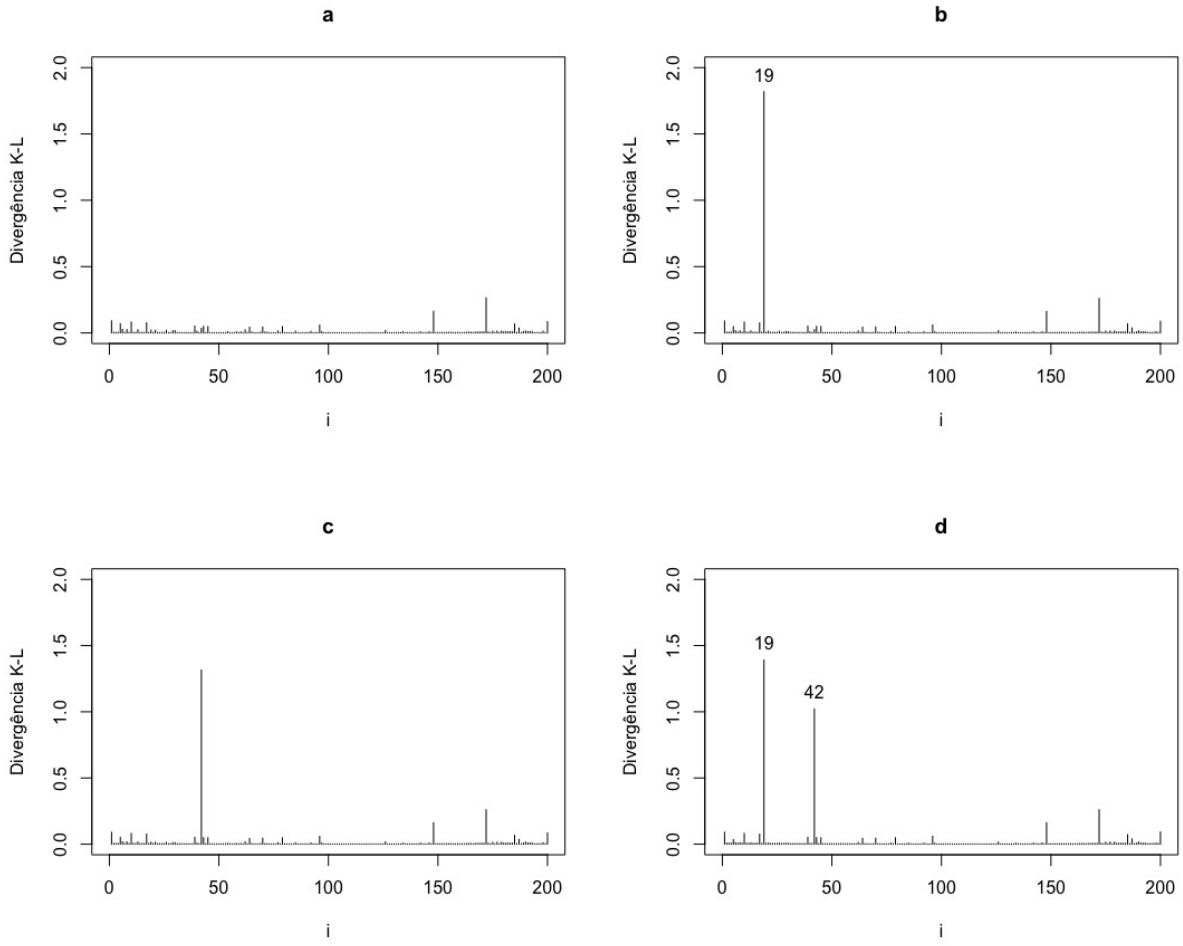

Figura 25 – Distância KL$(\pi, \pi_{(-i)})$ para os diferentes casos de pertubação, considerando o modelo 1−DP ajustado.

6.2 Dados Reais: Notificações de Óbitos Fetais

Nesta subseção apresentamos os resultados do modelo proposto aos dados de notificações de óbitos fetais em cidades do Estado da Bahia (Brasil) em 2014 descritos na Seção 1.1.

6.2.1 Análise do Conjunto de Dados Completo

Consideremos os dados de óbitos fetais descritos na Tabela 1 (página 27). Inicialmente, fizemos uma análise descritiva dos dados e encontramos evidências de um possível ponto discrepante, a observação y_{417}, que corresponde a 601 notificações de óbitos fetais na cidade de Salvador, a capital do Estado da Bahia. Então, removemos essa observação, resultando em uma amostra de tamanho $n = 416$.

Para este conjunto de dados ajustamos o modelo k-MP, considerando o IDH de cada cidade como uma variável explicativa para a estimação dos parâmetros μ_i e p_i (e ω_i) e, portanto, $\boldsymbol{x} \equiv \boldsymbol{z}$. Há uma alta frequência de observações zero no conjunto de dados e, assim, consideramos o ponto de modificação $k = 0$. Dessa forma, o modelo 0-MP foi ajustado aos dados considerando para ω as funções de ligação Logito, Complemento Log-log e Gumbel. Um sumário a *posteriori*, juntamente com os intervalos com 95% de credibilidade, estão na Tabela 55. Analisando os respectivos intervalos de credibilidade, é possível notar que nenhum deles contém o valor zero, indicando que esses parâmetros são significantes.

Tabela 55 – Sumário a *posteriori* e intervalo com 95% de credibilidade dos parâmetros β_{1i} e β_{2i}, $i = 0, 1$ do modelo $0-$MP ajustado aos dados de notificações de óbitos fetais em cidades da Bahia em 2014, considerando as funções de ligação Logito, Complemento Log-log e Gumbel.

Função de Ligação	Parâmetros	Média	Mediana	SD	IC (95%)
Logito	β_{10}	-9.806	-9.812	0.345	(-10.462; -9.122)
	β_{11}	18.507	18.517	0.534	(17.451; 19.518)
	β_{20}	-5.522	-5.466	1.852	(-9.360; -2.072)
	β_{21}	11.436	11.329	3.162	(5.592; 18.019)
Complemento Log-log	β_{10}	9.791	-9.785	0.325	(-10.419; -9.154)
	β_{11}	18.483	18.474	0.503	(17.495; 19.455)
	β_{20}	-3.385	-3.299	1.045	(-5.808; -1.530)
	β_{21}	6.381	6.231	1.763	(3.248; 10.434)
Gumbel	β_{10}	-9.822	-9.835	0.316	(-10.430; -9.207)
	β_{11}	18.531	18.551	0.488	(17.576; 19.464)
	β_{20}	-4.380	-4.363	1.529	(-7.401; -1.337)
	β_{21}	9.746	9.718	2.626	(4.512; 14.949)

Através dos critérios de seleção de modelos Bayesianos EAIC[1], EBIC[2], DIC[3] e B[4], o modelo escolhido foi aquele com a função de ligação Complemento Log-log, como pode ser visto na Tabela 56, e então, vamos apresentar nesta dissertação uma análise completa considerando esta função de ligação.

[1] Usaremos a sigla EAIC, do inglês *Expected Akaike Information Criterion*, referindo-se a Critério de Informação Akaike Esperado (AKAIKE, 1974).

[2] Usaremos a sigla EBIC, do inglês *Expected Bayesian Information Criterion, referindo-se a Critério de Informação Bayesiano Esperado (SCHWARZ, 1978)*.

[3] Usaremos a sigla DIC, do inglês *Deviance Information Criterion*, referindo-se a Critério de Informação Deviance (SPIEGELHALTER *et al.*, 2002).

[4] Usaremos a estatística B, equivalente ao critério do Logaritmo da Verossimilhança Pseudo Marginal, do inglês *Logarithm of the Pseudo Marginal Likelihood (LPML) (IBRAHIM; CHEN; SINHA, 2001)*.

Tabela 56 – Critérios de seleção de modelos Bayesianos para as funções de ligação Logito, Complemento Log-log e Gumbel, considerando os dados de notificações de óbitos fetais em cidades da Bahia em 2014.

Função de Ligação	EAIC	EBIC	DIC	B
Logito	2510.634	2526.757	2506.684	1268.384
Complemento Log-log	**2508.767**	**2524.890**	**2504.823**	**1265.841**
Gumbel	2511.039	2527.162	2506.893	1267.216

6.2.1.1 Resultados da amostra completa - função de ligação Complemento Log-log

Como a função de ligação Complemento Log-log foi escolhida por todos os critérios de seleção de modelos Bayesianos utilizados, fizemos uma análise detalhada do conjunto de dados de notificações de óbitos fetais considerando esta função de ligação. A Tabela 57 compara as estimativas dos parâmetros considerando a função de ligação Complemento Log-log em ambas abordagens Bayesiana e Clássica. Note que, ao considerar a abordagem Bayesiana, consideramos a média e mediana a *posteriori*. É possível notar que as estimativas clássica e Bayesiana estão próximas. Vale ressaltar que no caso clássico, o intervalo de confiança resultante foi obtido através do método *bootstrap*.

Tabela 57 – Sumário Bayesiano (e clássico) e os respectivos intervalos com 95% de credibilidade (confiança) dos parâmetros β_{1i} e β_{2i}, $i = 0,1$ do modelo $0-$MP ajustado aos dados de notificações de óbitos fetais em cidades da Bahia em 2014, considerando a função de ligação Complemento Log-log.

Abordagem	Parâmetros	Estimativas		SD	IC (95%)
		Média	Mediana		
Bayesiana	β_{10}	-9.791	-9.785	0.325	(-10.419; -9.154)
	β_{11}	18.483	18.474	0.503	(17.495; 19.455)
	β_{20}	-3.385	-3.299	1.045	(-5.808; -1.530)
	β_{21}	6.381	6.231	1.763	(3.248; 10.434)
Clássica	β_{10}	-9.800		0.370	(-10.523; -9.093)
	β_{11}	18.498		0.573	(17.398; 19.612)
	β_{20}	-3.548		0.970	(-4.562; -0.767)
	β_{21}	6.658		1.623	(1.138; 7.518)

Para verificar a existência de pontos influentes, consideramos o valor da distância KL e a sua calibração ρ_i. A Figura 26 **(A)** apresenta a distância KL, responsável por medir o efeito de cada observação na amostra. Através da Figura 26 **(B)** com as medidas já calibradas, observamos a existência de dois pontos influentes, y_{412} e y_{413}, que correspondem a 116 e 71 notificações de óbitos fetais nas cidades de Feira de Santana e Itabuna, respectivamente (valores de calibração $\rho_i > 0.950$). Já a Figura 26 **(C)** apresenta as notificações de óbitos fetais de acordo com o IDH, em que os pontos identificados como influentes encontram-se destacados.

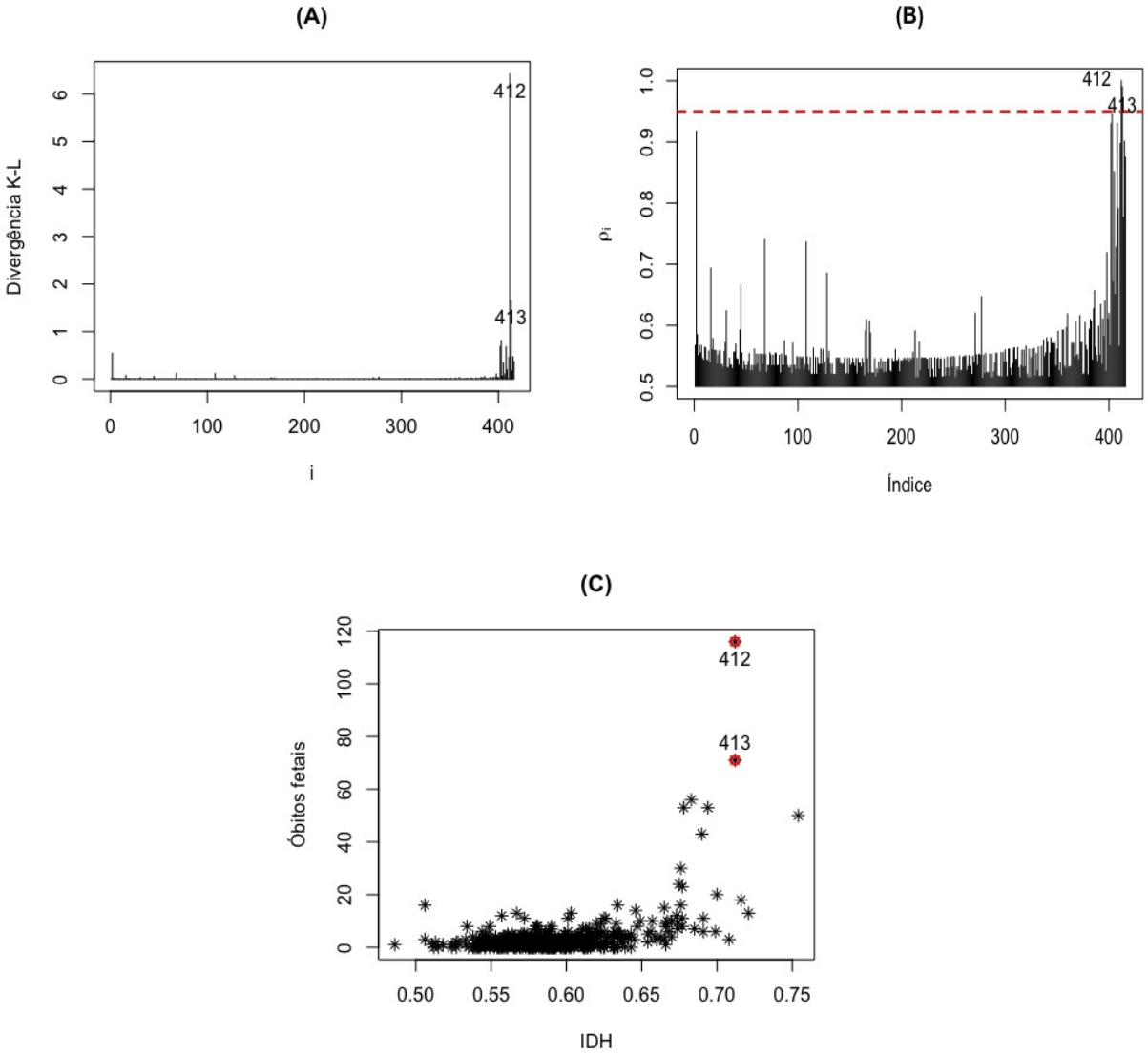

Figura 26 – Estudo de pontos influentes: **(A)** Plotagem de índices de $\mathrm{KL}(\pi, \pi_{(-i)})$. **(B)** Calibração. **(C)** Pontos influentes identificados.

Fonte: Elaborada pelo autor.

Para checar a influência de cada observação nas estimativas dos parâmetros do modelo, o processo inferencial Bayesiano foi repetido considerando três casos: removendo apenas a observação y_{412}, removendo apenas a observação y_{413} e finalmente removendo ambas observações. O resumo a *posteriori* para cada caso e o percentual de variação (com relação ao resumo a *posteriori* obtido a partir dos dados completos) estão apresentados na Tabela 58. Considerando a perda quadrática, analisando a média a *posteriori* como estimativas Bayesianas dos parâmetros para cada caso, observamos que a remoção da observação y_{412} referente à cidade Feira de Santana teve um maior impacto no ajuste do modelo, já que este ponto representa 116 óbitos fetais, que se destoa de forma significativa da massa dos dados e sua remoção diminui consideravelmente a média a *posteriori* ajustada das notificações de óbitos fetais em função do IDH.

Tabela 58 – Sumário a *posteriori* (e variação em %) e intervalos com 95% de credibilidade para β_{1i} e β_{2i}, $i = 0, 1$, do modelo 0−MP ajustado aos dados de notificações de óbitos fetais em cidades da Bahia em 2014, considerando a função de ligação Complemento Log-log e pontos influentes em diferentes casos.

Observação removida	Parâmetros	Média	Mediana	SD	IC (95%)
-	β_{10}	-9.791	-9.785	0.325	(-10.419; -9.154)
	β_{11}	18.483	18.474	0.503	(17.495; 19.455)
	β_{20}	-3.385	-3.299	1.045	(-5.808; -1.530)
	β_{21}	6.381	6.231	1.763	(3.248; 10.434)
[412]	β_{10}	-8.769(10.438%)	-8.780(10.271%)	0.339(4.308%)	(-9.441; -8.067)
	β_{11}	16.795(-9.133%)	16.813(-8.991%)	0.530(5.368%)	(15.701; 17.834)
	β_{20}	-3.631(-7.267%)	-3.672(-11.095%)	1.075(2.871%)	(-5.563; -1.420)
	β_{21}	6.796(6.504%)	6.860(10.095%)	1.815(2.949%)	(3.099; 10.113)
[413]	β_{10}	-9.353(4.473%)	-9.345(4.497%)	0.341(4.923%)	(-10.009; -8.701)
	β_{11}	17.758(-3.923%)	17.746(-3.941%)	0.531(5.567%)	(16.753; 18.780)
	β_{20}	-3.642(-7.592%)	-3.632(-10.094%)	1.088(4.115%)	(-5.833; -1.537)
	β_{21}	6.812(6.754%)	6.792(9.003%)	1.834(4.027%)	(3.290; 10.504)
[412, 413]	β_{10}	-8.908(9.018%)	-8.900(9.044%)	0.339(4.308%)	(-9.601; -8.244)
	β_{11}	17.022(-7.905%)	17.011(-7.964%)	0.531(5.567%)	(15.978; 18.103)
	β_{20}	-3.585(-5.908%)	-3.533(-7.093%)	0.969(-7.273%)	(-5.617; -1.789)
	β_{21}	6.715(5.234%)	6.633(6.452%)	1.638(-7.090%)	(3.653; 10.152)

A Figura 27 apresenta os envelopes para os resíduos dos dados considerando os diferentes casos vistos acima. Consideramos o resíduo Bayesiano studantizado (RBS)[5] dos dados, em que RBS = $(\boldsymbol{y} - \boldsymbol{\mu})/\sqrt{\boldsymbol{\sigma^2}}$, $\boldsymbol{y}$ é o vetor de dados, μ é o vetor de média e σ^2 é o vetor de variância dos dados, e através de simulações *bootstrap* de tamanho 1000, geramos novas amostras para obtenção das bandas do envelope, com intervalo interquantílico de 95%.

[5] Usaremos a sigla RBS referindo-se a Resíduo Bayesiano Studentizado (CORDEIRO; Lima Neto, 2006).

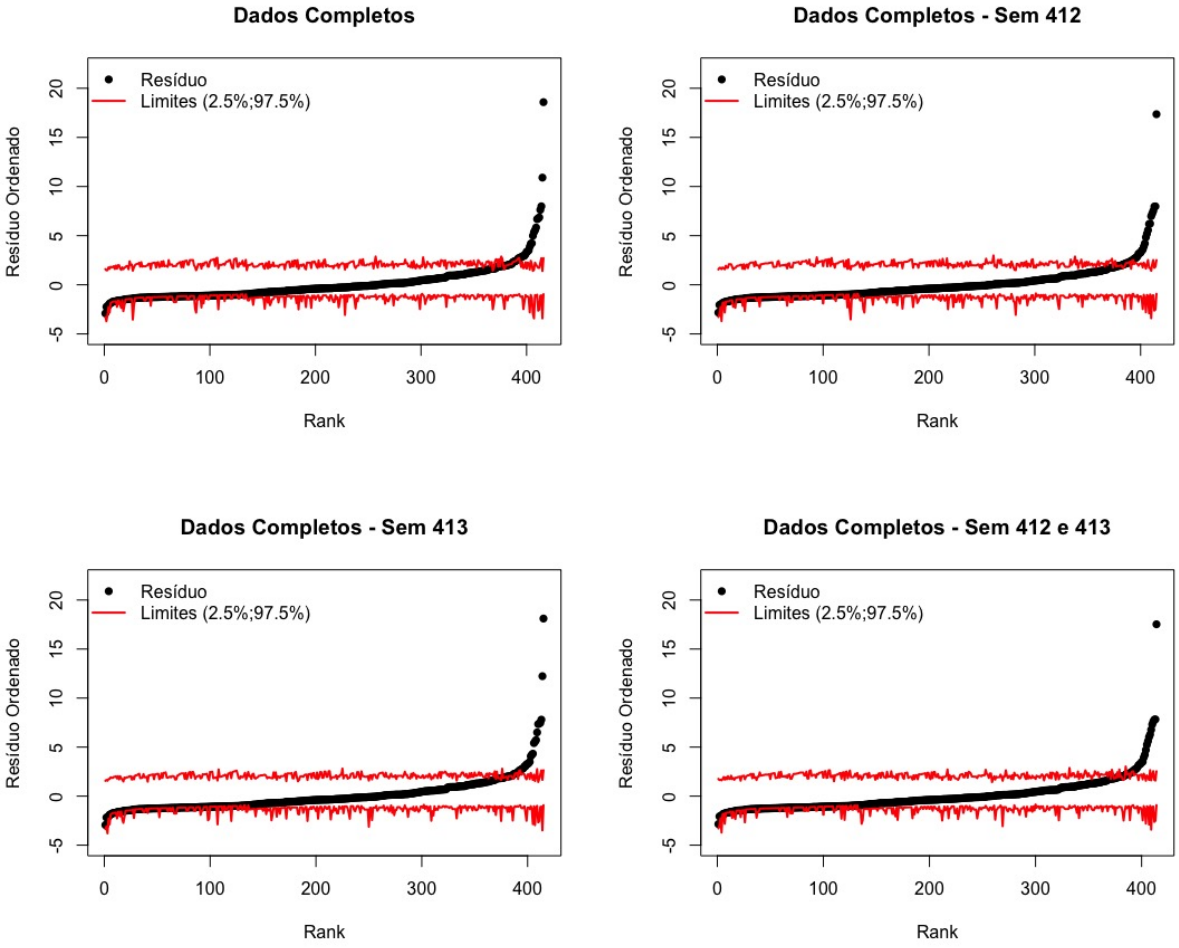

Figura 27 – Envelope dos resíduos considerando os diferentes casos.

Fonte: Elaborada pelo autor.

Através da Figura 27 acima, vemos que o comportamento dos resíduos são bem similares nos quatro casos, e portanto, utilizaremos o percentual variacional da Tabela 58 para concluir qual ponto mais influente.

Os gráficos apresentados na Figura 28 ilustram: (A) as estimativas Bayesianas e os intervalos com 95% de credibilidade da probabilidade zero, correspondendo a não-notificação de óbitos fetais, (B) as estimativas Bayesianas e os intervalos de 95% de credibilidade de $p(\boldsymbol{x})$, e (C) a média ajustada. Para cada gráfico, os dados completos (parte superior) e os dados sem a observação $y_{(412)}$ (parte inferior) foram considerados.

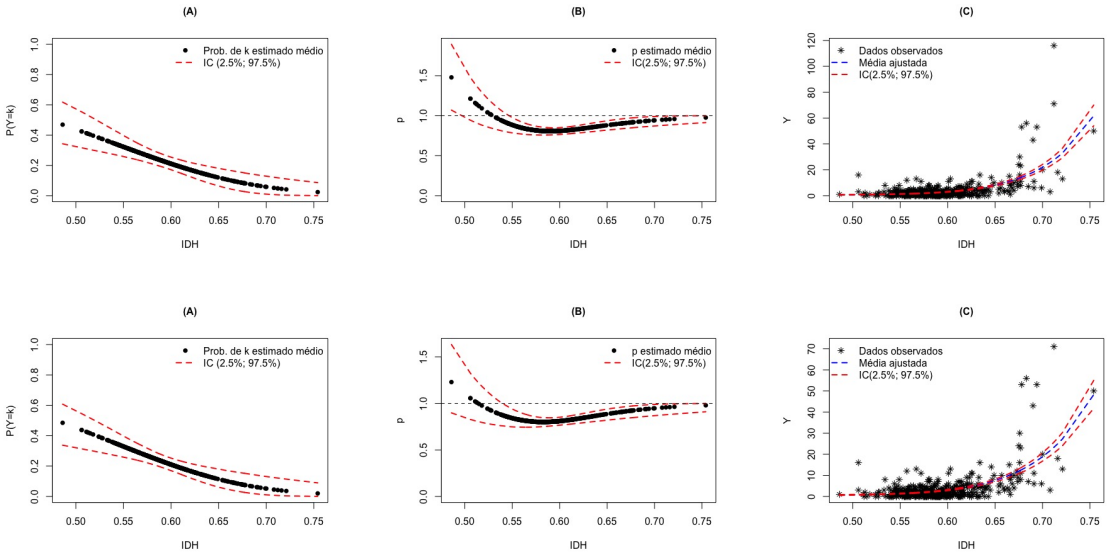

Figura 28 – Parte superior: dados completos; Parte inferior: sem a observação 412. **(A)** Estimativas Bayesianas (considerando a média a *posteriori*) e intervalos com 95% de credibilidade das probabilidades de não-notificações de óbitos fetais. **(B)** Estimativas Bayesianas (considerando a média a *posteriori*) e intervalos com 95% de credibilidade do parâmetro p. **(C)** Médias ajustadas, juntamente com os dados de notificações de óbitos fetais em função do IDH.

Fonte: Elaborada pelo autor.

Analisando o gráfico **(A)** na Figura 28, correspondendo à probabilidade de não-notificação de óbitos fetais ($\pi(k;\mu(\boldsymbol{x}),p(\boldsymbol{z}))=P(Y=k)$, com $k=0$ e $\boldsymbol{x}=\boldsymbol{z}$), observamos que estas probabilidades estão diminuindo com o aumento no IDH. Isto é, em cidades com alto IDH, a probabilidades de não-notificação de óbitos fetais é baixa. Novamente, é importante enfatizar que a ausência de notificações não significa ausência de óbitos fetais, mas evidências de precariedade no sistema de saúde para detectar esses casos. Na Figura 28 **(B)**, observamos que, para valores mais baixos no IDH, as estimativas de $p(\boldsymbol{z})$ tem intervalos de credibilidade que contém o valor 1 enquanto para IDH >0.56 temos $p(\boldsymbol{z})<1$. Como este parâmetro indica o tipo de modificação nos dados, concluímos que para baixos valores no IDH os dados são caracterizados com a distribuição de Poisson usual, enquanto que o conjunto de dados é caracterizado como inflacionado de 0's quando consideramos IDH >0.56, e isto não muda quando excluímos a observação $y_{(412)}$. Entretanto, este resultado requereria que os dados fossem divididos em dois subconjuntos para conseguir melhores estimativas para os parâmetros. A Figura 28 **(C)** (parte superior) mostra a média ajustada e é possível notar que ela está aumentando com o aumento no IDH, e a observação $y_{(412)}$ tem um alto impacto no padrão de crescimento da curva, dado que esta curva cresce menos significantemente quando removemos esta observação, o que pode ser visto na Figura 28 **(C)** (parte inferior).

6.2.2 Análise Considerando Cidades com IDH > 0.56

Para caracterizar o conjuntos de dados de forma única, podemos analisar as notificações de óbitos fetais exclusivamente em cidades com IDH maior que 0.56, já que ao considerar as funções de ligação Logito, Complemento Log-log e Gumbel, os subconjuntos com essas restrições no IDH foram sempre caracterizados como inflacionados no zero. Então, para as $n = 332$ cidades da Bahia com > 0.56, a distribuição de frequência das notifições de óbitos fetais é apresentada na Tabela 59:

Tabela 59 – Distribuição de frequência das notificações de óbitos fetais em cidades do Estado da Bahia em 2014, com IDH > 0.56.

y_i	0	1	2	3	4	5	6	7	8	9	10	11	12	13
f_i	72	63	43	34	32	27	11	9	6	5	4	6	1	3
y_i	14	15	16	18	20	23	24	30	43	50	53	56	71	116
f_i	1	1	2	1	1	1	1	1	1	1	2	1	1	1

Novamente, ajustamos o modelo k-MP para esse subconjunto com IDH > 0.56, escolhendo o IDH de cada cidade como uma variável explicativa para os parâmetros μ e p (e então, $\boldsymbol{x} = \boldsymbol{z}$). Como há uma alta frequência de observações zero no conjunto de dados, consideramos o ponto de modificação como $k = 0$, de forma que o modelo 0-MP foi ajustado aos dados com as funções de ligação Logito, Complemento Log-log e Gumbel. As estimativas Bayesianas (a média e mediana a *posteriori*), o desvio padrão (SD) e os intervalos com 95% de credibilidade são apresentados na Tabela 60. Analisando os respectivos intervalos de credibilidade, é possivel notar que nenhum deles contém o valor zero, indicando que os parâmetros são significantes.

Tabela 60 – Sumário a *posteriori* e intervalo com 95% de credibilidade dos parâmetros β_{1i} e β_{2i}, $i = 0,1$ do modelo $0-$MP ajustado aos dados de notificações de óbitos fetais em cidades da Bahia em 2014 com IDH > 0.56, considerando as funções de ligação Logito, Complemento Log-log e Gumbel.

Função de Ligação	Parâmetros	Média	Mediana	SD	IC (95%)
Logito	β_{10}	-11.092	-11.105	0.388	(-11.826; -10.334)
	β_{11}	20.421	20.439	0.592	(19.267; 21.547)
	β_{20}	-8.830	-8.559	3.023	(-15.477; -3.527)
	β_{21}	16.863	16.414	5.070	(8.005; 28.061)
Complemento Log-log	β_{10}	-11.082	-11.084	0.368	(-11.759; -10.335)
	β_{11}	20.405	20.411	0.562	(19.268; 21.437)
	β_{20}	-5.492	-5.478	1.408	(-8.491; -2.766)
	β_{21}	9.824	9.810	2.337	(5.287; 14.775)
Gumbel	β_{10}	-11.130	-11.125	0.382	(-11.878; -10.375)
	β_{11}	20.478	20.467	0.582	(19.326; 21.615)
	β_{20}	-7.701	-7.475	2.375	(-12.558; -3.475)
	β_{21}	15.212	14.834	4.003	(8.111; 23.373)

Assim, através dos critérios de seleção de modelos Bayesianos para as funções de ligação Logito, Complemento Log-log e Gumbel, concluímos novamente que a função de ligação

Complemento Log-log garante o melhor modelo, como pode ser visto na Tabela 61:

Tabela 61 – Critérios de seleção de modelos Bayesianos para as funções de ligação Logito, Complemento Log-log e Gumbel, considerando os dados de notificações de óbitos fetais em cidades da Bahia em 2014, com IDH > 0.56.

Função de Ligação	EAIC	EBIC	DIC	B
Logito	2073.371	2088.591	2069.526	1049.860
Complemento Log-log	**2071.321**	**2086.541**	**2067.219**	**1048.114**
Gumbel	2073.558	2088.778	2069.450	1049.934

6.2.2.1 *Resultados amostra incompleta - função de ligação Complemento Log-log*

Considerando novamente a função de ligação Complemento Log-log, a Tabela 62 apresenta as estimativas Bayesianas (média e mediana a *posteriori*) e clássicas, além do desvio padrão e intervalos com 95% de credibilidade e de confiança dos parâmetros do modelo. Vale ressaltar que, os intervalos de confiança foram estimados empiricamente através das amostras geradas (procedimento *bootstrap*).

Tabela 62 – Sumário Bayesiano (e clássico) e os respectivos intervalos com 95% de credibilidade (confiança) dos parâmetros β_{1i} e β_{2i}, $i = 0,1$ do modelo $0-$MP ajustado aos dados de notificações de óbitos fetais em cidades da Bahia em 2014 com IDH > 0.56, considerando a função de ligação Complemento Log-log.

Abordagem	Parâmetro	Estimativas		SD	IC (95%)
		Média	Mediana		
Bayesiana	β_{10}	-11.082	-11.084	0.368	(-11.759; -10.335)
	β_{11}	20.405	20.411	0.562	(19.268; 21.437)
	β_{20}	-5.492	-5.478	1.408	(-8.491; -2.766)
	β_{21}	9.824	9.810	2.337	(5.287; 14.775)
Clássica	β_{10}	-11.121		0.408	(-11.934; -10.323)
	β_{11}	20.466		0.622	(19.246; 21.705)
	β_{20}	-5.282		1.276	(-6.168; -1.140)
	β_{21}	9.478		2.096	(1.834; 10.074)

Uma análise para verificar a existência de pontos influentes é apresentada na Figura 29. Novamente, a Figura 29(A) ilustra a distância KL responsável por medir o efeito de cada observação na amostra. Nós consideramos que uma observação cuja distância tenha calibração ρ_i excedendo 0.95 é um ponto influente, e portanto, através da Figura (29(B)) com as medidas já calibradas, observamos a existência de três pontos influentes, $y_{(328)}$, $y_{(329)}$ e $y_{(332)}$, que correspondem a 116, 71 e 50 notificações de óbitos fetais nas cidades de Feira de Santana, Itabuna e Lauro de Freitas, respectivamente. A Figura 29(C) apresenta as notificações de óbitos fetais de acordo com o IDH e os pontos influentes estão destacados.

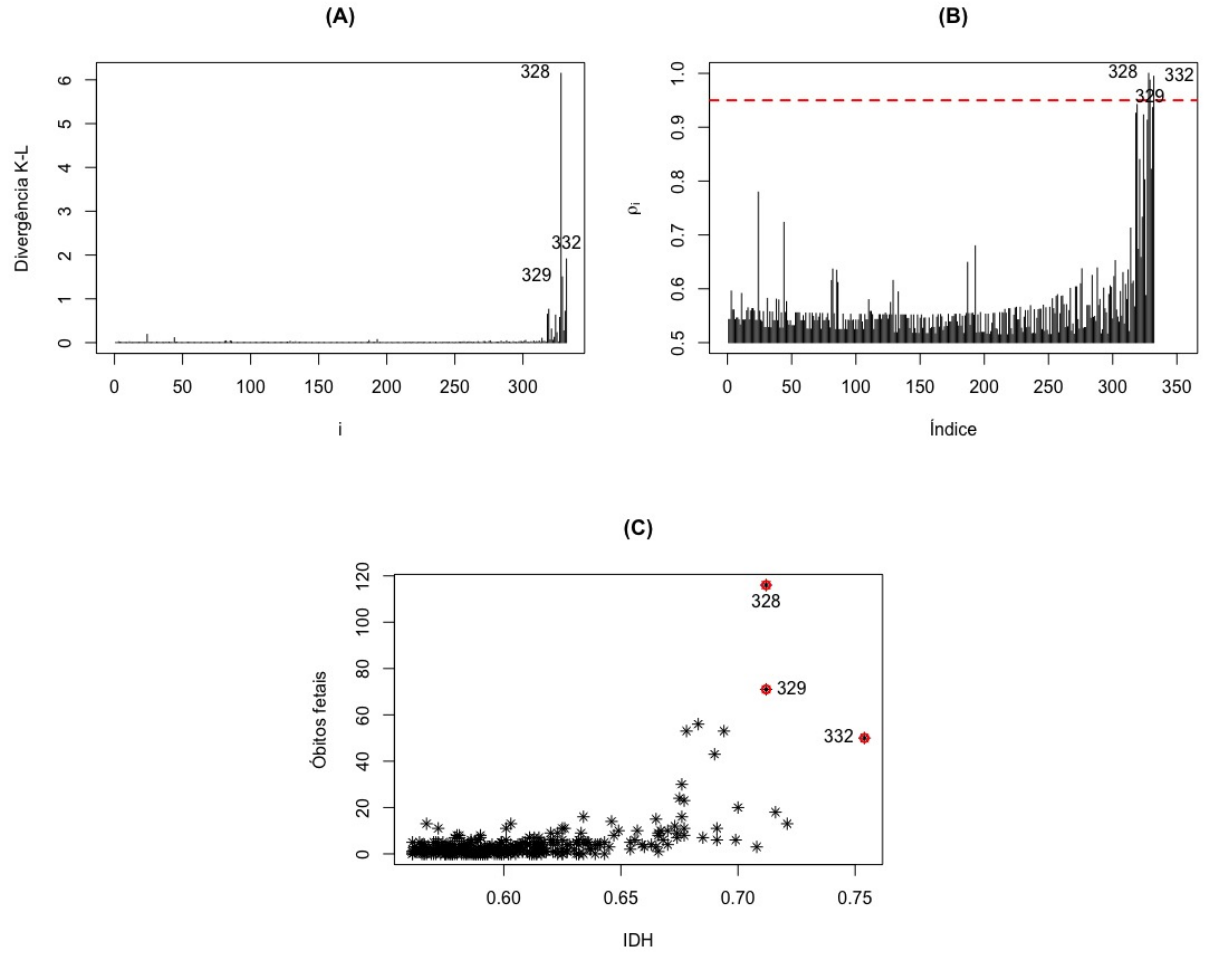

Figura 29 – Estudo de pontos influentes: **(A)** Plotagem dos índices de $\mathrm{KL}(\pi, \pi_{(-i)})$. **(B)** Calibração. **(C)** Pontos influentes identificados.

Fonte: Elaborada pelo autor.

Para checar a influência de cada observação destacada acima na inferência dos parâmetros do modelo, o processo inferencial Bayesiano foi repetido considerando agora sete casos: removendo apenas cada observação influente na amostra, removendo duas das três observações influentes por vez, e finalmente removendo todas as três observações influentes. O resumo a *posteriori* para cada caso e o percentual variacional, entre parênteses, com respeito às estimativas Bayesianas originais obtida dos dados com IDH > 0.56 estão presentes na Tabela 63. Analisando novamente as estimativas Bayesianas dos parâmetros para cada caso descrito, observamos que a retirada das observações $y_{(328)}$ e $y_{(329)}$, e a retirada de todas as três observações influentes (observações $y_{(328)}$, $y_{(329)}$ e $y_{(332)}$) tiveram maior impacto no ajuste do modelo.

Tabela 63 – Estimativas Bayesianas e intervalos com 95% de credibilidade para β_{1i} e β_{2i}, $i = 0, 1$, para os dados de notificações de óbitos fetais em cidades da Bahia com IDH > 0.56, considerando a função de ligação Complemento Log-log e pontos influentes.

Observação removida	Parâmetro	Média	Mediana	SD	IC (95%)
-	β_{10}	-11.082	-11.084	0.368	(-11.759; -10.335)
	β_{11}	20.405	20.411	0.562	(19.268; 21.437)
	β_{20}	-5.492	-5.478	1.408	(-8.491; -2.766)
	β_{21}	9.824	9.810	2.337	(5.287; 14.775)
[328]	β_{10}	-10.071(9.123%)	-10.073(9.121%)	0.362(-1.630%)	(-10.739; -9.301)
	β_{11}	18.752(-8.101%)	18.755(-8.113%)	0.558(-0.712%)	(17.569; 19.775)
	β_{20}	-5.419(1.329%)	-5.322(2.848%)	1.597(13.423%)	(-8.832; -2.595)
	β_{21}	9.702(1.242%)	9.528(2.874%)	2.651(13.436%)	(5.033; 15.311)
[329]	β_{10}	-10.674(3.682%)	-10.682(3.627%)	0.329(-10.598%)	(-11.284; -9.986)
	β_{11}	19.734(3.288%)	19.748(-3.248%)	0.505(-10.142%)	(18.680; 20.667)
	β_{20}	-5.503(-0.201%)	-5.516(-0.694%)	1.446(2.700%)	(-8.332; -2.501)
	β_{21}	9.841(0.173%)	9.853(0.483%)	2.399(2.653%)	(4.895; 14.546)
[332]	β_{10}	-11.758(-6.100%)	-11.766(-6.153%)	0.402(9.239%)	(-12.502; -10.984)
	β_{11}	21.476(5.249%)	21.488(5.277%)	0.617(7.117%)	(20.280; 22.616)
	β_{20}	-5.455(0.674%)	-5.382(1.752%)	1.531(8.736%)	(-8.517; -2.618)
	β_{21}	9.763(-0.621%)	9.645(-1.682%)	2.541(8.729%)	(5.096; 14.853)
[328,329]	β_{10}	-9.413(15.060%)	-9.428(14.925%)	0.389(5.707%)	(-10.170; -8.641)
	β_{11}	17.676(-13.374%)	17.701(-13.277%)	0.602(7.117%)	(16.481; 18.835)
	β_{20}	-5.522(0.546%)	-5.478(0.000%)	1.526(8.381%)	(-8.594; -2.750)
	β_{21}	9.875(-0.519%)	9.800(-0.102%)	2.533(8.387%)	(5.305; 14.951)
[328,332]	β_{10}	-10.397(6.181%)	-10.383(6.324%)	0.372(1.087%)	(-11.136; -9.665)
	β_{11}	19.270(-5.562%)	19.254(-5.669%)	0.576(2.491%)	(18.130; 20.408)
	β_{20}	-5.190(5.499%)	-5.121(6.517%)	1.439(2.202%)	(-8.126; -2.536)
	β_{21}	9.320(-5.130%)	9.207(-6.147%)	2.390(2.268%)	(4.892; 14.201)
[329,332]	β_{10}	-11.228(-1.317%)	-11.249(-4.702%)	0.395(7.337%)	(-11.952; -10.430)
	β_{11}	20.613(1.019%)	20.642(1.132%)	0.609(8.363%)	(19.383; 21.725)
	β_{20}	-5.332(2.913%)	-5.374(1.898%)	1.528(8.523%)	(-8.262; -2.305)
	β_{21}	9.559(-2.697%)	9.628(-1.855%)	2.538(8.601%)	(4.527; 14.406)
[328,329,332]	β_{10}	-9.465(14.591%)	-9.451(14.773%)	0.412(11.957%)	(-10.286; -8.671)
	β_{11}	17.758(-12.972%)	17.736(-13.106%)	0.642(14.235%)	(16.510; 19.036)
	β_{20}	-5.304(3.423%)	-5.264(3.907%)	1.607(14.134%)	(-8.458; -2.318)
	β_{21}	9.510(-3.196%)	9.445(-3.721%)	2.669(14.206%)	(4.582; 14.74)

A Figura 30 apresenta os envelopes dos resíduos considerando os todos os casos apresentados acima:

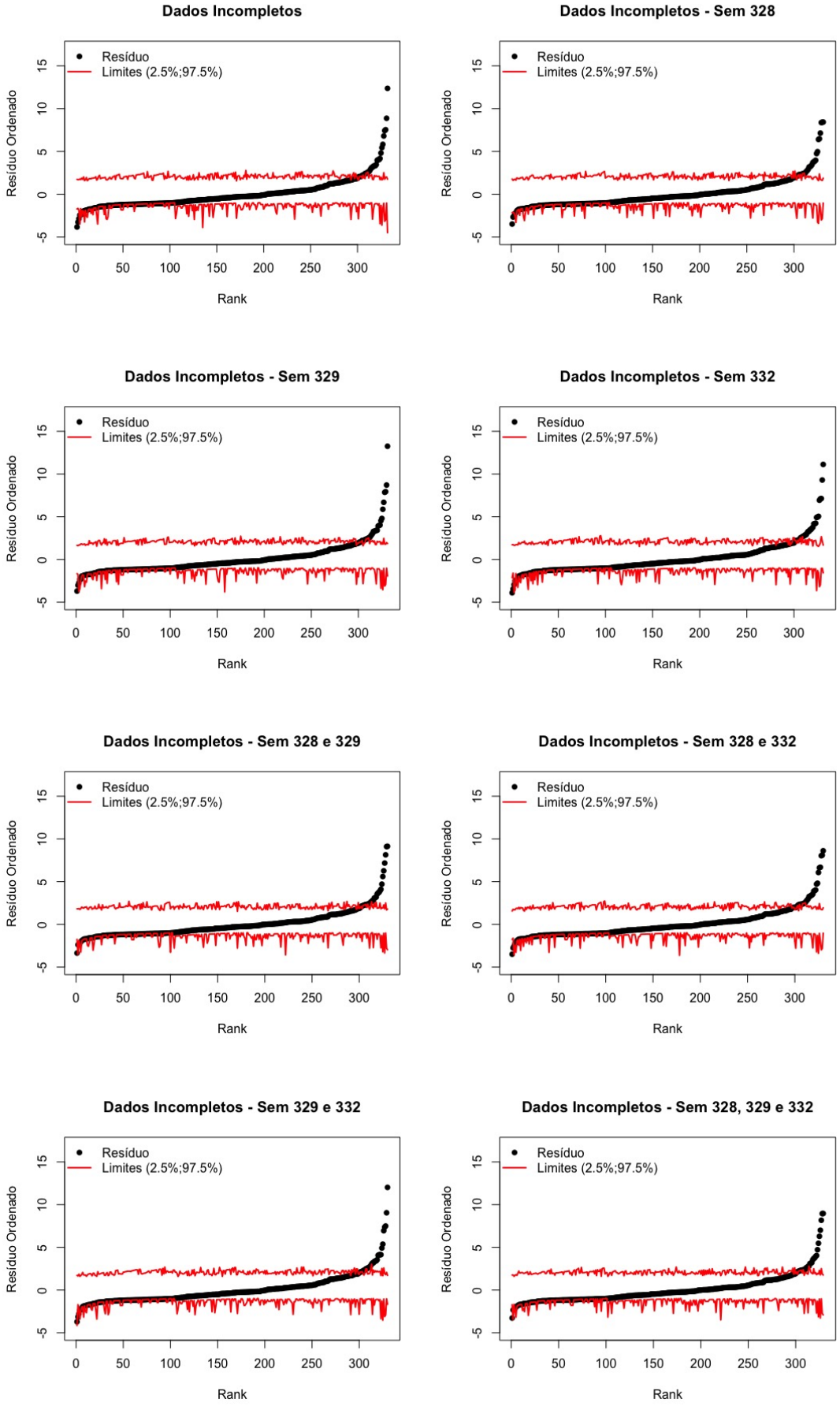

Figura 30 – Envelope para os diferentes casos, considerando IDH > 0.56.

Fonte: Elaborada pelo autor.

Os gráficos apresentados na Figura 31 ilustram: (A) as estimativas Bayesianas (considerando a média a *posteriori*) e intervalos com 95% de credibilidade de probabilidade zero, correspondendo a não-notificação de óbitos fetais, (B) as estimativas Bayesianas e intervalos com 95% de credibilidade de $p(\boldsymbol{x})$ e (C) a média ajustada. Para cada figura, os dados completos (parte superior), os dados sem as observações $y_{(328)}$ e $y_{(329)}$ (parte média), e os dados sem as três observações influentes (parte inferior) são considerados.

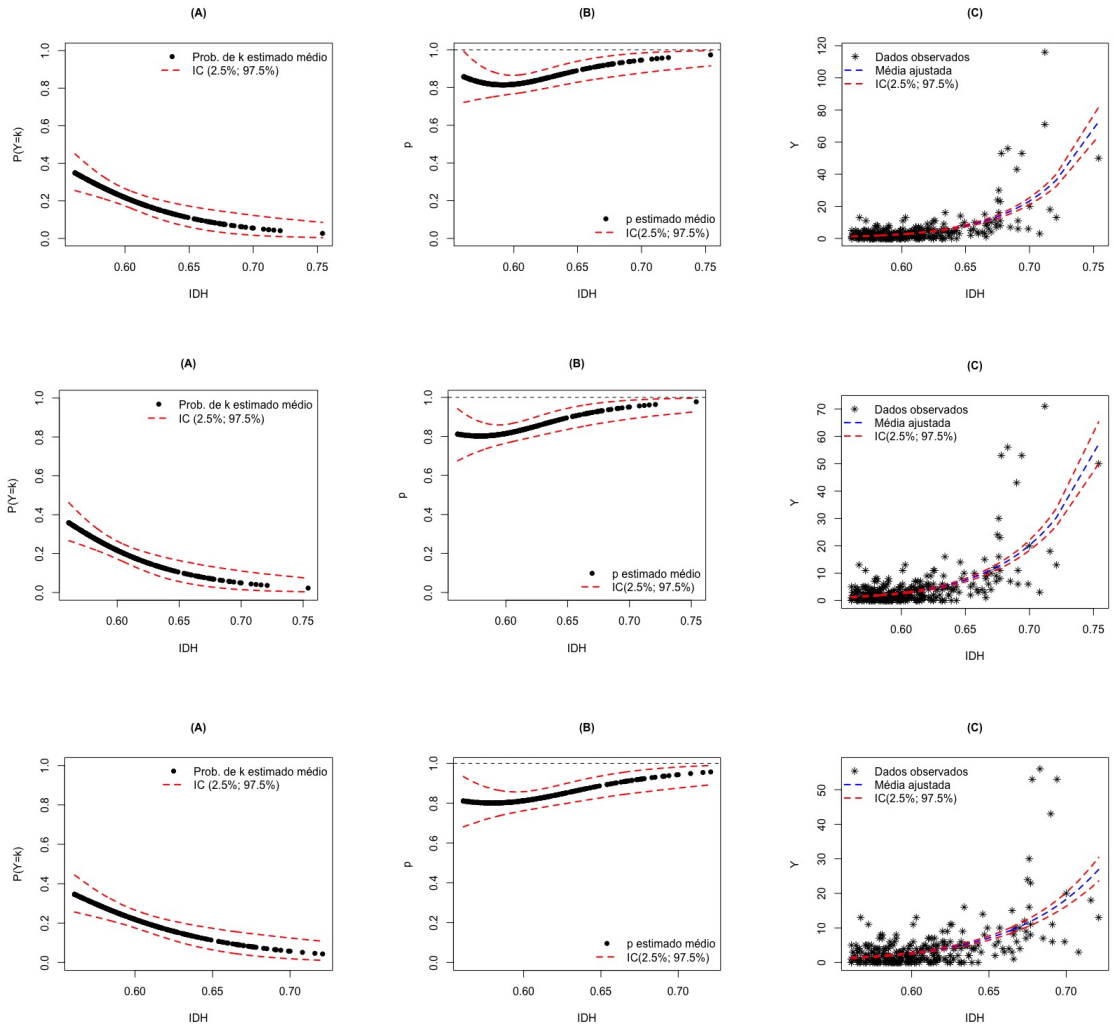

Figura 31 – Parte superior: dados completos, parte média: sem as observações 328 e 329, parte inferior: sem as observações 328,329 e 332. **(A)** Estimativas Bayesianas e intervalos com 95% de credibilidade da probabilidade de não-notificação de óbitos fetais. **(B)** Estimativas Bayesianas e intervalos com 95% de credibilidade do parâmetro p. **(C)** Médias ajustadas juntamente com os dados de notificações de óbitos fetais em função do IDH.

Fonte: Elaborada pelo autor.

Analisando o gráfico **(A)** na Figura 31, correspondendo à probabilidade de não-notificação de óbitos fetais ($\pi(k;\mu(\boldsymbol{x}),p(\boldsymbol{z}))=P(Y=k)$, com $k=0$ e $\boldsymbol{x}=\boldsymbol{z}$), observamos que estas probabilidades estão diminuindo com o aumento do IDH. Logo, em cidades com alto IDH a probabilidade de não-notificação de óbitos fetais é baixa. Novamente, enfatizamos que a ausência de notificações não significa ausência de óbitos fetais, mas evidencia a precariedade no sistema de saúde para detectar estes casos. Já através da Figura 31 **(B)**, observamos que o conjunto de dados é agora caracterizado como 0-Inflacionado, dado que $p<1$ e seu intervalo de credibilidade também, nos três casos analisados. A Figura 31 **(C)** (parte superior) mostra as médias ajustadas e é possível notar que elas estão aumentando à medida que o IDH cresce; as observações $y_{(328)}$ referente à cidade Feira de Santana e $y_{(329)}$ referente a Itabuna tem um alto impacto no padrão de crescimento da curva, já que elas possuem uma grande quantidade de notificações de óbitos fetais em função de seus IDHs e a remoção destas observações diminui significantemente o padrão da curva de médias ajustadas, o que pode ser visto na Figura 31 **(C)** (parte média); porém, ao remover todos os três pontos influentes, a curva de médias ajustadas sofre um alto impacto, crescendo de forma bem mais suave, pois a cidade de Lauro de Freitas, responsável pela observação $y_{(332)}$ possui um alto IDH e as suas notificações de óbitos fetais também tem alto impacto no padrão de crescimento da média ajustada.

CAPÍTULO

7

CONSIDERAÇÕES FINAIS

Neste trabalho, introduzimos a família de distribuições discretas com modificação em um ponto k (as distribuições k-MD), cujo caso particular é a família k-MPS. Esta família de distribuições $k-$modificadas é uma extensão da família de distribuições dicretas comumente utilizada para analisar problemas reais com dados de contagem. A família k-MD foi desenvolvida a partir da modificação nas probabilidades de ocorrência das observações das distribuições discretas tradicionais, para que discrepâncias (alta ou baixa frequência) na observação k possam ser explicadas (modeladas) adequadamente. Desta forma, esta família mostra-se bastante flexível para explicar o comportamento de conjuntos de dados de contagem com diferentes características, isto é, as distribuições $k-$MD podem ajustadas a conjuntos de dados k-inflacionados, conjuntos k-deflacionados, conjuntos com ausência de observações k ou ainda conjuntos pertencentes à família de distribuições discretas tradicionais.

Para a estimação pontual dos parâmetros das distribuições $k-$modificadas consideramos uma abordagem Clássica (utilizamos o método de máxima verossimilhança) e uma abordagem Bayesiana. Baseado em intervalos de confiança e de credibilidade, pudemos fazer inferências sobre os parâmetros, e observamos que nos dois procedimentos de estimação utilizados, as estimativas foram próximas. Ilustramos com algumas aplicações das distribuições da família k-MPS, utilizando conjuntos de dados reais. Através desses dados pudemos concluir que o modelo k-modificado se ajusta bem a conjuntos de dados que possuem discrepância em alguma observação k, sem a necessidade de conhecimentos prévios no tipo de k-modificação.

Ainda neste trabalho, apresentamos também uma extensão das distribuições $k-$MPS no contexto de modelos de regressão, de forma que tornou-se possível modelar o parâmetro de média μ e o parâmetro de modificação das probabilidades p em função de covariáveis. Para isso, consideramos novamente as abordagens Clássica e Bayesiana para a estimação dos parâmetros. Buscando identificar pontos influentes, consideramos a distância de Kullback-Leibler e a sua respectiva medida de calibração.

A fim de avaliar o desempenho dos estimadores dos parâmetros obtidos considerando a abordagem Bayesiana, realizamos um estudo de simulação com três diferentes distribuições da família $k-$MPS ($k-$MB, $k-$MG e $k-$MP) nos contextos de inflação e deflação de alguma observação k, e obtivemos bons resultados a partir deste estudo. Também utilizamos a distribuição k-MP para análise de um conjunto dados reais referente às notificaçõe de óbitos fetais no Estado da Bahia em 2014, considerando o IDH de cada cidade como variável explicativa. Analisando os resultados, foi possível caracterizar os dados como provindos de populações distintas, já que parte dos dados foram caracterizados como provenientes de uma distribuição de Poisson tradicional, e a outra parte deles foram caracterizados como inflacionados no ponto $k = 0$. Desta forma, analisamos de forma mais detalhada os dados que foram caracterizados como 0-inflacionados e pudemos identificar pontos influentes que estavam presentes. Assim, pudemos concluir que o modelo ajustou-se de forma adequada, permitindo a caracterização dos dados em função do IDH.

Finalmente, a família k-MD e, em particular a família k-MPS, podem ser consideradas alternativas interessantes para explicar o comportamento de dados de contagem, sem que haja preocupações com o tipo de discrepância na frequência da observação k.

Como propostas futuras, há diversas linhas de estudo que podem ser desenvolvidas a partir deste trabalho. Podemos propor os seguintes tópicos: utilizar as distribuições da família PS em conjuntos de dados com mais de uma observação discrepante (pontos k_1 e k_2); introduzir as distribuições da família de distribuições discretas (k_1, k_2)-modificadas no contexto de modelos de regressão; Considerar abordagens Clássicas e Bayesianas para a estimação de parâmetros, entre outros.

REFERÊNCIAS

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974. Citado na página 108.

ANGELL, J. K.; KORSHOVER, J. Estimate of global temperature variations in the 100 - 30 mb layer between 1958 and 1977. **Monthly Weather Review**, v. 106, n. 10, p. 1422–1432, 1978. Citado na página 48.

ATLAS. **Atlas do Desenvolvimento Humano no Brasil – IDH das cidades do estado da Bahia em 2014**. 2011. Disponível em: <http://www.atlasbrasil.org.br/2013/pt/consulta/>. Acesso em: 2017. Citado na página 28.

CANCHO, V. G.; DEY, D. K.; LACHOS, V. H.; ANDRADE, M. G. Bayesian nonlinear regression models with scale mixtures of skew-normal distributions: Estimation and case influence diagnostics. **Computational Statistics and Data Analysis**, v. 55, n. *, p. 588–602, 2011. Citado na página 83.

CARVALHO, S. O. Dissertação de Mestrado – Programa Interinstitucional de Pós-graduação em Estatística, **Distribuições k-modificadas da família série de potência uniparamétrica**. São Carlos – SP: [s.n.], 2017. Citado 7 vezes nas páginas 27, 28, 32, 34, 36, 37 e 45.

CHIB, S.; GREENBERG, E. Understanding the metropolis-hastings algorithm. **The American Statistician**, v. 49, n. 4, p. 327–335, 1995. Citado 2 vezes nas páginas 44 e 64.

CHO, H.; IBRAHIM, J. G.; SINHA, D.; ZHU, H. Bayesian case influence diagnostics for survival models. **Biometrics**, v. 65, n. 1, p. 441–459, 2009. Citado na página 65.

CONCEIÇÃO, K. S. Tese de Doutorado, **Modelos Séries de Potência Zero-Modificados**. São Carlos – SP: [s.n.], 2013. Citado 2 vezes nas páginas 37 e 55.

CONCEIÇÃO, K. S.; ANDRADE, M. G.; LOUZADA, F. Zero-modified poisson model: Bayesian approach, influence diagnostics, and an application to a brazilian leptospirosis notification data. **Biometrical Journal**, v. 55, n. 5, p. 661–678, 2013. Citado 4 vezes nas páginas 36, 37, 64 e 65.

CONOVER, W. **Practical Nonparametric Statistics**. [S.l.]: John Wiley & Sons, 1999. Third edition. Citado na página 46.

CONSUL, P. C. New class of location-parameter discrete probability distributions and their characterizations. **Communications in Statistics - Theory and Methods**, 1990. Citado 2 vezes nas páginas 25 e 31.

CORDEIRO, G. M.; ANDRADE, M. G.; CASTRO, M. d. Power. **Biometrics**, v. 65, n. 1, p. 441–459, 2009. Citado na página 32.

CORDEIRO, G. M.; Lima Neto, E. A. **Modelos Paramétricos**. [S.l.]: Recife: Universidade Federal Rural de Pernambuco, Departamento de Estatística e Informática, 2006. Citado na página 111.

DALRYMPLE, M. L.; HUDSON, I. L.; FORD, R. P. K. Finite mixture, zero-inflated poisson and hurdle models with application te sids. **Computational Statistics & Data Analysis**, v. 41, p. 491–504, 2003. Citado na página 38.

GELMAN, A.; RUBIN, D. Power series generalized nonlinear models. **Computation Statistics and Data Analysis**, v. 53, n. 4, p. 1155–1166, 2009. Citado 2 vezes nas páginas 65 e 68.

GUPTA, R. C. Modified power series distribution and some of its applications. **Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)**, v. 36, n. 3, p. 288–298, 1974. Citado 3 vezes nas páginas 25, 31 e 32.

HASTINGS, W. K. Monte carlo sampling methods using markov chains and their applications. **Biometrika**, v. 57, n. 1, p. 97–109, 1970. Citado 2 vezes nas páginas 44 e 64.

IBGE. **Óbitos fetais ocorridos e registrados no ano por duração da gestação da mãe**. 2016. Disponível em: <https://seriesestatisticas.ibge.gov.br/series.aspx?vcodigo=RC86&t=obitos-fetais-ocorridos-registrados-ano-duracao>. Acesso em: 2016. Citado na página 27.

IBRAHIM, J. G.; CHEN, M. H.; SINHA, D. **Bayesian Survival Analysis**. [S.l.]: New York: Springer-Verlag, 2001. 589 p. Citado na página 108.

JOHNSON, N. L.; KOTZ, S.; KEMP, A. W. **Univariate Discrete Distribution**. [S.l.]: Wiley, New York, 1992. Second edition. Citado 2 vezes nas páginas 25 e 31.

KHATRI, C. G. On certain properties of power-series distributions. **Biometrika**, v. 46, n. 3/4, p. 486–490, 1959. Citado 2 vezes nas páginas 25 e 31.

LAMBERT, D. Zero-inflated poisson regression, with an application to defects in manufacturing. **Technometrics**, [Taylor Francis, Ltd., American Statistical Association, American Society for Quality], v. 34, n. 1, p. 1–14, 1992. ISSN 00401706. Disponível em: <http://www.jstor.org/stable/1269547>. Citado na página 26.

MCCULLOCH, R. E. Local model influence. **Journal of the American Statistical Association**, v. 84, n. 406, p. 473–478, 1989. Citado na página 65.

MURAT, M.; SZYNAL, D. Non–zero inflated modified power series distributions. **Communications in Statistics - Theory and Methods – Taylor & Francis**, v. 27, n. 12, p. 3047–3064, 1998. Citado na página 39.

PANDEY, K. N. Generalized inflated poisson distribution. **Journal of Science and Research Banaraes Hindu University**, v. 15, n. 2, p. 157–162, 1965. Citado na página 26.

PATIL, G. P. Certain properties of the generalized power series distribution. **Annals of the Institute of Statistical Mathematics – Springer**, v. 14, n. *, p. 179–182, 1962. Citado 2 vezes nas páginas 25 e 31.

PLUMMER, M. **JAGS: Just Another Gibbs Sampler**. [S.l.], 2017. Disponível em: <https://sourceforge.net/projects/mcmc-jags/files/>. Citado 2 vezes nas páginas 65 e 68.

PNDU. **Desenvolvimento Humano e IDH**. 2016. Disponível em: <http://www.br.undp.org/content/brazil/pt/home/idh0.html>. Acesso em: 08/08/2018. Citado na página 28.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2015. Disponível em: <https://www.R-project.org/>. Citado 3 vezes nas páginas 64, 65 e 127.

ROSENQVIST, G.; ARIEN, S.-S.; SINTONEN, H. Modified cout data models with an application to demand for dental care. **Swedish School of Economics and Business Administration**, 1995. Citado na página 26.

SCHWARZ, G. Estimating the dimension of a model. **The Annals of Statistics**, v. 6, n. 2, p. 461–464, 1978. Citado na página 108.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. van der. Bayesian measures of model complexity and fit. **Journal of the Royal Statistical Society - Series B**, v. 64, n. 4, p. 583–639, 2002. Citado na página 108.

APÊNDICE

A

ALGUNS CÓDIGOS

Nesta seção apresentamos parte do código utilizado na linguagem do *Software* R Core Team (2015) para a aplicação dos modelos de regressão do Capítulo .

```
#############################################################################
#    Modelo Poisson k-Modificado
#############################################################################

# Funcoes de ligacao para w

pl= function(x,beta2)  exp(x%*%beta2)/(1+exp(x%*%beta2))    # Funcao Logito
#pl= function(x,beta2)  1-exp(-exp(x%*%beta2))           # Funcao C-Log-Log
#pl= function(x,beta2)  exp(-exp(-x%*%beta2))              # Funcao Gumbel

#---------------------------------------------------------------------------

n=200
k=1

#---------------------------------------------------------------------------

# Deflação

beta1=c(-0.5,0.8)   #k=0,1
beta2=c(1.5,2.5)
```

```
#-------------------------------------------------------------------------------

mu=c()
w=c()
p=c()
pk=c()
mu_kMP=c()
y=c()
X1= sort(runif(n))
x=z=matrix(c(rep(1,n),X1), nrow=n)

nk=0
while(nk==0){

  for(i in 1:n){
    mu[i]=exp(x[i,]%*%beta1)
    w[i]=pl(z[i,],beta2)
    p[i]=w[i]/(1-PI_Po(k,mu[i]))
    pk[i]=1-w[i]
    mu_kMP[i]=k*(1-p[i])+p[i]*mu[i]

    y[i]=rkmp(1,k,p[i],mu[i])
  }

  nk=length(y[y==k])
}

range(p)
plot(X1,p)



#-------------------------------------------------------------------------------

## Modelo Bayesiano

cat ("model
     {

     for(i in 1:n){
```

```
      d[i] <- equals(y[i],k)

      #link and linear predictors
      log(mu[i]) <- BETA1%*%x[i,]
      logit(omega[i]) <- BETA2%*%x[i,]
      #cloglog(omega[i]) <- BETA2%*%x[i,]
      #omega[i] <- exp(-exp(-(BETA2%*%x[i,])))

      ll[i] <- d[i]*log(1-omega[i]) + (1-d[i])*log(omega[i])
            + (1-d[i])*log(((exp(-mu[i])*pow(mu[i],y[i])/exp(loggam(y[i]+1)))
            *(1/(1-(exp(-mu[i])*pow(mu[i],k))/exp(loggam(k+1))))))

      phi[i]<-exp(ll[i])/C
      zeros[i]~dbern(phi[i])

      }

      #prior's distribuition
      BETA1 ~ dmnorm(EST_b1,tau1*He_b1)
      BETA2 ~ dmnorm(EST_b2,tau2*He_b2)
      C<- 10000000
      }",

      file = "Bayes.jags")

inicialization <- function(){
  list(BETA1 =EST_b1, BETA2 = EST_b2)
}


#-------------------------------------------------------------------------

tau1=0.5
tau2=0.5
zeros=rep(1,n)
data <- list("k", "n", "y", "x", "zeros", "EST_b1", "EST_b2",
             "He_b1", "He_b2", "tau1", "tau2")

model.params <- c("BETA1", "BETA2")
```

```
sim <- jags(data, inits = inicialization,
            parameters.to.save = model.params,
            n.chains = 2, n.iter = 50000, n.burnin = 10000, n.thin = 10,
            model.file = 'Bayes.jags')
print(sim)
```