UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelo geométrico de ordem k correlacionado

Roberta de Souza

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP
SÃO CARLOS

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

Roberta de Souza

Modelo geométrico de ordem k correlacionado

Tese 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 Doutora em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Carlos Alberto Ribeiro Diniz

USP – São Carlos
Junho de 2019

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

d719m

de Souza, Roberta
Modelo geométrico de ordem k correlacionado / Roberta de Souza; orientador Carlos Alberto Ribeiro Diniz. -- São Carlos, 2019.
124 p.

Tese (Doutorado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2019.

1. Distribuições discretas generalizadas. 2. Distribuição geométrica correlacionada. 3. Distribuição geométrica de ordem k. 4. Modelos de regressão. 5. Análise de diagnóstico. I. Ribeiro Diniz, Carlos Alberto, orient. II. Título.

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

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

Roberta de Souza

Correlated geometric model of order k

Doctoral 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 Doctorate Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Carlos Alberto Ribeiro Diniz

USP – São Carlos
June 2019

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 Tese de Doutorado da candidata Roberta de Souza, realizada em 29/08/2019:

Prof. Dr. Carlos Alberto Ribeiro Diniz
UFSCar

Prof. Dr. Nikolai Valtchev Kolev
USP

Prof. Dr. Luis Aparecido Milan
UFSCar

Profa. Dra. Teresa Cristina Martins Dias
UFSCar

Prof. Dr. Paulo Henrique Ferreira da Silva
UFBA

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Nikolai Valtchev Kolev e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Prof. Dr. Carlos Alberto Ribeiro Diniz

AGRADECIMENTOS

Sinto gratidão a tudo que, direta ou indiretamente, contribuiu para a realização deste trabalho. Menciono aqui os que foram imprescindíveis. Meus pais, José Carlos de Souza (*in memorian*) e Cláudia Olimpia de Souza, pelo apoio, amor incondicional e, acima de tudo, por terem me dado a vida. Minha irmã Gláucia de Souza e meus gatos, o Tot e a Hator, por todo amor envolvido. O orientador, Prof. Carlos Diniz, pelo conhecimento transmitido, compreensão e paciência. Os Professores do programa, em especial: José Galvão Leite, Ricardo Ehlers e Vicente Garibay Cancho pelos ensinamentos. Os amigos, Carlos Pimentel, João Ítalo Dias e Neale El-Dash e os que fiz neste programa, em especial: Clelto, Daiane, Edgar, Glauber, Jeremias, Leda, Lorena, Themis e Verônica. Os profissionais de saúde: Marcos Nogueira, Shanti Belda e Tereza Mendes. DEs-UFSCar e ICMC-USP pela oportunidade. A CAPES pelo suporte financeiro.

“It is generally recognized that the inferences of science and common sense differ from those of deductive logic and mathematics in a very important respect, namely, that, when the premisses are true and the reasoning correct, the conclusion is only probable.”

(Bertrand Russell)

RESUMO

SOUZA, R. **Modelo geométrico de ordem k correlacionado**. 2019. 124 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

Neste trabalho propomos a distribuição geométrica de ordem k correlacionada, $k \geq 1$, de parâmetros π e ρ; $\pi \in (0,1)$, $\max\{-1, -\frac{1-\pi}{\pi}\} \leq \rho < 1$, como uma extensão da generalização da distribuição geométrica proposta por Philippou e Muwafi (1980) e utilizando as idéias de Kolev, Minkova e Neytchev (2000) para generalizações de distribuições discretas provenientes de sequências de variáveis binárias. Sendo assim, é também uma releitura da distribuição geométrica de ordem k apresentada por Aki e Hirano (1993). Algumas propriedades da distribuição são demonstradas. Modelos de regressão foram desenvolvidos por ambos os métodos de estimação, clássico e bayesiano. Estudos de dados simulados mostram o comportamento das distribuições e algumas propriedades dos estimadores. A principal motivação em propor este modelo, além de contribuir para generalizações de distribuições discretas, é ter uma alternativa ainda mais adequada para análise de dados reais, pois considera-se o efeito da correlação individual existente pelo parâmetro ρ. Os ajustes dos modelos foram avaliados e análise de resíduos e de diagnóstico de influência ou divergência também é apresentada.

Palavras-chave: Distribuições discretas generalizadas, Distribuição geométrica correlacionada, Distribuição geométrica de ordem k, Modelos de regressão, Análise de diagnóstico.

ABSTRACT

SOUZA, R. **Correlated geometric model of order** k. 2019. 124 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

In this work we propose the correlated geometric distribution of order k, $k \geq 1$, with parameters π and ρ; $\pi \in (0,1)$, $\max\{-1, -\frac{1-\pi}{\pi}\} \leq \rho < 1$, as an extension of the generalized geometric distribution proposed by Philippou e Muwafi (1980) and considering the ideas of Kolev, Minkova e Neytchev (2000) for generalizations of discrete distributions by including an additional parameter ρ. Thus, it is also a re-reading of the geometric distribution of order k by Aki e Hirano (1993). Some properties of the proposed distribution are presented. Regression models are developed using classical and Bayesian estimation methods. Simulated data studies show the behavior of the distributions and some properties of the estimators. The main motivation in this research, besides contribute to generalizations of discrete distributions, is to propose an alternative analysis and even more suitable for real data, since the effect of the individual correlation is taken into account through the existence of the parameter. The fitted models are evaluated and the residual analysis and diagnosis of influence or divergence are also presented.

Keywords: Generalized discrete distributions, Correlated geometric distribution, Geometric distribution of order k, Regression models, Regression diagnostics.

LISTA DE ILUSTRAÇÕES

LISTA DE ALGORITMOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AIC	critério de informação de Akaike
AMIB	associação de medicina intensiva brasileira
BIC	critério de informação Bayesiano
CCP	cirurgia cardíaca prévia
CDC	crédito direto ao consumidor
CPO	ordenada preditiva condicional
DIC	*deviance information criterion*
DP	desvio padrão
EAIC	*expected Akaike information criterion*
EBIC	*expected Bayesian information criterion*
EMV	estimador de máxima verossimilhança
Geo	geométrica
HPD	*highest posterior density*
IC	intervalo de confiança
IGeo	geométrica correlacionada
IRC	insuficiência renal crônica
kIGeo	geométrica de ordem k correlacionada
LI	limite inferior
LPML	*log pseudo marginal likelihood*
LS	limite superior
MCMC	Monte Carlo via cadeia de Markov
MRG	modelo de regressão geométrico
MRGC	modelo de regressão geométrico correlacionado
MRGCk	modelo de regressão geométrico de ordem k correlacionado
MRGk	modelo de regressão geométrico de ordem k
MV	máxima verossimilhança
PC	probabilidade de cobertura
REQM	raiz quadrada do erro quadrático médio
SUS	sistema único de saúde
UTI	unidade de terapia intensiva

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Em análise de dados discretos a distribuição geométrica de parâmetro π ($Geo(\pi)$) pode ser utilizada quando se tem o interesse em avaliar a quantidade de falhas até que o sucesso ocorra, o que também podemos denominar de tempo de espera (discreto) até o sucesso, ou ainda, quando o interesse é determinar a probabilidade π de sucesso (ou $(1-\pi)$ de fracasso) nesta situação. Portanto, é comum sua aplicação nas diversas áreas de pesquisas, como na Medicina, em que se tem o interesse de investigar, por exemplo, qual o tempo (ou número de dias) de permanência de um paciente na unidade de terapia intensiva (UTI), o qual representa o tempo de espera até a alta do paciente. Em Economia ou Finanças, quando se tem o interesse em estimar a probabilidade de pagamento de um determinado produto pelo cliente após um período de atraso, em meses, da data de vencimento ou estimar os meses de atraso de pagamento pelo cliente.

Os exemplos acima e inúmeros outros casos podem ser modelados segundo uma distribuição geométrica, e ainda, por se tratar de um mesmo indivíduo, o sucesso ou fracasso ocorrido num determinado tempo pode estar relacionado com as respostas anteriores e haver, portanto, uma dependência entre as observações na sequência de respostas até o sucesso. Com a preocupação em considerar esta dependência, Kolev, Minkova e Neytchev (2000) incluíram um parâmetro de correlação ρ à distribuição geométrica de parâmetro π e propuseram a distribuição geométrica correlacionada com parâmetros π e ρ ($IGeo(\pi,\rho)$).

É de interesse no setor financeiro o conhecimento dos perfis de clientes relacionados a atrasos nos pagamentos de parcelas em operações de crédito bancárias, como atrasos no pagamento de consecutivas parcelas de empréstimo de crédito que tornam o cliente inadimplente. Algumas políticas internas bancárias consideram 90 dias em atraso (ou 3 meses) o tempo para inadimplência e possível transferência do cliente para uma empresa de cobranças, o que corresponde a três parcelas consecutivas não pagas. *Leasing* de veículos é outro produto de crédito bancário no qual parcelas consecutivas em atraso pelo cliente gera ação de busca e apreensão do veículo pelo banco. Estes são exemplos nos quais se pode observar as respostas

até que um número consecutivo de "sucessos" ocorra e, este número de respostas até k sucessos consecutivos pode ser modelado pela distribuição geométrica de ordem k, ou k-geométrica, de parâmetro π ($kGeo(\pi)$) proposta por Philippou e Muwafi (1980).

Além disso, o tempo até o primeiro sucesso (ou até os primeiros k consecutivos sucessos), o qual representa a variável com distribuição geométrica (ou geométrica de ordem k), pode estar relacionado com diferentes covariáveis. Nos exemplos acima apresentados, algumas variáveis como gênero, idade, peso, tipo de doença, podem ter influência na sequência de respostas de dias de permanência do paciente na UTI e; escolaridade, profissão e renda na sequência de pagamentos pelo cliente ao longo dos meses e, consequentemente, ter efeito nas probabilidades de sucesso π. Modelos de regressão podem ser usados para investigar associações entre covariáveis e variável resposta através de uma relação funcional entre estas covariáveis e, por exemplo, o parâmetro π da distribuição geométrica de interesse, de modo que mudanças nos valores das covariáveis sejam capazes de descrever o efeito em π (MCCULLAGH; NELDER, 1989).

Isto posto, nosso principal interesse é estender a proposta de Kolev, Minkova e Neytchev (2000) para os modelos geométricos de ordem k, isto é, incluir o parâmetro de correlação ρ na distribuição k-geométrica, ou seja, considerar a dependência na sequência de falhas e sucessos até a ocorrência de k consecutivos sucessos. Para isto, consideramos que a variável resposta k-geométrica é a quantidade de ensaios numa sequência de dados binários até que ocorram k sucessos consecutivos. Esta sequência pode ser vista como uma cadeia de Markov homogênea de dois estados conforme propõem Kolev, Minkova e Neytchev (2000). Consequentemente, fazemos uma releitura da distribuição geométrica de ordem k proposta por Aki e Hirano (1993) que também consideram a sequência de respostas binárias uma cadeia de Markov homogênea de dois estados e generalizam a distribuição proposta por Philippou e Muwafi (1980) incluindo as probabilidades de transição da cadeia na distribuição de probabilidade geométrica de ordem k e parâmetro π. Assim, sugerimos uma proposta para a distribuição geométrica de ordem k apresentada por Aki e Hirano (1993), construindo uma distribuição denominada distribuição geométrica de ordem k correlacionada, de dois parâmetros, π e ρ ($kIGeo(\pi,\rho)$). Também mostramos que as distribuições $Geo(\pi)$, $IGeo(\pi,\rho)$ e $kGeo(\pi)$ são casos particulares da distribuição proposta nesta tese. Algumas propriedades da nova distribuição são apresentadas, entre elas, a média e a variância.

A construção de modelos de regressão para os modelos probabilísticos $IGeo(\pi,\rho)$ e $kGeo(\pi)$ são discutidos em ambas as abordagens, clássica e Bayesiana. Na abordagem clássica, obtemos a estimação dos parâmetros dos modelos e intervalos de confiança e testes de hipóteses assintóticos para inferência dos parâmetros. Utilizamos técnicas de seleção de modelos para determinar a melhor função de ligação. Obtemos as estimativas dos parâmetros em estudos com dados simulados, verificando suas propriedades assintóticas. Avaliamos a qualidade dos ajustes dos modelos por uma análise de diagnóstico que verifica a adequação do ajuste dos modelos e pode identificar valores discrepantes e/ou observações influentes nas estimativas. Nesta avalia-

ção utilizamos medidas de distância apropriadas e técnicas gráficas considerando os resíduos quantílicos aleatorizados e medidas de influência global e local. Na abordagem Bayesiana, estimamos os coeficientes do modelo de regressão utilizando algoritmo de Metrópolis-Hastings. Utilizamos critérios Bayesianos de selecão para a escolha do melhor modelo e realizamos uma análise de diagnóstico para avaliar o ajuste do modelo Bayesiano através de resíduos quantílicos aleatorizados a *posteriori* e observações influentes por medidas de divergência-ψ.

A metodologia desenvolvida é ilustrada através de análise de dados reais. O modelo de regressão geométrico correlacionado $IGeo(\pi,\rho)$ é aplicado em um conjunto de dados reais de pacientes com doença de aorta para avaliação da probabilidade de alta do paciente da UTI, verificando efeitos de covariáveis e da correlação no tempo de internação na UTI após o paciente ter sido submetido a um procedimento cirúrgico. Os modelos geométricos de ordem k, $kGeo(\pi)$ e $kIGeo(\pi,\rho)$, são ajustados a dados reais de clientes de banco para avaliação da inadimplência em operações de crédito, verificando o efeito de covariáveis e correlação na probabilidade de atraso de pagamento da parcela pelo cliente.

1.1 Revisão da literatura

A literatura apresenta uma linha de trabalhos com a preocupação em considerar a dependência entre observações na sequência de dados discretos com distribuições clássicas conhecidas, como as de Poisson, binomial e a geométrica. Luceño (1995) e Luceño e Ceballos (1995) propuseram a distribuição binomial generalizada com parâmetros n, p e parâmetro de correlação ρ, e modelos parcialmente correlacionados binomial e Poisson. Conway e Maxwell (1962) generalizaram a distribuição de Poisson acrescentando um parâmetro, ν, que mede a extensão da associação entre variáveis com distribuição de Poisson de parâmetro λ; esta distribuição foi nomeada de Conway-Maxwell-Poisson com parâmetros λ e ν ($COMP(\lambda,\nu)$). Shmueli *et al.* (2005) estenderam este resultado para variáveis com distribuição binomial; e Kadane *et al.* (2016) definiram esta distribuição com dois parâmetros como uma generalização da distribuição binomial e a nomeia *COMB*. Kolev, Minkova e Neytchev (2000) propuseram a distribuição geométrica inflacionada (inflacionada ou deflacionada de zeros) e mostra sua equivalência com a distribuição geométrica correlacionada de dois parâmetros ($IGeo(\pi,\rho)$), generalizando a distribuição geométrica usual de parâmetro π ($Geo(\pi)$) pela inclusão do coeficiente de correlação ρ como um parâmetro adicional, considerando para isto a sequência de dados binários dependentes.

Pensando em ensaios até que um número consecutivo de sucessos ocorra, Philippou e Muwafi (1980) introduziram a distribuição geométrica de ordem k, também como uma generalização da distribuição geométrica usual, considerando como resposta a quantidade de ensaios (ou tempo de espera) até k sucessos consecutivos, o que pode ser observado na sequência de variáveis binárias ($0 = falha$, $1 = sucesso$) independentes até a ocorrência de k sucessos consecutivos. Generalizações de distribuições discretas usuais (de ordem 1) para distribuições de

ordem k foram feitas em grande parte pelo autor Andreas N. Philippou. Inicialmente, Philippou e Muwafi (1980) obtiveram a distribuição de probabilidade de uma variável aleatória que denota o número de ensaios até o k-ésimo sucesso consecutivo em ensaios de Bernoulli independentes com probabilidade de sucesso π, obtendo as possíveis combinações das sequências em termos de coeficientes multinomiais, baseando-se na sequência de Fibonacci de ordem k (GABAI, 1970). Motivados por estes resultados, dois anos depois Philippou, Georghiou e Philippou (1983) nomearam esta distribuição de geométrica de ordem k e parâmetro π ($kGeo(\pi)$), e derivaram algumas de suas propriedades, tais como função geradora de probabilidade, esperança e variância. Neste mesmo trabalho, os autores ainda obtiveram a distribuição binomial negativa de ordem k, de parâmetros π e r, pela soma de r variáveis aleatórias com distribuição $kGeo(\pi)$ independentes; e também a distribuição de Poisson de ordem k, demonstrando, sob certas condições de regularidade, que se refere a uma distribuição binomial negativa de ordem k quando $r \to \infty$. Em seguida, Philippou (1984) apresentaram a distribuição binomial negativa de ordem k e suas propriedades, e Philippou e Makri (1985) demonstraram uma função alternativa para a distribuição de probabilidade $kGeo(\pi)$, uma fórmula recursiva mais simples e bastante útil para propósitos computacionais. Philippou e Makri (1986) obtiveram a distribuição de probabilidade do número de sequências (corridas) de sucessos consecutivos de tamanho k e verificam que esta distribuição é binomial de ordem k. Hirano (1986) revisou e relacionou algumas distribuições generalizadas de ordem k derivadas da distribuição geométrica de ordem k e também avaliou algumas de suas propriedades. Aki e Hirano (1993) sugeriram a distribuição geométrica de ordem k considerando a sequência de variáveis binárias uma cadeia de Markov homogênea de dois estados e, portanto, dependência dentro da sequência. Após um ano, Aki e Hirano (1994) propuseram as distribuições dos números de falhas e de sucessos até k sucessos consecutivos em sequências de variáveis aleatórias binárias, tanto em sequência de repostas independentes e identicamente distribuídas como também em cadeia de Markov homogênea.

A relação funcional entre covariáveis e o parâmetro π em modelos correlacionados pode ser investigada por modelos de regressão, o que pode ser visto em diversos trabalhos da literatura, como em Williams (1982), que apresenta o modelo linear logístico considerando correlação comum entre as variáveis da sequência de Bernoulli dentro do *cluster*. Um modelo de regressão probito para modelar observações binárias equicorrelacionadas é descrito em Ochi e Prentice (1984). Diniz, Tutia e Leite (2010), numa abordagem Bayesiana, avaliaram o modelo binomial correlacionado generalizado por Luceño (1995) e Luceño e Ceballos (1995). Pires e Diniz (2012) propuseram abordagem Bayesiana para modelos de regressão binomial correlacionados.

1.2 Organização dos capítulos

A princípio, para melhor compreensão dos modelos, foram estudados os modelos probabilísticos geométricos generalizados por Kolev, Minkova e Neytchev (2000) e Philippou e Muwafi (1980), e desenvolvidos os respectivos modelos de regressão nas duas abordagens

(clássica e Bayesiana). Neste texto que mostra nosso trabalho, inicialmente apresentamos nossa proposta, o modelo probabilístico geométrico de ordem k correlacionado de parâmetros π e ρ e, posteriormente, os casos particulares como modelos de regressão. Os capítulos são apresentados na seguinte ordem:

- Capítulo 1: introdução, motivação, revisão da literatura, organização da apresentação e desenvolvimento de nosso trabalho e algumas definições e revisão de conceitos preliminares para modelos de regressão e análises de diagnóstico.
- Capítulo 2: apresentação dos modelos probabilísticos geométricos já mencionados, o modelo geométrico correlacionado e os modelos geométricos de ordem k, os quais serviram de base para nossa proposta apresentada na última seção deste capítulo, o modelo probabilístico geométrico de ordem k correlacionado, de parâmetros π e ρ.
- Capítulo 3: modelo de regressão geométrico correlacionado (MRGC) desenvolvido numa abordagem clássica: definição do modelo, estimação, intervalos de confiança e testes assintóticos, seleção de modelos, resíduos e medidas influentes, estudo de simulação, aplicação em dados reais, comparação com o modelo geométrico usual.
- Capítulo 4: modelo de regressão geométrico correlacionado (MRGC) desenvolvido numa abordagem Bayesiana: definição do modelo, estimação, seleção de modelos, resíduos e medidas influentes, estudo de simulação, aplicação em dados reais, comparação com o modelo geométrico usual de parâmetro π.
- Capítulo 5: modelo de regressão geométrico de ordem k (MRGk) desenvolvido numa abordagem clássica: definição do modelo, estimação, intervalos de confiança e testes assintóticos, seleção de modelos, resíduos e medidas influentes, estudo de simulação, aplicação em dados reais.
- Capítulo 6: modelo de regressão geométrico de ordem k correlacionado (MRGCk) desenvolvido numa abordagem Bayesiana: definição do modelo, estimação, seleção de modelos, resíduos e medidas influentes, estudo de simulação, aplicação em dados reais.
- Capítulo 7: conclusões e sugestões para pesquisas futuras.
- Referências Bibliográficas: bibliografia utilizada para estudo e desenvolvimento do trabalho.

1.3 Conceitos e resultados preliminares

Nesta seção apresentamos alguns conceitos e resultados utilizados para o desenvolvimento dos modelos propostos. A primeira subseção apresenta as funções de ligação e suas derivadas de primeira e segunda ordens utilizadas para as inferências dos modelos de regressão

frequentistas. Como parte da análise de diagnóstico dos modelos, o resíduo quantílico aleatorizado é introduzido na segunda subseção. Este resíduo é apropriado para dados discretos assimétricos, como é o caso de dados com distribuições geométricas.

1.3.1 Funções de ligação

Segundo Cordeiro e Demétrio (2007), a seleção de modelos é uma parte importante de toda pesquisa em modelagem estatística e envolve a procura de um modelo que descreva bem os dados observados. A escolha de um modelo adequado implica em melhores inferências sobre os parâmetros do modelo. Modelos de regressão envolvem um componente aleatório (variável resposta), um componente sistemático (variáveis explanatórias ou covariáveis) e uma amostra aleatória de m observações independentes. A ligação entre o componente aleatório e o componente sistemático se faz pela função de ligação.

Como também nosso objetivo é desenvolver modelos de regressão geométricos, e as respostas de variáveis com distribuições geométricas são obtidas em sequências de respostas binárias W_i, $W_i = \{0,1\}$, $i = 1,2,\ldots$, a qual denominamos 0 como $fracasso$ e 1 como *sucesso*, é de interesse principal modelar as probabilidades de sucesso $\pi's$. Nos modelos geométricos, portanto, os componentes aleatórios apresentam distribuições geométricas. As covariáveis descrevem a estrutura linear para os parâmetros (preditor linear) e, então, as $\pi's$ podem ser modeladas em função das covariáveis por meio de funções de ligação que garantam, para quaisquer valores dos parâmetros do preditor linear, um valor para as probabilidades de sucesso $\pi's$ no intervalo $(0,1)$.

Supondo $y_1, y_2, \ldots, y_m$ observações de m variáveis aleatórias independentes $Y_1, Y_2, \ldots, Y_m$ seguindo distribuições de probabilidade com parâmetros $\pi_i \in (0,1)$, $i = 1,2,\ldots,m$, respectivamente, $m \in \mathbb{N}^*$, modela-se π_i em função de covariáveis através da função de ligação $g(.)$, função estritamente monótona e duplamente diferenciável, de modo que, para cada observação da covariável x_{ri}, $r = 1,2,\ldots,R$, tem-se:

$$g(\pi_i) = \eta_i = \sum_{r=0}^{R} \beta_r x_{ri} = \boldsymbol{x}_i^T \boldsymbol{\beta}, \tag{1.1}$$

$$\pi_i = g^{-1}(\eta_i) = g^{-1}\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right), \tag{1.2}$$

em que $\boldsymbol{x}_i = (1, x_{1i}, \ldots, x_{Ri})^T$ e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_R)^T$. Para variáveis discretas provenientes de sequências de respostas binárias são úteis as funções de ligação logito, complementar log-log, log-log e probito propostas em McCullagh e Nelder (1989) e descritas na Tabela 1.

O método clássico de estimação por máxima verossimilhança é um dos mais habituais para se obter estimadores dos parâmetros de modelos. Isto é, para cada amostra $\boldsymbol{y}$ obtem-se os valores de $\boldsymbol{\pi}$ pelos quais a função de verossimilhança $\mathscr{L}(\boldsymbol{\pi}|\boldsymbol{y};\boldsymbol{x})$ tem valor máximo como função de $\boldsymbol{\pi}$, os quais são denominados estimadores de máxima verossimilhança (EMV) $\hat{\boldsymbol{\pi}}$. No modelo de regressão, temos que $\pi_i = g^{-1}\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)$, conforme (1.2). Pela propriedade de invariância dos

Tabela 1 – Funções de ligação para modelar π_i.

	$g(\pi_i)$	$g^{-1}(\eta_i)$
Logito	$\log\left(\frac{\pi_i}{1-\pi_i}\right)$	$\frac{\exp(\eta_i)}{1+\exp(\eta_i)}$
C log-log	$\log(-\log(1-\pi_i))$	$1-\exp(-\exp(\eta_i))$
Log-log	$-\log(-\log(\pi_i))$	$\exp(-\exp(-\eta_i))$
Probito	$\Phi^{-1}(\pi_i)$	$\Phi(\eta_i)$

Φ é a função de distribuição Normal padrão acumulada.

EMV, temos que se $\hat{\boldsymbol{\pi}}$ é EMV de $\boldsymbol{\pi}$, então para a função $g(\boldsymbol{\pi})$, $g(.)$ função bijetora, o EMV de $g(\boldsymbol{\pi})$ é $g(\hat{\boldsymbol{\pi}})$. Temos de (1.2) que $g(.)$ possui inversa, logo bijetora (LIMA, 2004). Analogamente, $\hat{\boldsymbol{\pi}} = g^{-1}(\hat{\boldsymbol{\eta}})$ é EMV de $\boldsymbol{\pi}$, pois $g^{-1}(.)$ possui inversa, com $\hat{\boldsymbol{\eta}} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$.

Para o modelo de regressão, a função de verossimilhança $\mathscr{L}(\boldsymbol{\pi}|\boldsymbol{y})$ é dada em termos dos parâmetros $\boldsymbol{\beta}$, $\mathscr{L}(\boldsymbol{\beta}|\boldsymbol{y};\boldsymbol{x})$, e então $\hat{\boldsymbol{\beta}}(\boldsymbol{Y}|\boldsymbol{x})$ são EMV de $\boldsymbol{\beta}$. Se a função de verossimilhança é diferenciável em β_r, uma condição necessária para que $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_R)$ seja EMV é que

$$\frac{\partial \mathscr{L}(\boldsymbol{\beta}|\boldsymbol{y};\boldsymbol{x})}{\partial \beta_r} = 0, \qquad r = 0, 1, \ldots, R. \tag{1.3}$$

Pontos em que as primeiras derivadas são nulas podem ser pontos de mínimo ou máximo, tanto locais como globais. Logo, se $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_R)$ satisfaz (1.3) e a matriz $\boldsymbol{H}$ de elementos das segundas derivadas da função de verossimilhança (matriz Hessiana) no ponto $\hat{\boldsymbol{\beta}}$ for definida negativa, ou seja, $\boldsymbol{z}'\boldsymbol{H}\boldsymbol{z} < 0$, $\forall \boldsymbol{z} \neq 0$, sendo cada elemento de $\boldsymbol{H}$ dado por:

$$h_{r,s} = \frac{\partial^2 \mathscr{L}(\boldsymbol{\beta}|\boldsymbol{y};\boldsymbol{x})}{\partial \beta_r \partial \beta_s}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}, \quad r, s = 0, 1, \ldots, R, \tag{1.4}$$

temos ponto de máximo e ainda se $\hat{\boldsymbol{\beta}}$ for solução única para (1.3) e o valor da função de verossimilhança em $\hat{\boldsymbol{\beta}}$ maior que para valores extremos de β_r ($\beta_r = \pm\infty$), temos um ponto de máximo global, e então EMV.

A função de verossimilhança é dada em termos dos parâmetros $\boldsymbol{\beta}$ como função de $g^{-1}(\boldsymbol{\eta})$. Logo, de (1.3) e (1.4), utilizamos as derivadas de primeira e segunda ordens de $g^{-1}(.)$ para a estimação dos parâmetros $\boldsymbol{\beta}$, as quais são apresentadas na Tabela 2 referentes às funções de ligação da Tabela 1.

Tabela 2 – Derivadas de primeira e segunda ordens das funções de ligação em relação aos parâmetros $\boldsymbol{\beta}$.

Função de ligação	$\Delta_i(\beta_r) = \partial g^{-1}(\eta_i)/\partial\beta_r$	$\Delta_i^2(\beta_r\beta_s) = \partial^2 g^{-1}(\eta_i)/\partial\beta_r\partial\beta_s$
Logito/Logística	$x_{ir}\exp(\eta_i)(1+\exp(\eta_i))^{-2}$	$x_{ir}x_{is}(1+\exp(\eta_i))^{-3}(1-\exp(\eta_i))$
Complemento log-log	$x_{ir}\exp(\eta_i-\exp(\eta_i))$	$x_{ir}x_{is}\exp(\eta_i-\exp(\eta_i))(1-\exp(\eta_i))$
Log-log	$x_{ir}\exp(-\eta_i-\exp(-\eta_i))$	$-x_{ir}x_{is}\exp(-\eta_i-\exp(-\eta_i))(1-\exp(-\eta_i))$
Probito/Normal Inversa	$x_{ir}\phi(\eta_i)$	$-x_{ir}x_{is}\eta_i\phi(\eta_i)$

ϕ é função densidade de probabilidade da distribuição Normal padrão.

1.3.2 Resíduo quantílico aletorizado

De acordo com Paula (2004), para a análise de um ajuste de modelo de regressão a um conjunto de dados é importante a verificação de possíveis afastamentos das suposições feitas para o modelo, especialmente para as partes aleatórias e sistemáticas do modelo, bem como a existência de observações extremas com alguma interferência desproporcional nos resultados do ajuste. Algum desvio sistemático entre os valores observados e os valores ajustados ou, um ou mais valores observados discrepantes em relação aos demais, podem resultar num ajuste insatisfatório. Escolhas inadequadas da função de variância, da função de ligação, da matriz do modelo ou ainda erro na definação da escala das variáveis do modelo podem ocasionar desvios sistemáticos. As discrepâncias isoladas podem ocorrer devido a valores nos extremos da amplitude de validade da covariável ou porque algum fator não controlado influenciou sua obtenção, ou ainda, porque realmente foram provenientes de uma leitura ou transcrição mal feitas (CORDEIRO; DEMÉTRIO, 2007).

A análise de diagnóstico envolve, portanto, algumas técnicas baseadas em exames visuais e medidas de distância para verificação do ajuste do modelo. A análise de resíduos pode detectar a presença de pontos extremos e avaliar a adequação da distribuição da variável resposta. Outra técnica é a detecção de observações influentes que exercem um peso desproporcional nas estimativas dos parâmetros. Muitas são as referências com definições e propostas de resíduos para modelos de regressão. Esta seção mostra um resíduo que se apresentou mais adequado para modelos geométricos, devido ao comportamento assimétrico da distribuição, o resíduo quantílico aleatorizado proposto por Dunn e Smyth (1996), pois ele corrige os efeitos de assimetria e curtose. Após o ajuste do modelo, as distâncias entre os valores ajustados (preditos) e os valores observados na amostra podem ser verificadas por estes resíduos. A visualização gráfica dos resíduos em função das observações nos permite observar o comportamento destas distâncias e assim identificar valores discrepantes.

Os resíduos quantílicos aleatorizados são obtidos pela função inversa da função de distribuição acumulada normal padrão para cada resposta e, assim, obtidos os respectivos quantis, portanto nomeados resíduos quantílicos. Seja $F(\boldsymbol{y};\boldsymbol{\pi})$ a função de distribuição acumulada de $\boldsymbol{Y}$. Os resíduos quantílicos r_{q_i} são dados por:

$$r_{q_i} = \Phi^{-1}\{F(y_i;\hat{\pi}_i)\}, \tag{1.5}$$

em que Φ é a função de distribuição acumulada normal padrão. Se F é contínua, então $F(y_i;\pi_i)$ é uniformemente distribuída no intervalo $(0,1)$ e r_{q_i} converge para a normal padrão se os $\hat{\pi}_i$'s forem estimadores consistentes. Se F não é contínua, o resíduo é mais geral, obtido aleatoriamente, nomeado resíduo quantílico aleatorizado. Seja $a_i = \lim_{y \uparrow y_i} F(y;\hat{\pi}_i)$ e $b_i = F(y_i;\hat{\pi}_i)$, o resíduo quantílico aleatorizado r_{q_i} é dado por:

$$r_{q_i} = \Phi^{-1}(u_i), \tag{1.6}$$

em que u_i é uma variável aleatória com distribuição uniforme no intervalo $(a_i, b_i]$ e, portanto, r_{q_i} tem distribuição normal padrão. Esta aleatorização tem como objetivo evitar massas de pontos sobrepostos na distribuição dos resíduos. Para distribuições discretas, como a geométrica, a_i e b_i são obtidos por:

$$a_i = F(y_i - 1; \hat{\pi}_i) \quad \text{e} \quad b_i = F(y_i; \hat{\pi}_i). \tag{1.7}$$

Desta forma, visualizando os resíduos em função das observações e/ou dos percentis de uma distribuição normal padrão ($QQplot$), é possível verificar a normalidade dos resíduos quantílicos aleatorizados e identificar distâncias discrepantes entre estimativas e observações.

CAPÍTULO

2

DISTRIBUIÇÕES GEOMÉTRICAS GENERALIZADAS

Neste capítulo apresentamos, nas primeiras seções, os modelos probabilísticos geométricos que nos serviram de base para o desenvolvimento do modelo proposto, os quais são generalizações da distribuição geométrica clássica já conhecida, o modelo geométrico correlacionado sugerido por Kolev, Minkova e Neytchev (2000) e os modelos geométricos de ordem k propostos por Philippou e Muwafi (1980) e Aki e Hirano (1993), respectivamente. Na terceira seção é proposto o modelo probabilístico geométrico de ordem k correlacionado de dois parâmetros, π e ρ, como extensão do modelo geométrico correlacionado e do modelo geométrico de ordem k, e, também, como uma releitura da distribuição geométrica de ordem k de variáveis aleatórias obtidas pela contagem de ensaios numa cadeia de Markov homogênea de dois estados apresentada por Aki e Hirano (1993). Mostramos que estas distribuições geométricas generalizadas são casos particulares da distribuição proposta. Algumas propriedades, como esperança, variância e função geradora de probabilidade da variável aleatória com distribuição geométrica de ordem k correlacionada ($kIGeo(\pi,\rho)$) também são apresentadas.

Sabe-se que a variável aleatória Y tem distribuição geométrica com parâmetro π, $\pi \in (0,1)$, quando Y representa o tempo de espera até o primeiro *sucesso*, em uma sequência de ensaios de Bernoulli independentes (ou sequência de variáveis binárias independentes) com probabilidade π de sucesso (portanto, $1-\pi$ de falha). A variável aleatória Y tem função de probabilidade dada por:

$$P(Y=y|\pi) = (1-\pi)^{y-1}\pi I_{\{1,2,\ldots\}}(y), \tag{2.1}$$

e média e variância dadas por:

$$E(Y) = \frac{1}{\pi} \quad \text{e} \quad Var(Y) = \frac{(1-\pi)}{\pi^2}. \tag{2.2}$$

A variável aleatória Y também pode representar a quantidade de $falhas$ até o primeiro

sucesso, o que corresponde ao tempo de espera até a última $falha$ que antecede o primeiro *sucesso* em uma sequência de ensaios de Bernoulli independentes com probabilidade de sucesso π. Desta forma, a função de probabilidade é dada por,

$$P(Y = y|\pi) = (1-\pi)^y \pi I_{\{0,1,2,\ldots\}}(y), \tag{2.3}$$

com média e variância, respectivamente,

$$E(Y) = \frac{1-\pi}{\pi} \quad e \quad Var(Y) = \frac{(1-\pi)}{\pi^2}. \tag{2.4}$$

2.1 Distribuição Geométrica Correlacionada

Kolev, Minkova e Neytchev (2000) sugerem uma extensão da distribuição geométrica incluindo o parâmetro adicional ρ à distribuição clássica, a qual tem uma interpretração direta em termos de "inflação de zeros" e então a nomeia de distribuição geométrica inflacionada ($IGeo(\pi,\rho)$). Neste caso, $\max\{-1, -\frac{1-\pi}{\pi}\} \leq \rho < 1$, com a "inflação de zeros" podendo ser positiva ou negativa (inflacionada ou deflacionada). Esta distribuição inflacionada é também interpretada como o tempo de espera até o fracasso que antecede o primeiro sucesso em uma sequência de ensaios de Bernoulli equicorrelacionados com probabilidade de sucesso π, $\pi \in (0,1)$, e coeficiente de correlação ρ, e então como distribuição geométrica correlacionada.

Assim, a variável aleatória Y tem distribuição geométrica correlacionada ($IGeo(\pi,\rho)$), quando Y é observada numa sequência de variáveis binárias equicorrelacionadas, $\{W_i\}_{i\geq 1}$, $W_i = 0,1$, com probabilidades, $1-\pi$ e π, respectivamente, $i = 1,2,\ldots$, e coeficientes de correlação constantes ρ, ou seja, $E(W_i) = \pi$, $\text{Var}(W_i) = \pi(1-\pi)$ e $\text{Corr}(W_i, W_j) = \rho$, para todo i, j, $i \neq j$. A distribuição de probabilidade de Y é dada por:

$$P(Y = y|\pi,\rho) = \pi I_{\{0\}}(y) + \pi(1-\pi)(1-\rho)[(1-\pi)(1-\rho)+\rho]^{y-1} I_{\{1,2,\ldots\}}(y) \tag{2.5}$$

e a média e variância de Y são dadas por:

$$E(Y) = \frac{1-\pi}{\pi(1-\rho)} \quad e \quad Var(Y) = \frac{(1-\pi)(1+\pi\rho)}{\pi^2(1-\rho)^2}. \tag{2.6}$$

Nota-se que, quando existe uma correlação negativa ($\rho < 0$) na sequência de variáveis de Bernoulli, o modelo clássico, o qual não considera esta dependência entre as variáveis ($\rho = 0$), pode superestimar a probabilidade de sucesso. O contrário ocorre com um coeficiente de correlação positivo ($\rho > 0$) em que o modelo clássico poderá então subestimar esta probabilidade.

2.2 Distribuições Geométricas de Ordem k

Philippou e Muwafi (1980) propuseram a distribuição geométrica de ordem k, $k \in \{1,2,\ldots\}$, a qual denotaremos por $kGeo(\pi)$, também como uma generalização da distribuição

geométrica clássica, considerando como resposta uma sequência de variáveis binárias ($0 = fracasso$, $1 = sucesso$) independentes até a ocorrência de k sucessos consecutivos. Portanto, a variável aleatória Y_k tem distribuição geométrica de ordem k ($kGeo(\pi)$) quando Y_k representa o número de ensaios até a ocorrência de k sucessos consecutivos, com probabilidade de sucesso π, $\pi \in (0,1)$, e $k \in \{1,2,...\}$. Também pode-se dizer que Y_k representa o tempo de espera até os primeiros k sucessos consecutivos. Sua função de probabilidade é dada por:

$$P(Y_k = y+k|\pi) = \sum_{y_1,\ldots,y_k} \binom{y_1+\ldots+y_k}{y_1,\ldots,y_k} (1-\pi)^{\sum_{i=1}^{k} y_i} \pi^{y+k-\sum_{i=1}^{k} y_i} I_{\{0,1,2,\ldots\}}(y), \tag{2.7}$$

sendo o somatório em relação a todos os termos $y_1,\ldots,y_k$ inteiros não negativos tais que $y_1+2y_2+\ldots+ky_k = y$. Pois, neste caso, um elemento do evento $\{"y+k"\}$ é uma combinação $\{"w_1w_2\ldots w_{y_1+\ldots+y_k}\underbrace{11\ldots1}_{k}"\}$, tal que y_1 de $w's$ são "0", y_2 de $w's$ são "10", y_3 de $w's$ são "110", ..., y_k de $w's$ são "$\underbrace{11\ldots1}_{k-1}$0", e $y_1+2y_2+\ldots+ky_k = y$. Para $y_1,y_2,\ldots,y_k$ fixados, o número de combinações possíveis é

$$\binom{y_1+\ldots+y_k}{y_1,\ldots,y_k}$$

e, para ensaios independentes, cada uma delas tem probabilidade

$$\begin{aligned} P\{"w_1w_2\ldots w_{y_1+\ldots+y_k}\underbrace{11\ldots1}_{k}"\} &= P\{"0"\}^{y_1}P\{"10"\}^{y_2}\ldots P\{"\underbrace{11\ldots1}_{k-1}0"\}^{y_k}P\{"\underbrace{11\ldots1}_{k}"\} \\ &= (1-\pi)^{\left(\Sigma_{i=1}^{k} y_i\right)}\pi^{\left(y+k-\Sigma_{i=1}^{k} y_i\right)}. \end{aligned}$$

Portanto, considerando todas as possíveis combinações e todos os inteiros não negativos $y_1,\ldots,y_k$ que satisfaçam $y_1+2y_2+\ldots+ky_k = y$, obtém-se a probabilidade em (2.7).

Philippou, Georghiou e Philippou (1983) mostram que a média e variância de Y_k são dadas por:

$$E(Y_k) = \frac{1-\pi^k}{(1-\pi)\pi^k}, \qquad Var(Y_k) = \frac{[1-(2k+1)(1-\pi)\pi^k - \pi^{2k+1}]}{(1-\pi)^2\pi^{2k}} \tag{2.8}$$

e função geradora de probabilidade $\varphi(t)$ para $|t| < 1$:

$$\varphi(t) = \frac{(\pi t)^k(1-\pi t)}{1-t+(1-\pi)\pi^k t^{k+1}}. \tag{2.9}$$

Nota-se que, para $k = 1$, tem-se distribuição, esperança e variância da variável aleatória com distribuição geométrica clássica.

Com o interesse em sequências dependentes de variáveis aleatórias de distribuições discretas, Aki e Hirano (1993) obtiveram a distribuição do tempo de espera até k sucessos consecutivos em sequências de uma cadeia de Markov homogênea de dois estados. Respostas (ensaios) independentes de variáveis com distribuições geométricas (clássica e de ordem k) são

obtidas em sequências de variáveis aleatórias independentes de valores $0,1$ (binárias), sendo assim, por cadeias de Markov, obtemos ensaios independentes de variáveis aleatórias discretas em sequências estacionárias dependentes. Com isto, é possível investigar distribuições geométricas relacionadas a eventos sucessivos em sequências estacionárias dependentes de variáveis aleatórias binárias (AKI; HIRANO, 1993), conforme também apresentado por Kolev, Minkova e Neytchev (2000) (Seção 2.1) para a distribuição geométrica (de ordem 1).

Aki e Hirano (1993), consideram uma cadeia de Markov homogênea de dois estados $\{X_n, n \geq 0\}$, $X_n = 0,1$, com probabilidades de estado inicial, $P(X_0 = 0) = p_0$ e $P(X_0 = 1) = p_1 = 1 - p_0$, $0 < p_0 < 1$; e com probabilidades de transição dos estados: $P(X_{n+1} = 0|X_n = 0) = p_{00}$, $P(X_{n+1} = 1|X_n = 0) = p_{01} = 1 - p_{00}$, $P(X_{n+1} = 0|X_n = 1) = p_{10}$ e $P(X_{n+1} = 1|X_n = 1) = p_{11} = 1 - p_{10}$, e propõem a seguinte distribuição de probabilidade de Y_k e respectiva função geradora de probabilidade $\varphi(t)$ para $|t| < 1$, com Y_k representando o número de transições até os primeiros k sucessos consecutivos na cadeia de Markov $\{X_n, n \geq 0\}$ apresentada:

$$\begin{aligned} P(Y_k = y+k) &= \{p_0 p_{01} p_{11}^{(k-1)} + p_1 p_{11}^k\} I_{\{0\}}(y) + \Bigg\{ p_0 \sum_{y_1,\ldots,y_k} \binom{y_1+\ldots+y_k}{y_1,\ldots,y_k} \\ &\quad p_{00}^{y_1} p_{01}^{y_2+\ldots y_k+1} p_{10}^{\sum_{i=2}^k y_i} p_{11}^{y_3+2y_4+\ldots+(k-2)y_k+k-1} + p_1 \sum_{j=1}^{k} p_{11}^{(j-1)} p_{10} \\ &\quad \sum_{x_1,\ldots,x_k} \binom{x_1+\ldots+x_k}{x_1,\ldots,x_k} p_{00}^{x_1} p_{01}^{x_2+\ldots+x_k+1} p_{10}^{\sum_{i=2}^k x_i} p_{11}^{x_3+\ldots+(k-2)x_k+k-1} \Bigg\} I_{\{1,2,\ldots\}}(y), \end{aligned} \tag{2.10}$$

para $k \geq 2$ e o somatório de todos os termos $y_1,\ldots,y_k$, inteiros não negativos, é tal que $y_1 + 2y_2 + \ldots + ky_k = y$ e, em relação a todos os termos $x_1,\ldots,x_k$, também inteiros não negativos, é tal que $x_1 + 2x_2 + \ldots + kx_k = y - j$, $j = 1,\ldots,k$.

$$\varphi(t) = \frac{(p_0 p_{01} + p_1 p_{11}) p_{11}^{k-1} t^k + p_1 (p_{01} p_{10} - p_{00} p_{11}) p_{11}^{k-1} t^{k+1}}{1 - (p_{00} t + p_{01} p_{10} t^2 + p_{01} p_{10} p_{11} t^3 + \ldots + p_{01} p_{10} p_{11}^{k-2} t^k)}. \tag{2.11}$$

Primeiro e segundo momentos não foram apresentados pelos autores para o cálculo de média e variância.

2.3 Distribuição Geométrica de Ordem k Correlacionada

Nesta seção propomos o modelo probabilístico geométrico de ordem k correlacionado como extensão dos modelos apresentados nas seções 2.1 e 2.2. Para isto, utilizamos a idéia de Kolev, Minkova e Neytchev (2000), que é obter as probabilidades de transição da cadeia de Markov em função dos parâmetros π e ρ, e com isto, reformulamos a distribuição de probabilidade de Aki e Hirano (1993).

Proposição 2.3.1. *Seja Y_k uma variável aleatória que representa o número de ensaios até a ocorrência de k sucessos consecutivos numa sequência de respostas binárias equicorrelacionadas, isto é, numa sequência de ensaios de Bernoulli equicorrelacionados com probabilidade de sucesso π, $\pi \in (0,1)$, e coeficiente de correlação ρ, $\max\{-1, -\frac{1-\pi}{\pi}\} \leq \rho < 1$, sendo k um número inteiro ≥ 1. Y_k tem função distribuição de probabilidade dada por:*

$$\begin{aligned} P(Y_k = y+k) &= \{(1-\pi)[(1-\rho)\pi][\pi(1-\rho)+\rho]^{(k-1)} + \pi[\pi(1-\rho)+\rho]^k\} I_{\{0\}}(y) + \\ &+ \left\{(1-\pi) \sum_{y_1,\ldots,y_k} \binom{y_1+\ldots+y_k}{y_1,\ldots,y_k} [(1-\pi)(1-\rho)+\rho]^{y_1} [(1-\rho)\pi]^{y_2+\ldots y_k+1}\right. \\ &\quad [(1-\pi)(1-\rho)]^{\sum_{i=2}^k y_i} [\pi(1-\rho)+\rho]^{y_3+2y_4+\ldots+(k-2)y_k+k-1} + \\ &+ \pi \sum_{j=1}^{k} [\pi(1-\rho)+\rho]^{(j-1)} [(1-\pi)(1-\rho)] \sum_{x_1,\ldots,x_k} \binom{x_1+\ldots+x_k}{x_1,\ldots,x_k} \\ &\quad [(1-\pi)(1-\rho)+\rho]^{x_1} [(1-\rho)\pi]^{x_2+\ldots+x_k+1} [(1-\pi)(1-\rho)]^{\sum_{i=2}^k x_i} \\ &\quad \left.[\pi(1-\rho)+\rho]^{x_3+\ldots+(k-2)x_k+k-1}\right\} I_{\{1,2,\ldots\}}(y), \end{aligned} \tag{2.12}$$

em que o somatório de todos os termos $y_1,\ldots,y_k$, inteiros não negativos, é tal que $y_1+2y_2+\ldots+ky_k = y$ e, em relação a todos os termos $x_1,\ldots,x_k$, também inteiros não negativos, é tal que $x_1+2x_2+\ldots+kx_k = y-j$, $j = 1,\ldots,k$.

Demonstração. Na Seção 2.1 foi visto que Kolev, Minkova e Neytchev (2000) propuseram a distribuição de probabilidade em (2.5) da variável aleatória Y que representa o número de falhas até o primeiro sucesso numa sequência de respostas binárias equicorrelacionadas, a distribuição geométrica correlacionada de dois parâmetros, com probabilidade de sucesso π, $\pi \in (0,1)$, e coeficiente de correlação ρ, $\max\{-1, -\frac{1-\pi}{\pi}\} \leq \rho < 1$, para a qual apresentam as seguintes interpretações: se considerarmos uma cadeia de Markov homogênea de dois estados $\{X_n, n \geq 0\}$, $X_n = 0,1$, com probabilidades de estado inicial, $P(X_0 = 0) = 1-\pi$ e $P(X_0 = 1) = \pi$; as probabilidades de transição desta cadeia correspondem às probabilidades condicionais de $W_{i+1}|W_i$ da sequência $\{W_i\}_{i\geq 1}$, $W_i = 0,1; i = 1,2,\ldots$ de variáveis binárias equicorrelacionadas, isto é, $\text{Corr}(W_i, W_j) = \rho$, para todo i, j, $i \neq j$ e $\max\{-1, -\frac{1-\pi}{\pi}\} \leq \rho < 1$. Desta forma, sejam W_1 e W_2 variáveis binárias equicorrelacionadas com coeficiente de correlação ρ e probabilidade de sucesso π. $W = W_1 + W_2$ tem distribuição binomial correlacionada sob a condição de que a distribuição de probabilidade de W dependa linearmente de ρ (LUCEÑO, 1995) e, assim, W representa a quantidade de sucessos em amostra de tamanho $n = 2$ de variáveis binárias equicorrelacionadas, $W \in \{0,1,2\}$. Luceño (1995) introduz a distribuição binomial correlacionada de parâmetros n, π, ρ ($CB(n,\pi,\rho)$), a qual é uma distribuição composta de uma variável com distribuição binomial de parâmetros n e π, com probabilidade $(1-\rho)$; e de uma variável com distribuição de Bernoulli modificada de parâmetro π, podendo assumir os valores 0 ou n; com probabilidade ρ, $0 < \rho < 1$. A função característica ϕ da variável com distribuição binomial correlacionada de parâmetros n, π e ρ, com n fixado, é dada por:

$$\phi_{CB(n,\pi,\rho)}(t) = [1-\pi+\pi e^{it}]^n (1-\rho) + [1-\pi+\pi e^{itn}]\rho,$$

para W,

$$\begin{aligned}\phi_{CB(2,\pi,\rho)}(t) &= [1-\pi+\pi e^{it}]^2(1-\rho)+[1-\pi+\pi e^{it2}]\rho \\ &= (1-\pi)[\rho+(1-\rho)(1-\pi)]+2(1-\pi)(1-\rho)\pi e^{it}+\pi[\rho+(1-\rho)\pi]e^{it2}\end{aligned}$$

e, por definição, $\phi_{CB(2,\pi,\rho)}(t) = E(e^{itW})$, e com isso se obtem a distribuição de probabilidade de W, pois,

$$E(e^{itW}) = P(W=0)+e^{it}P(W=1)+e^{it2}P(W=2)$$

Logo,

$$\begin{aligned}P(W=0) &= (1-\pi)[\rho+(1-\rho)(1-\pi)], \\ P(W=1) &= 2(1-\pi)(1-\rho)\pi \quad \text{e} \\ P(W=2) &= \pi[\rho+(1-\rho)\pi].\end{aligned}$$

Portanto,

$$\begin{aligned}P(W_2=0|W_1=0) &= \frac{P(W_2=0;W_1=0)}{P(W_1=0)} = \frac{P(W_2+W_1=0)}{P(W_1=0)} = \frac{P(W=0)}{P(W_1=0)}, \\ &= \frac{(1-\pi)[\rho+(1-\rho)(1-\pi)]}{(1-\pi)} = \rho+(1-\rho)(1-\pi), \end{aligned} \tag{2.13}$$

$$\begin{aligned}P(W_2=1|W_1=0) &= \frac{P(W_2=1;W_1=0)}{P(W_1=0)} = \frac{\frac{1}{2}P(W_2+W_1=1)}{P(W_1=0)} = \frac{\frac{1}{2}P(W=1)}{P(W_1=0)}, \\ &= \frac{1}{2}\frac{2(1-\pi)(1-\rho)\pi}{(1-\pi)} = (1-\rho)\pi, \end{aligned} \tag{2.14}$$

$$\begin{aligned}P(W_2=0|W_1=1) &= \frac{P(W_2=0;W_1=1)}{P(W_1=1)} = \frac{\frac{1}{2}P(W_2+W_1=1)}{P(W_1=1)}\frac{\frac{1}{2}P(W=1)}{P(W_1=1)}, \\ &= \frac{1}{2}\frac{2(1-\pi)(1-\rho)\pi}{\pi} = (1-\pi)(1-\rho) \end{aligned} \tag{2.15}$$

e

$$\begin{aligned}P(W_2=1|W_1=1) &= \frac{P(W_2=1;W_1=1)}{P(W_1=1)} = \frac{P(W_2+W_1=2)}{P(W_1=1)} = \frac{P(W=2)}{P(W_1=1)}, \\ &= \frac{\pi[\rho+(1-\rho)\pi]}{\pi} = \rho+(1-\rho)\pi. \end{aligned} \tag{2.16}$$

A distribuição de probabilidade em (2.10) da variável aleatória Y_k pode então ser reescrita em termos dos parâmetros π e ρ pelas probabilidades de transição obtidas em (2.13), (2.14), (2.15) e (2.16), isto é, Y_k representando o número de transições na cadeia de Markov homogênea de dois estados $\{X_n, n \geq 0\}$, $X_n = 0, 1$, com probabilidades de estado inicial, $p_0 = 1-\pi$ e $p_1 = \pi$,

$0 < \pi < 1$; probabilidades de transição dos estados $p_{00} = (1-\pi)(1-\rho)+\rho$, $p_{01} = (1-\rho)\pi$, $p_{10} = (1-\pi)(1-\rho)$ e $p_{11} = \pi(1-\rho)+\rho$, e, para $y_k = y+k \geq k$, $k \geq 1$. Kolev, Minkova e Neytchev (2000) mostram que, para estas probabilidades de transição, $\max\{-1, -\frac{1-\pi}{\pi}\} \leq \rho < 1$.

□

Denominamos a variável aleatória Y_k, com função de distribuição de probabilidade dada em (2.12), como sendo a variável aleatória com distribuição geométrica de ordem k correlacionada, $kIGeo(\pi,\rho)$, $k \geq 1$, a qual representa a quantidade de ensaios (ou tempo discreto de espera) até k sucessos consecutivos numa sequências de variáveis binárias dependentes com probabilidade de sucesso π, $\pi \in (0,1)$, e coeficiente de correlação constante ρ, sendo $\max\{-1, -\frac{1-\pi}{\pi}\} \leq \rho < 1$.

A função geradora de probabilidade da variável aleatória Y_k é também reescrita de (2.11) de acordo com as probabilidades de transição em termos de π e ρ, para $|t| < 1$:

Proposição 2.3.2. *A função geradora de probabilidade $\varphi_{Y_k}(t)$, $|t| < 1$, da variável aleatória $Y_k \sim kIGeo(\pi,\rho)$ é dada por:*

$$\varphi_{Y_k}(t) = \frac{\pi(1-\rho t)[1-\pi t-(1-\pi)\rho t][\pi+(1-\pi)\rho]^{k-1}t^k}{[1-(1-\pi)t-\pi\rho t][1-\pi t-(1-\pi)\rho t]-\pi(1-\pi)(1-\rho)^2t^2\{1-[\pi+(1-\pi)\rho]^{k-1}t^{k-1}\}}. \tag{2.17}$$

Demonstração. De (2.11), tem-se que

$$\varphi(t) = \frac{(P_0+P_1t)p_{11}^{k-1}t^k}{1-p_{00}t-P_2(t)}, \tag{2.18}$$

em que $P_0 = p_0p_{01}+p_1p_{11}$, $P_1 = p_1(p_{01}p_{10}+p_{00}p_{11})$ e $P_2(t) = p_{01}p_{10}t^2(1+p_{11}t+\ldots+p_{11}^{k-2}t^{k-2}) = p_{01}p_{10}t^2S_{k-2}$. S_{k-2} é a soma de uma progressão geométrica de razão $p_{11}t$, $|p_{11}t| < 1 \Rightarrow |t| < 1$, então,

$$S_{k-2} = \frac{1-(p_{11}t)^{k-1}}{1-p_{11}t}.$$

Sabendo que $p_0 = 1-\pi$, $p_1 = 1-p_0$ e p_{00}, p_{01}, p_{10} e p_{11} obtidas, respectivamente, em (2.13), (2.14), (2.15) e (2.16), obtém-se, após alguns cálculos, que $P_0 = \pi$, $P_1 = -\pi\rho$ e

$$P_2(t) = \pi(1-\pi)(1-\rho)^2t^2\frac{1-[\pi(1-\rho)+\rho]^{k-1}t^{k-1}}{1-[\pi(1-\rho)+\rho]}.$$

Desta forma, o numerador de $\varphi(t)$ é dado por:

$$\pi(1-\rho t)(1-[\pi(1-\rho)+\rho]t)[\pi(1-\rho)+\rho]^{k-1}t^k = \pi(1-\rho t)[1-\pi t-(1-\pi)\rho t][\pi+(1-\pi)\rho]^{k-1}t^k,$$

e o denominador por:

$$[1-(1-\pi)t-\pi\rho t][1-\pi t-(1-\pi)\rho t]-\pi(1-\pi)(1-\rho)^2t^2\{1-[\pi+(1-\pi)\rho]^{k-1}t^{k-1}\},$$

e, portanto, $\varphi(t) = \varphi_{Y_k}(t)$. □

Corolário 2.3.3. *A variável aleatória Y_k tem distribuição geométrica correlacionada, $IGeo(\pi,\rho)$, para $k=1$.*

Demonstração. Fazendo $k=1$ em (2.12), tem-se:

$$\begin{aligned}
P(Y_1=y+1) &= \{(1-\pi)[(1-\rho)\pi]+\pi[\pi(1-\rho)+\rho]\}I_{\{0\}}(y)+\\
&+ \left\{(1-\pi)\binom{y_1}{y_1}[(1-\pi)(1-\rho)+\rho]^{y_1}[(1-\rho)\pi]\right.\\
&+ \left.\pi[(1-\pi)(1-\rho)]\binom{x_1}{x_1}[(1-\pi)(1-\rho)+\rho]^{x_1}[(1-\rho)\pi]\right\}I_{\{1,2,\ldots\}}(y),\\
&= \{\pi(1-\rho)+\pi\rho]\}I_0(y)+\left\{(1-\pi)[(1-\pi)(1-\rho)+\rho]^{y_1}[(1-\rho)\pi]+\right.\\
&+ \left.\pi[(1-\pi)(1-\rho)][(1-\pi)(1-\rho)+\rho]^{y-1}[(1-\rho)\pi]\right\}I_{\{1,2,\ldots\}}(y)\\
P(Y_1=y+1) &= \pi I_{\{0\}}(y)+\{\pi(1-\pi)(1-\rho)[(1-\pi)(1-\rho)+\rho]^{y-1}\}I_{\{1,2,\ldots\}}(y),
\end{aligned}$$

a qual corresponde à distribuição de probabilidade da variável aleatória com distribuição geométrica correlacionada proposta por Kolev, Minkova e Neytchev (2000) (2.5), logo, a distribuição em (2.12) inclui $k=1$. □

Corolário 2.3.4. *A variável aleatória Y_k tem distribuição geométrica de ordem k, $kGeo(\pi)$, para $\rho=0$.*

Demonstração. Substituindo $\rho=0$ em (2.17), temos a função:

$$\begin{aligned}
\varphi_{Y_k}(t) &= \frac{(1-\pi t)(\pi t)^k}{(1-(1-\pi)t)(1-\pi t)-\pi(1-\pi)t^2[1-(\pi t)^{k-1}]}\\
&= \frac{(\pi t)^k(1-\pi t)}{(1-t+\pi t-\pi t+\pi(1-\pi)t^2-\pi(1-\pi)t^2+t(1-\pi)(\pi t)^k]},\\
&= \frac{(\pi t)^k(1-\pi t)}{1-t+(1-\pi)t(\pi t)^k},
\end{aligned}$$

que corresponde à função geradora de probabilidade em (2.9) da variável aleatória com distribuição geométrica de ordem k apresentada por Philippou e Muwafi (1980). □

Proposição 2.3.5. *A variável aleatória $Y_k \sim kIGeo(\pi,\rho)$ tem esperança dada por:*

$$E(Y_k)=\frac{[\pi(1-\rho)+\rho]^{1-k}-\pi[1+(1-\pi)\rho]}{\pi(1-\pi)(1-\rho)}. \tag{2.19}$$

Demonstração. Por definição, temos que a função geradora de momentos $M_{Y_k}(s)=E(e^{sY_k})$ desde que exista para algum $s\in(a,b)$, $a,b\in\mathbb{R}$, $a<b$. $M_{Y_k}(s)$ pode ser vista também como função geradora de probabilidade $\varphi_{Y_k}(t)$ em $t=\exp(s)$. Os momentos de ordem n da variável

aleatória com distribuição geométrica correlacionada de ordem k são obtidos pelas derivadas de ordem n da função geradora de momentos em relação a s no ponto $s = 0$:

$$M_{Y_k}^{(n)}(s) = \frac{d^n M_{Y_k}(s)}{ds^n} \quad \text{e} \quad M_{Y_k}^{(n)}(0) = E(Y_k^n).$$

Logo, os momentos são também funções das derivadas da função geradora de probabilidade (2.17) em relação a t no ponto $t = 1$. Desta forma, a esperança da variável aleatória Y_k é obtida por:

$$E(Y_k) \quad = \quad M'_{Y_k}(s)_{s=0} = \frac{d\varphi_{Y_k}(t)}{dt}\frac{dt}{ds} = t\varphi'_{Y_k}(t)_{t=1} = \varphi'_{Y_k}(1). \tag{2.20}$$

Para menor dificuldade dos cálculos, vamos considerar $\varphi_{Y_k}(t)$ em (2.18):

$$\varphi_{Y_k}(t) = \frac{(1-\rho t)\pi p_{11}^{k-1} t^k}{1 - p_{00}t - P_2(t)} = \frac{f(t)}{g(t)}, \tag{2.21}$$

em que $P_2(t) = p_{01}p_{10}t^2(1-(p_{11}t)^{k-1})/(1-p_{11}t)$ e p_{00}, p_{01}, p_{10} e p_{11} obtidas, respectivamente, em (2.13), (2.14), (2.15) e (2.16). Logo, para encontrar $E(Y_k)$ por (2.20), precisamos obter

$$\varphi'_{Y_k}(t) = \frac{f'(t)g(t) - f(t)g'(t)}{g^2(t)}.$$

De (2.21), temos que:

$$\begin{aligned}
f(t) &= (1-\rho t)\pi(p_{11}t)^{k-1} \quad \Rightarrow \quad f(1) = \pi(1-\rho)p_{11}^{k-1}, && (2.22)\\
g(t) &= 1 - p_{00}t - P_2(t) \quad \Rightarrow \quad g(1) = 1 - p_{00} - P_2(1) = p_{01} - P_2(1), && (2.23)\\
P_2(t) &= \frac{p_{01}p_{10}t^2(1-(p_{11}t)^{k-1})}{1-p_{11}t} \quad \Rightarrow \quad P_2(1) = p_{01}(1-p_{11}^{k-1}), && (2.24)\\
&\Rightarrow \quad g(1) = p_{01}p_{11}^{k-1} = \pi(1-\rho)p_{11}^{k-1} = f(1), && (2.25)
\end{aligned}$$

$$\begin{aligned}
f'(t) &= [k-(k+1)\rho t]\pi(p_{11}t)^{k-1} \quad \Rightarrow \quad f'(1) = [k-(k+1)\rho]\pi p_{11}^{k-1}, && (2.26)\\
g'(t) &= -p_{00} - P_2'(t) \quad \Rightarrow \quad g'(1) = -p_{00} - P_2'(1), && (2.27)\\
P_2'(t) &= \frac{p_{01}p_{10}t[2-(k+1)(p_{11}t)^{k-1}] + p_{11}P_2(t)}{1-p_{11}t}\\
&\Rightarrow \quad P_2'(1) = \frac{p_{01}p_{10}[2-(k+1)p_{11}^{k-1}] + p_{11}P_2(1)}{1-p_{11}},
\end{aligned}$$

De (2.24),

$$P_2'(1) = \frac{p_{01}}{p_{10}}[1 + p_{10} - (1+kp_{10})p_{11}^{k-1}], \tag{2.28}$$

e assim, de (2.25), tem-se:

$$\varphi'_{Y_k}(1) \quad = \quad \frac{f'(1)g(1) - g(1)g'(1)}{g^2(1)} = \frac{f'(1) - g'(1)}{g(1)}, \tag{2.29}$$

e, de (2.27) e (2.28),

$$g'(1) = p_{01} - 1 - \frac{p_{01}}{p_{10}}[1 + p_{10} - (1 + kp_{10})p_{11^{k-1}}]. \tag{2.30}$$

Sabendo que $p_{00} + p_{01} = p_{10} + p_{11} = 1$, de (2.26) e (2.30) e, substituindo p_{01} por (2.14) e p_{10} por (2.15), após alguns cálculos, obtém-se,

$$\begin{aligned} f'(1) - g'(1) &= [k(1-\rho) - \rho]\pi p_{11}^{k-1} - \frac{[k\pi(1-\pi)(1-\rho) + \pi]p_{11}^{k-1} - 1}{1-\pi}, \\ &= \frac{1 - \pi[1 + (1-\pi)\rho]p_{11}^{k-1}}{1-\pi}. \end{aligned} \tag{2.31}$$

De (2.22), (2.23), (2.25), (2.29) e (2.31) :

$$\varphi'_{Y_k}(1) = \frac{1 - \pi[1 + (1-\pi)\rho]p_{11}^{k-1}}{\pi(1-\pi)(1-\rho)p_{11}^{k-1}}, \tag{2.32}$$

sendo $p_{11} = \pi(1-\rho) + \rho$ (de (2.16)). □

Pode-se então constatar que, para $\rho = 0$,

$$E(Y_k) = \frac{1 - \pi^k}{(1-\pi)\pi^k},$$

e então temos a esperança em (2.8). Se também $k = 1$, temos a esperança $1/\pi$ da variável geométrica clássica, conforme visto em (2.2). E, para $k = 1$ e $\rho \neq 0$,

$$E(Y_1) = \frac{1 - \pi\rho}{\pi(1-\rho)}. \tag{2.33}$$

A esperança em (2.33) é da variável aleatória geométrica correlacionada Y_1 que representa o número de ensaios até o sucesso. A esperança da variável aleatória Y que representa o número de falhas até o sucesso é obtida da relação entre as duas variáveis, $Y_1 = Y + 1$, logo, $E(Y_1) = E(Y) + 1 \Rightarrow E(Y) = E(Y_1) - 1$, e, de (2.33):

$$E(Y) = \frac{1 - \pi\rho}{\pi(1-\rho)} - 1 = \frac{1 - \pi\rho - \pi + \pi\rho}{\pi(1-\rho)} = \frac{1-\pi}{\pi(1-\rho)},$$

a qual corresponde à esperança dada em (2.6).

Proposição 2.3.6. *A variável aleatória $Y_k \sim kIGeo(\pi, \rho)$ tem variância dada por:*

$$\begin{aligned} Var(Y_k) &= k\Big(k - \frac{1+\rho}{1-\rho}\Big) + \frac{2[\pi(1-\rho) + \rho]^{1-k} - \{2 + (k-1)(1-\pi)(1-\rho)[2 + k(1-\pi)(1-\rho)]\}}{(1-\pi)^2(1-\rho)^2} + \\ &+ E(Y_k)\Big\{1 - E(Y_k) + 2\frac{[\pi(1-\rho) + \rho]^{1-k} - \pi\big[k(1-\pi)(1-\rho) + 1\big]}{\pi(1-\pi)(1-\rho)}\Big\}, \end{aligned} \tag{2.34}$$

com $E(Y_k)$ dada em (2.19).

Demonstração. Sabemos que $Var(Y_k) = E(Y_k^2) - E^2(Y_k)$ e que,

$$\begin{aligned} E(Y_k^2) &= M''_{Y_k}(s)_{s=0} = \frac{d\varphi_{Y_k}(t)}{dt}\frac{d^2t}{ds^2} + \frac{d^2\varphi_{Y_k}(t)}{dt^2}\left(\frac{dt}{ds}\right)^2 \\ &= t\varphi'_{Y_k}(t)_{t=1} + t^2\varphi''_{Y_k}(t)_{t=1} = \varphi'_{Y_k}(1) + \varphi''_{Y_k}(1). \end{aligned}$$

Logo,

$$Var(Y_k) = \varphi''_{Y_k}(1) + \varphi'_{Y_k}(1)(1 - \varphi'_{Y_k}(1)). \tag{2.35}$$

Temos $\varphi'_{Y_k}(1) = E(Y_k)$ de (2.32) e agora precisamos obter:

$$\varphi''_{Y_k}(t) = \frac{f''(t)g(t) - f(t)g''(t) - \varphi'_{Y_k}(t)(g^2)'(t)}{g^2(t)}. \tag{2.36}$$

Para isto, temos:

$$\begin{aligned} (g^2)'(t) &= 2g(t)g'(t) \quad \Rightarrow \quad (g^2)'(1) = 2g(1)g'(1), && (2.37)\\ f''(t) &= [k(k-1) - (k+1)k\rho t]\pi p_{11}^{k-1}t^{k-2} \\ &\Rightarrow \quad f''(1) = [k(k-1) - (k+1)k\rho]\pi p_{11}^{k-1}, && (2.38)\\ g''(t) &= -P_2''(t) \quad \Rightarrow \quad g''(1) = -P_2''(1), && (2.39)\\ P_2''(t) &= \frac{p_{01}p_{10}[2 - (k+1)kp_{11}^{k-1}t^{k-2}] + 2p_{11}P_2'(t)}{1 - p_{11}t} \\ &\Rightarrow \quad P_2''(1) = \frac{p_{01}p_{10}[2 - (k+1)kp_{11}^{k-1}] + 2p_{11}P_2'(1)}{1 - p_{11}}, && (2.40) \end{aligned}$$

e assim, de (2.25), (2.36) e (2.37), tem-se:

$$\varphi''_{Y_k}(1) = \frac{f''(1)g(1) - g(1)g''(1) - \varphi'_{Y_k}(1)2g'(1)g(1)}{g^2(1)} = \frac{f''(1) - g''(1) - 2\varphi'_{Y_k}(1)g'(1)}{g(1)}. \tag{2.41}$$

E, de (2.35), (2.39) e (2.41),

$$Var(Y_k) = \frac{f''(1) + P_2''(1)}{g(1)} + E(Y_k)\left[1 - \frac{2g'(1)}{g(1)} - E(Y_k)\right]. \tag{2.42}$$

Desta forma, obtemos:

De (2.40), (2.28), (2.14) e (2.15),

$$\begin{aligned} P_2''(1) &= \frac{p_{01}}{p_{10}}\Big\{p_{10}[2-(k+1)kp_{11}^{k-1}]+2\frac{p_{11}}{p_{10}}[1+p_{10}-(1+kp_{10})p_{11}^{k-1}]\Big\} \\ &= \frac{p_{01}}{p_{10}^2}\{2-[k(k-1)p_{10}^2+2(1-p_{10}+kp_{10})]p_{11}^{k-1}\} \\ &= \frac{p_{01}}{p_{10}^2}\{2-[2+(k-1)p_{10}(kp_{10}+2)]p_{11}^{k-1}\} \\ &= \frac{\pi\{2-[2+(k-1)(1-\pi)(1-\rho)(2+k(1-\pi)(1-\rho))]p_{11}^{k-1}\}}{(1-\pi)^2(1-\rho)}. \end{aligned} \tag{2.43}$$

De (2.31) e (2.25),

$$2\frac{g'(1)}{g(1)} = 2\frac{[k(1-\pi)(1-\rho)+1]\pi p_{11}^{k-1}-1}{\pi(1-\pi)(1-\rho)p_{11}^{k-1}}. \tag{2.44}$$

Logo, substituindo (2.43) e (2.44) em (2.42), com $g(1)$ de (2.23), $f''(1)$ de (2.38) e $E(Y_k)$ de (2.19), temos a variância em (2.34), sendo $p_{11}=\pi(1-\rho)+\rho$ (de (2.16)). □

Para $\rho = 0$,

$$Var(Y_k) = \frac{[1-(2k+1)(1-\pi)\pi^k-\pi^{2k+1}]}{(1-\pi)^2\pi^{2k}}$$

e, para $k=1$ e $\rho \neq 0$,

$$Var(Y_1) = \frac{(1-\pi)(1+\pi\rho)}{\pi^2(1-\rho)^2},$$

as quais correspondem às variâncias das variáveis aleatórias com distribuição geométrica de ordem k e distribuição geométrica correlacionada, conforme visto em (2.8) e (2.6), respectivamente.

Na Figura 1 visualizamos o comportamento da distribuição de probabilidade da variável aleatória geométrica de ordem k correlacionada para diferentes valores de k, π e ρ.

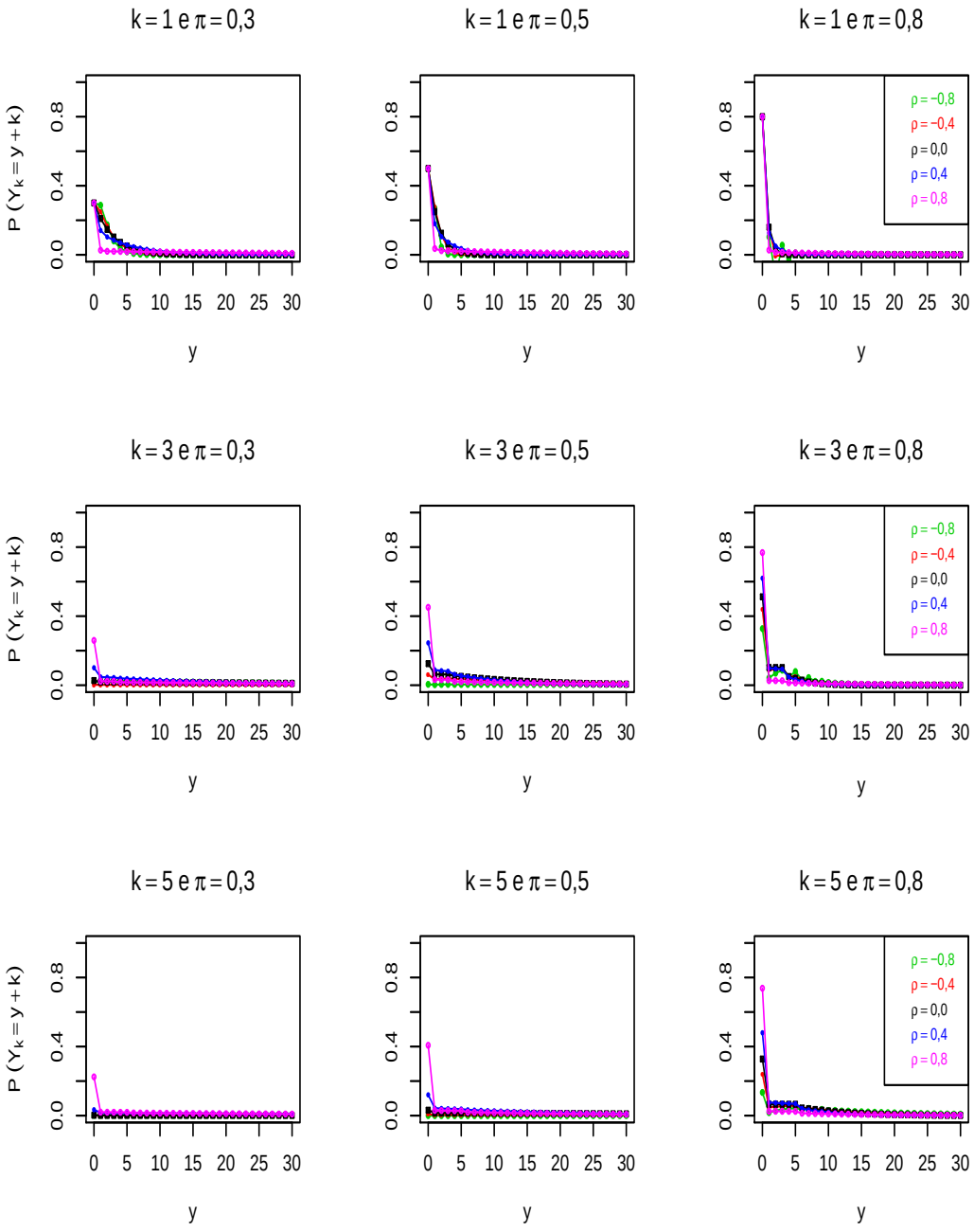

Figura 1 – Distribuição $kIGeo(\pi, \rho)$ para diferentes valores de k, π e ρ.

CAPÍTULO

3

MODELO DE REGRESSÃO GEOMÉTRICO CORRELACIONADO CLÁSSICO

Neste capítulo apresentamos o modelo de regressão considerando uma abordagem clássica para o caso particular do modelo geométrico de ordem k correlacionado apresentado na Seção 2.1 do Capítulo 2, em que $k = 1$, o qual corresponde ao modelo geométrico correlacionado proposto por Kolev, Minkova e Neytchev (2000). A estrutura de regressão pode ser modelada por diferentes funções de ligação e o ajuste obtido por maximização da função de verossimilhança, via processo numérico iterativo. A escolha dos melhores ajustes foi baseada em medidas de critério de seleção de modelos. Testes de hipóteses e intervalos de confiança assintóticos foram propostos para a inferência estatística sobre os parâmetros. Para análise de diagnóstico utilizamos medidas de resíduo quantílico aleatorizado, distância de Cook generalizada e distância de verossimilhança para verificar as suposições iniciais do modelo e identificar valores discrepantes e/ou observações influentes nas estimativas dos parâmetros. Estudos de simulação ilustram os métodos propostos e as propriedades assintóticas dos estimadores de máxima verossimilhança. Uma aplicação do modelo envolvendo dados reais de pacientes com doença é apresentada. Uma comparação com o modelo geométrico usual, que não possui o parâmetro de correlação ρ, inteira este capítulo. Mostramos o melhor ajuste do modelo proposto. Ademais, o modelo geométrico usual superestima as probabilidades de alta dos pacientes na aplicação, e também apresenta inferências distintas às do modelo geométrico correlacionado.

3.1 Modelo de regressão clássico MRGC

Na Seção 2.1 do Capítulo 2 vimos que a variável aleatória Y tem distribuição geométrica correlacionada, com parâmetros π, $\pi \in (0,1)$ e ρ, $\max\{-1, -\frac{1-\pi}{\pi}\} \leq \rho < 1$ ($IGeo(\pi,\rho)$), quando Y representa a quantidade de $falhas$ que antecede o primeiro *sucesso* em uma sequência de ensaios de Bernoulli dependentes com probabilidade de sucesso π, coeficiente de correlação

ρ e distribuição de probabilidade dada por (2.5):

$$P(Y=y|\pi,\rho)=\pi I_{\{0\}}(y)+\pi(1-\pi)(1-\rho)[(1-\pi)(1-\rho)+\rho]^{y-1}I_{\{1,2,\ldots\}}(y),$$

com média e variância dadas por (2.6):

$$E(Y)=\frac{1-\pi}{\pi(1-\rho)} \quad e \quad Var(Y)=\frac{(1-\pi)(1+\pi\rho)}{\pi^2(1-\rho)^2}.$$

Suponha $Y_1,Y_2,\ldots,Y_m$ variáveis aleatórias independentes em que Y_i tem distribuição $IGeo(\pi_i,\rho_i)$ dada em (2.5). Assumindo que $y_1,y_2,\ldots,y_m$ são observações de $Y_1,Y_2,\ldots,Y_m$, respectivamente, pode-se modelar π_i em função de covariáveis através da função de ligação $g^{-1}(\eta_i)$, g função estritamente monótona e duplamente diferenciável. Deste modo, para $x_{i1},x_{i2},\ldots,x_{iR}$, observações das covariáveis X_{ir}, $r=1,2,\ldots,R$, define-se o modelo de regressão geométrico correlacionado em que π_i, $i=1,\ldots,m$, satisfazem as relações funcionais:

$$\eta_i=g(\pi_i)=\sum_{r=0}^{R}\beta_r x_{ir}=\boldsymbol{x}_i^T\boldsymbol{\beta}, \tag{3.1}$$

(3.2)

sendo β_r parâmetros desconhecidos, $r=1,\ldots,R$.

Em nossa análise, o coeficiente de correlação ρ não é modelado em função das covariáveis, assim temos $\boldsymbol{\beta}=(\beta_0,...,\beta_R)^T$ o vetor de coeficientes de regressão com $\boldsymbol{\beta}\in\mathbb{R}^p$, $p=R+1$, tal que $p+1<m$, η o preditor linear e $\boldsymbol{x}_i=(1,x_{i1},\ldots,x_{iR})$ o vetor de covariáveis de ordem p. Assume-se que a matriz de planejamento $\boldsymbol{X}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m)^T$, de ordem $m\times p$, tem posto completo p. Sendo a função de ligação neste caso $g:(0,1)\longrightarrow\mathbb{R}$, é comum utilizar as funções monótonas e diferenciáveis apresentadas em McCullagh e Nelder (1989), conforme já vistas na Tabela 1: logito, $g(\pi_i)=\log\left(\frac{\pi_i}{1-\pi_i}\right)$; complemento log-log, $g(\pi_i)=\log[-\log(1-\pi_i)]$; log-log, $g(\pi_i)=-\log[-\log(\pi_i)]$; e probito, $g(\pi_i)=\Phi^{-1}(\pi_i)$, em que Φ é a função de distribuição Normal padrão acumulada.

3.2 Estimação

O método clássico de máxima verossimilhança (ou máxima log-verossimilhança) obtém estimadores para os parâmetros do modelo de modo que as estimativas para os parâmetros são tais que a probabilidade conjunta dos valores observados é máxima. Considerando a função de ligação em (1.1), a função de log-verossimilhança de $\boldsymbol{\theta}=(\boldsymbol{\beta},\rho)$ dado a amostra observada $\mathscr{D}=(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\delta})$, é dada por:

$$l(\boldsymbol{\theta}|\mathscr{D})=\sum_{i=1}^{m}\log\left\{g^{-1}(\eta_i)^{\delta_i}\{g^{-1}(\eta_i)(1-g^{-1}(\eta_i))(1-\rho)[(1-g^{-1}(\eta_i))(1-\rho)+\rho]^{y_i-1}\}^{1-\delta_i}\right\}, \tag{3.3}$$

sendo $\delta_i = 1$ se $y_i = 0$ e $\delta_i = 0$ se $y_i > 0$; $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)^T$; com $\eta_i = g(\pi_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$, $\max\{-1, -\frac{1-\pi_i}{\pi_i}\} \leq \rho < 1$, $i = 1, \ldots, m$, sendo $\boldsymbol{\theta}$ o vetor com $p+1$ parâmetros desconhecidos.

Para cada amostra $\boldsymbol{y}$ os valores de $\boldsymbol{\theta}$ pelos quais $l(\boldsymbol{\theta}|\mathscr{D})$ tem valor máximo como função de $\boldsymbol{\theta}$, denominado $\hat{\boldsymbol{\theta}}$, são estimadores de máxima verossimilhança (EMV) e, neste caso, a log-verossimilhança (3.3) pode não apresentar solução analítica explícita e, assim, recorre-se a métodos numéricos, como o processo iterativo de Newton-Raphson em que $\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^n + [\boldsymbol{J}(\boldsymbol{\theta}^{(n)})]^{-1}\boldsymbol{U}(\boldsymbol{\theta}^{(n)})$, $n = 0, 1, \ldots$, e, para $|\boldsymbol{\theta}^{(n+1)} - \boldsymbol{\theta}^{(n)}| < \varepsilon$, ε arbitrado, temos $\boldsymbol{\theta}^{(n+1)}$ a estimativa de máxima verossimilhança, sendo $\boldsymbol{J}(\boldsymbol{\theta}) = -\partial \boldsymbol{U}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ a matriz de informação observada de Fisher, em que $\boldsymbol{U}(\boldsymbol{\theta}) = \partial l(\boldsymbol{\theta}|\mathscr{D})/\partial \boldsymbol{\theta}$ é a função escore.

Sendo $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\rho})$ e $\Delta_i(\boldsymbol{\beta}) = \partial g^{-1}(\eta_i)/\partial \boldsymbol{\beta}$, temos a função escore $\boldsymbol{U}^T(\boldsymbol{\theta}) = \big(\big(\frac{\partial l((\boldsymbol{\theta}|\mathscr{D})}{\partial \boldsymbol{\beta}}\big)^T, \frac{\partial l(\boldsymbol{\theta}|\mathscr{D})}{\partial \rho}\big)$, em que,

$$\partial l(\boldsymbol{\theta}|\mathscr{D})/\partial \boldsymbol{\beta} = \sum_{i=1}^{m} \Delta_i(\boldsymbol{\beta}) g^{-1}(\eta_i)^{-\delta_i} \Big[\frac{1 - 2g^{-1}(\eta_i)}{g^{-1}(\eta_i)(1 - g^{-1}(\eta_i))} - \frac{(y_i - 1)(1-\rho)}{(1 - g^{-1}(\eta_i))(1-\rho) + \rho}\Big]^{1-\delta_i}, \tag{3.4}$$

$$\partial l(\boldsymbol{\theta}|\mathscr{D})/\partial \rho = \sum_{i=1}^{m} \Big[\frac{(y_i - 1) g^{-1}(\eta_i)}{(1 - g^{-1}(\eta_i))(1-\rho) + \rho} - \frac{1}{1-\rho}\Big]^{1-\delta_i}. \tag{3.5}$$

As expressões para as derivadas de primeira e segunda ordens das funções de ligação (Tabela 1) em relação aos parâmetros $\boldsymbol{\beta}$ estão na Tabela 2 da Seção 1.3.1 do Capítulo 1. Nestes casos o vetor escore $\boldsymbol{U}(\boldsymbol{\theta})$ tem dimensão $R+2$. A matriz de informação observada de Fisher pra o modelo geométrico correlacionado $\boldsymbol{J}(\boldsymbol{\theta})$ tem dimensão $(R+2)\times(R+2)$, $\boldsymbol{J}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{J}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \boldsymbol{J}_{\boldsymbol{\beta}\rho} \\ \boldsymbol{J}_{\boldsymbol{\beta}\rho}^T & J_{\rho\rho} \end{pmatrix}$, em que $\boldsymbol{J}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ é uma matriz $(R+1)\times(R+1)$; $\boldsymbol{J}_{\boldsymbol{\beta}\rho}$ um vetor de tamanho $R+1$, $\boldsymbol{J}_{\boldsymbol{\beta}\rho}^T = (J_{\beta_0\rho}\ J_{\beta_1\rho}\ \ldots\ J_{\beta_R\rho})^T$, e $J_{\rho\rho}$ um elemento. Os elementos de $\boldsymbol{J}(\boldsymbol{\theta})$ são dados por :

$$\begin{aligned} J_{\beta_r\beta_s} = & - \sum_{i=1}^{m} \Big[\frac{\Delta_i^2(\beta_r\beta_s)}{g^{-1}(\eta_i)} - \frac{\Delta_i(\beta_r)\Delta_i(\beta_s)}{(g^{-1}(\eta_i))^2}\Big]^{\delta_i} \Big\{\Delta_i^2(\beta_r\beta_s)\Big(\frac{1 - 2g^{-1}(\eta_i)}{g^{-1}(\eta_i)(1 - g^{-1}(\eta_i))} + \\ & - \frac{(y_i - 1)(1-\rho)}{(1 - g^{-1}(\eta_i))(1-\rho) + \rho}\Big) - \Delta_i(\beta_r)\Delta_i(\beta_s)\Big[\Big(\frac{1}{g^{-1}(\eta_i)}\Big)^2 + \\ & + \Big(\frac{1}{g^{-1}(1 - g^{-1}(\eta_i))}\Big)^2 + (y_i - 1)\Big(\frac{1-\rho}{(1 - g^{-1}(\eta_i))(1-\rho) + \rho}\Big)^2\Big]\Big\}^{1-\delta_i}, \end{aligned} \tag{3.6}$$

$$J_{\beta_r\rho} = -\sum_{i=1}^{m} \Big\{\frac{\Delta_i(\beta_r)(y_i - 1)}{[(1 - g^{-1}(\eta_i))(1-\rho) + \rho]^2}\Big\}^{1-\delta_i}, \tag{3.7}$$

$$J_{\rho\rho} = -\sum_{i=1}^{m} \Big\{(y_i - 1)\Big[\frac{g^{-1}(\eta_i)}{(1 - g^{-1}(\eta_i))(1-\rho) + \rho}\Big]^2 - \Big(\frac{1}{1-\rho}\Big)^2\Big\}^{1-\delta_i}, \tag{3.8}$$

$r, s = 0, 1, \ldots, R$, $\Delta_i(\beta_r)$ e $\Delta_i^2(\beta_r\beta_s)$ as derivadas de primeira e segunda ordens, respectivamente, das funções de ligação em relação aos parâmetros $\boldsymbol{\beta}$, conforme Tabela 2.

3.3 Critérios de seleção de modelos

O critério de informação de Akaike (*AIC*) (AKAIKE, 1974) e o critério de informação Bayesiano (*BIC*) (SCHWARZ *et al.*, 1978) podem ser utilizados para selecionar o melhor modelo de regressão. Menores valores de *AIC* e/ou *BIC* indicam melhores modelos:

$$\begin{aligned} AIC &= -2l(\hat{\boldsymbol{\theta}}|\mathscr{D}) + p + 1 \\ &= -2\sum_{i=1}^{m} \log\left\{ g^{-1}(\hat{\eta}_i)^{\delta_i} \left\{ g^{-1}(\hat{\eta}_i)(1-g^{-1}(\hat{\eta}_i))(1-\hat{\rho}) \left[(1-g^{-1}(\hat{\eta}_i))(1-\hat{\rho}) + \hat{\rho} \right]^{y_i - 1} \right\}^{1-\delta_i} \right\} \\ &\quad + p + 1, \end{aligned} \tag{3.9}$$

$$\begin{aligned} BIC &= -2l(\hat{\boldsymbol{\theta}}|\mathscr{D}) + (p+1)\log(m) \\ &= -2\sum_{i=1}^{m} \log\left\{ g^{-1}(\hat{\eta}_i)^{\delta_i} \left\{ g^{-1}(\hat{\eta}_i)(1-g^{-1}(\hat{\eta}_i))(1-\hat{\rho}) \left[(1-g^{-1}(\hat{\eta}_i))(1-\hat{\rho}) + \hat{\rho} \right]^{y_i - 1} \right\}^{1-\delta_i} \right\} \\ &\quad + (p+1)\log(m), \end{aligned} \tag{3.10}$$

em que $\mathscr{D} = (\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\delta})$ é a amostra observada; $\delta_i = 1$ se $y_i = 0$ e $\delta_i = 0$ se $y_i > 0$; $l(\hat{\boldsymbol{\theta}}|\mathscr{D})$ é o valor da função de log-verossimilhança avaliada no estimador de máxima verossimilhança; $\hat{\eta}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$; $p+1$ é o número de parâmetros no modelo, neste caso $p+1 = R+2$, em que R é a quantidade de covariáveis; e m é o número de observações.

3.4 Intervalos de confiança e testes de hipóteses assintóticos

Para grandes amostras e função distribuição de probabilidade satisfazendo as condições de regularidade apresentadas no Apêndice A, intervalos de confiança podem ser aproximados para cada um dos parâmetros θ_r com coeficiente de confiança de $100(1-\alpha)\%$:

$$IC(\theta_r; \alpha) = \hat{\theta}_r \pm z_{\alpha/2} (J_r^{-1}(\hat{\boldsymbol{\theta}}))^{1/2}, \tag{3.11}$$

em que $z_{\alpha/2}$ é o quantil $\alpha/2$ da distribuição $N(0,1)$ e $J_r^{-1}(\hat{\boldsymbol{\theta}})$ é o termo de posição r da diagonal $\boldsymbol{J}^{-1}(\boldsymbol{\theta})$ avaliada em $\hat{\boldsymbol{\theta}}$, correspondente a variância estimada do estimador de máxima verossimilhança, $r = 1, \ldots, R+2$.

Para testar a hipótese $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ (hipótese nula), em que $\boldsymbol{\theta}_0$ é um vetor p-dimensional de valores fixados, contra a hipótese alternativa $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, podemos utilizar as três estatísticas de teste empregadas para testar H_0 a seguir:

1. Estatística de Wald (W):

$$W = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \boldsymbol{J}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0). \tag{3.12}$$

2. Estatística de Wilks ou da razão de verossimilhança (RV):

$$\begin{aligned} RV &= 2\{l(\hat{\boldsymbol{\theta}}|\mathscr{D}) - l(\boldsymbol{\theta}_0|\mathscr{D})\} \\ &= 2\sum_{i=1}^{m} \log\left\{\left[\frac{g^{-1}(\hat{\eta}_i)}{g^{-1}(\eta_{0i})}\right]^{\delta_i}\left[\frac{g^{-1}(\hat{\eta}_i)(1-g^{-1}(\hat{\eta}_i))(1-\hat{\rho})}{g^{-1}(\eta_{0i})(1-g^{-1}(\eta_{0i})(1-\rho_0)}\right]^{1-\delta_i}\right\} + \\ &\quad +2\sum_{i=1}^{m}(1-\delta_i)(y_i-1)\log\left[\frac{(1-g^{-1}(\hat{\eta}_i))(1-\hat{\rho})+\hat{\rho}}{(1-g^{-1}(\eta_{0i}))(1-\rho_0)+\rho_0}\right]. \end{aligned} \tag{3.13}$$

3. Estatística de Rao ou dos escores eficientes (EE):

$$EE = \boldsymbol{U}(\boldsymbol{\theta}_0)^T \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1} \boldsymbol{U}(\boldsymbol{\theta}_0). \tag{3.14}$$

Em que $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\rho})$ é o EMV de $\boldsymbol{\theta} = (\boldsymbol{\beta}; \rho)$; $\boldsymbol{\beta} = (\beta_0\ \beta_1\ \ldots\ \beta_R)^T$, $\boldsymbol{x}_i = (1\ x_{i1}\ \ldots\ x_{iR})^T$, $\hat{\eta}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ e $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$ a matriz de informação observada de Fisher cujos elementos são obtidos por (3.6), (3.7) e (3.8), avaliados em $\hat{\boldsymbol{\theta}}$. $\boldsymbol{U}(\boldsymbol{\theta}_0)$ é a função escore obtida por (3.4) e (3.5), avaliada em $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0; \rho_0)$, e $\eta_{0i} = \boldsymbol{x}_i^T\boldsymbol{\beta}_0$, os quais não dependem do EMV.

Estas três estatísticas, sob a hipótese nula, têm distribuição assintótica qui-quadrado com p graus de liberdade, $\chi^2_{(p)}$, em que p é a quantidade de parâmetros. Em nosso modelo, $p = R+2$. As demonstrações destes resultados estão apresentadas em Leite e Singer (1990) (p. 123-125).

Quando há o interesse em testar hipóteses de apenas um subconjunto do vetor de parâmetros, considera-se $\boldsymbol{\theta} = [\boldsymbol{\beta}^T\rho]^T = [\boldsymbol{\beta}_1^T\boldsymbol{\beta}_2^T\rho]^T = [\boldsymbol{\theta}_1^T\boldsymbol{\theta}_2^T]^T$ uma partição do vetor de parâmetros $\boldsymbol{\theta}$, em que $\boldsymbol{\theta}_1 = \boldsymbol{\beta}_1$ é q-dimensional e $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}_2^T, \rho)$ é $p-q$-dimensional, sendo $\boldsymbol{\theta}_1$ o vetor de interesse. Deste modo, tem-se também a partição do vetor escore: $\boldsymbol{U}(\boldsymbol{\theta}) = [\boldsymbol{U}_1(\boldsymbol{\theta})^T\ \boldsymbol{U}_2(\boldsymbol{\theta})^T]^T$, e da matriz de informação observada de Fisher: $\boldsymbol{J}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{J}_{11}(\boldsymbol{\theta}) & \boldsymbol{J}_{12}(\boldsymbol{\theta}) \\ \boldsymbol{J}_{21}(\boldsymbol{\theta}) & \boldsymbol{J}_{22}(\boldsymbol{\theta}) \end{bmatrix}$. A matriz de variâncias e covariâncias assintóticas de $\hat{\boldsymbol{\theta}}_1$ é dada por (DEMÉTRIO, 2001):

$$Cov(\hat{\boldsymbol{\theta}}_1) = [\boldsymbol{J}_{11}(\boldsymbol{\theta}) - \boldsymbol{J}_{12}(\boldsymbol{\theta})\boldsymbol{J}_{22}(\boldsymbol{\theta})^{-1}\boldsymbol{J}_{21}(\boldsymbol{\theta})]^{-1}, \tag{3.15}$$

em que $\boldsymbol{J}_{12} = \boldsymbol{J}_{21}^T$, e pode ser estimada por :

$$\widehat{Cov}(\hat{\boldsymbol{\theta}}_1) = [\boldsymbol{J}_{11}(\hat{\boldsymbol{\theta}}) - \boldsymbol{J}_{12}(\hat{\boldsymbol{\theta}})\boldsymbol{J}_{22}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{J}_{21}(\hat{\boldsymbol{\theta}})]^{-1}, \tag{3.16}$$

em que $\hat{\boldsymbol{\theta}} = [\hat{\boldsymbol{\theta}}_1{}^T\ \hat{\boldsymbol{\theta}}_2{}^T]^T$ é o EMV para $\boldsymbol{\theta}$ sem restrição.

Neste caso, considera-se testar a hipótese $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$ contra a hipótese $H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_{10}$, em que $\boldsymbol{\theta}_{10}$ é um vetor q-dimensional de valores fixados e $\hat{\boldsymbol{\theta}}_0 = [\boldsymbol{\theta}_{10}^T\ \ \hat{\boldsymbol{\theta}}_{20}^T]^T$, $\hat{\boldsymbol{\theta}}_{20}$ é o EMV para $\boldsymbol{\theta}_2$ sob H_0. As três estatísticas de teste, sob a hipótese nula, têm distribuição assintótica qui-quadrado com q graus de liberdade, $\chi^2_{(q)}$, e são dadas por:

$$W = (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})^T[\widehat{Cov}(\hat{\boldsymbol{\theta}}_1)]^{-1}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}), \tag{3.17}$$

$$RV = 2\{\log\mathscr{L}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2|\mathscr{D}) - \log\mathscr{L}(\boldsymbol{\theta}_{10}, \hat{\boldsymbol{\theta}}_{20}|\mathscr{D})\}, \tag{3.18}$$

$$EE = \boldsymbol{U}_1(\hat{\boldsymbol{\theta}}_0)^T\widehat{Cov}_0(\hat{\boldsymbol{\theta}}_1)\boldsymbol{U}_1(\hat{\boldsymbol{\theta}}_0), \tag{3.19}$$

em que $\widehat{Cov}_0(\hat{\boldsymbol{\theta}}_1)$ é a matriz de variâncias e covariâncias estimada dada em (3.16) sob H_0, $\hat{\boldsymbol{\theta}}_0 = [\boldsymbol{\theta}_{10}^T \ \hat{\boldsymbol{\theta}}_{20}^T]^T$.

3.5 Diagnóstico

A análise de diagnóstico envolve algumas técnicas baseadas em exames visuais e medidas de distância para verificação do ajuste do modelo, como a verificação das suposições iniciais e identificação de valores discrepantes e/ou observações influentes nas estimativas dos parâmetros.

3.5.1 Resíduo

A análise de resíduos pode detectar a presença de pontos extremos e avaliar a adequação da distribuição da variável resposta. Conforme visto na Seção 1.3.2 do Capítulo 1, Dunn e Smyth (1996) propõem o resíduo quantílico aleatorizado como medida de distância entre os valores ajustados e os valores observados na amostra, mesmo quando os dados apresentam distribuição assimétrica, que é o caso de distribuições geométricas. Estes resíduos são obtidos pela função inversa da função de distribuição acumulada normal padrão para cada resposta (1.6) e, assim, obtidos os respectivos quantis. Para o modelo geométrico correlacionado temos que $a_i = F(y_i - 1; \hat{\pi}_i, \hat{\rho})$ e $b_i = F(y_i; \hat{\pi}_i, \hat{\rho})$, o resíduo quantílico aleatorizado r_{q_i} é obtido por:

$$r_{q_i} = \Phi^{-1}(u_i), \tag{3.20}$$

em que F é a função de distribuição acumulada, de 2.5 tem-se:

$$F(y_i; \pi_i, \rho) = \begin{cases} \pi_i \ \text{ se } \ y_i = 0, \\ \pi_i + \sum_{j=1}^{y_i} \pi_i(1-\pi_i)(1-\rho)[(1-\pi_i)(1-\rho)+\rho]^{j-1} \ \text{ se } \ y_i \geq 1. \end{cases}$$

Para $y_i \geq 1$, $F(y_i; \pi_i, \rho)$ é função da soma de uma progressão geométrica de razão $(1-\pi_i)(1-\rho)+\rho$, logo:

$$F(y_i; \pi_i, \rho) = \begin{cases} \pi_i \ \text{ se } \ y_i = 0, \\ \pi_i + (1-\pi_i)[1 - ((1-\pi_i)(1-\rho)+\rho)^{y_i}] \ \text{ se } \ y_i \geq 1. \end{cases} \tag{3.21}$$

$\hat{\pi}_i$ e $\hat{\rho}$ são os EMV de π_i e ρ, respectivamente, $\hat{\pi}_i = g^{-1}(\hat{\eta}_i) = g^{-1}(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$, u_i é uma variável aleatória com distribuição uniforme no intervalo $(a_i, b_i]$ e, portanto, r_{q_i} tem distribuição normal padrão. Esta aleatorização tem como objetivo evitar massas de pontos sobrepostos na distribuição dos resíduos. Desta forma, visualizando os resíduos em função das observações e/ou dos percentis de uma distribuição normal padrão, é possível avaliar a normalidade dos resíduos quantílicos aleatorizados e identificar distâncias discrepantes entre estimativas e observações.

3.5.2 Influência

Observações influentes podem ser detectadas por deleção de pontos pois, assim, o impacto da retirada de uma observação particular nas estimativas de regressão pode ser avaliado. Para a análise da influência global são apresentadas as medidas de distância de Cook generalizada (COOK, 1977) e distância de verossimilhança (COOK; PEÑA; WEISBERG, 1988). A retirada individual de pontos pode deixar de detectar pontos conjuntamente influentes. Para avaliação da influência conjunta de observações, Cook (1986) propõe verificar o impacto de pequenas perturbações no modelo ao invés da avaliação pela retirada de pontos, o que se denomina de influência local.

A distância de Cook generalizada mede o impacto da i-ésima observação no estimador de máxima verossimilhança de $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\rho})$ e é obtida por:

$$C_{(i)} = [\hat{\boldsymbol{\theta}}_{(-i)} - \hat{\boldsymbol{\theta}}]^T \boldsymbol{J}(\hat{\boldsymbol{\theta}})[\hat{\boldsymbol{\theta}}_{(-i)} - \hat{\boldsymbol{\theta}}], \tag{3.22}$$

em que $\hat{\boldsymbol{\theta}}_{(-i)}$ é o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ com a i-ésima observação retirada, $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$ é a matriz de informação de Fisher observada cujos elementos são obtidos por (3.6), (3.7) e (3.8).

A distância da verossimilhança, que mede a diferença entre $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_{(-i)}$ a partir do afastamento das funções de (log) verossimilhança, avaliando, portanto, o impacto nas estimativas de máxima verossimilhança dos parâmetros, é obtida por:

$$\begin{aligned} LD_i &= 2\{l(\hat{\boldsymbol{\theta}}|\mathscr{D}) - l(\hat{\boldsymbol{\theta}}_{(-i)}|\mathscr{D}\} \\ &= 2\left\{ \sum_{i=1}^{m} \delta_i \log\left[\frac{g^{-1}(\hat{\eta}_i)}{g^{-1}(\hat{\eta}_{(-i)})}\right] + \sum_{i=1}^{m}(1-\delta_i)\log\left[\frac{g^{-1}(\hat{\eta}_i)(1-g^{-1}(\hat{\eta}_i))(1-\hat{\rho})}{g^{-1}(\hat{\eta}_{(-i)})(1-g^{-1}(\hat{\eta}_{(-i)})(1-\hat{\rho}_{(-i)})}\right] + \right. \\ &\quad \left. + \sum_{i=1}^{m}(1-\delta_i)(y_i-1)\log\left[\frac{(1-g^{-1}(\hat{\eta}_i))(1-\hat{\rho})+\hat{\rho}}{(1-g^{-1}(\hat{\eta}_{(-i)}))(1-\hat{\rho}_{(-i)})+\hat{\rho}_{(-i)}}\right]\right\}, \end{aligned} \tag{3.23}$$

em que $\hat{\eta}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ e $\hat{\eta}_{(-i)} = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(-i)}$; $\hat{\boldsymbol{\beta}}$ e $\hat{\rho}$ os EMVs dado a amostra observada, $\hat{\boldsymbol{\theta}}|\mathscr{D}$, com $\hat{\boldsymbol{\beta}}_{(-i)}$ e $\hat{\rho}_{(-i)}$ EMVs dado a amostra observada com a i-ésima observação excluída, $\hat{\boldsymbol{\theta}}_{(-i)}|\mathscr{D}_{(-i)}$.

Assintoticamente, as medidas de distância apresentam aproximadamente distribuição qui-quadrado com p graus de liberdade $(\chi^2_p(\alpha))$, em que p é a quantidade de parâmetros do modelo. Ou seja, a i-ésima observação é considerada como influente se o valor da distância de Cook generalizada, $C_{(i)}$, ou o valor da distância da verossimilhança, LD_i, é maior que um valor crítico d_α, sendo α tal que $P(C_{(i)}$ ou $LD_i > d_\alpha) = \alpha$. Para este nosso modelo, $p = R+2$. Por gráficos dos componentes das distâncias em relação aos índices observados é possível visualizar os valores das distâncias identificando os possíveis pontos influentes nas estimativas.

Para verificar o impacto de conjuntos de pontos, Cook (1986) propõe avaliar pequenas perturbações no modelo, o que corresponde à analise da influência local. Este impacto é avaliado perturbando-se elementos da função de log-verossimilhança, covariáveis ou a variável resposta e

verificando a distância entre a função de log-verossimilhança em relação ao EMV do modelo sem perturbação, $\hat{\boldsymbol{\theta}}$, de $\boldsymbol{\theta}$, e a função de log-verossimilhança em relação ao EMV do modelo perturbado, $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}, \hat{\rho}_{\boldsymbol{\gamma}})$, de $\boldsymbol{\theta}$, isto é:

$$\begin{aligned} LD_{\boldsymbol{\gamma}} &= 2\{l(\hat{\boldsymbol{\theta}}|\mathscr{D}) - l(\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}|\mathscr{D}\} \\ &= 2\Bigg\{ \sum_{i=1}^{m} \delta_i \log\left[\frac{g^{-1}(\hat{\eta}_i)}{g^{-1}(\hat{\eta}_{\boldsymbol{\gamma}_i})}\right] + \sum_{i=1}^{m}(1-\delta_i)\log\left[\frac{g^{-1}(\hat{\eta}_i)(1-g^{-1}(\hat{\eta}_i))(1-\hat{\rho})}{g^{-1}(\hat{\eta}_{\boldsymbol{\gamma}_i})(1-g^{-1}(\hat{\eta}_{\boldsymbol{\gamma}_i})(1-\hat{\rho}_{\gamma})}\right] + \\ &+ \sum_{i=1}^{m}(1-\delta_i)(y_i-1)\log\left[\frac{(1-g^{-1}(\hat{\eta}_i))(1-\hat{\rho})+\hat{\rho}}{(1-g^{-1}(\hat{\eta}_{\boldsymbol{\gamma}_i}))(1-\hat{\rho}_{\boldsymbol{\gamma}})+\hat{\rho}_{\boldsymbol{\gamma}}}\right]\Bigg\}, \end{aligned} \tag{3.24}$$

em que $\hat{\eta}_{\gamma_i} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{\gamma}$, $\boldsymbol{\gamma}$ é um vetor de perturbação m-dimensional cujos elementos γ_i são tipos de perturbação, definidos tal que $0 \leq \gamma_i \leq 1$, $i = 1, \ldots, m$.

Para obtermos $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}$ precisamos da função de log-verossimilhança do modelo perturbado, $l_{\boldsymbol{\gamma}}(\boldsymbol{\theta}|\mathscr{D}, \boldsymbol{\gamma})$. Na perturbação de ponderação de casos, perturba-se elementos da função de log-verossimilhança do modelo não perturbado:

$$l_{\boldsymbol{\gamma}}(\boldsymbol{\theta}|\mathscr{D}, \boldsymbol{\gamma}) = \sum_{i=1}^{m} \gamma_i l(\boldsymbol{\theta}|\mathscr{D}_i), \tag{3.25}$$

em que $\mathscr{D}_i = (y_i; \boldsymbol{x}_i; \delta_i)$ é a i-ésima observação da amostra $\mathscr{D}$, $l(\boldsymbol{\theta}|\mathscr{D}_i)$ o i-ésimo elemento da função de log-verossimilhança em (3.3) (i-ésima parcela). Quando $\gamma_i = 1$, para todo $i = 1, \ldots, m$, não há perturbação no modelo, denominaremos esse vetor $\boldsymbol{\gamma} = \mathbf{1}$ de $\boldsymbol{\gamma}_0$. Quando $\gamma_i = 0$ tem-se que a i-ésima observação foi excluída.

A perturbação também pode ocorrer na covariável ou na variável resposta, nestes casos, a covariável r perturbada $\boldsymbol{x}_{r\boldsymbol{\gamma}} = \boldsymbol{x}_r + \boldsymbol{\gamma}$, $r = 1, \ldots, R$, a variável resposta perturbada $\boldsymbol{y}_{\boldsymbol{\gamma}} = \boldsymbol{y} + \boldsymbol{\gamma}$, $\boldsymbol{\gamma}_0 = \mathbf{0}$ e a função de log-verossimilhança do modelo perturbado é obtida por:

$$l_{\boldsymbol{\gamma}}(\boldsymbol{\theta}|\mathscr{D}, \boldsymbol{\gamma}) = \sum_{i=1}^{m} l(\boldsymbol{\theta}|\mathscr{D}_{\boldsymbol{\gamma}_i}), \tag{3.26}$$

com $\mathscr{D}_{\boldsymbol{\gamma}} = (\boldsymbol{y}; \boldsymbol{x}_{\boldsymbol{\gamma}}; \boldsymbol{\delta})$ na perturbação de covariável ou $\mathscr{D}_{\boldsymbol{\gamma}} = (\boldsymbol{y}_{\boldsymbol{\gamma}}; \boldsymbol{x}; \boldsymbol{\delta})$ na perturbação da variável reposta e $l(\boldsymbol{\theta}|\mathscr{D}_{\boldsymbol{\gamma}_i})$ a i-ésima parcela da função de log-verossimilhança em (3.3) considerando $\mathscr{D} = \mathscr{D}_{\boldsymbol{\gamma}}$

Desta forma, Cook (1986) sugere avaliar a influência estudando a maior variação da $LD_{\boldsymbol{\gamma}}$ em torno de $\boldsymbol{\gamma}_0$, o que corresponde a maximizar $|\boldsymbol{d}^T \boldsymbol{A} \boldsymbol{d}|$ obtendo o maior valor absoluto do autovalor da matriz $\boldsymbol{A}$ sendo $\boldsymbol{d}_{max}$ o autovetor correspondente que contém os valores da influência local das observações, em que $t\boldsymbol{d} = \boldsymbol{\gamma}_0 - \boldsymbol{\gamma}$, $t \in \mathbb{R}$, $\boldsymbol{d}^T \boldsymbol{d} = 1$ e $\boldsymbol{A} = \boldsymbol{\Gamma}' J(\hat{\theta})^{-1} \boldsymbol{\Gamma}$, sendo $\boldsymbol{\Gamma} = \partial^2 l(\boldsymbol{\theta}|\mathscr{D}_{\boldsymbol{\gamma}})/\partial \boldsymbol{\theta}' \partial \boldsymbol{\gamma}$ e $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$ a matriz de informação observada de Fisher cujos elementos são

obtidos por (3.6), (3.7) e (3.8), ambas avaliadas em $\hat{\boldsymbol{\theta}}$ e $\boldsymbol{\gamma}_0$. Temos que $|\boldsymbol{d}_{max}| \in [0,1]$ e o gráfico de seus valores em relação à ordem das observações pode revelar os pontos com maior influência na vizinhança de $\boldsymbol{LD}_{\boldsymbol{\gamma}}$.

3.6 Estudos de simulação

Para exemplificar a metodologia clássica apresentada, simulamos dados associados ao ajuste de modelo geométrico correlacionado. Os algoritmos para geração das variáveis com distribuição geométrica correlacionada, ajuste dos modelos de regressão e obtenção das medidas para análise de diagnóstico foram desenvolvidos no *software* estatístico R (R Core Team, 2018) e estão apresentados no Apêndice B. As estimativas de máxima verossimilhança para os parâmetros dos modelos foram obtidas e verificadas as propriedades frequentistas destes estimadores. As medidas de diagnóstico foram construídas da simulação de um conjunto de dados perturbados com o objetivo de avaliar a capacidade de detecção de observações discrepantes e/ou influentes.

Conjuntos de dados com distribuição geométrica correlacionada foram simulados considerando $R = 2$, $\beta_0 = 2$, $\beta_1 = -1$, $\beta_2 = -0,5$ e $\rho = 0,3$. Duas covariáveis foram consideradas, X_{i1} e X_{i2}, com distribuições $U(0,2)$ e $N(0,4)$, respectivamente; $i = 1,\ldots,m$. O procedimento numérico iterativo de Newton-Raphson foi aplicado para a estimação dos parâmetros por maximização da função de verossimilhança com o auxílio do pacote e função *maxLik* do *software* R. Podemos dividir este estudo de simulação em três etapas.

Na primeira etapa (Etapa I), conjuntos de dados foram gerados para verificar as funções de ligação (Tabela 1) que melhor relacionaram as estimativas das probabilidades de sucesso com as covariáveis de acordo com as medidas para critério de seleção de modelos, (3.9) e (3.10). Os conjuntos de dados também foram gerados com as funções de ligação apresentadas na Tabela 1.

Na segunda etapa (Etapa II), conjuntos de dados com as funções de ligação adequadas foram simulados e então verificadas as propriedades frequentistas dos EMVs como distribuição normal e intervalos de confiança assintóticos (3.11).

Na terceira etapa (Etapa III), um conjunto com dados perturbados foi simulado para verificação dos resíduos quantílicos aleatorizados (3.20), das medidas de influência global, (3.22) e (3.23), e local (3.24).

3.6.1 Seleção de Modelos

Nesta etapa, quatro conjuntos de dados foram gerados utilizando as quatro funções de ligação $g(.)$; logito, complemento log-log, log-log e probito, conforme descritas na Tabela 1, aqui as denominamos ligações de referência. Cada um desses quatro conjuntos de dados corresponde a 100 amostras de tamanho $m = 100$ de Y_i com distribuição $IGeo(\pi_i, \rho)$, simulados com $\pi_i = g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})$, $i = 1,\ldots,100$; $\beta_0 = 2$, $\beta_1 = -1$, $\beta_2 = -0,5$ e $\rho = 0,3$, para

todo i. Após a geração destas variáveis aleatórias, modelos de regressão foram ajustados com estas quatro funções de ligação. Assim, os ajustes dos modelos foram comparados de acordo com a função de ligação. Esta comparação é feita por seleção do melhor modelo dois a dois, comparação esta entre o modelo com a ligação de referência, isto é, com a mesma função de ligação da geração das variáveis, e cada um dos demais modelos ajustados com as outras três funções de ligação. Os critérios de seleção utilizados foram *AIC* e *BIC* e, então, feita a contagem do números de vezes que o modelo com a ligação de referência foi pior que os demais, pior significa apresentar maiores valores de *AIC* e *BIC*. A Tabela 3 apresenta estes resultados e, como exemplo, tomamos a primeira linha desta tabela, a qual mostra que, de 100 modelos ajustados, o modelo logito de referência apresentou 19 dos ajustes piores do que os ajustes do modelo complemento log-log, 07 piores que do modelo log-log e 20 piores que do modelo probito. Nota-se que o modelo logito de referência apresentou as menores frequências de piores ajustes que os demais modelos. O modelo log-log mostrou a pior performance entre os demais. Estes resultados indicam ser logito a função de ligação mais adequada para ajustes de modelos geométricos correlacionados.

Tabela 3 – Percentual de rejeição do modelo ajustado com a função de ligação referência em relação a cada modelo ajustado com as demais funções de ligação.

	Ligação utilizada no ajuste			
Ligação de referência	Logito	C.log-log	Log-log	Probito
Logito	–	19	07	20
C.log-log	33	–	13	27
Log-log	65	47	–	54
Probito	26	42	20	–

3.6.2 Propriedades assintóticas dos EMVs

Quatro conjuntos de dados referentes a quatro tamanhos amostrais, $m = 50$, 100, 200 e 400, foram gerados e modelos foram ajustados para avaliação das estimativas de máxima verossimilhança dos parâmetros em $n = 1000$ replicações em cada conjunto.

A Tabela 4 apresenta as estimativas médias de máxima verossimilhança para os parâmetros do modelo de regressão geométrico correlacionado ajustado com função ligação logito, o qual apresentou melhor desempenho de acordo com os critérios de seleção de modelos observados na Etapa I.

Para verificar o comportamento assintótico dos EMVs, medidas estimadas de viés e raiz do erro quadrático médio (REQM) também foram calculadas. Neste caso estima-se o viés por $\hat{\theta}_p - \theta_p$ e REQM $= \sqrt{\frac{\sum_{i=1}^{n}(\theta_{pi}-\hat{\theta}_p)^2}{n}}$, em que $\hat{\theta}_p = \sum_{i=1}^{n} \frac{\theta_{pi}}{n}$, θ_p o p-ésimo componente do vetor de parâmetros real, em nosso estudo $p = 1,\ldots,4$ e $n = 1000$ o número de réplicas. Nota-se que conforme o tamanho da amostra aumenta, os viéses e REQMs diminuem, sinalizando a propriedade de consistência dos EMVs. Também ocorre o aumento das probabilidades de

Tabela 4 – Estimativas dos parâmetros do modelo para diferentes tamanhos de amostra (m).

	m	$\hat{\beta}_0$ (2,0)	$\hat{\beta}_1$ (-1,0)	$\hat{\beta}_2$ (-0,5)	$\hat{\rho}$ (0,3)	m	$\hat{\beta}_0$ (2,0)	$\hat{\beta}_1$ (-1,0)	$\hat{\beta}_2$ (-0,5)	$\hat{\rho}$ (0,3)
Média	50	2,075	-1,046	-0,507	0,328	200	1,970	-0,982	-0,502	0,289
Viés		0,075	-0,046	-0,008	-0,028		-0,029	0,017	-0,002	0,011
REQM		0,783	0,553	0,117	0,616		0,369	0,252	0,045	0,090
PC		0,731	0,819	0,779	0,849		0,895	0,901	0,913	0,933
Média	100	2,056	-1,013	-0,507	0,316	400	1,988	-1,000	-0,501	0,292
Viés		0,056	-0,013	-0,007	-0,016		-0,011	0,000	-0,001	0,008
REQM		0,518	0,368	0,075	0,126		0,257	0,166	0,030	0,071
PC		0,825	0,863	0,834	0,878		0,918	0,936	0,946	0,954

cobertura (PC) com o aumento dos tamanhos de amostra, as quais se aproximam de 95%. As PC são as proporções dos intervalos de confiança assintóticos de nível $\alpha = 5\%$ ($IC(95\%)$) (3.11) que contêm o verdadeiro valor do parâmetro. A Figura 2 também mostra indicativos de comportamento assintótico normal dos EMVs, pois observa-se redução da assimetria dos histogramas e aproximação dos quantis aos quantis da distribuição normal padrão no QQ-plot. Portanto, embora as condições de regularidade para as propriedades dos estimadores de máxima verossimilhança não tenham sido verificadas analiticamente, observa-se neste estudo de simulação a convergência assintótica para amostras moderadas.

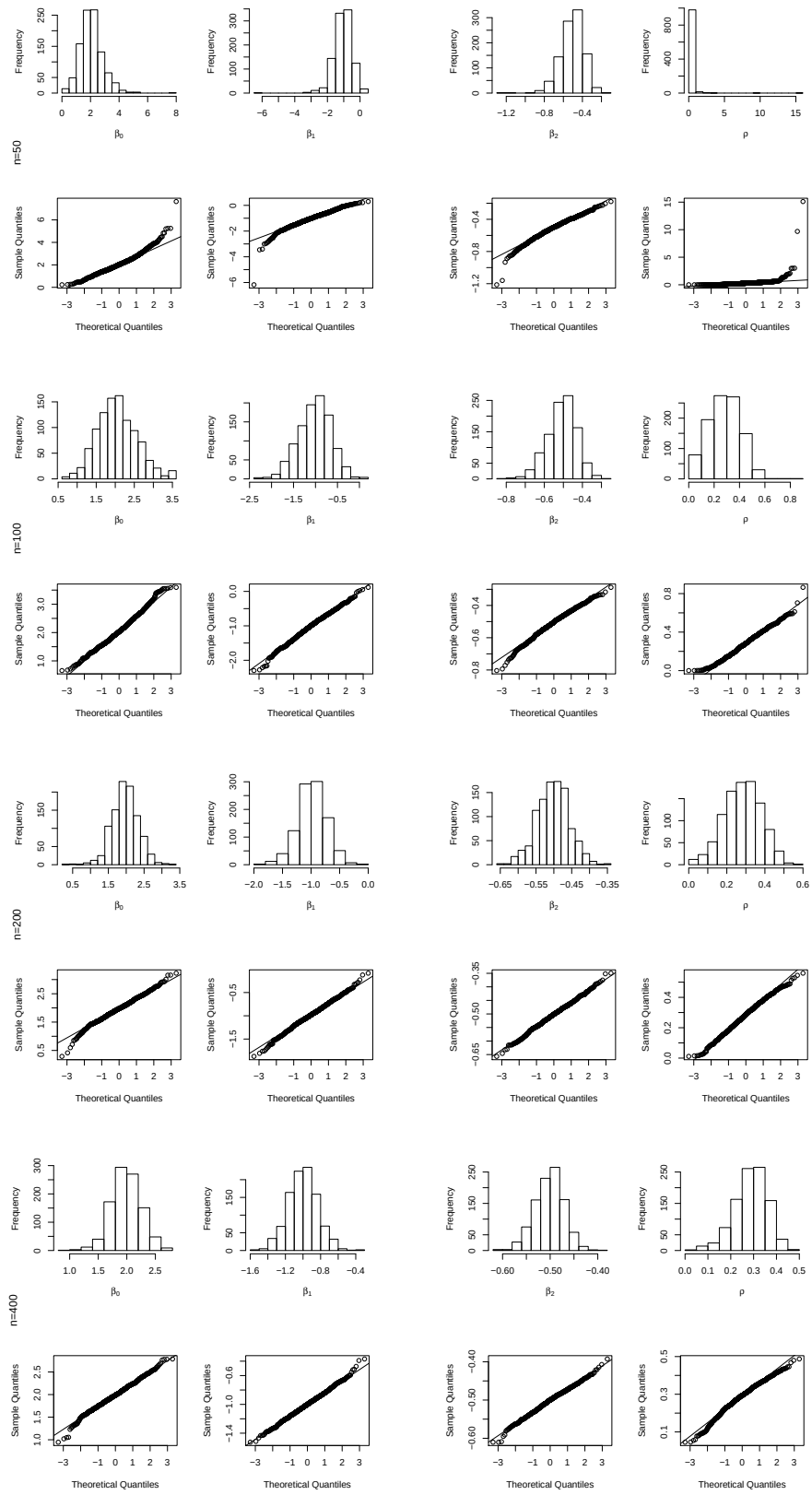

Figura 2 – Histograma e QQ-plot normal das estimativas de máxima verossimilhança.

3.6.3 Análise de Diagnóstico

Nesta etapa, uma amostra de tamanho 200 foi gerada de um modelo de regressão geométrico correlacionado também de parâmetros $\beta_0 = 2$, $\beta_1 = -1$, $\beta_2 = -0,5$, $\rho = 0,3$ e função de ligação logito. O modelo de regressão foi ajustado e as estimativas de máxima verossimilhança obtidas. Esta amostra foi perturbada com alterações em duas observações. A primeira alteração foi feita nas covariáveis da 100-ésima observação, em que x_{r100} recebeu seu valor $x_{r100} + 3max_i(x_{ri})$, $r = 1,2$ e $i = 1 \ldots m$. Uma segunda perturbação foi feita na variável resposta na observação 128 em que $y_{128} = y_{128} + 2max_i(y_i)$. As estimativas de máxima verossimilhança estão na Tabela 5. Embora verifica-se maiores alterações nas estimativas de β_0 e β_1 obtidas com a amostra perturbada em relação à amostra sem perturbação (maiores mudanças relativas), a estimativa do $IC95\%$ da amostra perturbada não inclui o verdadeiro valor do parâmetro β_2 e inclui o zero para ρ.

Para o conjunto de dados perturbados foi feita análise de diagnóstico para verificarmos o comportamento dos resíduos quantílicos aleatorizados e das medidas de distância que avaliam influência global e local das observações, verificando, portanto, se é possível detectar a influência destas observações discrepantes nas estimativas dos parâmetros. As medidas de diagnóstico podem ser visualizadas na Figura 3. Nota-se que os resíduos identificam a observação 100 como discrepante. As medidas de influência global evidenciam as duas observações, 100 e 128, como influentes, o que é evidente graficamente como também os valores dessas distâncias são maiores que o valor crítico $d_{0,05}$ de uma distribuição χ^2_4, $d_{0,05} = 9,5$, exceto para distância de Cook da observação 128. As medidas de influência local sugerem maior afastamento da observação 100 pela ponderação de perturbação de variável resposta ou covariável.

Tabela 5 – Estimativas dos coeficientes dos modelos com e sem perturbação.

	Amostra				Amostra perturbada			
	Valor real	*EMV*	*IC*95%		*EMV*	*IC*95%		Mudança Relativa (%)
			LI	*LS*		*LI*	*LS*	
β_0	2,000	1,886	1,669	2,525	2,153	1,529	2,777	14,2
β_1	−1,000	−0,924	−1,485	−0,971	−1,229	−1,542	−0,917	33,0
β_2	−0,500	−0,439	−0,507	−0,475	−0,409	−0,445	−0,372	−6,8
ρ	0,300	0,294	0,022	0,601	0,310	−0,097	0,716	5,3

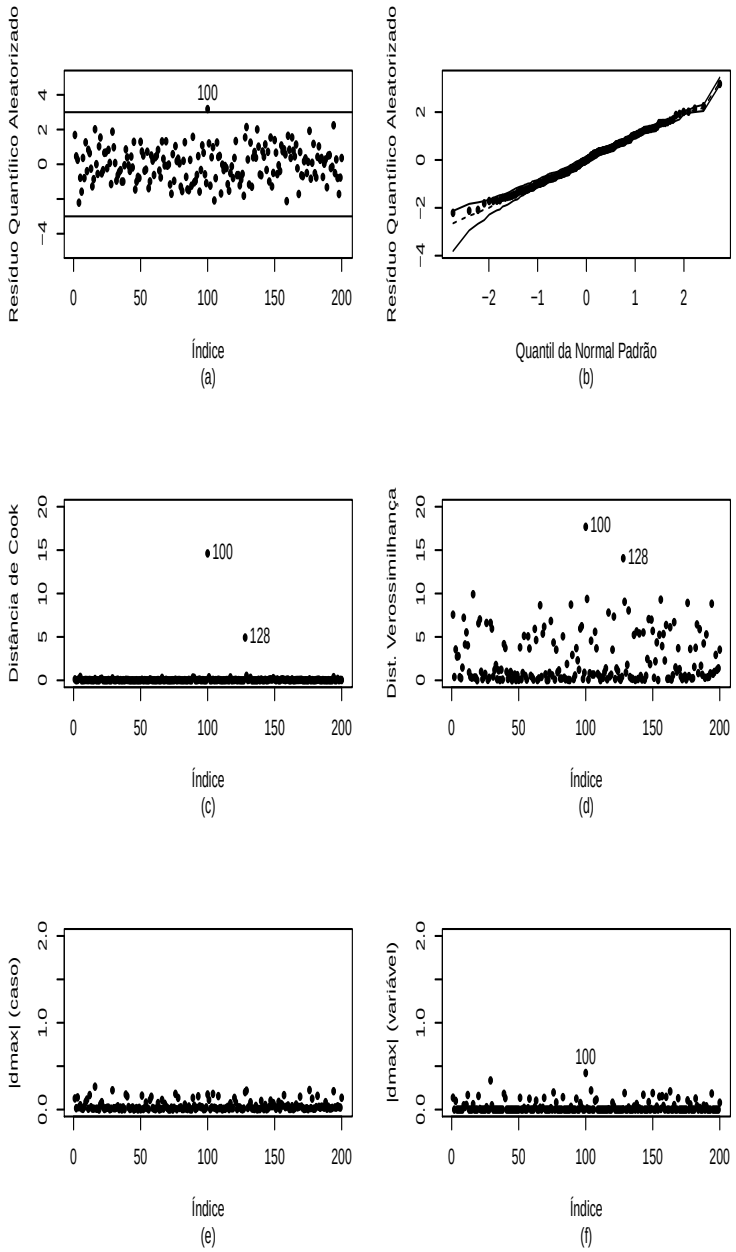

Figura 3 – (a) resíduo quantílico aleatorizado, (b) QQ-plot normal com envelope, (c) distância de Cook, (d) distância de verossimilhança, (e) influência local sob perturbação de casos, (f) influência local sob perturbação de respostas.

3.7 Aplicação: *Dados de internação na UTI*

A unidade de terapia intensiva, UTI, é uma unidade de extrema importância para um hospital. Um levantamento do Conselho Federal de Medicina do Brasil em 2018 mostra crescente falta de leitos nestas unidades. Menos de 10% dos municípios brasileiros possuem leitos de UTI pelo Sistema Único de Sáude (SUS). A Associação de Medicina Intensiva Brasileira (AMIB) aponta como ideal 1 a 3 leitos de UTI para cada 10 mil habitantes e, no ano de 2018, 17 unidades federais estavam abaixo deste índice para oferta de leitos de UTI pelo SUS. Na gestão do SUS,

é de responsabilidade dos estados e municípios a execução de serviços e organização da rede de assistência. Com isto, algumas ações governamentais podem contribuir para amenizar este problema da falta de leitos, entre elas estão o aprimoramento da rede de referências e contrarreferências do sistema público de saúde (rede de acesso ao sistema secundário de saúde, como os prontos-socorros); a regionalização dos cuidados especializados; a divulgação e educação eficientes dos conceitos de terminalidade e cuidados ao final da vida; e a desospitalização. Sendo assim, para a identificação de fatores significativos no tempo até a alta do paciente da UTI, tanto visando o processo de recuperação do paciente após cirurgia como também uma logística dos leitos, um modelo de regressão geométrico correlacionado pode ser bem apropriado. Para isto temos um conjunto de dados com 257 pacientes de um hospital público de cardiologia do estado de São Paulo, os quais apresentavam doença de aorta e foram submetidos a um procedimento cirúrgico. Os dias de permanência na UTI após cirurgia foram observados com o objetivo de estimar o efeito das covariáveis idade (anos completos), peso (kg), diagnóstico do tipo da doença (o qual pode ser de dois tipos), gênero, presença de insuficiência renal crônica (IRC) e presença de cirurgia cardíaca prévia (CCP) na probabilidade de alta dos pacientes, desta forma tem-se, para $i = 1, \ldots, 257$:

y_i : dias de UTI (média = 4,3; mediana = 3; min = 0, max =39);

x_{1i} : idade (em anos) (média = 57,7; mediana = 59,; min = 18; max = 81);

x_{2i} : peso (em kg) (média = 72,2; mediana = 72; min = 41; max =115);

x_{3i} : gênero (masculino ou feminino (34,7%));

x_{4i} : diagnóstico do tipo da doença (I ou II (75,1%));

x_{5i} : insuficiência renal crônica (ausência ou presença (8,2%));

x_{6i} : cirurgia cardíaca prévia (ausência ou presença (21,4%)).

A distribuição de frequências dos dias de permanência na UTI pode ser vista na Figura 4 e a relação destes dias com as covariáveis na Figura 5.

Como foi visto em (3.1), o modelo relaciona a probabilidade π_i de alta do paciente i, $i = 1, \ldots, 257$; com as covariáveis $\boldsymbol{x}_i = (1, x_{1i}, \ldots, x_{6i})^T$ por meio da função de ligação g, com $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_6)^T$, dada por:

$$g(\pi_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_6 x_{6i},\ i = 1, \ldots, 257.$$

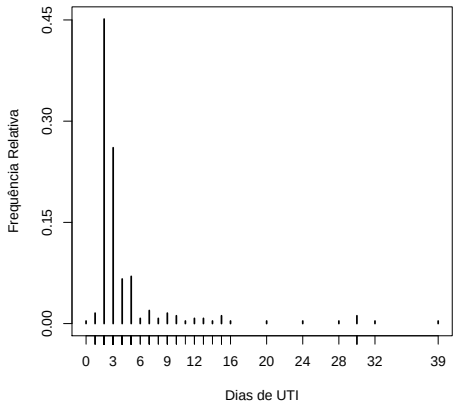

Figura 4 – Frequência relativa dos dias de permanência na UTI.

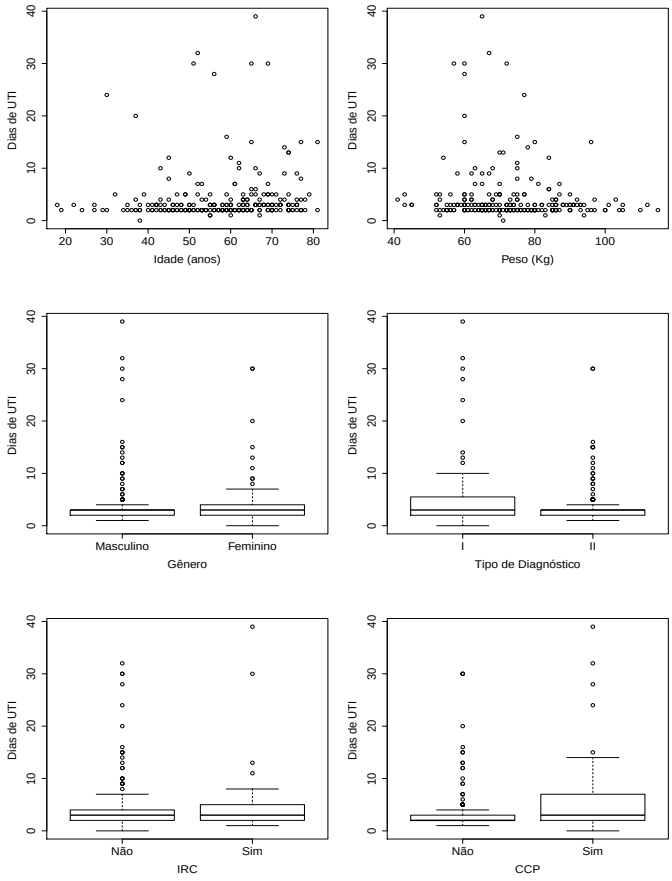

Figura 5 – Dias de permanência na UTI em função das covariáveis.

Inicialmente, os efeitos foram avaliados num modelo de regressão geométrico clássico. Como a permanência ou não na UTI, a cada dia, se refere a um mesmo indivíduo, pode existir uma dependência entre estas respostas na sequência de dias. Esta pode então ser medida pelo coeficiente de correlação da sequência de dias que ele permanece na UTI até sua alta, utilizando

modelo com dados de distribuição geométrica correlacionada. Os respectivos resultados estão na Tabela 6.

A Tabela 7 apresenta os resultados dos ajustes dos modelos incluindo somente os efeitos significativos (valor de $p < 0,10$). Para isto, foi utilizado o método *stepwise* para seleção de variáveis, sendo este um método iterativo que adiciona e remove as covariáveis a cada passo, de acordo com um critério, neste caso o critério de remoção foi valor de p maior que $0,10$ em parceria com os valores de *AIC* e *BIC*.

Tabela 6 – Estimativas de máxima verossimilhança para os efeitos nos Dias de permanência na UTI.

Parâmetro	EMV	Erro Padrão	Valor de p
MRG			
β_0	-1,788	0,573	0,002
$\beta_{1\ Idade}$	-0,088	0,057	0,121
$\beta_{2\ Peso}$	0,096	0,059	0,102
$\beta_{3\ G\ Masculino}$	-0,015	0,167	0,930
$\beta_{4\ Diag\ I}$	0,468	0,164	0,004
$\beta_{5\ IRC}$	-0,446	0,254	0,079
$\beta_{6\ CCP}$	-0,435	0,177	0,014
AIC = 1292, BIC = 1324			
MRGC			
β_0	-2,418	0,517	0,000
$\beta_{1\ Idade}$	-0,097	0,051	0,054
$\beta_{2\ Peso}$	0,096	0,052	0,063
$\beta_{3\ G\ Masculino}$	-0,015	0,150	0,920
$\beta_{4\ Diag\ I}$	0,480	0,151	0,001
$\beta_{5\ IRC}$	-0,437	0,235	0,063
$\beta_{6\ CCP}$	-0,436	0,163	0,007
ρ	-0,999	0,241	0,000
AIC = 1226, BIC = 1269			

Os critérios *AIC* e *BIC* avaliam melhor ajuste do modelo geométrico correlacionado (menores valores). Para corroborar com os resultados da Tabela 7, avaliamos se ρ contribui de forma significativa em explicar a variabilidade dos dados pelas estatísticas de teste apresentadas na Seção 3.4, como a estatística da razão de verossimilhanças (RV) (3.13). Assim, testamos a hipótese nula de interesse $H_0 : \rho = 0$. O valor da estatística de teste foi $67,54$ com valor de p= $2,22\text{x}10^{-16}$, confirmando forte contribuição significativa da correlação nos dias de UTI. Neste modelo, observa-se efeito significativo de uma forte correlação negativa. Além do efeito significativo do coeficiente de correlação, pode-se concluir que maiores idades, diagnóstico da doença de tipo 1, insuficiência renal crônica e cirurgia cardíaca prévia tiveram efeito negativo significante na probabilidade de alta dos pacientes sugerindo associações com maiores dias na UTI. Exceto peso que, quanto maior, aumenta a probabilidade de alta do paciente, sugerindo associação com menores dias na UTI.

O gráfico de resíduos em função dos índices estão na Figura 6. Nota-se, em ambos os casos, a presença de uma única observação com resíduo quantílico acima de 3, correspondente à

Tabela 7 – Estimativas de máxima verossimilhança para os efeitos nos Dias de permanência na UTI do modelo reduzido.

Parâmetro	EMV	Erro Padrão	Valor de p
MRG			
β_0	-2,343	0,413	0,000
$\beta_{2\ Peso}$	0,105	0,054	0,053
$\beta_{4\ Diag\ I}$	0,433	0,162	0,007
$\beta_{5\ IRC}$	-0,529	0,247	0,032
$\beta_{6\ CCP}$	-0,394	0,170	0,020
AIC = 1293, BIC = 1315			
MRGC			
β_0	-2,440	0,479	0,000
$\beta_{1\ Idade}$	-0,098	0,050	0,052
$\beta_{2\ Peso}$	0,099	0,047	0,036
$\beta_{4\ Diag\ I}$	0,480	0,150	0,001
$\beta_{5\ IRC}$	-0,437	0,235	0,063
$\beta_{6\ CCP}$	-0,431	0,157	0,006
ρ	-0,999	0,241	0,000
AIC = 1225, BIC = 1225			

observação 238. As medidas de diagnóstico na Figura 7 identificaram como possíveis pontos influentes as observações 165, 195 e 238 pela distância de verossimilhança LD, as quais também estão acima de um valor crítico d_α da distribuição χ^2_8 ($p = 8$) para $\alpha = 0,10$ ($d_{0,10} = 13,3$). Embora também podemos observar valores da distância de Cook mais afastados que a maioria, estes são muito pequenos (menores que 10^{-4}). Na análise da influência local, somente a ponderação por perturbação de caso (verossimilhança) sugere as observações 195, 215 e 238 como influentes. Com estes resultados, nota-se que estas observações podem influenciar nas estimativas dos parâmetros por causar impacto na função de verossimilhança. Portanto, as estimativas de máxima verossimilhança (MV) foram também obtidas na ausência dessas observações para a avaliação deste impacto. Nota-se que a exclusão dessas observações pode mudar as inferências sobre os parâmetros, pois o nível descritivo de alguns testes (valor de p) muda, como vemos para β_1 (idade), β_2 (peso) e β_5 (IRC) (Tabela 8). Porém, as estimativas de MV dos parâmetros não apresentaram grandes variações nos valores, com mudanças relativas menores que $0,01\%$.

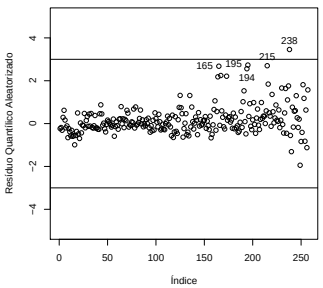

Figura 6 – Resíduos quantílicos aleatorizados: Dias de permanência na UTI.

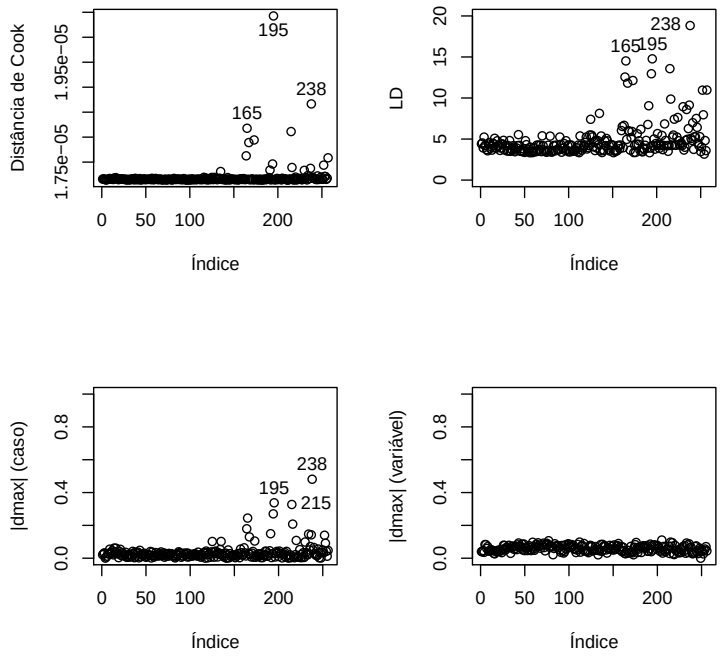

Figura 7 – Distância de Cook, distância de verossimilhança (LD), influência local sob perturbação de casos e sob perturbação de respostas ($|\boldsymbol{d}_{max}|$).

Para avaliação do modelo final, considera-se todas as observações porém tendo conhecimento das observações influentes (Tabela 8) nos resultados da análise. O modelo logístico final estimado é dado por:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_0 + \hat{\beta}_{1Peso}\, x_{1i} + \hat{\beta}_{2Idade}\, x_{2i} + \hat{\beta}_{4Diagnstico}\, x_{4i} + \hat{\beta}_{5IRC}\, x_{5i} + \hat{\beta}_{6CCP}\, x_{6i}, \quad (3.27)$$

$i = 1, \ldots, 257$; com $\hat{\boldsymbol{\beta}}$ dado na Tabela 7.

A Figura 8 mostra a distribuição das probabilidades de alta estimadas e a relação com as covariáveis (à direita). O mesmo foi feito para o MRG-logístico clássico (à esquerda), o qual também não avalia significativo o efeito de gênero, porém avalia a idade sem significância. Pela figura observa-se estimativas menores das probabilidades de alta pelo MRGC em relação ao MRG, mostrando que o modelo geométrico clássico pode, neste caso, superestimar a probabilidade de alta dos pacientes.

Resumindo, os resultados mostram que, após procedimento cirúrgico, as covariáveis idade, tipo de doença, IRC, CCP e peso tem efeitos significativos na probabilidade de alta do paciente da UTI. De forma mais detalhada, maiores idades, doença tipo I, IRC, CCP e menores pesos estão associados a maior permanência na UTI, ou seja, a menores probabilidades de alta. No MRGC observa-se efeito significativo de uma forte correlação negativa. É importante salientar que, caso o modelo clássico ($\rho = 0$) fosse ajustado, a conclusão seria diferente, isto é, a covariável idade não teria efeito significativo e as estimativas das probabalidades de alta

Tabela 8 – Estimativas de máxima verossimilhança para os efeitos nos Dias de permanência na UTI do MRGC na ausência de observações influentes.

Exclusão		ρ	β_0	β_1	β_2	β_3	β_4	β_5	β_6
Nenhuma	EMV	-0,999	-2,418	-0,097	0,096	-0,015	0,480	-0,437	-0,436
	Erro Padrão	0,241	0,517	0,051	0,052	0,150	0,151	0,235	0,163
	Valor de p	0,000	0,000	0,054	0,063	0,920	0,001	0,063	0,007
{165}	EMV	-0,999	-2,418	-0,097	0,096	-0,015	0,480	-0,437	-0,436
	Erro Padrão	0,240	0,522	0,050	0,052	0,149	0,152	0,235	0,164
	Valor de p	0,000	0,000	0,038	0,103	0,871	0,008	0,062	0,004
{195}	EMV	-0,999	-2,418	-0,097	0,096	-0,015	0,480	-0,437	-0,436
	Erro Padrão	0,239	0,512	0,050	0,051	0,150	0,151	0,242	0,164
	Valor de p	0,000	0,000	0,055	0,090	0,678	0,001	0,517	0,003
{215}	EMV	-0,999	-2,418	-0,097	0,096	-0,015	0,480	-0,437	-0,436
	Erro Padrão	0,241	0,517	0,051	0,052	0,150	0,151	0,235	0,163
	Valor de p	0,000	0,000	0,054	0,063	0,920	0,001	0,063	0,007
{238}	EMV	-0,999	-2,418	-0,097	0,096	-0,015	0,480	-0,437	-0,436
	Erro Padrão	0,238	0,511	0,050	0,052	0,149	0,151	0,235	0,162
	Valor de p	0,000	0,000	0,074	0,105	0,708	0,001	0,045	0,006
{165, 195}	EMV	-0,999	-2,418	-0,097	0,096	-0,015	0,480	-0,437	-0,436
	Erro Padrão	0,238	0,516	0,050	0,051	0,149	0,151	0,241	0,165
	Valor de p	0,000	0,000	0,039	0,141	0,876	0,006	0,504	0,001
{165, 195}	EMV	-0,999	-2,418	-0,097	0,096	-0,015	0,480	-0,437	-0,436
	Erro Padrão	0,238	0,531	0,050	0,052	0,149	0,151	0,235	0,163
	Valor de p	0,000	0,000	0,061	0,036	0,941	0,006	0,044	0,003
{165, 238}	EMV	-0,999	-2,418	-0,097	0,096	-0,015	0,480	-0,437	-0,436
	Erro Padrão	0,237	0,515	0,049	0,052	0,148	0,151	0,235	0,163
	Valor de p	0,000	0,000	0,051	0,172	0,507	0,004	0,044	0,003
{195, 215}	EMV	-0,999	-2,418	-0,097	0,096	-0,015	0,480	-0,437	-0,436
	Erro Padrão	0,241	0,517	0,051	0,052	0,150	0,151	0,235	0,163
	Valor de p	0,000	0,000	0,054	0,063	0,920	0,001	0,063	0,007
{195, 238}	EMV	-0,999	-2,418	-0,097	0,096	-0,015	0,480	-0,437	-0,436
	Erro Padrão	0,241	0,517	0,051	0,052	0,150	0,151	0,235	0,163
	Valor de p	0,000	0,000	0,054	0,063	0,920	0,001	0,063	0,007
{215, 238}	EMV	-0,999	-2,418	-0,097	0,096	-0,015	0,480	-0,437	-0,436
	Erro Padrão	0,237	0,515	0,049	0,052	0,148	0,151	0,235	0,163
	Valor de p	0,000	0,000	0,051	0,172	0,507	0,004	0,044	0,003
{165, 195, 215}	EMV	-0,999	-2,418	-0,097	0,096	-0,015	0,480	-0,437	-0,436
	Erro Padrão	0,236	0,525	0,050	0,052	0,148	0,151	0,241	0,164
	Valor de p	0,000	0,000	0,064	0,052	0,812	0,004	0,422	0,001
{165, 195, 238}	EMV	-0,999	-2,418	-0,097	0,096	-0,015	0,480	-0,437	-0,436
	Erro Padrão	0,235	0,510	0,049	0,051	0,147	0,151	0,241	0,164
	Valor de p	0,000	0,000	0,052	0,228	0,721	0,003	0,413	0,001
{165, 215, 238}	EMV	-0,999	-2,418	-0,097	0,096	-0,015	0,480	-0,437	-0,436
	Erro Padrão	0,234	0,524	0,049	0,052	0,147	0,151	0,235	0,162
	Valor de p	0,000	0,000	0,079	0,066	0,558	0,003	0,031	0,002
{195, 215, 238}	EMV	-0,999	-2,418	-0,097	0,096	-0,015	0,480	-0,437	-0,436
	Erro Padrão	0,234	0,515	0,050	0,052	0,147	0,150	0,241	0,162
	Valor de p	0,000	0,000	0,114	0,052	0,984	0,000	0,351	0,002
{165,195, 215, 238}	EMV	-0,999	-2,418	-0,097	0,096	-0,015	0,480	-0,437	-0,436
	Erro Padrão	0,232	0,519	0,049	0,052	0,146	0,151	0,241	0,163
	Valor de p	0,000	0,000	0,083	0,090	0,783	0,002	0,343	0,001

dos pacientes seriam maiores. Os critérios *AIC* e *BIC* para seleção de modelos avaliaram como melhor ajuste o modelo geométrico correlacionado, corroborado pela rejeição da hipótese nula $H_0 : \rho = 0$, o qual também mostrou satisfazer as suposições iniciais pelos resíduos quantílicos aleatorizados. Este modelo, portanto, para esta aplicação discutida, se apresentou mais adequado.

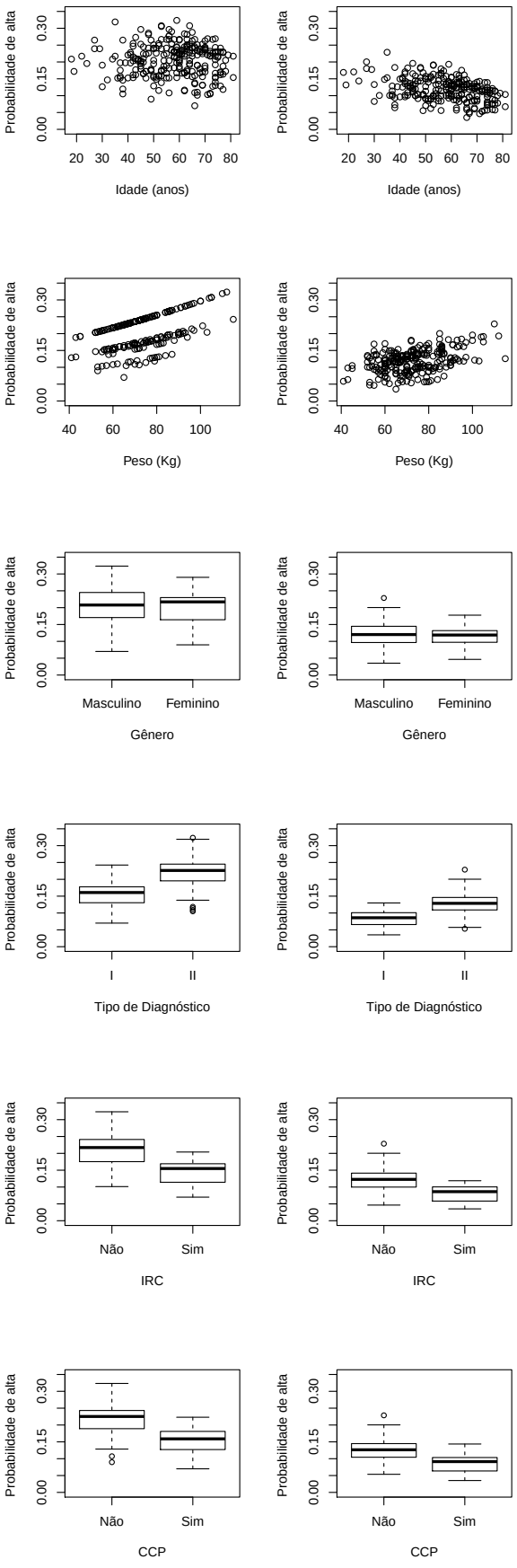

Figura 8 – Probabilidade de alta da UTI: estimativas do MRG (à esquerda) e do MRGC (à direita).

CAPÍTULO

4

MODELO DE REGRESSÃO GEOMÉTRICO CORRELACIONADO BAYESIANO

Neste capítulo apresentamos o modelo de regressão geométrico correlacionado (MRGC) considerando uma abordagem Bayesiana. Nesta abordagem, o principal objetivo é obter as estimativas dos coeficientes do modelo através de suas distribuições de probabilidade a *posteriori*. Estas ditribuições foram então adquiridas por métodos iterativos de simulação estocástica de Monte Carlo via cadeias de Markov (MCMC) utilizando algoritmo de Metropolis-Hastings via amostrador de Gibbs. As estimativas Bayesianas são obtidas através da minimização da perda esperada pela distribuição a *posteriori* para uma função de perda de interesse. No MRGC, as covariáveis se relacionam com a probabilidade do sucesso por meio de uma função de ligação adequada. Critérios de seleção Bayesianos foram utilizados para a escolha do melhor modelo de acordo com a função de ligação. Uma análise de diagnóstico avalia o ajuste do modelo por resíduos quantílicos aleatorizados a *posteriori* e observações influentes por medidas de divergência-ψ. Estudos de simulação ilustram os métodos propostos. Os dados reais de pacientes, já apresentados na Seção 3.7 do Capítulo 3, foram também analisados utilizando este modelo Bayesiano. A comparação com os resultados obtidos com o modelo geométrico usual é apresentada na Seção 4.6.

4.1 Modelo de regressão Bayesiano

Conforme já visto nos Capítulos 2 e 3, nas Seções 2.1 e 3.1, respectivamente, a variável aleatória Y tem distribuição geométrica correlacionada com parâmetros π, $\pi \in (0,1)$ e ρ, $\max\{-1, -\frac{1-\pi}{\pi}\} \leq \rho < 1$ ($IGeo(\pi,\rho)$), quando Y representa a quantidade de $falhas$ que antecede o primeiro *sucesso* em uma sequência de ensaios de Bernoulli dependentes com probabilidade de sucesso π, coeficiente de correlação ρ e distribuição de probabilidade dada por (2.5), com média e variância dadas por (2.6).

Seja $\boldsymbol{y} = (y_1, \ldots, y_m)$ m observações de $\boldsymbol{Y}$ em que cada $Y_i \sim IGeo(\pi_i, \rho)$, $i = 1, \ldots, m$. O modelo de regressão geométrico correlacionado (MRGC) é tal que as probabilidades de sucesso π_i satisfazem a relação funcional em (1.1), também apresentada em (3.1). Neste capítulo, as funções de ligação utilizadas $g : (0,1) \longrightarrow \mathbb{R}$, estritamente monótonas e duplamente diferenciáveis, foram as já apresentadas na Tabela 1.

4.2 Estimação

Considerando a função de ligação em (3.1), a função de verossimilhança de $\boldsymbol{\theta} = (\boldsymbol{\beta}, \rho)$ dado a amostra observada $\mathscr{D} = (\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\delta})$ é obtida por:

$$\mathscr{L}(\boldsymbol{\theta}|\mathscr{D}) = \prod_{i=1}^{m} (g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}))^{\delta_i} \{g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})(1-g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}))(1-\rho)[(1-g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}))(1-\rho)+\rho]^{y_i-1}\}^{1-\delta_i}, \tag{4.1}$$

sendo $\delta_i = 1$ se $y_i = 0$ e $\delta_i = 0$ se $y_i > 0$; $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)^T$.

Uma parte importante da análise Bayesiana é especificar as distribuições a *priori* para todos os parâmetros desconhecidos do modelo. Para garantir distribuições a *posteriori* próprias, adota-se distribuições a *priori* próprias para todas as quantidade desconhecidas do modelo, isto é, distribuições a *priori* integráveis. Assim, considera-se as seguintes distribuições a *priori* para os parâmetros do MRGC:

- $\beta_j \sim N(0, \sigma_j^2)$, $\sigma_j^2 < \infty$ são conhecidos, com $j = 1, \ldots, p$;
- $\rho|\boldsymbol{\beta} \sim U(\max\{-1, -\tau\}, 1)$, com $\tau = \frac{1-g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta})}{g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta})}$;

em que $N(\mu, \sigma^2)$ é a distribuição normal com média μ e variância σ^2, e $U(\max\{-1, -\tau\}, 1)$ a distribuição uniforme no intervalo $(\max\{-1, -\tau\}, 1)$. Sob essas condições, a densidade conjunta de $\boldsymbol{\beta}$ e ρ é dada por:

$$P(\boldsymbol{\theta}) = P(\boldsymbol{\beta}, \rho) = P(\boldsymbol{\beta})P(\rho|\boldsymbol{\beta}) \propto \frac{\exp\left\{-\sum_{j=1}^{p} \frac{\beta_j^2}{2\sigma_j^2}\right\}}{1 - \max\{-1, -\tau\}}. \tag{4.2}$$

Nota-se que, no caso em que $\sigma_j \longrightarrow \infty$ $(j = 1, \ldots, p)$, a função densidade a *priori* conjunta é imprópria, pois $P(\boldsymbol{\theta}) \propto \frac{1}{1-\max\{-1,-\tau\}}$, em que $-1 < \max(-1, -\tau) < 0$, então $0 < 1 - \max\{-1, -\tau\} < 2$ e $P(\boldsymbol{\theta}) \propto C$; C constante e o intervalo de variação de $\boldsymbol{\beta}$ é ilimitado.

Combinando a função de verossimilhança (4.1) e a função densidade a *priori* conjunta (4.2), pelo Teorema de Bayes tem-se que a função densidade conjunta a *posteriori* de $\boldsymbol{\theta}|\mathscr{D} = (\boldsymbol{\beta}, \rho)|\mathscr{D}$ é obtida por:

$$P(\boldsymbol{\theta}|\mathscr{D}) \propto \prod_{i=1}^{m} (g^{-1}(\eta_i))^{\delta_i} \{g^{-1}(\eta_i)(1-g^{-1}(\eta_i))(1-\rho)[(1-g^{-1}(\eta_i))(1-\rho)+\rho]^{y_i-1}\}^{1-\delta_i} P(\boldsymbol{\theta}), \tag{4.3}$$

em que $\mathscr{D}$ é o conjunto de dados observados, $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ e $P(\boldsymbol{\theta})$ obtida por (4.2).

Observa-se que a função de densidade conjunta a *posteriori* (4.3) não tem forma conhecida e, então, recorre-se a métodos aproximados para a avaliação das funções de densidades marginais a *posteriori* como os métodos de simulação estocástica MCMC. Para isto, o algoritmo Metropolis-Hastings pode ser utilizado. Neste caso, tem-se $\boldsymbol{\theta} = (\boldsymbol{\beta}, \rho)$, e, de (4.2), $P(\boldsymbol{\theta}) = P(\boldsymbol{\beta})P(\rho|\boldsymbol{\beta}) \propto \frac{\exp\left\{-\sum_{j=1}^{p}\frac{\beta_j^2}{2\sigma_j^2}\right\}}{1-\max\{-1,-\tau\}}$ e fica mais vantajoso computacionalmente utilizar o caso particular deste algoritmo, o amostrador de Gibbs (CASELLA; GEORGE, 1992), para uma convergência mais rápida da distribuição estacionária que é a densidade a *posteriori* de interesse. No amostrador de Gibbs a cadeia irá sempre se mover, as transições de um estado para o outro são feitas de acordo com as distribuições condicionais completas $P(\theta_k|\boldsymbol{\theta}_{(-k)})$, $k = 1,\ldots,(p+1)$, sendo $\boldsymbol{\theta}_{(-k)} = (\theta_1,\ldots,\theta_{k-1},\theta_{k+1},\ldots,\theta_{p+1})^T$, conforme algoritmo:

(1) iniciar com contador de iteração $j = 0$ e qualquer valor $\boldsymbol{\theta}^{(0)} = (\beta_1^{(0)},\ldots,\beta_p^{(0)},\rho^{(0)})^T$,

(2) obter um novo valor de $\boldsymbol{\theta}^{(j)}$ a partir de $\boldsymbol{\theta}^{(j-1)}$ através das gerações sucessivas de valores :

$$\begin{aligned}
\beta_1^{(j)} &\sim P\left(\beta_1|\beta_2^{(j-1)},\beta_3^{(j-1)},\ldots,\beta_p^{(j-1)},\rho^{(j-1)}\right)\\
\beta_2^{(j)} &\sim P\left(\beta_2|\beta_1^{(j)},\beta_3^{(j-1)},\ldots,\beta_p^{(j-1)},\rho^{(j-1)}\right)\\
&\vdots\\
\beta_p^{(j)} &\sim P\left(\beta_2|\beta_1^{(j)},\beta_2^{(j)},\ldots,\beta_{p-1}^{(j)},\rho^{(j-1)}\right)\\
\rho^{(j)} &\sim P\left(\rho|\beta_1^{j},\beta_2^{(j)},\ldots,\beta_p^{(j)}\right)
\end{aligned}$$

(3) incrementar o contador de j para $j+1$ e voltar ao passo (2).

Isto é, para se obter uma amostra a *posteriori* de $\boldsymbol{\beta}$ e ρ, o amostrador de Gibbs se baseia em sucessivas gerações das distribuições condicionais completas $P(\boldsymbol{\beta}|\rho,\mathscr{D})$ e $P(\rho|\boldsymbol{\beta},\mathscr{D})$ em cada iteração. Da densidade a *posteriori* conjunta dada em (4.3), tem-se as densidades condicionais completas a *posteriori* para o algoritmo de Gibbs, dadas por:

$$\begin{aligned}
P(\boldsymbol{\beta}|\rho,\mathscr{D}) &\propto \prod_{i=1}^{m} g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})(1-g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}))^{1-\delta_i}[(1-g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}))(1-\rho)+\rho]^{(y_i-1)(1-\delta_i)}\,P(\boldsymbol{\beta}),\\
P(\rho|\boldsymbol{\beta},\mathscr{D}) &\propto (1-\rho)^{m-d}\prod_{i=1}^{m}[(1-g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}))(1-\rho)+\rho]^{(y_i-1)(1-\delta_i)}\,P(\rho|\boldsymbol{\beta}),
\end{aligned} \tag{4.4}$$

sendo $d = \sum_{i=1}^{m}\delta_i$ a quantidade de valores de Y observados iguais a zero.

A informação contida na distribuição a *posteriori* também pode ser resumida por um único ponto (estimação pontual) ou por intervalo (intervalo de credibilidade ou intervalo de confiança bayesiano).

Na estimação pontual, a regra de decisão λ para um único valor para o parâmetro $\theta \in \boldsymbol{\theta} = (\boldsymbol{\beta}, \rho)$ está associada a uma perda $L(\lambda, \theta)$ de tal modo que quanto maior a distância entre a regra de decisão e θ maior a perda. Neste caso, o risco de uma regra de decisão, denotado por $R(\lambda)$, é a perda esperada a *posteriori*: $R(\lambda) = E_{\theta|\boldsymbol{y}}[L(\lambda, \theta)]$. A regra de decisão ótima (regra de Bayes) é a que apresenta menor risco, consiste portanto em escolher a estimativa que minimiza esta perda esperada. Para a função de perda quadrática, $L(\lambda, \theta) = (\lambda - \theta)^2$, o estimador de Bayes é a média a *posteriori* $E(\theta|\boldsymbol{y})$. Para a função de perda absoluta, $L(\lambda, \theta) = |\lambda - \theta|$, é a mediana a *posteriori*. Para a função de perda $0-1$, $L(\lambda, \theta) = \begin{cases} 1 \text{ se } |\lambda - \theta| > \varepsilon \\ 0 \text{ se } |\lambda - \theta| \leq \varepsilon \end{cases}$, para todo $\varepsilon > 0$, o estimador de Bayes é a moda da distribuição a *posteriori*. É importante também associar alguma informação de precisão desta estimação pontual. Para a média a *posteriori*, as medidas de incerteza mais usuais são variância ou coeficiente de variação. Para a mediana a *posteriori*, é a distância entre quartis. Para a moda a *posteriori*, a medida de informação observada de Fisher.

Na estimação por intervalo, C é um intervalo de credibilidade de $100(1-\alpha)\%$ para θ se $P(\theta \in C) \geq 1 - \alpha$. C é um intervalo de credibilidade de comprimento mínimo quando θ é de máxima densidade a *posteriori*, intervalo HPD (*highest posterior density*), sendo assim, $C = \{\theta \in \Theta : p(\theta|\boldsymbol{x}) \geq c(\alpha)\}$ em que $c(\alpha)$ é a maior constante tal que $P(\theta \in C) \geq 1 - \alpha$ e Θ é o espaço paramétrico.

4.3 Critérios de comparação de modelos

Sabemos que mais de um modelo pode ser ajustado para um mesmo conjunto de dados. Existem alguns métodos de comparação de modelos para selecionar o melhor. Apresentamos quatro critérios Bayesianos de comparação de modelos : o *deviance information criterion* (DIC), proposto por Spiegelhalter *et al.* (2002) e também discutido mais atualmente em Spiegelhalter *et al.* (2014); o critério de informação de Akaike esperado (*expected Akaike information criterio*, EAIC) (BROOKS, 2002); o critério de informação Bayesiana esperado (EBIC) (CARLIN; LOUIS, 2001); e o critério log pseudo verossimilhança marginal (*log pseudo marginal likelihood*, LPML) que é derivado da ordenada preditiva condicional (CPO) (GELFAND; DEY; CHANG, 1992).

Seja $\mathscr{D}$ o conjunto de dados completos e $\mathscr{D}^{(-i)}$ o conjunto de dados com a i-ésima observação excluída. Seja $\boldsymbol{\theta}$ o vetor de parâmetros e distribuição a *posteriori* de $\boldsymbol{\theta}|\mathscr{D}^{(-i)} = P(\boldsymbol{\theta}|\mathscr{D}^{(-i)})$, $i = 1, \ldots, m$. Para a i-ésima observação, CPO_i é dada por:

$$CPO_i = \int_{\theta \in \Theta} f((y_i)|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathscr{D}^{(-i)}) d\boldsymbol{\theta} = \left\{ \int_{\theta \in \Theta} \frac{P(\boldsymbol{\theta}|\mathscr{D})}{f((y_i)|\boldsymbol{\theta})} d\boldsymbol{\theta} \right\}^{-1}, \tag{4.5}$$

em que $P(\boldsymbol{\theta}|\mathscr{D})$ é a distribuição de probabilidade a *posteriori* em (4.3) e $f(.)$ a função distribuição de probabilidade da variável do modelo a ser comparado. Assim, para comparar o MRGC

proposto em (2.5) e (3.1) com modelos alternativos, para um conjunto de dados observados, tem-se que:

$$f(y_i|\boldsymbol{\theta}) = \begin{cases} g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}), & y_i = 0 \\ g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})(1-g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}))(1-\rho)[(1-g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}))(1-\rho)+\rho]^{y_i-1}, & y_i = 1,2\ldots. \end{cases} \tag{4.6}$$

Neste caso, a *CPO* não apresenta uma forma analítica fechada e, sendo assim, uma estimativa de MCMC para CPO_i pode ser obtida usando uma amostra da distribuição a *posteriori*. Seja $\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(Q)}$ uma amostra aleatória de tamanho Q de $P(\boldsymbol{\theta}|\mathscr{D})$, uma aproximação de Monte Carlo para CPO_i é dada por:

$$\widehat{CPO}_i = \left\{ \frac{1}{Q}\sum_{q=1}^{Q} \frac{1}{f(y_i|\boldsymbol{\theta}^{(q)})} \right\}^{-1}. \tag{4.7}$$

Para comparação de modelos o critério *LPML* é definido por $LPML = \sum_{i=1}^{m} \log(\widehat{CPO}_i)$. A medida de desvio (ou *deviance*) D é dada por $D(\boldsymbol{\theta}^{(q)}) = -2\sum_{i=1}^{m} \log(f(y_i|\boldsymbol{\theta}^{(q)})$. Os critérios *EAIC*, *EBIC* e *DIC* podem então ser estimado por:

$$\widehat{EAIC} = \bar{D} + 2(p+1),\ \ \widehat{EBIC} = \bar{D} + 2(p+1)\log(m)\ \text{ e }\ \widehat{DIC} = 2\bar{D} - \hat{D},$$

em que $p+1$ é o número de parâmetros do modelo, $\bar{D} = \frac{1}{Q}\sum_{q=1}^{Q} D(\boldsymbol{\theta}^{(q)})$ e $\hat{D} = D(\bar{\boldsymbol{\theta}}^{(q)})$; com $\boldsymbol{\theta}^{(q)} = \left(\beta_0^{(q)},\ldots,\beta_p^{(q)},\rho^{(q)}\right)^T$ e $\bar{\boldsymbol{\theta}}^{(q)} = D\left(\frac{1}{Q}\sum_{q=1}^{Q}\beta_0^{(q)},\ldots,\frac{1}{Q}\sum_{q=1}^{Q}\beta_p^{(q)},\frac{1}{Q}\sum_{q=1}^{Q}\rho^{(q)}\right)^T$.

Na comparação entre modelos, maior valor de *LPML* indica o melhor modelo enquanto que para os critérios *EAIC*, *EBIC* e *DIC*, são os modelos que apresentam valores menores desses critérios.

4.4 Diagnóstico

Após a modelagem é importante verificar as suposições iniciais do modelo, como independência entre as variáveis resposta e sua distribuição de probabilidade, e a presença de observações discrepantes e/ou influentes que possam causar alguma distorção nos resultados da análise.

4.4.1 Resíduos

Como visto na Seção 1.3.2, após o ajuste do modelo, as distâncias entre os valores ajustados (preditos) e os valores observados na amostra podem ser verificadas pelos resíduos. A visualização gráfica dos resíduos em função das observações nos permite verificar o comportamento destas distâncias e, assim, identificar valores discrepantes.

Os resíduos quantílicos aleatorizados propostos por Dunn e Smyth (1996), definido na Seção 1.3.2, se mostram apropriados para distribuições discretas assimétricas como as distribuições geométricas. Novamente, seja $F(\boldsymbol{y};\boldsymbol{\pi},\rho)$ a função de distribuição acumulada de $\boldsymbol{Y}$; $a_i = F(y_i-1;\hat{\pi}_i,\hat{\rho})$ e $b_i = F(y_i;\hat{\pi}_i,\hat{\rho})$, o resíduo quantílico aleatorizado r_{q_i} é dado por:

$$r_{q_i} = \Phi^{-1}(u_i),$$

em que u_i é uma variável aleatória com distribuição uniforme no intervalo $(a_i, b_i]$ e, portanto, r_{q_i} tem distribuição normal padrão. E, conforme visto em (3.21) no Capítulo 3,

$$F(y_i;\hat{\pi}_i,\hat{\rho}) = \begin{cases} \hat{\pi}_i \ \text{ se } \ y_i = 0, \\ \hat{\pi}_i + (1-\hat{\pi}_i)\{1-[(1-\hat{\pi}_i)(1-\hat{\rho})+\hat{\rho}]^{y_i}\} \ \text{ se } \ y_i \geq 1. \end{cases}$$

Neste caso, $\hat{\pi}_i = g^{-1}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$ conforme definido em (3.1); $\hat{\boldsymbol{\beta}}$ e $\hat{\rho}$ são as estimativas Bayesianas de $\boldsymbol{\beta}$ e ρ que minimizam a perda esperada pela distribuição a *posteriori* (4.3) para uma função de perda de interesse conforme visto na Seção 4.2.

4.4.2 Influência

Observações influentes podem ser detectadas por deleção de pontos pois, assim, o impacto da retirada de uma observação particular, ou de um conjunto de obervações, nas estimativas de regressão pode ser avaliado (COOK, 1977). Peng e Dey (1995) apresentam medidas de divergência para detectar observações influentes no ajuste de modelos de regressão Bayesianos baseadas na distribuição a *posteriori*. Esta avaliação é feita para detectar o impacto de cada observação nas distribuições a *posteriori*, conjuntas e marginais, através da retirada de cada observação do conjunto de dados.

Seja $D_\psi(P,P_{(-i)})$ a divergência-ψ entre P e $P_{(-i)}$, em que P é a distribuição a *posteriori* de $\boldsymbol{\theta}|\mathscr{D}$ e $P_{(-i)}$ a distribuição a *posteriori* de $\boldsymbol{\theta}|\mathscr{D}^{(-i)}$ (excluindo a i-ésima observação). A divergência-ψ é então obtida por:

$$D_\psi(P,P_{(-i)}) = \int P(\boldsymbol{\theta}|\mathscr{D})\ \psi\left(\frac{P(\boldsymbol{\theta}|\mathscr{D})}{P(\boldsymbol{\theta}|\mathscr{D}^{(-i)})}\right)d\boldsymbol{\theta} = E_{\boldsymbol{\theta}|\mathscr{D}}\left[\psi\left(\frac{CPO_i}{f(y_i|\boldsymbol{\theta})}\right)\right], \tag{4.8}$$

em que CPO é a ordenada preditiva condicional definida em (4.5) e $E_{\boldsymbol{\theta}|\mathscr{D}}$ a esperança em relação à *posteriori* conjunta $P(\boldsymbol{\theta}|\mathscr{D})$. Pode-se obter $D_\psi(P,P_{(-i)})$ por amostragem a partir da distribuição a *posteriori* de $\boldsymbol{\theta}$ por método MCMC: seja $\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(Q)}$ uma amostra de tamanho Q de $P(\boldsymbol{\theta}|\mathscr{D})$, a estimativa de Monte Carlo é dada por:

$$\widehat{D_\psi}(P,P_{(-i)}) = \frac{1}{Q}\sum_{q=1}^{Q}\psi\left(\frac{\widehat{CPO_i}}{f(y_i|\boldsymbol{\theta}^{(q)})}\right). \tag{4.9}$$

$D_\psi(P,P_{(-i)})$ mede o tamanho da retirada da i-ésima observação dos dados completos na distribuição a *posteriori* de $\boldsymbol{\theta}$. Dependendo da escolha da função ψ, tem-se diferentes medidas de

divergência (DEY; BIRMIWAL, 1994): se $\psi(u) = -\log(u)$ a divergência Kullback-Leibler (K-L *divergence*); se $\psi(u) = (u-1)\log(u)$ a divergência J-*distance*; para $\psi(u) = 0,5|u-1|$ tem-se a norma L_1 e para $\psi(u) = (u-1)^2$ a divergência qui-quadrado (χ^2-*divergence*). Logo, de (4.8), a divergência Kullback-Leibler é dada por: $D_{K-L}(P, P_{(-i)}) = -\log(CPO_i) + E_{\boldsymbol{\theta}|\mathscr{D}}\left[\log f(y_i|\boldsymbol{\theta})\right]$ e a estimativa MCMC por: $\widehat{D_{K-L}}(P, P_{(-i)}) = -\log(\widehat{CPO}_i) + \frac{1}{Q}\sum_{q=1}^{Q} \log f(y_i|\boldsymbol{\theta}^{(q)})$.

Pode-se considerar o i-ésimo caso influente quando o valor d de $D_\psi(-i)$ estiver acima de um ponto de corte. Este ponto de corte pode ser avaliado obtendo a medida de divergência-ψ entre uma moeda não viciada (probabilidade de sucesso $p = 0,5$) e uma moeda bastante viciada com probabilidade de sucesso p muito alta (ou muito baixa), como $p = 0,8$ (ou $p = 0,2$), por exemplo. Neste caso, com distribuição de probabilidade Bernoulli, a medida de divergência-ψ é simétrica em torno de seu valor mínimo que é quando $p = 0,5$ e com as medidas de divergência-ψ considerando a probabilidade de sucesso da moeda viciada, neste caso $0,8$ (ou $0,2$), obtém-se os pontos de corte para cada divergência-ψ : $d_{K-L} > 0,22$, $d_{J-distance} > 0,416$, $d_{L_1}(0,80) > 0,30$, $d_{\chi^2} > 0,562$. Maiores detalhes sobre os cálculos da divergência-ψ utilizando o vício de moedas para a obtenção dos pontos de corte podem ser vistos em Yiqi (2012).

4.5 Estudo de simulação

Nesta seção, um estudo de simulação foi realizado com o objetivo de avaliar as propriedades frequentistas dos estimadores Bayesianos baseados na perda quadrática e na perda absoluta dos parâmetros do MRGC. Estas duas perdas foram consideradas por exigir pouco desempenho computacional para seus cálculos e de suas medidas de incerteza associadas, variância e distância entre quartis, respectivamente. E também para que possamos verificar nas estimativas Bayesianas se existem diferenças relevantes entre considerar uma ou outra perda.

Neste estudo é considerado o modelo de regressão geométrico correlacionado com função de ligação logito, ou seja,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i},\ i = 1,\ldots,m,$$

em que $\beta_0 = 2$, $\beta_1 = -1$, $\beta_2 = -0,5$ e $\rho = 0,3$. A covariável x_1 foi gerada de uma distribuição normal $N(0,16)$ e a covariável x_2 foi de uma distribuição uniforme $U(0,2)$. Diferentes tamanhos amostrais (m = 50, 100, 200 e 400) foram considerados no estudo. Foram feitas 1000 simulações para cada tamanho de amostra e foram obtidos média, desvio padrão (DP), viés, raiz quadrada do erro quadrático médio (REQM) das estimativas Bayesianas sob perda quadrática e absoluta (média e mediana a *posteriori*) e a probabilidade de cobertura do intervalo HPD de 95%, para cada parâmetro do MRGC.

Para cada simulação foi considerada distribuição a *priori* dada em (4.2) com $\beta_j \sim N(0,100)$ $(j = 0,1,2)$ e $\rho|\boldsymbol{\beta} \sim U(\max\{-1,-\tau\},1)$, com $\tau = \frac{1-g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta})}{g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta})}$. A partir das densidades condicionais completas foram geradas cadeias com 15000 iterações cada utilizando o

caso especial do algoritmo de Metropolis-Hastings, o amostrador de Gibbs. Com a finalidade de diminuir o efeito dos pontos iniciais foram descartadas as primeiras 5000 iterações e, para evitar o problema de autocorrelação das séries, espaçamentos de tamanho 5 foram fixados, conduzindo a uma amostra a *posteriori* de tamanho 2000 para cada cadeia.

Tabela 9 – Médias das estimativas Bayesianas sob perda quadrática e absoluta, DP, REQM, viés e a probabilidade de cobertura (PC) do intervalo HPD de credibilidade de 95%.

		Perda quadrática				Perda absoluta				
m	Parâmetro	Média	DP	Viés	REQM	Média	DP	Viés	REQM	PC
50	β_0	2,162	0,701	0,162	0,719	2,141	0,696	0,141	0,710	0,965
	β_1	-1,020	0,081	-0,020	0,083	-1,016	0,080	-0,016	0,082	0,963
	β_2	-0,499	0,436	0,001	0,436	-0,490	0,435	0,010	0,435	0,958
	ρ	0,354	0,101	0,054	0,114	0,355	0,112	0,055	0,125	0,976
100	β_0	2,064	0,461	0,064	0,465	2,055	0,459	0,055	0,462	0,956
	β_1	-1,006	0,058	-0,006	0,058	-1,004	0,057	-0,004	0,057	0,947
	β_2	-0,501	0,295	-0,001	0,295	-0,502	0,295	-0,002	0,294	0,967
	ρ	0,320	0,100	0,020	0,102	0,321	0,108	0,021	0,110	0,966
200	β_0	2,033	0,309	0,033	0,311	2,027	0,309	0,027	0,310	0,959
	β_1	-1,005	0,042	-0,005	0,042	-1,004	0,042	-0,004	0,042	0,947
	β_2	-0,507	0,202	-0,007	0,202	-0,506	0,202	-0,006	0,202	0,940
	ρ	0,292	0,086	-0,008	0,087	0,293	0,091	-0,007	0,091	0,961
400	β_0	2,016	0,231	0,016	0,231	2,013	0,231	0,013	0,231	0,953
	β_1	-1,001	0,031	-0,001	0,031	-1,001	0,031	-0,001	0,031	0,955
	β_2	-0,501	0,151	-0,001	0,151	-0,501	0,151	-0,001	0,151	0,944
	ρ	0,296	0,077	-0,004	0,077	0,298	0,079	-0,002	0,079	0,934

A Tabela 9 apresenta as médias das estimativas Bayesianas sob perda quadrática e absoluta (média e mediana a *posteriori*) e respectivos desvios padrão, raiz quadrada do erro quadrático médio, viés e a probabilidade de cobertura do intervalo HPD de 95% de credibilidade para cada parâmetro do MRGC. Observa-se que o viés e a REQM diminuem a medida que o tamanho da amostra aumenta, se aproximando dos verdadeiros valores dos parâmetros. Isto ocorre para ambas as perdas, as quais também apresentam estimativas bem próximas. Além disso, as probabilidades de cobertura se aproximam do nível de credibilidade de 95%. Pode-se observar a similaridade entre os resultados obtidos com as simulações do método clássico apresentadas no capítulo anterior e os resultados desta análise Bayesiana.

4.6 Aplicação: *Dados de internação na UTI*

Nesta seção é apresentada a aplicação do modelo de regressão geométrico correlacionado (MRGC) Bayesiano aos dados já analisados no Capitulo 3 por estimação clássica. Foi visto na Seção 3.7 que o conjunto de dados é de um estudo observacional com 257 pacientes com doença de aorta que foram submetidos a um procedimento cirúrgico num hospital público de cardiologia do estado de São Paulo. Nele observa-se como resposta os dias de permanência do paciente na UTI após realização de procedimento cirúrgico (dias até alta da UTI). O objetivo principal é avaliar o efeito da idade, do peso, do diagnóstico do tipo da doença, do gênero, da insuficiência renal crônica e da cirurgia cardíaca prévia na probabilidade de alta dos pacientes.

O MRGC foi ajustado aos dados com todas as covariáveis, assumindo que a probabilidade de alta de cada paciente i (π_i) do MRGC se relaciona com as covariáveis por meio da função de ligação dada em (3.1), com $\boldsymbol{x}_i = (1, x_{1i}, \ldots, x_{6i})^T$ e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_6)^T$, ou seja,

$$g(\pi_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_6 x_{6i},\ i = 1, \ldots, 257,$$

especificamente considera-se as funções de ligação logito, complemento log-log, log-log e probito.

Na análise Bayesiana foram consideradas as distribuições a *priori* para os parâmetros dadas em (4.2) com $\beta_j \sim N(0, 100)$ $(j = 0, 1 \ldots 6)$ e $\rho|\boldsymbol{\beta} \sim U(\max\{-1, -\tau\}, 1)$, com $\tau = \frac{1-g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta})}{g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta})}$ e as amostras a *posteriori* são então obtidas da mesma forma que na Seção 4.5. Assim, o algoritmo MCMC foi implementado no *software* R (R Core Team, 2018) e, a partir das densidades condicionais completas foram geradas cadeias com 45000 iterações cada, utilizando o amostrador de Gibs, e com isso estimadas as distribuições a *posteriori* dos coeficientes do modelo. Com a finalidade de diminuir o efeito dos pontos iniciais foram descartadas as primeiras 5000 iterações e, para evitar o problema de autocorrelação das séries, fixados espaçamentos de tamanho 10, conduzindo a uma amostra a *posteriori* de 4000 iterações para cada cadeia.

Na Tabela 10 são apresentados os critérios de seleção de modelos Bayesianos DIC, EAIC, EBIC e LPML. Embora os valores estimados estejam bem próximos, todos os critérios indicam que o MRGC com função de ligação logito e complemento log-log são os que melhores ajustam esses dados. Pela melhor interpretabilidade da ligação logito, este será considerado nesta análise, o MRGC logístico.

Tabela 10 – Estimativas dos critérios Bayesianos de seleção de modelos para o ajuste dos dados de pacientes com doença de aorta.

Ligação	DIC	EAIC	EBIC	LPML
Logito	**1233,053**	**1242,232**	**1270,625**	**-617,072**
Complemento log-log	1233,603	1242,649	1271,042	-617,096
Log-log	1236,399	1245,953	1274,346	-618,652
Probito	1233,648	1242,453	1270,845	-617,259

A Tabela 11 mostra as estimativas de Bayes dos coeficientes de regressão e do coeficiente de correlação ρ do modelo geométrico correlacionado, estimativas baseadas na perda quadrática e absoluta e intervalos HPD de 95% de credibilidade. As estimativas Bayesianas dos coeficientes de correlação resultaram em -0,948 e -0,962, respectivamente, para perda quadrática e absoluta. Com um intervalo HPD de credibilidade de 95% de (-1,000 , -0,879), o qual não contém o zero, indicando, portanto, efeito significativo não nulo deste coeficiente. A distribuição a *posteriori* do parâmetro ρ pode ser vista na Figura 9. As estimativas de Monte Carlo dos critérios Bayesianos de comparação entre modelos de regressão correlacionado e não correlacionado estão na Tabela 12 e todos eles indicam melhor ajuste do MRGC.

Tabela 11 – Média, mediana, desvio padrão a *posteriori* e o intervalo HPD de 95% de credibilidade dos parâmetros dos MRG e MRGC logísticos.

	Não Correlacionado					Correlacionado				
				Intervalo HPD (95%)					Intervalo HPD (95%)	
Parâmetro	Média	Mediana	DP	LI	LS	Média	Mediana	DP	LI	LS
β_0	-1,827	-1,825	0,571	-2,803	-0,958	-2,414	-2,417	0,487	-3,226	-1,628
β_1	-0,086	-0,087	0,059	-0,190	0,006	-0,094	-0,095	0,051	-0,179	-0,008
β_2	0,098	0,100	0,058	0,008	0,198	0,096	0,097	0,051	0,015	0,175
β_3	-0,002	0,006	0,158	-0,274	0,237	-0,024	-0,024	0,150	-0,291	0,207
β_4	0,472	0,466	0,160	0,241	0,754	0,477	0,477	0,155	0,233	0,736
β_5	-0,459	-0,451	0,249	-0,880	-0,072	-0,466	-0,463	0,222	-0,861	-0,134
β_6	-0,444	-0,446	0,184	-0,725	-0,143	-0,445	-0,444	0,161	-0,684	-0,157
ρ	-	-	-	-	-	-0,948	-0,962	0,051	-1,000	-0,879

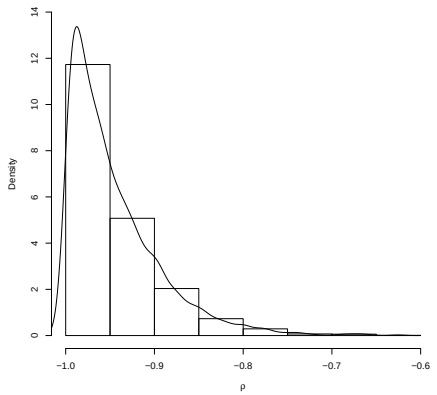

Figura 9 – Distribuição a *posteriori* de ρ para o modelo final dos dados de pacientes com doença de aorta.

Tabela 12 – Estimativas dos critérios Bayesianos para seleção do modelo geométrico logístico (correlacionado ou usual)

Modelo	DIC	EAIC	EBIC	LPML
Não-Correlacionado	1298,944	1306,032	1330,876	-649,100
Correlacionado	1233,053	1242,232	1270,625	-617,072

Os gráficos dos resíduos quantílicos aleatorizados a *posteriori* em função dos índices estão na Figura 10. Nota-se nenhuma observação discrepante com valor absoluto de resíduo acima de 3 e todos os valores se encontram dentro das bandas de confiança no gráfico de envelope, verifica-se algumas observações com valores de resíduos entre 2 e 3 cujos índices estão identificados na figura.

Com as amostras a *posteriori* também foram obtidas as estimativas de Monte Carlo das quatro medidas de divergência-ψ para os modelos de regressão geométricos correlacionado e não correlacionado (Tabela 11). Para o MRGC, todos os critérios indicam os casos 173, 195, 215 e 238 como possíveis observações influentes. Outros casos influentes são observados no modelo

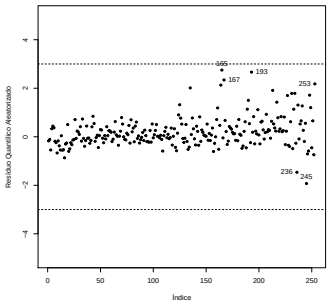

Figura 10 – Resíduos quantílicos aleatorizados a *posteriori*: Dias até alta da UTI.

não correlacionado, o que pode sugerir melhor robustez do MRGC em relação ao MRG.

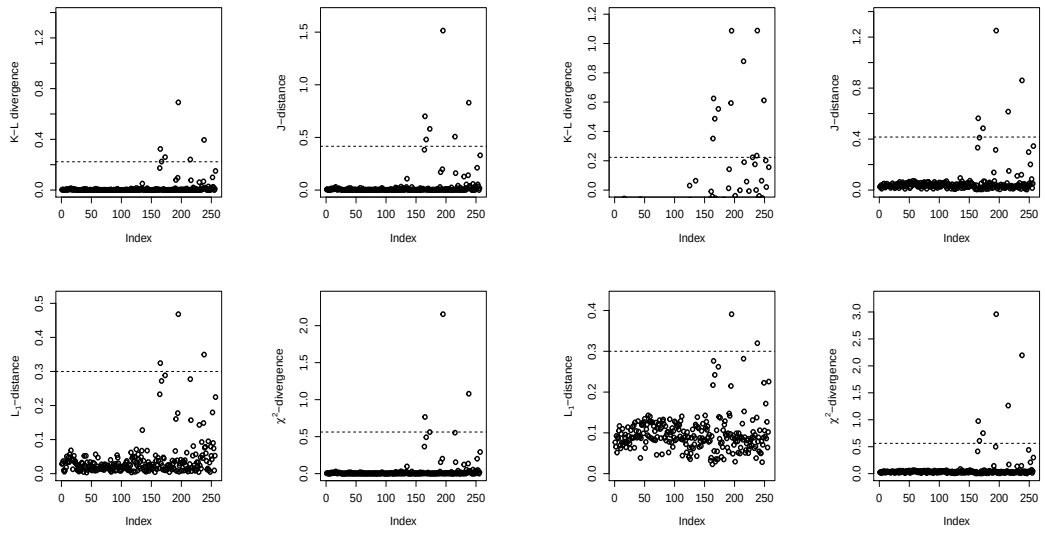

Figura 11 – Medidas de divergência-ψ do MRGC-logístico (à direita) e MRG-logístico (à esquerda) para dados de pacientes com doença de aorta.

Para avaliar o impacto da exclusão destas observações nas estimativa Bayesianas baseadas na perda quadrática foi feita uma análise reajustando o MRGC logístico sob a retirada uma a uma e conjuntamente dessas observações possivelmente influentes. Em cada caso se obteve a mudança relativa absoluta (MRA) para cada parâmetro estimado, isto é, $[|\hat{\theta}_j - \hat{\theta}_{j(i)}|/\hat{\theta}_j] \times 100$, em que $\hat{\theta}_{j(i)}$ é a estimativa de Bayes de θ_j com a retirada da i-ésima observação. A Tabela 13 apresenta essas mudanças relativas na qual observa-se que, com a retirada das observações 173, 195, 215 e 238, pelo menos conjunta 2-a-2, a covariável x_5 (IRC) deixa de ser significante em alguns casos. Também pela Tabela 13 observa-se a sensibilidade do modelo proposto a possíveis pontos influentes dos dados analisados. Considerando o critério LPML (Tabela 13) para os modelos reajustados, o melhor ajuste ocorreu na ausência das observações 173,195, 215 e 238, o qual não identificou efeito significante das covariáveis x_3 (gênero) e x_5 (IRC). Este modelo

reduzido resulta portanto em:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_0 + \hat{\beta}_{1,idade}x_{1i} + \hat{\beta}_{2,peso}x_{2i} + \hat{\beta}_{4,diagnóstico}x_{4i} + \hat{\beta}_{6,cirurgia}x_{6i}, \;\; i = 1,\ldots,253, \tag{4.10}$$

e as estimativas de Bayes baseadas em perda quadrática e absoluta dos parâmetros do MRGC logístico reduzido estão na Tabela 14 com respectivos intervalos HPD de 95% de credibilidade.

Tabela 13 – Mudança relativa absoluta da média a *posteriori* e intervalo HPD de 95% de credibilidade, excluindo as observações 173, 195, 215 e 238

Exclusão		β_0	β_1	β_2	β_3	β_4	β_5	β_6	ρ	LPML
Nenhum	Média	-2,406	-0,089	0,094	-0,013	0,449	-0,492	-0,373	-0,945	-617,07
	LI	-3,203	-0,165	0,009	-0,250	0,196	-0,877	-0,637	-1,000	
	LS	-1,546	-0,005	0,179	0,229	0,688	-0,112	-0,099	-0,873	
{173}	Média	-2,349	-0,089	0,087	-0,031	0,444	-0,470	-0,384	-0,947	-611,75
	MRA(%)	2,37	0,00	7,45	138,46	1,11	4,47	2,95	0,21	
	LI	-3,203	-0,174	0,000	-0,272	0,174	-0,865	-0,639	-1,000	
	LS	-1,577	-0,013	0,172	0,223	0,687	-0,096	-0,097	-0,876	
{195}	Média	-2,273	-0,099	0,083	-0,074	0,496	-0,198	-0,504	-0,946	-608,60
	MRA(%)	5,53	11,24	11,7	469,23	10,47	59,76	35,12	0,11	
	LI	-3,032	-0,181	-0,003	-0,320	0,231	-0,586	-0,794	-1,000	
	LS	-1,465	-0,023	0,164	0,179	0,725	0,231	-0,234	-0,879	
{215}	Média	-2,607	-0,088	0,120	-0,032	0,492	-0,488	-0,465	-0,946	-610,17
	MRA(%)	8,35	1,12	27,66	146,15	9,58	0,81	24,66	0,11	
	LI	-3,348	-0,171	0,040	-0,279	0,252	-0,909	-0,739	-1,000	
	LS	-1,694	-0,006	0,209	0,206	0,732	-0,145	-0,195	-0,876	
{238}	Média	-2,338	-0,090	0,082	0,038	0,510	-0,514	-0,460	-0,947	-610,62
	MRA(%)	22,83	1,12	12,77	392,31	13,59	4,47	23,32	0,21	
	LI	-3,148	-0,164	-0,006	-0,204	0,275	-0,923	-0,727	-1,000	
	LS	-1,496	-0,007	0,167	0,285	0,777	-0,142	-0,179	-0,880	
{173,195}	Média	-2,193	-0,116	0,087	-0,063	0,476	-0,182	-0,460	-0,949	-609,96
	MRA(%)	8,85	30,34	7,45	384,62	6,01	63,01	23,32	0,42	
	LI	-2,982	-0,198	-0,003	-0,312	0,231	-0,568	-0,723	-1,000	
	LS	-1,363	-0,033	0,167	0,161	0,724	0,214	-0,201	-0,882	
{195,215}	Média	-2,562	-0,085	0,113	-0,080	0,514	-0,223	-0,499	-0,946	-601,27
	MRA(%)	6,48	4,49	20,21	515,38	14,48	54,67	33,78	0,53	
	LI	-3,396	-0,163	0,021	-0,324	0,280	-0,594	-0,778	-1,000	
	LS	-1,694	-0,009	0,196	0,174	0,786	0,194	-0,240	-0,876	
{195,238}	Média	-2,294	-0,087	0,074	0,008	0,531	-0,223	-0,500	-0,948	-601,89
	MRA(%)	4,66	2,25	21,28	161,54	18,26	54,67	34,05	0,32	
	LI	-3,117	-0,168	-0,010	-0,260	0,272	-0,597	-0,742	-1,000	
	LS	-1,439	-0,004	0,161	0,244	0,761	0,176	-0,207	-0,883	
{173,195,238}	Média	-2,508	-0,076	0,095	-0,009	0,555	-0,232	-0,511	-0,948	-602,34
	MRA(%)	4,24	14,61	1,06	30,77	23,61	52,85	37,00	0,32	
	LI	-3,324	-0,153	0,011	-0,271	0,296	-0,579	-0,765	-1,000	
	LS	-1,732	0,008	0,177	0,216	0,793	0,205	-0,253	-0,882	
{195,215,238}	Média	-2,508	-0,076	0,095	-0,009	0,555	-0,232	-0,511	-0,948	-590,91
	MRA(%)	4,24	14,61	1,06	30,77	23,61	52,85	37	0,32	
	LI	-3,324	-0,153	0,011	-0,271	0,296	-0,579	-0,765	-1,000	
	LS	-1,732	0,008	0,177	0,216	0,793	0,205	-0,253	-0,882	
{173,195,215,238}	Média	-2,407	-0,096	0,103	-0,006	0,507	-0,240	-0,469	-0,950	-585,00
	MRA(%)	0,04	7,87	9,57	53,85	12,92	51,22	25,74	0,53	
	LI	-3,225	-0,181	0,022	-0,246	0,255	-0,630	-0,717	-1,000	
	LS	-1,551	-0,011	0,186	0,238	0,738	0,145	-0,190	-0,887	

Tabela 14 – Estimativas bayesianas dos parâmetros do MRGC logístico reduzido.

				Intervalo HPD (95%)	
Parâmetro	Média	Mediana	Desvio padrão	LI	LS
β_0	-2,377	-2,373	0,454	-3,114	-1,609
$\beta_{1,idade}$	-0,106	-0,105	0,049	-0,189	-0,031
$\beta_{2,peso}$	0,103	0,102	0,046	0,029	0,180
$\beta_{4,diagnóstico}$	0,519	0,515	0,147	0,287	0,763
$\beta_{6,cirurgia}$	-0,495	-0,492	0,158	-0,739	-0,230
ρ	-0,950	-0,964	0,046	-1,000	-0,888

Desta forma, avalia-se efeitos significativos de idade, tipo de doença, insuficiência renal crônica, cirurgia cardíaca prévia e peso na probabilidade de alta do paciente da UTI, de modo que maiores idades, doença tipo I, IRC, CCP e menores pesos sugerem mais dias de permanência na UTI (menores probabilidades de alta). No MRGC reduzido, com a retirada de valores influentes, a IRC não tem efeito significante na probabilidade de alta do paciente. Em ambas as situações, observa-se efeito significativo de uma forte correlação negativa e, juntamente com o melhor ajuste do modelo correlacionado, avalia-se o MRGC como o mais adequado para avaliação dos dados de pacientes com doença de aorta com internação pós cirúrgica na UTI. Vale ressaltar ainda, que o modelo clássico superestima a probabilidade de alta do paciente e, consequentemente, subestima o tempo de UTI.

Concluindo, o procedimento inferencial desenvolvido por métodos Bayesianos, baseados em simulção MCMC via algoritmo de Metropolis-Hastings por amostrador de Gibbs, mostrou-se adequado para o modelo proposto. A aplicação do modelo em conjunto de dados de pacientes na UTI mostrou efeitos significativos de idade, tipo de doença, IRC, CCP e peso na probabilidade de alta do paciente, de modo que maiores idades, doença tipo I, IRC, CCP e menores pesos estão associados a uma maior permanência na UTI (menores probabilidades de alta). No MRGC, observa-se efeito significativo de uma forte correlação negativa. Os critérios Bayesianos de seleção de modelos avaliaram melhor ajuste do modelo geométrico correlacionado, o qual também se mostrou satisfazer as suposições inciais pelos resíduos quantílicos aleatorizados a *posteriori*, e, portanto, para esta aplicação nosso MRGC se apresentou mais adequado. Também foi possível identificar observações influentes pelas medidas de divergência -ψ, as quais causam impacto na inferência dos dados. Neste caso, elas influenciam no efeito da insuficiência renal crônica para um maior tempo de UTI e a ausência dessas observações influentes sugerem também ausência do efeito da IRC.

CAPÍTULO

5

MODELO DE REGRESSÃO GEOMÉTRICO DE ORDEM *K* CLÁSSICO

Neste capítulo apresentamos o modelo de regressão considerando uma abordagem clássica para o caso particular do modelo geométrico de ordem k correlacionado apresentado na Seção 2.2 do Capítulo 2, em que $\rho = 0$, o qual corresponde ao modelo geométrico de ordem k ($kGeo(\pi)$) proposto por Philippou e Muwafi (1980), $k \in \{1,2,...\}$. A estrutura de regressão foi modelada por diferentes funções de ligação e o ajuste obtido por maximização da função de verossimilhança via processo numérico iterativo. A escolha dos melhores ajustes foi baseada em medidas de critério de seleção de modelos. Testes de hipóteses e intervalos de confiança assintóticos foram propostos para a inferência estatística sobre os parâmetros. Para análise de diagnóstico, medidas de resíduos quantílicos aleatorizados, de distância de Cook generalizada e de distância de verossimilhança foram utilizadas. Estudos de simulação ilustram os métodos propostos e as propriedades assintóticas dos estimadores de máxima verossimilhança. Uma aplicação envolvendo conjunto de dados reais de clientes inadimplentes na operação de crédito direto ao consumidor (CDC) enteira o modelo proposto.

5.1 Modelo de regressão *k*-geométrico - MRGk

Considera-se novamente o modelo geométrico de ordem k proposto por Philippou e Muwafi (1980), $kGeo(\pi)$, com função de probabilidade definida em (2.7):

$$P(Y_k = y+k|\pi) = \sum_{y_1,\ldots,y_k} \binom{y_1+\ldots+y_k}{y_1,\ldots,y_k} (1-\pi)^{\Sigma_{i=1}^k y_i} \pi^{y+k-\Sigma_{i=1}^k y_i} I_{\{0,1,2,\ldots\}}(y),$$

sendo o somatário em relação a todos os termos $y_1,\ldots,y_k$, inteiros não negativos, tais que $y = y_1 + 2y_2 + \ldots + ky_k$, $k = 1,2,\ldots$. A média e variância de Y_k, conforme visto em (2.8), são

dadas por:

$$E(Y_k) = \frac{(1-\pi^k)}{(1-\pi)\pi^k} \quad \text{e} \quad Var(Y_k) = \frac{[1-(2k+1)(1-\pi)\pi^k - \pi^{2k+1}]}{(1-\pi)^2\pi^{2k}}.$$

Sejam $Y_{k1}, Y_{k2}, \ldots Y_{km}$ variáveis aleatórias independentes em que Y_{ki} segue distribuição geométrica de ordem k com probabilidade de sucesso π_i, $kGeo(\pi_i)$, $i = 1, \ldots, m$. E, sejam $y_1+k, y_2+k, \ldots, y_m+k$ observações de $Y_{k1}, Y_{k2}, \ldots, Y_{km}$, respectivamente. O modelo de regressão geométrico de ordem k resulta da probabilidade de sucesso π_i satisfazer a relação funcional dada por (1.1). As funções de ligação discutidas anteriormente são utilizadas nesta análise.

5.2 Estimação

A função de log-verossimilhança do vetor $\boldsymbol{\beta}$ de coeficientes de regressão do modelo geométrico de ordem k, considerando a amostra observada $\mathscr{D} = (\boldsymbol{z}, \boldsymbol{x})$, em que $\boldsymbol{z} = (y_1+k, y_2+k, \ldots, y_m+k)^T$ (observações de $\boldsymbol{Y}_k$), é dada por:

$$l(\boldsymbol{\beta}|\mathscr{D}) = \sum_{i=1}^{m} \log\left\{ \sum_{y_{1i},\ldots,y_{ki}} \binom{S_{ki}}{y_{1i}\ldots,y_{ik}} [1-g^{-1}(\eta_i)]^{S_{ki}} g^{-1}(\eta_i)^{z_i - S_{ki}} I_{\{k,k+1,\ldots\}}(z_i) \right\}, \quad (5.1)$$

em que $z_i = y_i + k$, $y_i = y_{1i} + 2y_{2i} + \ldots + ky_{ki}$, $S_{ki} = y_{1i} + \ldots + y_{ki}$; $\eta_i = \sum_{r=0}^{R} \beta_r x_{ri}$ em que β_r são parâmetros desconhecidos, $i = 1, \ldots, m$, $r = 0, \ldots, R$.

A estimativa do vetor $\boldsymbol{\beta}$ não apresenta uma forma explicitamente conhecida e, assim, recorremos a métodos numéricos como o processo iterativo de Newton-Raphson, em que, no passo $n+1$, $\boldsymbol{\beta}^{(n+1)} = \boldsymbol{\beta}^{(n)} + [\boldsymbol{J}(\boldsymbol{\beta}^{(n)})]^{-1}\boldsymbol{U}(\boldsymbol{\beta}^{(n)})$, $n = 0, 1, \ldots$, e para $|\boldsymbol{\beta}^{(n+1)} - \boldsymbol{\beta}^{(n)}| < \varepsilon$, ε arbitrado, temos $\boldsymbol{\beta}^{(n+1)}$ a estimativa de máxima verossimilhança, sendo $\boldsymbol{J}(\boldsymbol{\beta}) = -\partial U(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ a matriz de informação observada de Fisher, em que $\boldsymbol{U}(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{x})/\partial\boldsymbol{\beta}$ é a função escore. Para $y_i \geq 0$ e $\Delta_i(\boldsymbol{\beta}) = \partial g^{-1}(\boldsymbol{\eta}_i)/\partial\boldsymbol{\beta}$, temos a função escore $\boldsymbol{U}(\boldsymbol{\beta})$ cujos elementos são:

$$U_{\beta_r} = \sum_{i=1}^{m} \left\{ \frac{\sum_{y_{i1},\ldots,y_{ik}} \left\{ \binom{S_{ki}}{y_{1i}\ldots,y_{ki}} \Delta_i(\beta_r)(1-g^{-1}(\eta_i))^{S_{ki}-1} g^{-1}(\eta_i)^{z_i-1-S_{ki}} [(z_i)[1-g^{-1}(\eta_i)] - S_{ki}] \right\}}{\sum_{y_{i1},\ldots,y_{ik}} \left\{ \binom{S_{ki}}{y_{1i}\ldots,y_{ki}} [1-g^{-1}(\eta_i)]^{S_{ki}} g^{-1}(\eta_i)^{z_i-S_{ki}} \right\}} \right\}. \quad (5.2)$$

Para o modelo geométrico de ordem k, $\boldsymbol{J}(\boldsymbol{\beta})$ tem dimensão $(R+1)\times(R+1)$, cujos elementos $J_{\beta_r\beta_s}$ são dados por : $J_{\beta_r\beta_s} = \sum_{i=1}^{m} J_{\beta_r\beta_s}(i)$ e $J_{\beta_r\beta_s}(i)$ é obtido por:

$$\frac{\sum_{y_{i1},\ldots,y_{ik}} \binom{S_{ki}}{y_{i1}\ldots,y_{ik}} \{\delta_i[(z_i)(1-g^{-1}(\eta_i)) - S_{ki}] - \Delta_i(\beta_r)\Delta_i(\beta_s) g^{-1}(\eta_i) S_{ki}\} (1-g^{-1}(\eta_i))^{S_{ki}-2} g^{-1}(\eta_i)^{z_i-2-S_{ki}}}{\sum_{y_{i1},\ldots,y_{ik}} \binom{S_{ki}}{y_{i1}\ldots,y_{ik}} (1-g^{-1}(\eta_i))^{S_{ki}} g^{-1}(\eta_i)^{z_i-S_{ki}}} +$$

$$- \left[\frac{\sum_{y_{i1},\ldots,y_{ik}} \binom{S_{ki}}{y_{i1}\ldots,y_{ik}} [(z_i)(1-g^{-1}(\eta_i)) - S_{ki}] \Delta_i(\beta_r)(1-g^{-1}(\eta_i))^{S_{ki}-1} g^{-1}(\eta_i)^{z_i-1-S_{ki}}}{\sum_{y_{i1},\ldots,y_{ik}} \binom{S_{ki}}{y_{i1}\ldots,y_{ik}} (1-g^{-1}(\eta_i))^{S_{ki}} g^{-1}(\eta_i)^{z_i-S_{ki}}} \right]^2, \quad (5.3)$$

em que $\delta_i = \Delta_i^2(\beta_r\beta_s)(1-g^{-1}(\eta_i))g^{-1}(\eta_i)+\Delta_i(\beta_r)\Delta_i(\beta_s)[(1-g^{-1}(\eta_i))(z_i-1)-1]$; $\Delta_i(\beta_r)$ e $\Delta_i^2(\beta_r\beta_s)$ as derivadas de primeira e segunda ordens, respectivamente, das funções de ligação em relação aos parâmetros $\boldsymbol{\beta}$, conforme descritas na Tabela 2 da Seção 1.3.1 do Capítulo 1, $r,s = 0,1,\ldots,R$ e $i = 1,\ldots,m$.

5.2.1 Seleção de modelos

O critério de informação de Akaike (*AIC*) (AKAIKE, 1974) e o critério de informação Bayesiano (*BIC*) (SCHWARZ *et al.*, 1978) podem ser utilizados para selecionar o melhor modelo de regressão sendo que os menores valores destes critérios indicam o melhor modelo: $AIC = -2l(\hat{\boldsymbol{\beta}}|\mathscr{D}) + p$ e $BIC = -2l(\hat{\boldsymbol{\beta}}|\mathscr{D}) + p\log(m)$, em que $l(\hat{\boldsymbol{\beta}}|\mathscr{D})$ é o valor da função de log-verossimilhança dada em (5.1) avaliada no estimador de máxima verossimilhança $(\hat{\boldsymbol{\beta}}|\mathscr{D})$, p é o número de parâmetros no modelo, neste caso $p = R+1$ e m é o número de observações.

5.2.2 Intervalos de confiança e testes de hipóteses assintóticos

Testes de hipóteses e intervalos de confiança sobre os parâmetros do modelo de regressão geométrico de ordem k podem ser feitos considerando a normalidade assintótica dos estimadores de máxima verossimilhança. Assim, sob certas condições de regularidade, pode-se demonstrar que o EMV dos parâmetros do modelo de regressão geométrico de ordem k, $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$, tem distribuição assintótica normal multivariada (Teorema no Apêndice A) (LEITE; SINGER, 1990) (p.112), dentre estas condições de regularidade tem-se, basicamente, que o suporte $S(\boldsymbol{z}) = \{z; f(\boldsymbol{z}|\boldsymbol{\beta}) > \mathbf{0}\}$ deve ser independente de $\boldsymbol{\beta}$ e que a troca das ordens das operações de derivação e de integracão sob a distribuicão da variável aleatória $\boldsymbol{Y}_k$ seja possível (BOLFARINE; SANDOVAL, 2001). Então:

$$\sqrt{m\boldsymbol{I}(\boldsymbol{\beta})}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, I_p),$$

em que $\boldsymbol{I}(\boldsymbol{\beta})$ é a matriz de informação esperada de Fisher e I_p a matriz identidade de ordem p. A distribuição assintótica dos EMVs continua válida se a matriz de informação de Fisher for substituída pela matriz de informação observada avaliada no EMV, $\boldsymbol{J}(\widehat{\boldsymbol{\beta}})$. Os elementos da matriz de informação observada são dados em (5.3) e também podem ser obtidos por aproximação numérica.

Da normalidade assintótica dos EMV, os intervalos de confiança podem ser aproximados para cada um dos parâmetros β_r com coeficiente de confiança de $100(1-\alpha)\%$:

$$IC(\beta_r;\alpha) = \hat{\beta}_r \pm n_{\alpha/2}(J_r^{-1}(\hat{\boldsymbol{\beta}}))^{1/2} \tag{5.4}$$

em que $n_{\alpha/2}$ é o quantil $\alpha/2$ da distribuição $N(0,1)$ e $J_r^{-1}(\hat{\boldsymbol{\beta}})$ é o termo de posição r da diagonal da matriz inversa $\boldsymbol{J}^{-1}(\hat{\boldsymbol{\beta}})$, correspondente à variância assintótica estimada do estimador de máxima verossimilhança, $r = 0,1,\ldots,R$.

Para testar a hipótese $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ (hipótese nula), em que $\boldsymbol{\beta}_0$ é um vetor p-dimensional de valores fixados, contra a hipótese alternativa $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, utiliza-se as três estatísticas de teste empregadas para testar H_0 :

1. Estatística de Wald (W):

$$W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \boldsymbol{J}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0). \tag{5.5}$$

2. Estatística de Wilks ou da razão de verossimilhança (RV):

$$\begin{aligned} RV &= 2\{l(\hat{\boldsymbol{\beta}}|\mathscr{D}) - l(\boldsymbol{\beta}_0|\mathscr{D})\} \\ &= 2\sum_{i=1}^{m} \log\left\{ \left[\frac{1-g^{-1}(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)}{1-g^{-1}(\boldsymbol{\beta}_0^T \boldsymbol{x}_i)}\right]^{S_{ki}} \left[\frac{g^{-1}(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)}{g^{-1}(\boldsymbol{\beta}_0^T \boldsymbol{x}_i)}\right]^{z_i - S_{ki}} I_{\{k,k+1,\ldots\}}(z_i) \right\}. \end{aligned} \tag{5.6}$$

3. Estatística de Rao ou dos escores eficientes (EE):

$$EE = \boldsymbol{U}(\boldsymbol{\beta}_0)^T \boldsymbol{J}(\boldsymbol{\beta}_0)^{-1} \boldsymbol{U}(\boldsymbol{\beta}_0), \tag{5.7}$$

em que $\mathscr{D} = (\boldsymbol{z}, \boldsymbol{x})$, $\hat{\boldsymbol{\beta}}$ é o EMV de $\boldsymbol{\beta}$; $\boldsymbol{\beta} = (\beta_0\ \beta_1\ \ldots\ \beta_R)^T$, $\boldsymbol{x}_i = (1,\ x_{1i},\ \ldots, x_{Ri})$ e $J(\hat{\boldsymbol{\beta}})$ a matriz de informação observada de Fisher cujos elementos são obtidos por (5.3) avaliados em $\hat{\boldsymbol{\beta}}$. $\boldsymbol{U}(\boldsymbol{\beta}_0)$ é a função escore cujos elementos são obtidos por (5.2) avaliada em $\boldsymbol{\beta}_0$, a qual não depende do EMV. Estas três estatísticas, sob a hipótese nula, têm distribuição assintótica qui-quadrado com p graus de liberdade, χ^2_p, em que p é a quantidade de parâmetros. Em nosso modelo $p = R+1$.

Para o teste de hipóteses de apenas um subconjunto do vetor de parâmetros consideramos as estatísticas de teste em (5.8), (5.9) e (5.10), para $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T\ \ \boldsymbol{\beta}_2^T]^T$ uma partição do vetor de parâmetros $\boldsymbol{\beta}$, $\boldsymbol{\beta}_1$ q-dimensional e $\boldsymbol{\beta}_2$ $p-q$-dimensional em que $\boldsymbol{\beta}_1$ é o vetor de interesse. Neste caso considera-se testar a hipótese $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10}$ contra a hipótese $H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_{10}$, em que $\boldsymbol{\beta}_{10}$ é um vetor q-dimensional de valores fixados e $\hat{\boldsymbol{\beta}}_0 = [\boldsymbol{\beta}_{10}^T\ \ \hat{\boldsymbol{\beta}}_{20}^T]^T$, $\hat{\boldsymbol{\beta}}_{20}$ é o EMV para $\boldsymbol{\beta}_2$ sob H_0. A três estatísticas de teste, sob a hipótese nula, têm distribuição assintótica qui-quadrado com q graus de liberdade, $\chi^2_{(q)}$.

$$W = (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})^T [\widehat{Cov}(\hat{\boldsymbol{\beta}}_1)]^{-1} (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}), \tag{5.8}$$

$$RV = 2\{\log \mathscr{L}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2|\boldsymbol{y};\boldsymbol{x}) - \log \mathscr{L}(\boldsymbol{\beta}_{10}, \hat{\boldsymbol{\beta}}_{20}|\boldsymbol{y};\boldsymbol{x})\}, \tag{5.9}$$

$$EE = \boldsymbol{U}_1(\hat{\boldsymbol{\beta}}_0)^T \widehat{Cov}_0(\hat{\boldsymbol{\beta}}_1) \boldsymbol{U}_1(\hat{\boldsymbol{\beta}}_0), \tag{5.10}$$

em que $\widehat{Cov}_0(\hat{\boldsymbol{\beta}}_1)$ é a matriz de variâncias e covariâncias estimada sob H_0, em que, $\widehat{Cov}(\hat{\boldsymbol{\beta}}_1)$ é dada por:

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}_1) = [\boldsymbol{J}_{11}(\hat{\boldsymbol{\beta}}) - \boldsymbol{J}_{12}(\hat{\boldsymbol{\beta}}) \boldsymbol{J}_{22}(\hat{\boldsymbol{\beta}})^{-1} \boldsymbol{J}_{21}(\hat{\boldsymbol{\beta}})]^{-1}, \tag{5.11}$$

e $\hat{\boldsymbol{\beta}} = [\hat{\boldsymbol{\beta}}_1^T\ \ \hat{\boldsymbol{\beta}}_2^T]^T$ é o EMV para $\boldsymbol{\beta}$ sem restrição.

5.3 Diagnóstico

Como visto na Seção 3.5.1 do Capítulo 3, a análise de diagnóstico envolve algumas técnicas baseadas em exames visuais e medidas de distância para verificação do ajuste do modelo.

A análise de resíduos pode detectar a presença de pontos extremos e avaliar a adequação da distribuição da variável resposta. Para dados que apresentam distribuição assimétrica, como dados com distribuições geométricas, utilizamos o resíduo quantílico aleatorizado, r_{q_i}, proposto por Dunn e Smyth (1996), conforme já visto em (1.6), isto é, $r_{q_i} = \Phi^{-1}(u_i)$, em que Φ é função de distribuição acumulada normal padrão e u_i é uma variável aleatória com distribuição uniforme no intervalo $(a_i, b_i]$. Para o modelo geométrico de ordem k, temos que $a_i = F(y_i - 1; \hat{\pi}_i)$ e $b_i = F(y_i; \hat{\pi}_i)$, sendo F a função de distribuição acumulada de Y_k dada por:

$$F(y_i;\pi_i) = \sum_{l=0}^{y_i} \Big\{ \sum_{l_1,\ldots,l_k} \binom{l_1+\ldots+l_k}{l_1,\ldots,l_k} (1-\pi_i)^{\Sigma_{j=1}^k l_j} \pi_i^{l+k-\Sigma_{j=1}^k l_j} \Big\}, \tag{5.12}$$

em que $l = l_1 + 2l_2 + \ldots + kl_k$, $y_i = 0, 1, \ldots$, $\hat{\pi}_i$ é EMV de π_i e $\hat{\pi}_i = g^{-1}(\hat{\eta}_i) = g^{-1}(\sum_{r=0}^R \hat{\beta}_r x_{ri})$.

Conforme Seção 3.5.2 do Capítulo 3, a análise de influência global é feita por deleção de pontos para avaliação do impacto da retirada de uma observação particular nas estimativas dos parâmetros do modelo. Para isto, utilizamos as distâncias de Cook generalizada $\boldsymbol{C}$ (COOK, 1977) e de verossimilhança $\boldsymbol{LD}$ (COOK; PEÑA; WEISBERG, 1988):

$$C_{(i)} = [\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}}]^T \boldsymbol{J}(\hat{\boldsymbol{\beta}})[\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}}], \tag{5.13}$$

$$\begin{aligned} LD_i &= 2\{l(\hat{\boldsymbol{\beta}}|\mathscr{D}) - l(\hat{\boldsymbol{\beta}}_{(-i)}|\mathscr{D})\}, \\ &= 2\sum_{i=1}^{m} \log\left\{ \left[\frac{1-g^{-1}(\hat{\eta}_i)}{1-g^{-1}(\hat{\eta}_{(-i)})}\right]^{S_{ki}} \left[\frac{g^{-1}(\hat{\eta}_i)}{g^{-1}(\hat{\eta}_{(-i)})}\right]^{z_i - S_{ki}} I_{\{k,k+1,\ldots\}}(z_i) \right\} \end{aligned} \tag{5.14}$$

em que $\hat{\eta}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$ e $\hat{\eta}_{(-i)} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(-i)}$; $\hat{\boldsymbol{\beta}}$ o EMV dado a amostra observada, $\hat{\boldsymbol{\beta}}|\mathscr{D}$, e $\hat{\boldsymbol{\beta}}_{(-i)}$ EMV dado a amostra observada com a i-ésima observação excluída, $\hat{\boldsymbol{\beta}}_{(-i)}|\mathscr{D}_{(-i)}$. $\boldsymbol{J}(\hat{\boldsymbol{\beta}})$ é a matriz de informação de Fisher observada cujos elementos são obtidos pela soma das parcelas em (5.3). A distância de Cook generalizada mede o impacto da i-ésima observação no estimador de máxima verossimilhança de $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, e a distância da verossimilhança a diferença entre $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\beta}}_{(-i)}$ pelo impacto nas estimativas da função de máxima verossimilhança.

Assintoticamente, as distâncias de Cook generalizada e de verossimilhança apresentam aproximadamente distribuição qui-quadrado com p graus de liberdade (χ^2_p), em que p é a quantidade de parâmetros do modelo. Ou seja, a i-ésima observação é considerada como influente se $C_{(i)}$ ou LD_i for maior que um valor crítico d_α, sendo α tal que $P(C_{(i)}$ ou $LD_i > d_\alpha) = \alpha$. Para este nosso modelo, $p = R + 1$. Por gráficos dos componentes das distâncias em relação aos índices observados é possível visualizar os valores das distâncias identificando os possíveis pontos influentes nas estimativas.

Uma análise da influência local verifica o impacto de conjuntos de pontos nas estimativas dos parâmetros. Para isto, Cook (1986) propõe pequenas perturbações no modelo, conforme visto na Seção 3.5.2 do Capítulo 3. O impacto é avaliado perturbando-se elementos da função de verossimilhança, covariáveis ou a variável resposta e com isto verifica-se a distância entre a função de verossimilhança em relação ao EMV de $\boldsymbol{\beta}$ do modelo sem perturbação, $\hat{\boldsymbol{\beta}}$, e a função de verossimilhança em relação ao EMV de $\boldsymbol{\beta}$ do modelo perturbado, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$:

$$\begin{aligned} LD_{\boldsymbol{\gamma}} &= 2\{l(\hat{\boldsymbol{\beta}}|\mathscr{D}) - l(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}|\mathscr{D})\} \\ &= 2\sum_{i=1}^{m} \log\left\{\left[\frac{1-g^{-1}(\hat{\eta}_i)}{1-g^{-1}(\hat{\eta}_{\boldsymbol{\gamma}_i})}\right]^{S_{ki}}\left[\frac{g^{-1}(\hat{\eta}_i)}{g^{-1}(\hat{\eta}_{\boldsymbol{\gamma}_i})}\right]^{z_i-S_{ki}} I_{\{k,k+1,\ldots\}}(z_i)\right\}, \end{aligned} \tag{5.15}$$

em que $\hat{\eta}_{\gamma i} = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{\gamma}$, $\boldsymbol{\gamma}$ é um vetor de perturbação m-dimensional cujos elementos γ_i são tipos de perturbação, definidos tal que $0 \leq \gamma_i \leq 1$, $i = 1,\ldots,m$.

Para obtermos $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ precisamos da função de log-verossimilhança do modelo perturbado, $l_{\boldsymbol{\gamma}}(\boldsymbol{\beta}|\mathscr{D},\boldsymbol{\gamma})$. Na perturbação de ponderação de casos, perturba-se elementos da função de log-verossimilhança do modelo não perturbado:

$$l_{\boldsymbol{\gamma}}(\boldsymbol{\beta}|\mathscr{D},\boldsymbol{\gamma}) = \sum_{i=1}^{m}\gamma_i l(\boldsymbol{\beta}|\mathscr{D}_i), \tag{5.16}$$

em que $\mathscr{D}_i = (z_i,\boldsymbol{x}_i)$ é a i-ésima observação da amostra $\mathscr{D}$, $l(\boldsymbol{\beta}|\mathscr{D}_i)$ o i-ésimo elemento da função de log-verossimilhança em (5.1). Quando $\gamma_i = 1$, para todo $i = 1,\ldots,m$, não há perturbação no modelo, denominaremos esse vetor $\boldsymbol{\gamma} = \mathbf{1}$ de $\boldsymbol{\gamma}_0$. Quando $\gamma_i = 0$, tem-se que a i-ésima observação foi excluída.

A perturbação também pode ocorrer na covariável ou na varivável resposta, nestes casos, a covariável r perturbada $\boldsymbol{x}_{r\boldsymbol{\gamma}} = \boldsymbol{x}_r + \boldsymbol{\gamma}$, $r = 1,\ldots,R$, a variável resposta perturbada $\boldsymbol{z}_{\boldsymbol{\gamma}} = \boldsymbol{z} + \boldsymbol{\gamma}$, $\boldsymbol{\gamma}_0 = \mathbf{0}$ e a função de log-verossimilhança do modelo perturbado é obtida por:

$$l_{\boldsymbol{\gamma}}(\boldsymbol{\beta}|\mathscr{D},\boldsymbol{\gamma}) = \sum_{i=1}^{m} l(\boldsymbol{\beta}|\mathscr{D}_{\boldsymbol{\gamma}_i}), \tag{5.17}$$

com $\mathscr{D}_{\boldsymbol{\gamma}} = (\boldsymbol{z},\boldsymbol{x}_{\boldsymbol{\gamma}})$ na perturbação de covariável ou $\mathscr{D}_{\boldsymbol{\gamma}} = (\boldsymbol{z}_{\boldsymbol{\gamma}},\boldsymbol{x})$ na perturbação da variável resposta ou ainda $\mathscr{D}_{\boldsymbol{\gamma}} = (\boldsymbol{z}_{\boldsymbol{\gamma}},\boldsymbol{x}_{\boldsymbol{\gamma}})$ e $l(\boldsymbol{\beta}|\mathscr{D}_{\boldsymbol{\gamma}_i})$ a i-ésima parcela da função de log-verossimilhança em (5.1) considerando $\mathscr{D} = \mathscr{D}_{\boldsymbol{\gamma}}$

Desta forma, Cook (1986) sugere avaliar a influência estudando a maior variação da $LD_{\boldsymbol{\gamma}}$ em torno de $\boldsymbol{\gamma}_0$, o que corresponde a maximizar $|\boldsymbol{d}^T\boldsymbol{A}\boldsymbol{d}|$ obtendo o maior valor absoluto do autovalor da matrix $\boldsymbol{A}$ sendo $\boldsymbol{d}_{max}$ o autovetor correspondente que contém os valores da influência local das observações, em que $t\boldsymbol{d} = \boldsymbol{\gamma}_0 - \boldsymbol{\gamma}$, $t \in \mathbb{R}$, $\boldsymbol{d}^T\boldsymbol{d} = 1$ e $\boldsymbol{A} = \boldsymbol{\Gamma}^T J(\hat{\beta})^{-1}\boldsymbol{\Gamma}$, sendo

$\boldsymbol{\Gamma} = \partial^2 l(\boldsymbol{\beta}|\mathscr{D}_{\boldsymbol{\gamma}})/\partial\boldsymbol{\beta}^T\partial\boldsymbol{\gamma}$ e $\boldsymbol{J}(\hat{\boldsymbol{\beta}})$ a matriz de informação observada de Fisher cujos elementos são obtidos pela soma de 5.3, ambas avaliadas em $\hat{\boldsymbol{\beta}}$ e $\boldsymbol{\gamma}_0$. O gráfico de $|\boldsymbol{d}_{max}|$ em relação à ordem das observações pode revelar os pontos com maior influência na vizinhança de $\boldsymbol{LD}_{\boldsymbol{\gamma}}$. Temos que $|\boldsymbol{d}_{max}| \in [0,1]$ e o gráfico de seus valores em relação à ordem das observações pode revelar os pontos com maior influência na vizinhança de $\boldsymbol{LD}_{\boldsymbol{\gamma}}$.

5.4 Estudos de simulação

Para exemplificar a metodologia clássica apresentada, simulamos dados com distribuição geométrica de ordem $k = 3$. Os algoritmos e funções para geração das variáveis aleatórias, ajuste dos modelos de regressão e obtenção das medidas para análise de diagnóstico foram desenvolvidos no *software* estatístico R e estão apresentados no Apêndice B. As estimativas de máxima verossimilhança para os parâmetros dos modelos foram obtidas e verificadas as propriedades frequentistas destes estimadores. As medidas para análise de diagnóstico foram construídas da simulação de um conjunto de dados perturbados com o objetivo de avaliar a capacidade de detecção de observações discrepantes e/ou influentes.

Os conjuntos de dados com distribuição geométrica de ordem k foram simulados considerando $R = 2$, $\beta_0 = 2$, $\beta_1 = -1$ e $\beta_2 = -0,5$. Duas covariáveis foram consideradas, X_{i1} e X_{i2}, com distribuições Bernoulli $(0,5)$ e uniforme no intervalo $(0,1)$, $U(0,1)$, respectivamente; $i = 1,\ldots,m$. O procedimento numérico iterativo de Newton-Raphson foi aplicado para a estimação dos parâmetros por maximização da função de verossimilhança com o auxílio do pacote e função *maxLik* do *software* R. Podemos subdividir este estudo de simulação em três etapas de acordo com as subseções seguintes.

Na primeira etapa (Etapa I), conjuntos de dados foram gerados para verificar as funções de ligação (Tabela 1) que melhor relacionaram as estimativas das probabilidades de sucesso com as covariáveis de acordo com as medidas para critério de seleção de modelos (Subseção 5.2.1). Os conjuntos de dados também foram gerados com as funções de ligação apresentadas na Tabela 1.

Na segunda etapa (Etapa II), conjuntos de dados com as funções de ligação adequadas foram simulados e então verificadas as propriedades frequentistas dos EMVs, como distribuição normal e intervalos de confiança assintóticos (5.4).

Na terceira etapa (Etapa III), um conjunto com dados perturbados foi simulado para verificação dos resíduos quantílicos aleatorizados, (1.6) e (5.12), e das medidas de influência global, (5.13) e (5.14), e local ($|d_{max}|$).

5.4.1 Seleção de modelos

Nesta etapa, quatro conjuntos de dados foram gerados utilizando as quatro funções de ligação $g(.)$: logito, complemento log-log, log-log e probito, conforme descritas na Tabela 1, aqui as denominamos ligações de referência. Cada um desses quatro conjuntos de dados corresponde a 100 amostras de tamanho $m = 100$ de Y_{ki} com distribuição $kGeo(\pi_i)$, simulados com $\pi_i = g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})$, $i = 1, \ldots, 100$, e $\beta_0 = 2$, $\beta_1 = -1$ e $\beta_2 = -0,5$. Após a geração destas variáveis aleatórias, modelos de regressão foram ajustados com estas quatro funções de ligação. Assim, os ajustes dos modelos foram comparados entre si de acordo com a função de ligação. Esta comparação é feita por seleção do melhor modelo dois a dois, entre o modelo com a ligação de referência, isto é, com a mesma função de ligação da geração das variáveis e cada um dos demais modelos ajustados com as outras três funções de ligação. Os critérios de seleção utilizados foram *AIC* e *BIC* e, então, feita a contagem do números de vezes que o modelo com a ligação referência foi pior que os demais, pior significa apresentar maiores valores de *AIC* e *BIC*. A Tabela 15 apresenta estes resultados e, como exemplo, tomamos a primeira linha da tabela, a qual mostra que, de 100 modelos ajustados, o modelo logito de referência apresentou 65 dos ajustes piores do que o ajuste do modelo complemento log-log; nenhum ajuste pior que do modelo log-log e 74 piores que do modelo probito. Nota-se que o modelo complementar log-log de referência não apresentou nenhum ajuste pior que dos demais modelos, indicando ser a melhor função de ligação para modelos geométricos de ordem 3.

Tabela 15 – Percentual de rejeição do modelo ajustado com a função de ligação referência em relação a cada modelo ajustado com as demais funções de ligação.

	Ligação utilizada no ajuste			
Ligação de referência	Logito	C.log-log	Log-log	Probito
Logito	–	65	0	74
C.log-log	0	–	0	0
Log-log	74	48	–	67
Probito	14	68	0	–

5.4.2 Propriedades assintóticas do EMV

Para avaliar as propriedades assintóticas dos EMV, quatro conjunto de dados referentes a cinco tamanhos amostrais, $m = 30$, 50, 100, 200 e 400, foram gerados e modelos foram ajustados para avaliação das estimativas de máxima verossimilhança dos parâmetros em $n = 1000$ réplicas simuladas em cada conjunto. Neste estudo, foi considerado o modelo de regressão geométrico de ordem $k = 3$ (MRG3) com função de ligação complemento log-log, a qual apresentou melhor desempenho de acordo com os critérios de seleção de modelos observados na Etapa I. O MRG3 é então obtido por:

$$\log[-\log(1-\pi_i)] = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i},\ i = 1, \ldots, m,$$

em que $\beta_0 = 2$, $\beta_1 = -1$ e $\beta_2 = -0,5$; x_{1i} gerada de uma distribuição Bernoulli com probabilidade de sucesso $0,5$ e x_{2i} gerada de uma distribuição uniforme no intervalo $(0,1)$; para todo i.

Para verificar o comportamento assintótico dos EMVs, medidas estimadas de média, DP, viés e REQM foram calculadas, neste caso, estima-se o viés por $\hat{\theta}_p - \theta_p$ e REQM $= \sqrt{\frac{\sum_{i=1}^{n}(\theta_{pi}-\hat{\theta}_p)^2}{n}}$, em que $\hat{\theta}_p = \sum_{i=1}^{n} \frac{\theta_{pi}}{n}$, θ_p o p-ésimo componente do vetor de parâmetros, em nosso estudo, $p = 1,2,3$, e $n = 1000$ o número de réplicas. Nota-se em geral que, conforme o tamanho da amostra aumenta, os vieses e REQMs diminuem, sinalizando a propriedade de consistência dos EMVs. Assim como a média do erro padrão (EP) fica próximo ao desvio padrão à medida que o tamanho de amostra aumenta. Também ocorre o aumento das probabilidades de cobertura (PC) com o aumento dos tamanhos de amostra, as quais se aproximam de 95%. As PC são as proporções dos intervalos de confiança assintóticos de nível $\alpha = 5\%$ ($IC(95\%)$) (5.4) que contêm o verdadeiro valor do parâmetro. Indicativos de comportamento assintótico normal dos EMVs também são observados com a redução da assimetria dos histogramas e aproximação dos quantis aos quantis da distribuição normal padrão no QQ-plot. Estes resultados podem ser vistos na Tabela 16 e Figura 12, os quais mostram válidas as propriedades assintóticas, como consistência, eficiência e normalidade assintótica dos estimadores para amostras de tamanho moderado (acima de 100).

Tabela 16 – Estimativas dos coeficientes de regressão para diferentes tamanhos de amostra (m).

m	Parâmetro (valor real)	Média	Média EP	DP	VIÉS	REQM	PC
30	β_0 (2,0)	2,059	0,581	0,591	0,059	0,593	0,961
	β_1 (−1,0)	-0,990	0,518	0,552	0,010	0,552	0,956
	β_2 (−0,5)	-0,428	1,135	1,125	0,072	1,126	0,965
50	β_0 (2,0)	2,031	0,445	0,450	0,031	0,451	0,956
	β_1 (−1,0)	-0,991	0,366	0,373	0,009	0,372	0,959
	β_2 (−0,5)	-0,464	0,770	0,749	0,036	0,749	0,965
100	β_0 (2,0)	2,016	0,316	0,311	0,016	0,311	0,956
	β_1 (−1,0)	-0,993	0,249	0,244	0,007	0,244	0,958
	β_2 (−0,5)	-0,502	0,462	0,451	-0,002	0,451	0,958
200	β_0 (2,0)	2,017	0,211	0,213	0,017	0,214	0,952
	β_1 (−1,0)	-1,001	0,173	0,169	-0,001	0,169	0,959
	β_2 (−0,5)	-0,513	0,294	0,293	-0,013	0,294	0,954
400	β_0 (2,0)	2,008	0,151	0,153	0,008	0,153	0,943
	β_1 (−1,0)	-1,005	0,124	0,121	-0,005	0,121	0,958
	β_2 (−0,5)	-0,502	0,200	0,199	-0,002	0,199	0,948

5.4.3 Análise de diagnóstico

Nesta etapa uma amostra de tamanho 100 foi gerada de um modelo de regressão geométrico de ordem 3 também de parâmetros $\beta_0 = 2$, $\beta_1 = -1$, $\beta_2 = -0,5$ e função de ligação complemento log-log. O modelo de regressão foi ajustado e as estimativas de máxima verossimilhança dos coeficientes obtidas. Esta amostra foi perturbada com alterações em duas observações. A primeira alteração foi feita na observação 01 da variável resposta, em que $y_3[01]$ recebeu seu valor $y_3[01] + 4\max_i(y_3[i])$. E uma segunda perturbação foi feita na observação 60 da covariável

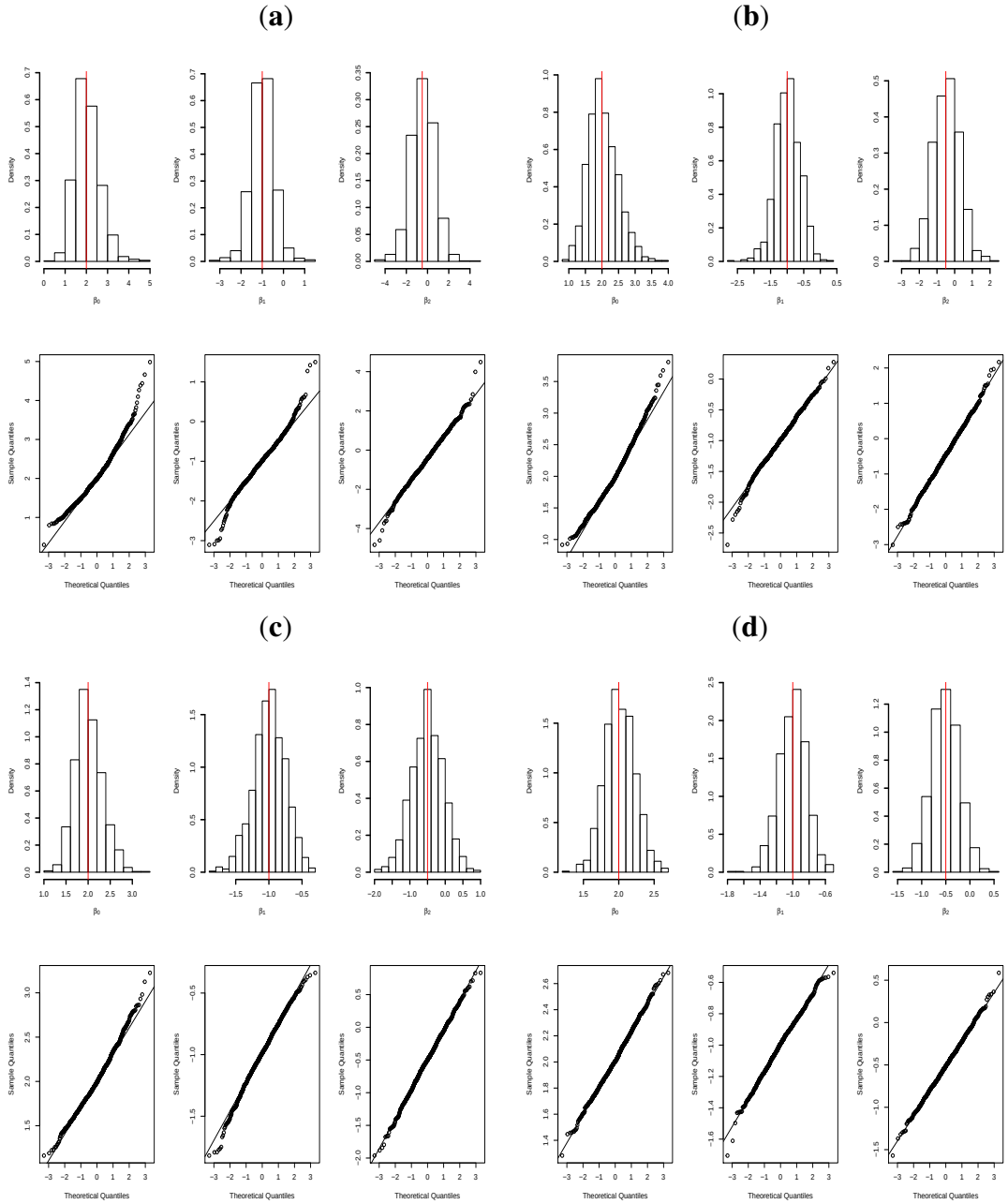

Figura 12 – Histogramas e respectivos gráficos de quantis das estimativas de máxima verossimilhança dos parâmetros do MRG3 para: (a) $m = 30$, (b) $m = 50$, (c) $m = 100$ e (d) $m = 200$.

x_2 da mesma forma, $x_2[60] = x_2[60] + 4\max_i x_2[i]$, $i = 1, \ldots, 100$. As estimativas de máxima verossimilhança dos parâmetros dos modelos ajustados com os dados sem nenhuma perturbação e com as duas perturbações estão na Tabela 17. Nota-se que os intervalos de confiança com dados sem perturbação contêm os verdadeiros valores dos parâmetros do modelo, enquanto que, para os dados perturbados, somente o intervalo de confiança para β_1 contêm o verdadeiro valor. Pelos critérios AIC e BIC, tem-se o melhor ajuste do modelo com os dados sem perturbação. As medidas de diagnóstico podem ser visualizadas na Figura 13.

As medidas de influência global evidenciam as duas observações, 01 e 60, como influ-

Tabela 17 – Estimativas dos coeficientes de regressão e respectivos IC 95% (limites) para os dados sem perturbação e com perturbação.

	Amostra sem perturbação			Amostra com perturbação		
Parâmetros	EMV	LI	LS	EMV	LI	LS
β_0	1,766	1,170	2,362	1,480	1,131	1,829
β_1	-1,218	-1,702	-0,733	-0,780	-1,212	-0,347
β_2	0,154	-0,756	1,065	-0,124	-0,417	0,170
AIC	359,501			401,615		
BIC	367,317			409,430		

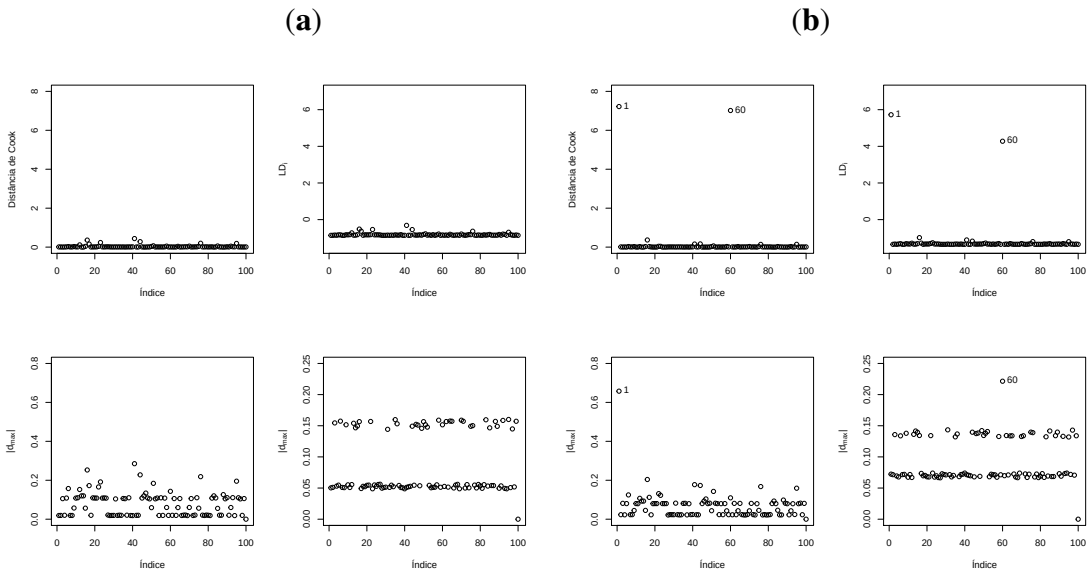

Figura 13 – Medidas de influência global: distância de Cook e distância de verossimilhança (LD). Local: $|d_{max}|$ sob perturbação de ponderação de casos (gráficos à esquerda) e $|d_{max}|$ sob perturbação da variável resposta ou covariável (gráficos à direita). (a) dados sem perturbação e (b) dados perturbados.

entes. Note que, para $p = 3$, o valor crítico $\chi^2_3(0,05) = 7,8$ e, para este tamanho de distância, somente a distância de Cook sugere as observações 01 e 60 como influentes. Portanto, mesmo quando o valor crítico é maior que as distâncias de influência obtidas, a vizualização gráfica pode auxiliar na decisão sobre a influência de observações. As medidas de influência local detectam a observação 01 (alteração na variável resposta) pelo esquema de perturbação de casos e a observação 60 (alteração na covariável) é detectada no esquema de perturbação da variável resposta ou covariável. Nos casos de uma maior variabilidade nos valores das distâncias, um afastamento moderado ou abaixo do valor crítico não deve ser considerado como indicativo de influência. Porém, em situações em que a maior parte das observações apresenta distâncias muito próximas de zero, um afastamento moderado pode ser estudado como uma possível observação influente, assim como ocorreu neste nosso exemplo estudado.

5.5 Aplicação: *Dados de inadimplência*

Segundo Diniz e Louzada (2013), a partir da publicação do primeiro volume da revista Econometrica, em 1933, se intensificou o desenvolvimento de métodos estatísticos para, dentre outros objetivos, dar suporte à concessão de crédito. A concessão de crédito ganhou força na rentabilidade das empresas do setor financeiro, se tornando uma das principais fontes de receita. Por isso, rapidamente este setor percebeu a necessidade de se aumentar o volume de recursos concedidos sem perder a agilidade e a qualidade dos empréstimos e, nesse ponto, a contribuição da modelagem estatística foi e continua sendo essencial.

Os conceitos e idéias da análise de escore de crédito surgiram com Durand *et al.* (1941), que utilizou a análise discriminante desenvolvida por Fisher (1936) para discriminar bons e maus clientes tomadores de créditos. Markowitz (1952) foi um dos pioneiros na criação de um modelo estatístico para o uso financeiro, o qual foi utilizado para medir o efeito da diversificação no risco total de uma carteira de ativos. Black e Scholes (1973) desenvolveram um modelo clássico para a precificação de uma opção, uma das mais importantes fórmulas usadas no mercado financeiro. Uma revisão dos métodos de classificação estatística de escore de crédito ao consumidor pode ser vista em Hand e Henley (1997).

Diferentes tipos de modelos ou procedimentos são então utilizados no problema de crédito, com o intuito de alcançar melhorias na redução do risco e/ou no aumento da rentabilidade, entre eles pode-se citar redes neurais (PERRODO, 1993), árvores de decisão (BREIMAN *et al.*, 1984), regressão logística (BENSIC; SARLIJA; ZEKIC-SUSAC, 2005), lógica fuzzy (WANG; WANG; LAI, 2005), algoritmos genéticos (ABDOU, 2009), redes Bayesianas (BAESENS *et al.*, 2002), mineração de dados (HUANG; CHEN; WANG, 2007) e análise de sobrevivência (COX; OAKES, 1984; TONG; MUES; THOMAS, 2012).

Nesta Seção apresentamos a aplicação do modelo proposto na Seção 5.1 ao conjunto de dados de 380 clientes inadimplentes de um banco brasileiro, tomadores de empréstimo na modalidade crédito direto ao consumidor (CDC), sob tempo de contrato de 36 meses. O tempo observado (em meses) refere-se ao tempo até o cliente ser classificado como inadimplente, o que ocorre quando o pagamento de três parcelas consecutivas do empréstimo não são realizadas. Portanto, o modelo geométrico de ordem $k = 3$ contribui para a análise de crédito, pois é apropriado para o tipo de resposta deste problema. O evento de interesse (sucesso) é o atraso do pagamento mensal da parcela de empréstimo, logo a probabilidade π_i de sucesso de cada cliente i, $i = 1, \ldots, 380$, é a probabilidade de atraso do pagamento da parcela pelo cliente. O objetivo principal desta análise é poder identificar fatores que influenciam significativamente nesta probabilidade de atraso. Para isto, as seguintes variáveis foram observadas para cada cliente da amostra:

- y_i: meses até inadimplência (min=3, max=30, mediana =6, média=8,4, desvio padrão =5,7);

- x_{1i}: idade em anos (min=18, max=93, mediana =37,4, média=40,7, desvio padrão=14,9);
- x_{2i}: sexo (0= feminino (37%), 1=masculino(63%);
- x_{3i}: renda (0= renda$<$ 3.000 (47%), 1= renda $\geq$ 3.000 (53%)).

Inicialmente, verifica-se na Figura 14 a distribuição de frequências do tempo até inadimplência (meses) e observa-se a relação com as covariáveis na Figura 15.

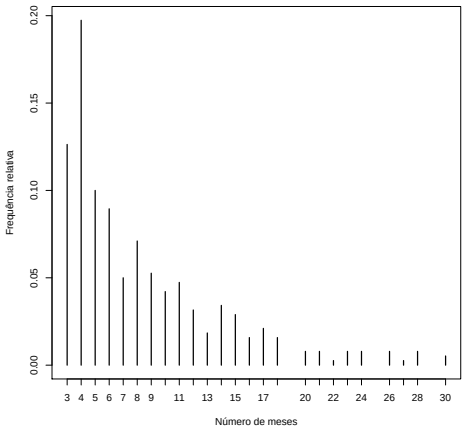

Figura 14 – Frequência relativa do número de meses até a inadimplência.

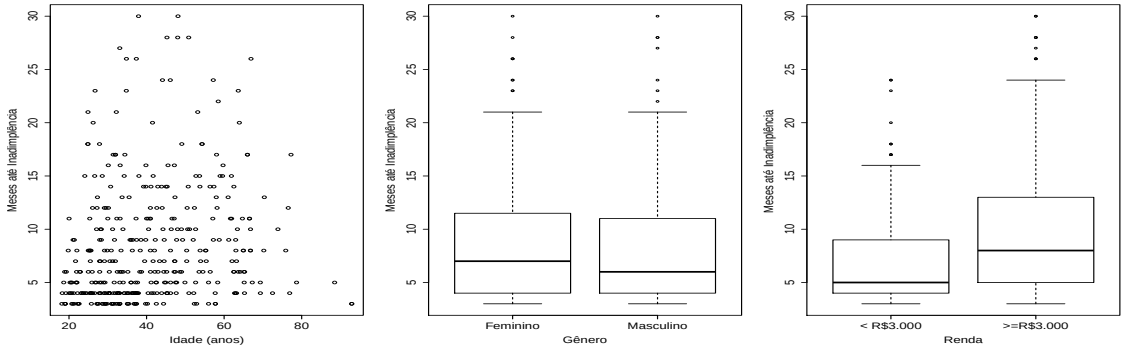

Figura 15 – Meses até a inadimplência em relação às covariáveis.

Os dados foram ajustados ao modelo de regressão geométrico de ordem $k = 3$, com todas as covariáveis, assumindo que a probabilidade de atraso no pagamento mensal (π_i) do modelo se relaciona com as covariáveis por meio da função de ligação $g(.)$. Neste caso, $\boldsymbol{x}_i = (1, x_{1i}, x_{2i}, x_{3i})^T$ e $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$, e então,

$$g(\pi_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i},\ i = 1, \ldots, 380,$$

consideraremos as funções de ligação estudadas.

Tabela 18 – Critérios AIC e BIC.

Ligação	$\ell(\hat{\boldsymbol{\beta}})$	AIC	BIC
Logito	-1065,803	2139,607	2155,367
Complemento log-log	**-1065,549**	**2139,099**	**2154,860**
Log-log	-1065,985	2139,970	2155,731
Probito	-1065,762	2139,523	2155,284

Na Tabela 18 estão apresentados os critérios AIC e BIC para as estimativas de máxima verossimilhança e todos eles sugerem que o MRG3 com função de ligação complemento log-log foi o que melhor se ajustou aos dados. A Tabela 19 apresenta as estimativas de máxima verossimilhança, os erros padrão, estatística de teste e valor de p para os parâmetros do modelo. Os resultados mostram que gênero (β_2) não teve efeito significativo na probabilidade de atraso no pagamento e, o sinal negativo dos coeficientes de regressão associados à idade (β_1) e renda (β_3), indicam que esta probabilidade diminui com o aumento da idade e maior nível de renda.

Tabela 19 – Estimativas dos coeficientes do modelo.

Parâmetro	EMV	Erro Padrão	Estatística	p-valor
β_0	0,196	0,092	2,125	0,034
β_1	-0,047	0,022	-2,156	0,031
β_2	0,071	0,059	1,198	0,231
β_3	-0,162	0,061	-2,654	0,008

Para verificar o ajuste do modelo e a existência de possíveis valores discrepantes e pontos influentes, as medidas para análise de diagnóstico descritas na Seção 5.3 foram obtidas e os gráficos dos resíduos quantílicos aleatorizados estão na Figura 16. A Figura 17 mostra as medidas de influência global (distâncias de Cook e de verossimilhança) e de influência local $|d_{max}|$ sob perturbação de ponderação de casos e da variável resposta e covariável. Nenhuma observação discrepante com valor absoluto de resíduo acima de 3 foi observada. As medidas de influência global mostram que as observações 288 e 362 se encontram um pouco afastadas das demais. Porém estes afastamentos são muito pequenos, menores que $0,1$, bem abaixo do valor crítico $9,5$ ($\chi^2_4(0,05)$). Para avaliação da influência local sob perturbação de ponderação de casos e da variável resposta ou covariável, os gráficos de $|d_{max}|$ mostram as observações 353, 253 e 308 com valores acima das demais, porém abaixo de 0,1. E também não são casos isolados, existe uma variabilidade de outras observações com distâncias menores, porém bem próximas. Assim, conclui-se por não haver medidas que possam influenciar nos resultados da análise do ajuste do modelo.

Observa-se na Tabela 19 que somente a variável gênero não foi significativa na probabilidade de atraso da parcela ou no tempo até inadimplência. Com isto, também pode-se avaliar se esta covariável x_2 (gênero) contribui de forma significativa em explicar a variabilidade dos dados pela estatística de teste da razão de verossimilhanças (RV) (5.6) e, assim, testar a hipótese nula de

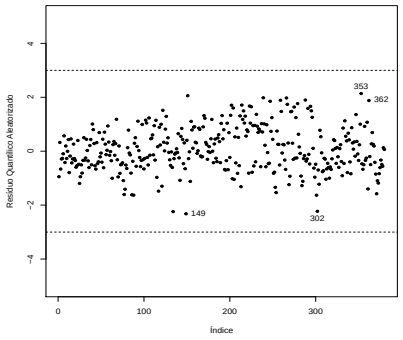

Figura 16 – Resíduos quantílicos aleatorizados: Meses até inadimplência

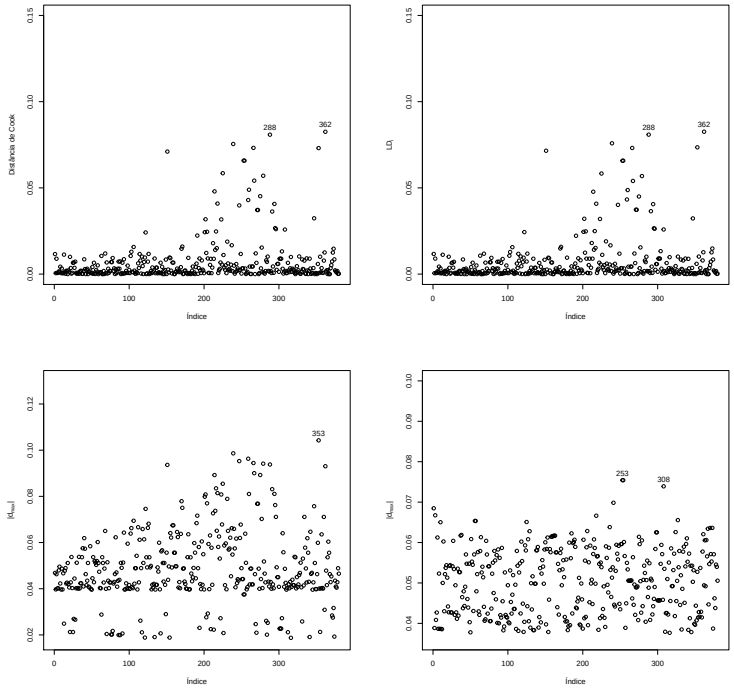

Figura 17 – Medidas de influência global: distância de Cook e distância de verossimilhança (LD); e medidas de influência local: $|d_{max}|$ sob perturbação de ponderação de casos (à esquerda) e de variável resposta (à direita).

interesse $H_0 : \beta_2 = 0$. O valor da estatística de teste foi $1,44$ com valor de p= 0,231, confirmando não haver contribuição significativa do gênero na probabilidade de atraso de pagamento pelo cliente. As estimativas de máxima verossimilhança do modelo reduzido (sem a covariável x_2) e respectivos desvios padrão, foram:

$$\hat{\beta}_0 = 0,226(0,088),\ \hat{\beta}_1 = -0,044(0,022) \text{ e } \hat{\beta}_3 = -0,163(0,061).$$

E o modelo final estimado fica dado por:

$$\log\left[-\log(1-\hat{\pi}_i)\right] = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_3 x_{3i},\ i = 1,\ldots,380. \tag{5.18}$$

De (5.18) obtém-se as estimativas de máxima verossimilhança das probabilidades de

atraso no pagamento da parcela para cada cliente, as quais podem ser visualizadas na Figura 18 pelos *boxplots* para cada categoria de renda e por dispersão entre as idades estratificadas por renda. A figura mostra claramente a relação negativa entre a probabilidade de atraso e idade como também as diferenças destas probabilidades entre as categorias de renda. As estimativas do tempo médio até inadimplência (meses) também estão apresentadas nesta figura pelos mesmos gráficos; estes tempos foram estimados por (2.8): $\widehat{E(Y_3)} = (1-\hat{\pi}^3)/(1-\hat{\pi})\hat{\pi}^3$. Aqui, nota-se o aumento do número de meses até inadimplência com maior renda e maiores idades.

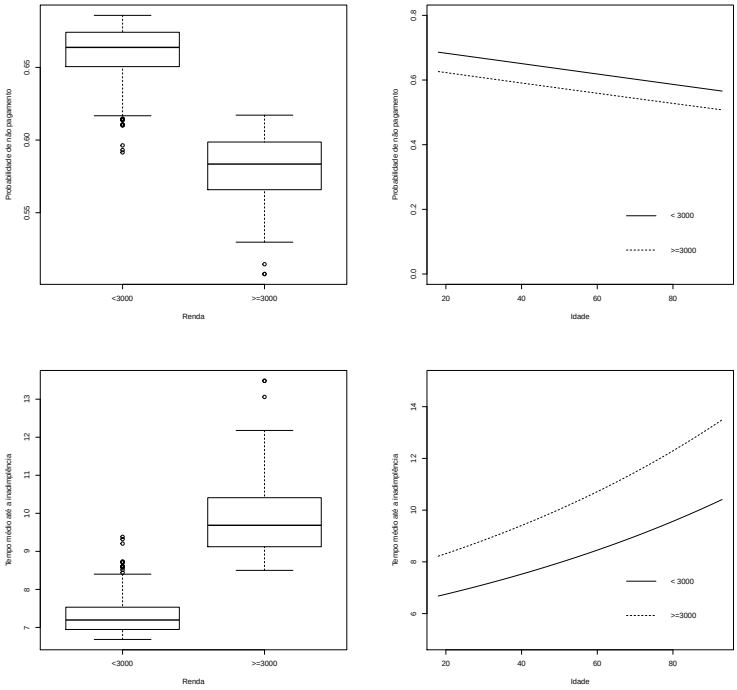

Figura 18 – Estimativas das probabilidades de atraso no pagamento e dos tempos médios até inadimplência (meses) por categorias de renda e por idade (anos) estratificada por renda.

Concluindo o capítulo, embora as condições de regularidade para as propriedades dos estimadores de máxima verossimilhança não foram verificadas analiticamente, um estudo de simulação foi realizado com esse propósito, no qual se observa a convergência assintótica para amostras de tamanho moderado. A aplicação do modelo em conjunto de dados reais de clientes inadimplentes em operação de crédito, em que temos como resposta a quantidade de meses (ou parcelas) até a inadimplência, mostrou que a probabilidade de atraso da prestação diminui com o aumento da idade e em clientes com renda superior a R$ 3.000,00 reais e, por consequência, o tempo médio até a inadimplência é maior. Em outras palavras, a inadimplência ocorre mais rápido em clientes mais jovens e com renda abaixo de R$ 3,000,00 reais. A análise de diagnóstico sugeriu ajuste adequado do modelo e também não detectou observações discrepantes ou influentes nos resultados.

CAPÍTULO

6

MODELO DE REGRESSÃO GEOMÉTRICO DE ORDEM *K* CORRELACIONADO BAYESIANO

Neste capítulo estudamos o modelo de regressão geométrico de ordem k correlacionado numa abordagem Bayesiana. As propriedades frequentistas dos estimadores Bayesianos foram avaliadas em maiores detalhes para o caso particular em que $\rho = 0$ (MRGk), ilustradas por estudos de simulação. O modelo geométrico de ordem k correlacionado (MRGCk) proposto nesta tese foi estudado pela aplicação do modelo de regressão (MRGCk) ao conjunto de dados reais de clientes inadimplentes já apresentado na Seção 5.5 do Capítulo 5, neste caso $k = 3$. Com isto, uma verificação das estimativas dos coeficientes entre os dois modelos, MRGk e MRGCk, é apresentada no ajuste dos dados reais. O principal objetivo da análise Bayesiana é obter as estimativas dos coeficientes do modelo através de suas distribuições de probabilidade a *posteriori*. Estas distribuições foram então adquiridas por métodos iterativos de simulação estocástica MCMC utilizando algoritmo de Metropolis-Hastings (MH). Nos casos gerais, considerando $\max\{-1, -\frac{1-\pi}{\pi}\} \leq \rho < 1$, em que $\pi \in (0,1)$, o algoritmo de MH se mostrou mais vantajoso computacionalmente via amostrador de Gibbs. As estimativas Bayesianas são obtidas através da minimização da perda esperada pela distribuição a *posteriori* para uma função de perda de interesse. No MRGCk, as covariáveis se relacionam com a probabilidade do sucesso por meio de uma função de ligação adequada. Critérios de seleção Bayesianos foram utilizados para a escolha do melhor modelo de acordo com a função de ligação. Uma análise de diagnóstico avalia o ajuste dos modelos por resíduos quantílicos aleatorizados a *posteriori* e observações influentes por medidas de divergência-ψ.

6.1 Modelos de regressão

Considera-se novamente os modelos probabilísticos geométricos de ordem k correlacionado, $kIGeo(\pi,\rho)$ e o caso particular em que $\rho = 0$, $kGeo(\pi)$, propostos no Capítulo 2 nas Seções 2.3 e 2.2, respectivamente. Novamente, apresentamos as funções de distribuição de probabilidade da variável aleatória Y_k, $k \geq 1$, que representa a quantidade de ensaios em sequências de respostas binárias até a ocorrência de k sucessos consecutivos, com probabilidade de sucesso π, $\pi \in (0,1)$ e coeficiente de correlação ρ, $\max\{-1, -\frac{1-\pi}{\pi}\} \leq \rho < 1$, as funções propostas em (2.12) e (2.7), respectivamente:

$$\begin{aligned} P(Y_k = y+k|\pi,\rho) &= \{(1-\pi)[(1-\rho)\pi][\pi(1-\rho)+\rho]^{(k-1)} + \pi[\pi(1-\rho)+\rho]^k\} I_{\{0\}}(y) + \\ &+ \left\{(1-\pi)\sum_{y_1,\ldots,y_k}\binom{y_1+\ldots+y_k}{y_1,\ldots,y_k}[(1-\pi)(1-\rho)+\rho]^{y_1}[(1-\rho)\pi]^{y_2+\ldots y_k+1}\right. \\ &\quad [(1-\pi)(1-\rho)]^{\sum_{i=2}^k y_i}[\pi(1-\rho)+\rho]^{y_3+2y_4+\ldots+(k-2)y_k+k-1} + \\ &+ \pi\sum_{j=1}^{k}[\pi(1-\rho)+\rho]^{(j-1)}[(1-\pi)(1-\rho)]\sum_{x_1,\ldots,x_k}\binom{x_1+\ldots+x_k}{x_1,\ldots,x_k} \\ &\quad [(1-\pi)(1-\rho)+\rho]^{x_1}[(1-\rho)\pi]^{x_2+\ldots+x_k+1}[(1-\pi)(1-\rho)]^{\sum_{i=2}^k x_i} \\ &\quad \left.[\pi(1-\rho)+\rho]^{x_3+\ldots+(k-2)x_k+k-1}\right\} I_{\{1,2,\ldots\}}(y), \end{aligned} \tag{6.1}$$

e para $\rho = 0$:

$$P(Y_k = y+k|\pi,\rho=0) = \sum_{y_1,\ldots,y_k}\binom{y_1+\ldots+y_k}{y_1,\ldots,y_k}(1-\pi)^{\sum_{i=1}^k y_i}\pi^{y+k-\sum_{i=1}^k y_i} I_{\{0,1,2,\ldots\}}(y), \tag{6.2}$$

sendo o somatório em relação a todos os termos, $y_1,\ldots,y_k$ inteiros não negativos, tais que $y = y_1 + 2y_2 + \ldots + ky_k$ e, em relação a todos os termos $x_1,\ldots,x_k$, também inteiros não negativos, tais que $x_1 + 2x_2 + \ldots + kx_k = y - j$, $y \geq k$.

A esperança e variância da variável aleatória Y_k estão no Capítulo 2, dadas em (2.19) e (2.34) e, para $\rho = 0$, são dadas por (2.8).

Sejam $Y_{k1}, Y_{k2}, \ldots Y_{km}$ variáveis aleatórias independentes em que Y_{ki} segue distribuição geométrica de ordem k correlacionada com probabilidade $kIGeo(\pi_i,\rho)$, $\rho_i = \rho$, para todo $i = 1,\ldots,m$. E, sejam $y_1 + k, y_2 + k, \ldots, y_m + k$ observações de $Y_{k1}, Y_{k2}, \ldots, Y_{km}$, respectivamente. O MRGCk é tal que as probabilidades de sucesso π_i satisfazem a seguinte relação funcional:

$$\eta_i = g(\pi_i) = \boldsymbol{x}_i^T\boldsymbol{\beta},\ i = 1,\ldots,m, \tag{6.3}$$

em que $\boldsymbol{\beta} = (\beta_0,\beta_1,...,\beta_R)^T$ é o vetor de coeficientes de regressão com $\boldsymbol{\beta} \in \mathbb{R}^p$, $p = R+1$, tal que $p < m$; η_i o preditor linear, $\boldsymbol{x}_i = (1, x_{1i},\ldots,x_{Ri})$ o vetor de covariáveis de ordem p, $\boldsymbol{X}$ a matriz de planejamento $\boldsymbol{X} = (\boldsymbol{1},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m)^T$ de ordem $m \times p$ de posto completo p. As funções de ligação discutidas anteriormente são utilizadas nesta análise.

6.2 Estimação do modelo de regressão

Considerando a função de ligação em (6.3), a função de verossimilhança de $\boldsymbol{\theta} = (\boldsymbol{\beta}, \rho)$, dado a amostra observada $\mathscr{D} = (\boldsymbol{y} + k, \boldsymbol{x})$, é obtida por:

$$\mathscr{L}(\boldsymbol{\beta}, \rho | \mathscr{D}) = \prod_{i=1}^{m} P(Y_{ki} = y_i + k | g^{-1}(\eta_i), \rho) = \prod_{i=1}^{m} f(y_i + k | \boldsymbol{\beta}, \rho), \tag{6.4}$$

em que $y_i = y_{1i} + 2y_{2i} + \ldots + ky_{ki}$, η_i o preditor linear e $g(.)$ a função de ligação conforme o modelo de regressão em (6.3). Logo, substituindo a probabilidade (6.1) (ou (6.2) para $\rho = 0$) em (6.4) temos a função de verossimilhança do modelo em função de $\boldsymbol{\beta}$ e ρ, pois no modelo de regressão $\pi_i = g^{-1}(\eta_i)$, ou seja:

$$\begin{aligned} f(y_i + k | \boldsymbol{\beta}, \rho) \quad &= \{(1 - g^{-1}(\eta_i))[(1-\rho)g^{-1}(\eta_i)][g^{-1}(\eta_i)(1-\rho) + \rho]^{(k-1)} + \\ &\quad g^{-1}(\eta_i)[g^{-1}(\eta_i)(1-\rho) + \rho]^k\} I(y_i = 0) + \Bigg\{ (1 - g^{-1}(\eta_i)) \sum_{y_{1i}, \ldots, y_{ki}} C_{ki}(y_i) \\ &\quad [(1 - g^{-1}(\eta_i))(1-\rho) + \rho]^{y_{1i}} [(1-\rho)g^{-1}(\eta_i)]^{y_{2i} + \ldots y_{ki} + 1} \\ &\quad [(1 - g^{-1}(\eta_i))(1-\rho)]^{\sum_{j=2}^{k} y_{ji}} [g^{-1}(\eta_i)(1-\rho) + \rho]^{y_{3i} + 2y_{4i} + \ldots + (k-2)y_{ki} + k - 1} + \\ &\quad g^{-1}(\eta_i) \sum_{l=1}^{k} [g^{-1}(\eta_i)(1-\rho) + \rho]^{(l-1)} g[(1 - g^{-1}(\eta_i))(1-\rho)] \sum_{x_{1i}, \ldots, x_{ki}} C_{ki}(x_i) \\ &\quad [(1 - g^{-1}(\eta_i))(1-\rho) + \rho]^{x_{1i}} [(1-\rho)g^{-1}(\eta_i)]^{x_{2i} + \ldots + x_{ki} + 1} \\ &\quad [(1 - g^{-1}(\eta_i))(1-\rho)]^{\sum_{j=2}^{k} x_{ji}} [g^{-1}(\eta_i)(1-\rho) + \rho]^{x_3 + \ldots + (k-2)x_k + k - 1} \Bigg\} I(y_i > 0), \end{aligned} \tag{6.5}$$

em que $x_i = y_i - l$ e, de modo geral, $C_{ki}(v) = \binom{v_1 + \ldots + v_k}{v_1, \ldots, v_k}$, sendo $v_1, v_2, \ldots, v_k$ tais que $v_1 + 2v_2 + \ldots + kv_k = v$; $i = 1, \ldots, m$.

Assim, considera-se distribuições a *priori* próprias ou integráveis independentes para os parâmetros do MRGCk: $\beta_j \sim N(0, \sigma_j^2)$, $\sigma_j^2 < \infty$ são conhecidos, com $j = 1, \ldots, p$; $\rho | \boldsymbol{\beta} \sim U(\max\{-1, -\tau\}, 1)$, com $\tau = \frac{1 - g^{-1}(\boldsymbol{x}^T \boldsymbol{\beta})}{g^{-1}(\boldsymbol{x}^T \boldsymbol{\beta})}$; em que $N(\mu, \sigma^2)$ é a distribuição normal com média μ e variância σ^2, e $U(\max\{-1, -\tau\}, 1)$ a distribuição uniforme no intervalo $(\max\{-1, -\tau\}, 1)$. Sob essas condições a densidade conjunta de $\boldsymbol{\beta}$ e ρ é dada por:

$$P(\boldsymbol{\theta}) = P(\boldsymbol{\beta}, \rho) = P(\boldsymbol{\beta}) P(\rho | \boldsymbol{\beta}) \propto \frac{\exp\left\{ -\sum_{j=1}^{p} \frac{\beta_j^2}{2\sigma_j^2} \right\}}{1 - \max\{-1, -\tau\}}. \tag{6.6}$$

Combinando a função de verossimilhança (6.4) e a função de densidade a *priori* conjunta (6.6), pelo Teorema de Bayes, tem-se que a função densidade conjunta a *posteriori* de $(\boldsymbol{\beta}, \rho) | \mathscr{D}$:

$$P(\boldsymbol{\beta}, \rho | \mathscr{D}) \propto \mathscr{L}(\boldsymbol{\beta}, \rho | \mathscr{D}) \frac{\exp\left\{ -\sum_{j=1}^{p} \frac{\beta_j^2}{2\sigma_j^2} \right\}}{1 - \max\{-1, -\tau\}}. \tag{6.7}$$

Quando a função de densidade conjunta a *posteriori* não tem forma conhecida, recorre-se a métodos aproximados para a avaliação das funções de densidade marginais a *posteriori*, como os métodos de simulação estocástica MCMC. Para isto, o algoritmo Metropolis-Hastings pode ser utilizado por se tratar de um algoritmo menos restritivo em relação à distribuição a *posteriori*, não havendo necessidade de conhecimento completo das distribuições condicionais. Neste algoritmo, um valor para $\boldsymbol{\theta}$ é gerado de uma distribuição auxiliar e aceito com uma dada probabilidade $\alpha(\boldsymbol{\theta},\boldsymbol{\theta}')$, isto é, se a cadeia está no estado $\boldsymbol{\theta}$ e um valor $\boldsymbol{\theta}'$ é gerado de uma distribuição proposta $q(.|\boldsymbol{\theta})$, ele então é aceito com probabilidade

$$\alpha(\boldsymbol{\theta},\boldsymbol{\theta}') = \min\left(1, \frac{\mathscr{L}(\boldsymbol{\theta}'|\boldsymbol{y};\boldsymbol{x})P(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\mathscr{L}(\boldsymbol{\theta}|\boldsymbol{y};\boldsymbol{x})P(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right). \tag{6.8}$$

Este procedimento garante a convergência da cadeia para a distribuição de equilíbrio. O algoritmo de MH pode ser implementado pelos seguintes passos:

(1) iniciar com contador de iteração $j = 0$ e qualquer valor $\boldsymbol{\theta}^{(0)}$;

(2) gerar um novo valor $\boldsymbol{\theta}'$ da distribuição $q(.|\boldsymbol{\theta})$;

(3) calcular a probabilidade de aceitação α (6.8) e gerar $u \sim U(0,1)$;

(4) Se $u \leq \alpha$, aceitar o novo valor ($\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}'$); caso contrário, rejeitar ($\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)}$);

(5) incrementar o contador de j para $j+1$ e voltar ao passo (2)
(repetir os passos (2), (3) e (4) até conseguir a distribuição estacionária).

No modelo correlacionado, tem-se $\boldsymbol{\theta} = (\boldsymbol{\beta},\rho)$ e $P(\boldsymbol{\theta}) = P(\boldsymbol{\beta})P(\rho|\boldsymbol{\beta})$. Da densidade a *posteriori* conjunta, tem-se as densidades condicionais completas por:

$$P(\boldsymbol{\beta}|\rho,\mathscr{D}) \propto \mathscr{L}(\boldsymbol{\beta},\rho|\mathscr{D})\exp\left\{-\sum_{j=1}^{p}\frac{\beta_j^2}{2\sigma_j^2}\right\} \tag{6.9}$$

$$P(\rho|\boldsymbol{\beta},\mathscr{D}) \propto \mathscr{L}(\boldsymbol{\beta},\rho|\mathscr{D})/(1-\max\{-1,-\tau\}), \tag{6.10}$$

$\mathscr{L}(\boldsymbol{\beta},\rho|\mathscr{D})$ obtida por (6.4); e fica mais vantajoso computacionalmente utilizar o amostrador de Gibbs para uma convergência mais rápida da distribuição estacionária, que é a densidade a *posteriori* de interesse. No amostrador de Gibbs, a cadeia irá sempre se mover pois a probabilidade de aceitação $\alpha(\boldsymbol{\theta},\boldsymbol{\theta}')$ do novo valor, em (6.8), é 1 e as transições de um estado para o outro são feitas de acordo com as distribuições condicionais completas $P(\theta_k|\boldsymbol{\theta}_{(-k)})$, $k = 1,\ldots,(p+1)$, sendo $\boldsymbol{\theta}_{(-k)} = (\theta_1,\ldots,\theta_{k-1},\theta_{k+1},\ldots,\theta_{p+1})^T$, e os itens (2), (3) e (4) do algoritmo são substituidos por um único item (2) que é obter um novo valor de $\boldsymbol{\theta}^{(j)}$ a partir de $\boldsymbol{\theta}^{(j-1)}$ através das

gerações sucessivas de valores, isto é:

$$\begin{aligned}
\beta_1^{(j)} &\sim P\big(\beta_1|\beta_2^{(j-1)},\beta_3^{(j-1)},\ldots,\beta_p^{(j-1)},\rho^{(j-1)}\big)\\
\beta_2^{(j)} &\sim P\big(\beta_2|\beta_1^{(j)},\beta_3^{(j-1)},\ldots,\beta_p^{(j-1)},\rho^{(j-1)}\big)\\
&\vdots\\
\beta_p^{(j)} &\sim P\big(\beta_p|\beta_1^{(j)},\beta_2^{(j)},\ldots,\beta_{p-1}^{(j)},\rho^{(j-1)}\big)\\
\rho^{(j)} &\sim P\big(\rho|\beta_1^{(j)},\beta_2^{(j)},\ldots,\beta_p^{(j)}\big)
\end{aligned}$$

Ou seja, para se obter uma amostra a *posteriori* de $\boldsymbol{\beta}$ e ρ, o amostrador de Gibbs se baseia em sucessivas gerações das distribuições condicionais completas, $P(\boldsymbol{\beta}|\rho,\mathscr{D})$ e $P(\rho|\boldsymbol{\beta},\mathscr{D})$, em cada iteração.

A informação contida na distribuição a *posteriori* também pode ser resumida por um único ponto (estimação pontual) ou por intervalo (intervalo de credibilidade ou intervalo de confiança Bayesiano), conforme vimos na Seção 4.2 do Capítulo 4. O objetivo é obter estimativas que minimizam uma perda esperada. Apresentaremos em nossas análises os estimadores Bayesianos para perda quadrática e absoluta, os quais correspondem à média e mediana a *posteriori*, respectivamante. Também obteremos em nossas análises intervalos HPD para estimação intervalar.

Observando com maior detalhe o modelo particular em que $\rho = 0$ (MRGk), a função de verossimilhança é obtida por:

$$\mathscr{L}(\boldsymbol{\beta}|\mathscr{D}) = \prod_{i=1}^{m}\left\{\sum_{y_{1i},\ldots,y_{ki}}\binom{S_{ki}}{y_{1i}\ldots,y_{ik}}[1-g^{-1}(\eta_i)]^{S_{ki}}g^{-1}(\eta_i)^{z_i-S_{ki}}I_{\{k,k+1,\ldots\}}(z_i)\right\}, \quad (6.11)$$

em que $z_i = y_i + k$; $S_{ki} = y_{1i} + \ldots + y_{ki}$. E assim obtemos a densidade conjunta a *posteriori* de $\boldsymbol{\beta}$:

$$P(\boldsymbol{\beta}|\mathscr{D}) \propto \prod_{i=1}^{m}\sum_{y_{1i},\ldots,y_{ki}}\binom{S_{ki}}{y_{1i},\ldots,y_{ki}}(1-g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}))^{S_{ki}}(g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}))^{z_i-S_{ki}}\exp\left\{-\frac{1}{2}\boldsymbol{\beta}_\sigma^T\boldsymbol{\beta}_\sigma\right\}. \quad (6.12)$$

sendo $\boldsymbol{\beta}_\sigma = (\frac{\beta_1}{\sigma_1},\ldots,\frac{\beta_p}{\sigma_p})^T$.

6.3 Critérios de comparação de modelos

Como mais de um modelo pode ser ajustado para um mesmo conjunto de dados, os quatro critérios Bayesianos de comparação de modelos, apresentados anteriormente no Capítulo 4, serão utizados para a seleção do melhor modelo: o DIC,o EAIC, o EBIC e o LPML. Lembramos que, para obtenção destes critérios, utilizamos as estimativas MCMC da ordenada preditiva condicional (CPO) e medida de desvio (*deviance*) D, as quais, para a i-ésima observação, são estimadas por:

$$\widehat{CPO}_i = \left\{ \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{f(z_i|\boldsymbol{\theta}^{(q)})} \right\}^{-1} \text{ e } D(\boldsymbol{\theta}^{(q)}) = -2 \sum_{i=1}^{m} \log(f(z_i|\boldsymbol{\theta}^{(q)}),$$

em que $z_i = y_i + k$, $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(Q)}$ é uma amostra aleatória de tamanho Q da distribuição a *posteriori* $P(\boldsymbol{\theta}|\mathscr{D})$. Neste caso, $f(.)$ é a função distribuição de probabilidade da variável do modelo a ser comparado, para os MRGCk $f(z_i|\boldsymbol{\theta})$ é a função em (6.5). Com isto temos:

$$LPML = \sum_{i=1}^{m} \log(\widehat{CPO}_i), \ \widehat{EAIC} = \bar{D} + 2p, \ \widehat{EBIC} = \bar{D} + 2p\log(m) \text{ e } \widehat{DIC} = 2\bar{D} - \hat{D},$$

em que p é o número de parâmetros do modelo, $\bar{D} = \frac{1}{Q} \sum_{q=1}^{Q} D(\boldsymbol{\theta}^{(q)})$, $\hat{D} = D(\bar{\boldsymbol{\theta}}^{(q)})$; com $\boldsymbol{\theta}^{(q)} = \left(\beta_0^{(q)}, \ldots, \beta_R^{(q)}, \rho^{(q)}\right)^T$ e $\bar{\boldsymbol{\theta}}^{(q)} = D\left(\frac{1}{Q} \sum_{q=1}^{Q} \beta_0^{(q)}, \ldots, \frac{1}{Q} \sum_{q=1}^{Q} \beta_R^{(q)}, \sum_{q=1}^{Q} \rho^{(q)}\right)^T$. Na comparação entre modelos, maior valor de *LPML* indica o melhor modelo, enquanto para os critérios *EAIC*, *EBIC* e *DIC*, são os menores valores.

6.4 Diagnóstico

As suposições iniciais do modelo, como independência entre as variáveis resposta e sua distribuição de probabilidade, e também a presença de observações discrepantes e/ou influentes que possam causar alguma distorção nos resultados da análise, podem ser avaliadas por resíduos e medidas de divergência. Devido ao comportamento assimétrico das distribuições geométricas, utilizaremos o resíduo quantílico aleatorizado $\boldsymbol{r}_q$, já visto nas análises anteriores. Para cada observação i, $r_{q_i} = \Phi^{-1}(u_i)$, e, portanto, r_{q_i} tem distribuição normal padrão. u_i é uma variável aleatória com distribuição uniforme no intervalo $(a_i, b_i]$, sendo $a_i = F(z_i - 1; \hat{\pi}_i, \hat{\rho})$ e $b_i = F(z_i; \hat{\pi}_i, \hat{\rho})$, e $F(.)$ a função de distribuição acumulada da variável aleatória Y_k. Neste caso, $\hat{\pi}_i = g^{-1}(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$ conforme definido em (6.3); $\hat{\boldsymbol{\beta}}$ e $\hat{\rho}$ são as estimativas Bayesianas de $\boldsymbol{\beta}$ e ρ que minimizam a perda esperada pela distribuição a *posteriori* (6.7) para uma função de perda de interesse.

Observações influentes podem ser detectadas por deleção de pontos e, assim, o impacto com a retirada de uma observação particular, ou de um conjunto de observações, nas estimativas de regressão pode ser avaliado. As medidas de divergência, apresentadas na Subseção 4.4.2 do Capítulo 4, foram utilizadas para detectar observações influentes no ajuste de modelos de regressão Bayesianos baseadas na distribuição a *posteriori*, de modo que esta avaliação é feita para detectar o impacto de cada observação nas distribuições a *posteriori*, conjuntas e marginais, através da retirada de cada observação do conjunto de dados. Novamente, seja $D_{\psi}(P, P_{(-i)})$ a divergência-ψ entre P e $P_{(-i)}$ em que P é a distribuição a *posteriori* de $\boldsymbol{\theta}|\mathscr{D}$ e $P_{(-i)}$ a distribuição a *posteriori* de $\boldsymbol{\theta}|\mathscr{D}^{(-i)}$ (excluindo a i-ésima observação). A divergência-ψ é então obtida por

amostragem a partir da distribuição a *posteriori* de $\boldsymbol{\theta}$ por método MCMC: seja $\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(Q)}$ uma amostra de tamanho Q de $P(\boldsymbol{\theta}|\mathscr{D})$, a estimativa de Monte Carlo é dada por:

$$\widehat{D_{\psi}}(P,P_{(-i)}) = \frac{1}{Q}\sum_{q=1}^{Q}\psi\left(\frac{\widehat{CPO_i}}{f(z_i|\boldsymbol{\theta}^{(q)})}\right). \tag{6.13}$$

$D_{\psi}(P,P_{(-i)})$ mede o tamanho da retirada da i-ésima observação dos dados completos na distribuição a *posteriori* de $\boldsymbol{\theta}$. Dependendo da escolha da função ψ, tem-se diferentes medidas de divergência: Kullback-Leibler (K-L *divergence*), J-*distance*, norma L_1 e qui-quadrado (χ^2-*divergence*). Pode-se considerar o i-ésimo caso influente quando o valor d de $D_{\psi}(-i)$ estiver acima de um ponto de corte: $d_{K-L} > 0,22$, $d_{J-distance} > 0,416$, $d_{L_1}(0,80) > 0,30$, $d_{\chi^2} > 0,562$ conforme mencionado na Subseção 4.4.2.

6.5 Estudo de simulação

O estudo de simulação foi realizado para o modelo particular do MRGCk em que $\rho = 0$, o MRGk. O objetivo é avaliar as propriedades frequentistas dos estimadores Bayesianos baseados na perda quadrática e perda absoluta.

Para este estudo de simulação foi considerado o modelo de regressão geométrico de ordem 3 com função de ligação logito, isto é:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i},\ i = 1,\ldots,m$$

em que $\beta_0 = 2$, $\beta_1 = -1$ e $\beta_2 = -0,5$. A covariável x_1 foi gerada de uma distribuição Bernoulli com probabilidade de sucesso $0,5$ e a covariável x_2 gerada de uma distribuição uniforme no intervalo $(0,1)$. Diferentes tamanhos amostrais (m = 15, 30, 50, 100 e 200) foram considerados no estudo e realizadas 1000 simulações para cada tamanho de amostra, obtendo-se as estatísticas de média, DP, viés e REQM das estimativas Bayesianas e a probabilidade de cobertura do intervalo HPD de 95%, para cada parâmetro do MRG3.

Para cada simulação foi considerada distribuição a *priori* dada em (6.6) com $\beta_j \sim N(0,100)$ ($j = 0,1,2$). Uma cadeia de Markov com 15000 iterações foi gerada utilizando o método de Monte Carlo via algoritmo de MH, conforme descrito na Seção de estimação (6.2) e distribuição a *posteriori* em (6.12). Com a finalidade de diminuir o efeito dos pontos iniciais foram descartadas as primeiras 5000 iterações e, para evitar o problema de autocorrelação das séries, espaçamentos de tamanho 5 foram fixados, conduzindo a uma amostra a *posteriori* de tamanho 2000.

A Tabela 20 mostra as médias de Monte Carlo das estimativas Bayesianas sob perda quadrática e absoluta (média e mediana a *posteriori*) e os respectivos desvios padrão, raiz quadrada do erro quadrático médio, viés e a probabilidade de cobertura do intervalo HPD de credibilidade de 95% para as estimativas de cada parâmetro do MRG3. Observa-se, em geral,

Tabela 20 – Médias, DP, REQM, viés PC do intervalo HPD de 95% de credibilidade para as estimativas de cada parâmetro do modelo.

		Perda quadrática				Perda absoluta				
m	Parâmetro	Média	DP	Viés	REQM	Média	DP	Viés	REQM	PC
15	β_0	2,910	4,854	0,910	4,936	2,794	4,195	0,794	4,267	0,928
	β_1	-0,594	8,435	0,406	8,441	-0,640	7,171	0,360	7,176	0,934
	β_2	-0,923	2,200	-0,423	2,240	-0,872	2,103	-0,372	2,134	0,948
30	β_0	2,230	0,673	0,230	0,711	2,204	0,662	0,204	0,693	0,952
	β_1	-1,085	0,550	-0,085	0,557	-1,076	0,539	-0,076	0,544	0,934
	β_2	-0,553	1,039	-0,053	1,040	-0,546	1,026	-0,046	1,027	0,960
50	β_0	2,079	0,460	0,079	0,466	2,068	0,456	0,068	0,461	0,945
	β_1	-1,026	0,387	-0,026	0,387	-1,017	0,381	-0,017	0,382	0,943
	β_2	-0,471	0,568	0,029	0,569	-0,469	0,566	0,031	0,566	0,947
100	β_0	2,061	0,317	0,061	0,323	2,055	0,316	0,055	0,321	0,952
	β_1	-1,036	0,268	-0,036	0,271	-1,032	0,267	-0,032	0,269	0,950
	β_2	-0,497	0,414	0,003	0,414	-0,496	0,413	0,004	0,412	0,950
200	β_0	2,023	0,211	0,023	0,212	2,021	0,210	0,021	0,211	0,956
	β_1	-1,019	0,182	-0,019	0,183	-1,017	0,181	-0,017	0,182	0,948
	β_2	-0,492	0,267	0,008	0,267	-0,492	0,267	0,008	0,267	0,956

que o viés e a REQM se aproximam de zero e à medida que o tamanho da amostra aumenta, em consequência, as estimativas se aproximam dos verdadeiros valores dos parâmetros. Além disso, as probabilidades de cobertura estão próximas ao nível de credibilidade de 95%. Observa-se que os resultados de simulação clássicos (Capítulo 5) e Bayesianos são similares.

6.6 Aplicação: *Dados de inadimplência*

Nesta seção é feita a aplicação dos modelos propostos ao conjunto de dados de 380 clientes inadimplentes de um banco brasileiro, tomadores de empréstimo na modalidade CDC, sob tempo de contrato de 36 meses; dados estes já apresentados na Seção 5.5 do capítulo anterior. O tempo observado (em meses) refere-se ao tempo até o cliente ser classificado como inadimplente, o que ocorre quando o pagamento de três parcelas consecutivas do empréstimo não são realizadas, O sucesso de interesse é o atraso do pagamento mensal da parcela de empréstimo. Com isto, o modelo geométrico correlacionado de ordem $k = 3$ para modelar o número de meses até a inadimplência, com probabilidade de sucesso sendo a probabilidade de atraso do pagamento da parcela, foi ajustado aos dados.

O objetivo principal desta análise é identificar quais fatores influenciam significativamente na probabilidade de atraso do pagamento pelo cliente.

Os dados foram ajustados por modelos de regressão geométricos de ordem 3, correlacionado e não correlacionado, com todas as covariáveis, assumindo que a probabilidade de atraso na parcela (π_i) se relaciona com as covariáveis por meio da função de ligação g, isto é, para $\boldsymbol{x}_i = (1, x_{1i}, x_{2i}, x_{3i})^T$ e $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_3)^T$ tem-se:

$$g(\pi_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i},\ i = 1, \ldots, 380,$$

considera-se as funções de ligação usuais: logito, complemento log-log, log-log e probito, já mencionadas anteriormente. As distribuições a *priori* consideradas foram: $\beta_j \sim N(0,100)$ $(j=0,1,2,3)$ e $\rho|\boldsymbol{\beta} \sim U\left(\max\{-1,-\tau\},1\right)$, com $\tau = \frac{1-g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta})}{g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta})}$. As amostras a *posteriori* foram obtidas por método de simulação MCMC implementado no *software* R. Cadeias de Markov de 25000 iterações foram geradas via amostrador de Gibbs, conforme descrito na seção de estimação (6.2), e densidades condicionais completas em (6.9) e (6.10). Com a finalidade de diminuir o efeito dos pontos iniciais foram descartadas as primeiras 5000 iterações e, para evitar o problema de autocorrelação da série, foi fixado um espaçamento de tamanho 10, conduzindo a uma amostra a *posteriori* de tamanho 2000. Nas Tabelas 21 e 22 são apresentadas as estimativas dos critérios Bayesianos DIC, EAIC, EBIC e LPML para seleção de modelos, tanto para o MRG3 como para o MRGC3. Todos eles sugerem que o modelo de regressão geométrico de ordem 3 com função de ligação complemento log-log é o que melhor ajusta o conjunto de dados, isto é, $g(\pi_i) = \log\left(-\log(1-\pi_i)\right)$. Os critérios também selecionam o modelo correlacionado como melhor ajuste em relação ao não correlacionado. Assim, considera-se o MRGC3 complemento log-log o mais adequado para avaliação dos dados de clientes inadimplentes.

Tabela 21 – Estimativas de DIC, EAIC, EBIC e LPML para MRG3

Ligação	DIC	EAIC	EBIC	LPML
Logito	2140,334	2144,253	2160,014	-1069,501
Complemento log-log	**2138,977**	**2143,040**	**2158,800**	**-1068,835**
Log-log	2139,847	2143,914	2159,675	-1069,277
Probito	2139,655	2143,591	2159,352	-1069,168

Tabela 22 – Estimativas de DIC, EAIC, EBIC e LPML para MRGC3

Ligação	DIC	EAIC	EBIC	LPML
Logito	2127,552	2132,406	2152,107	-1062,548
Complemento log-log	**2126,936**	**2131,856**	**2151,557**	**-1062,282**
Log-log	2127,406	2132,568	2152,268	-1062,547
Probito	2127,098	2132,180	2151,881	-1062,354

A Tabela 23 mostra as estimativas de Bayes dos parâmetros dos modelos e os intervalos HPD 95% do MRGC3 e também do caso particular em que $\rho = 0$, isto é, do MRG3. Os resultados de ambos os ajustes mostram que idade e renda do cliente tiveram efeito significativo no tempo até a inadimplência, indicando que este tempo diminui com o aumento da idade e da renda. Somente a covariável gênero não foi significativa, pois os intervalos HPD para o coeficiente de regressão associado contém o zero, sugerindo que esta covariável não contribui de forma significativa na probabilidade de atraso da parcela. Para o MRGC3, o coeficiente de correlação ρ também foi significativo e positivo, no modelo não correlacionado o peso deste efeito também foi para a média geral β_0, significante neste modelo, porém perde a significância e podendo ser considerada nula no modelo correlacionado. O coeficiente de correlação ρ, embora significativo, apresentou estimativas menores que $0,3$, o que sugere uma correlação fraca. Esta correlação positiva indica que a probabilidade de atraso estimada é maior que a estimada pelo modelo não

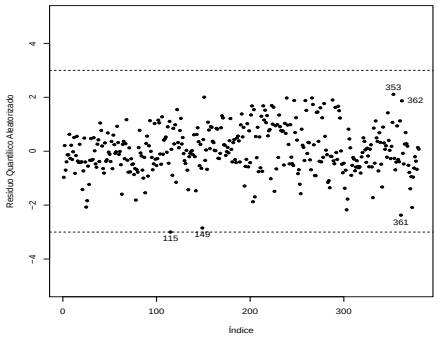

Figura 19 – Resíduos quantílicos aleatorizados a *posteriori*: Meses até inadimplência.

correlacionado. Neste caso, os ajustes de ambos os modelos, correlacionado e não correlacionado, não alterou as inferências sobre os dados. Mas como o MRGC3 foi selecionado por melhores critérios, seguimos nossa análise com o MRGC3 complementar log-log.

Tabela 23 – Média, mediana e desvio padrão a *posteriori* e o intervalo HPD de 95% de credibilidade dos parâmetros dos MRG3 e MRGC3 log-complementares.

	Não Correlacionado					Correlacionado				
				Intervalo HPD (95%)					Intervalo HPD (95%)	
Parâmetro	Média	Mediana	DP	LI	LS	Média	Mediana	DP	LI	LS
β_0	0,194	0,196	0,091	0,026	0,338	0,094	0,096	0,124	-0,161	0,337
β_1	-0,047	-0,047	0,022	-0,088	-0,004	-0,058	-0,057	0,029	-0,115	-0,000
β_2	0,071	0,072	0,058	-0,039	0,181	0,089	0,091	0,075	-0,057	0,233
β_3	-0,164	-0,163	0,060	-0,284	-0,048	-0,206	-0,206	0,080	-0,363	-0,057
ρ	-	-	-	-	-	0,286	0,289	0,062	0,167	0,400

A análise de resíduos para o MRGC3 é visualizada no gráfico dos resíduos quantílicos aleatorizados a *posteriori* na Figura 19. Nota-se nenhuma observação discrepante com valor absoluto de resíduo acima de 3 e todos distribuídos aleatoriamente.

Com as amostras a *posteriori*, obtém-se as estimativas de Monte Carlo das quatro medidas de divergência-ψ para avaliar a presença de observações influentes na distribuição a *posteriori* dos parâmetros do modelo. As medidas de divergência não detectaram observações influentes na distribuição a *posteriori* (Figura 20).

Como verifica-se efeito de idade, tomamos clientes com idades em diferentes décadas e obtemos a estimativa de Bayes das probabilidades de atraso da parcela pelos clientes tomadores de empréstimo CDC com idades 22, 47 e 83 anos, para cada categoria de renda. A Tabela 24 mostra estas estimativas com o DP a *posteriori* e intervalo HPD de 95%, em que observa-se que o tempo até a inadimplência diminui com o aumento da idade e, que para clientes com renda < 3000 reais, este tempo também é menor que de clientes com renda ≥ 3000 reais, o que implica em maiores probabilidades de atraso das parcelas. A Figura 21 mostra o *boxplot* da probabilidade de atraso por nível de renda, em que observa-se que a mediana da probabilidade

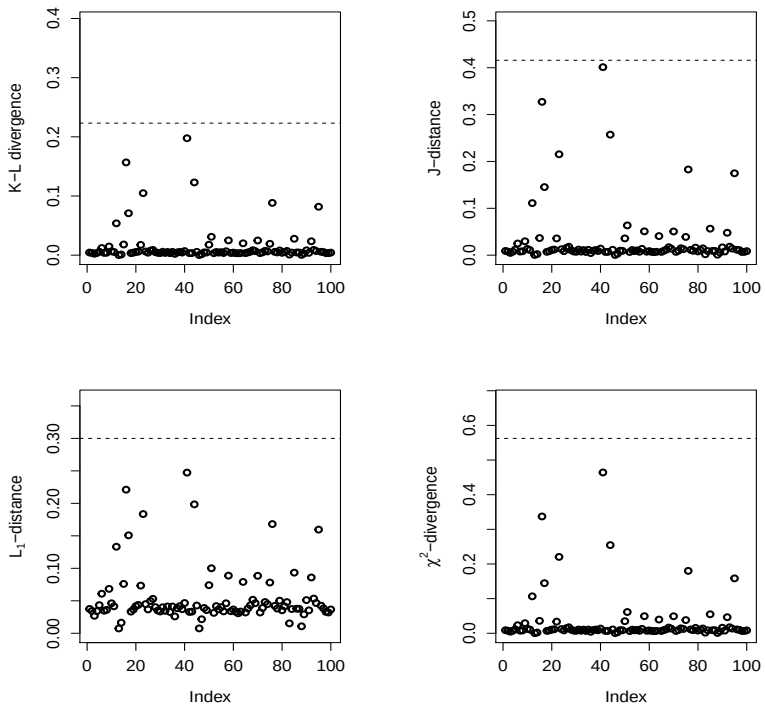

Figura 20 – Estimativas MCMC das medidas de divergência-ψ.

de atraso de clientes com renda inferior a R$3000,00 reais é de 65,9%, enquanto para os clientes com renda maiores ou iguais a R$ 3000,00 reais é de 58,1%. Além disso, a Figura 21 apresenta as densidades a *posteriori* das probabilidades de atraso de clientes com idades de 22, 47 e 83 anos para se observar a diferença entre as probabilidades de atraso por idade e nível de renda.

Tabela 24 – Estimativa de Bayes da probabilidade de atraso para clientes com 22, 47 e 83 anos.

					Intervalo (95%)	
Idade (anos)	Renda (em reais)	Média	Mediana	Desvio Padrão	L	U
22	<3000	0,679	0,679	0,019	0,648	0,708
	$>=3000$	0,620	0,620	0,024	0,580	0,658
47	<3000	0,638	0,639	0,018	0,612	0,670
	$>=3000$	0,579	0,579	0,014	0,557	0,601
83	<3000	0,580	0,578	0,040	0,519	0,649
	$>=3000$	0,522	0,521	0,031	0,473	0,576

Com as amostras a *posteriori*, obtém-se a distribuição a *posteriori* do tempo médio (em meses) até a inadimplência ($E[Y_k]$). A Figura 22 apresenta a estimativa de Monte Carlo desses tempos médios e poemos observar que os clientes com renda menor que R$ 3000,00 apresenta tempo médio até inadimplência menor queclientes com renda maior ou igual a R$ 3000,00. A Tabela 25 mostra as estimativas do tempo médio até a inadimplência para clientes com idades de 22, 47 e 83 anos e a Figura 22 apresenta a densidade a *posteriori* aproximada correspondente, onde é possivel visualizar as diferenças dos tempos médios até inadimplência entre clientes de diferentes idades para cada categoria de renda.

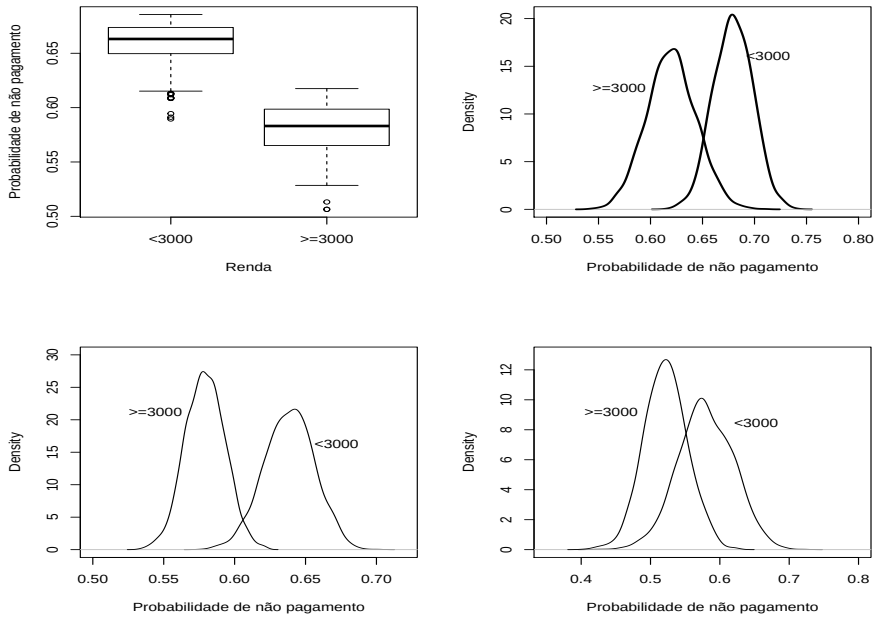

Figura 21 – *Boxplot* da probabilidade de atraso para todos os clientes que participaram do estudo estratificado por nível de renda; e as densidades a *posteriori* da probabilidade de atraso de clientes com idades de 22 anos, 47 anos e 83 anos, respectivamente.

Tabela 25 – Estimativa de Bayes do tempo médio até inadimplência para clientes com idades de 22, 47 e 83 anos.

					Intervalo (95%)	
Idade (anos)	Renda (em reais)	Média	Mediana	Desvio Padrão	L	U
22	< 3000	6.85	6.83	0.43	6.16	7.52
	$>= 3000$	8.46	8.41	0.75	7.15	9.61
47	< 3000	7.89	7.85	0.51	6.99	8.63
	$>= 3000$	9.88	9.86	0.56	9.03	10.83
83	< 3000	10.04	9.88	1.70	7.29	12.38
	$>= 3000$	12.83	12.67	1.90	9.59	15.51

Concluindo, o tempo (em meses) até a inadimplência de clientes que contrataram o empréstimo CDC por um período de 36 meses, avaliado pelo modelo de regressão geométrico correlacionado de ordem 3, mostrou efeito significativo de idade e renda na probabilidade de atraso da parcela e, consequentemente, no tempo até inadimplência, de modo que menores rendas e menores idades foram fatores signicativos para uma maior probabilidade de atraso, logo, para um menor tempo médio até inadimplência. A análise de diagnóstico mostrou ajuste adequado do modelo e ausência de medidas influentes.

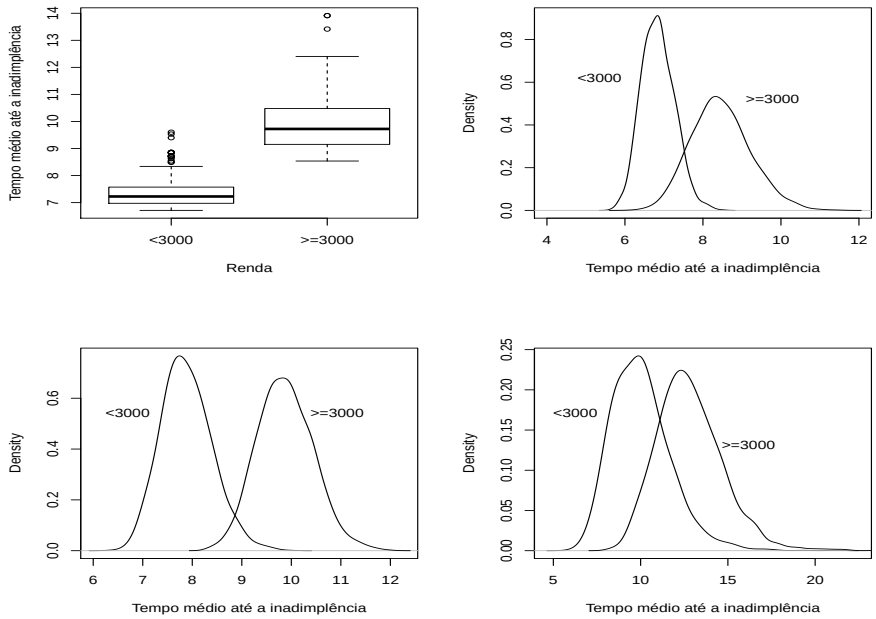

Figura 22 – *Boxplot* do tempo médio até a inadimplência de todos os clientes por categoria de renda; e densidades a *posteriori* do tempo até inadimplência de clientes com idades de 22 anos, 47 anos e 83 anos, respectivamente.

CAPÍTULO

7

CONSIDERAÇÕES FINAIS E CONTINUIDADE DE PESQUISA

Neste trabalho foi proposto o modelo geométrico de ordem k correlacionado de dois parâmetros, π e ρ, como uma extensão dos modelos geométricos já existentes: os modelos geométricos clássico e de ordem k, ambos de parâmetros π, e o modelo geométrico correlacionado de parâmetros π e ρ. A principal motivação em propor este modelo, além de contribuir para generalizações de distribuições discretas, é a proposta de uma análise alternativa e ainda mais adequada para dados reais, pois considera-se o efeito da correlação individual existente pelo parâmetro ρ. Os modelos de regressão complementam este trabalho. Os estudos de simulação e análise de dados reais mostraram melhores ajustes do modelo proposto. Para ambas as abordagens, nas aplicações da distribuição proposta, foi possível constatar que efeitos de covariáveis nas probabilidades de sucesso π_i podem ser superestimados ou subestimados quando modeladas sem considerar o efeito da dependência no modelo, isto é, sem incluir o parâmetro de correlação ρ, pois este efeito pode ser distribuído entre as covariáveis do modelo e então causar divergências nas inferências sobre os coeficientes de regressão.

As probabilidades de sucesso π_i também podem ser modeladas em função das covariáveis por uma nova classe de funções de ligação mais flexíveis, as funções de ligação potência simétrica. Ligações simétricas, como a logito e probito, ou que apresentam assimetria fixa, como as ligações log-log e sua complementar, podem não ser flexíveis o suficiente para incorporar a assimetria existente na distribuição dos dados. Jiang *et al.* (2013) propõem uma nova classe de funções de ligação, as funções de ligação potência simétrica. Esta classe consiste na combinação de uma função de ligação base potência simétrica com sua função reflexo. Isto é, um parâmetro de potência p_o é introduzido em funções de distribuições acumulada de variáveis com função densidade de probabilidade simétrica em relação a zero. Estas funções de ligação potência simétrica permitem maior flexibilidade no ajuste da assimetria pois, conforme os valores de p_o vão diminuindo do valor 1, assimetrias à esquerda vão sendo alcançadas. O contrário ocorre na

função reflexo em que, conforme p_o se distância para maiores valores que 1, vai se atingindo assimetrias à direita. Isto é representado da seguinte forma:

Seja F_0^{-1} uma função de ligação base com função distribuição acumulada F_0, cuja função densidade de probabilidade é simétrica em relação a zero. A função de ligação potência simétria F é então obtida por:

$$F(x,p_o) = F_0^{p_o}(x/p_o)I_{(0,1]}(p_o) + [1 - F_0^{\frac{1}{p_o}}(-p_o x)]I_{(1,+\infty]}(p_o), \tag{7.1}$$

e assim se obtém maior flexibilidade para acomodar a assimetria em ambas as direções, tanto positiva como negativa, simetricamente em relação a função base.

Desta forma, Jiang *et al.* (2013) sugerem três famílias de funções de ligação segundo a função base F_0, a família de ligação logito potência simétrica (*splogit*), a família de ligação *t* potência simétrica (*spt*) e a família de ligação potência exponencial potência simétrica (*spep*). Nós estudamos a família de ligação logito potência simétrica por metódo de inferência clássica, para as quais devemos obter as derivadas de 1a. e 2a. ordens para a estimação dos parâmetros de regressão. A função de ligação *splogit* é dada por:

$$\begin{aligned} g(\pi_i) &= p_o \log\left(\frac{\pi_i^{1/p_o}}{1-\pi_i^{1/p_o}}\right) I_{(0,1]}(p_o) + \left[-\frac{1}{p_o}\log\left(\frac{(1-\pi)^{p_o}}{1-(1-\pi_i)^{p_o}}\right)\right] I_{[1,+\infty)}(p_o), \\ g^{-1}(\eta_i) &= \left[\frac{\exp(\eta_i/p_o)}{1+\exp(\eta_i/p_o)}\right]^{p_o} I_{(0,1]}(p_o) + \left[1 - \left(\frac{\exp(-p_o\eta_i)}{1+\exp(-p_o\eta_i)}\right)^{1/p_o}\right] I_{[1,+\infty)}(p_o), \end{aligned} \tag{7.2}$$

sendo $\eta_i = g(\pi_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$; $i = 1,\ldots,m$.

As derivadas de primeira ordem da função de ligação *splogit* (7.2) em relação aos parâmetros $\boldsymbol{\beta}$, $\Delta_i(\beta_r) = \partial g^{-1}(\eta_i)/\partial\beta_r$, e, em relação ao parâmetro p_o, $\Delta_i(p_o) = \partial g^{-1}(\eta_i)/\partial p_o$; $r = 1,\ldots,p$, são dadas por:

$$\Delta_i(\beta_r) = \frac{x_{ir}\exp(\eta_i)}{[1+\exp(\eta_i/p_o)]^{p_o}} I_{(0,1]}(p_o) + \frac{x_{ir}\exp(-\eta_i)}{[1+\exp(-p_o\eta_i)]^{\frac{1}{p_o}}} I_{[1,+\infty)}(p_o), \tag{7.3}$$

$$\begin{aligned} \Delta_i(p_o) = {} & \left\{\frac{\eta_i}{p_o}\left[\frac{\exp(\eta_i/p_o)}{1+\exp(\eta_i/p_o)}\right] - \log(1+\exp(\eta_i/p_o))\right\}\left[\frac{\exp(\eta_i/p_o)}{1+\exp(\eta_i/p_o)}\right]^{p_o} I_{(0,1]}(p_o) \\ & \left\{\frac{\eta_i}{p_o}\left[\frac{\exp(-\eta_i p_o)}{1+\exp(-\eta_i p_o)}\right] + \frac{1}{p_o^2}\log(1+\exp(-\eta_i p_o))\right\}\left[\frac{\exp(-\eta_i p_o)}{1+\exp(-\eta_i p_o)}\right]^{\frac{1}{p_o}} I_{[1,+\infty)}(p_o). \end{aligned} \tag{7.4}$$

As derivadas de segunda ordem da função de ligação *splogit* em relação aos parâmetros $\boldsymbol{\beta}$ e p_o, isto é, $\Delta_i^2(\beta_r\beta_s) = \partial^2 g^{-1}(\eta_i)/\partial\beta_r\partial\beta_s$, $\Delta_i^2(p_o\beta_r) = \partial^2 g^{-1}(\eta_i)/(\partial p_o\partial\beta_s)$ e $\Delta_i^2(p_o^2) =$

$\partial g^{-1}(\eta_i)/(\partial^2 p_o)^2$, $r,s = 1,\ldots,p$, são dadas por :

$$\begin{aligned}\Delta_i^2(\beta_r\beta_s) &= \frac{x_{ir}x_{is}\exp(\eta_i)}{(1+\exp(\eta_i/p_o))^{p_o+1}}\left[1-\frac{(1+p_o)\exp(\eta_i/p_o)}{p_o(1+\exp(\eta_i/p_o))}\right]I_{(0,1]}(p_o)+\\ &- \frac{x_{ir}x_{is}\exp(\eta_i)}{(1+\exp(-p_o\eta_i))^{\frac{1}{p_o}+1}}\left[1-\frac{(1+p_o)\exp(-p_o\eta_i)}{(1+\exp(-p_o\eta_i))}\right]I_{[1,+\infty)}(p_o),\end{aligned} \tag{7.5}$$

$$\begin{aligned}\Delta_i^2(p_o\beta s) &= \frac{x_{is}}{p_o}\left[\frac{(p_o+1)\eta_i}{p_o(1+\exp(\eta_i/p_o))}-\frac{p_o\log(1+\exp(\eta_i/p_o))}{\exp(\eta_i/p_o)}\right]\left[\frac{\exp(\eta_i/p_o)}{1+\exp(\eta_i/p_o)}\right]^{p_o+1}I_{(0,1]}(p_o)+\\ &- \frac{x_{is}}{p_o}\left[\frac{(p_o+1)\eta_i}{(1+\exp(-p_o\eta_i))}+\frac{\log(1+\exp(-p_o\eta_i))}{p_o\exp(-p_o\eta_i)}\right]\left[\frac{\exp(-p_o\eta_i)}{1+\exp(-p_o\eta_i)}\right]^{\frac{1}{p_o}+1}I_{[1,+\infty)}(p_o),\end{aligned} \tag{7.6}$$

$$\begin{aligned}\Delta_i^2(p_o^2) &= \left\{-\frac{\eta_i^2}{p_o^3(1+\exp(\eta_i/p_o))}\left[\frac{\exp(\eta_i/p_o)}{1+\exp(\eta_i/p_o)}\right]+\left\{\frac{\eta_i}{p_o}\left[\frac{\exp(\eta_i/p_o)}{1+\exp(\eta_i/p_o)}\right]+\right.\right.\\ &- \left.\left.\log(1+\exp(\eta_i/p_o))\right\}^2\right\}\left[\frac{\exp(\eta_i/p_o)}{1+\exp(\eta_i/p_o)}\right]^{p_o}I_{(0,1]}(p_o)+\\ &+ \left\{-\frac{\eta_i}{p_o}\left(\frac{2}{p_o}+\frac{\eta_i}{1+\exp(-\eta_i p_o)}\right)\left[\frac{\exp(-\eta_i p_o)}{1+\exp(-\eta_i p_o)}\right]-\frac{2\log(1+\exp(-\eta_i p_o))}{p_o^3}+\right.\\ &- \left.\left\{\frac{\eta_i}{p_o}\left[\frac{\exp(-\eta_i p_o)}{1+\exp(-\eta_i p_o)}\right]+\frac{1}{p_0^2}\log(1+\exp(-\eta_i p_o))\right\}^2\right\}\left[\frac{\exp(-\eta_i p_o)}{1+\exp(-\eta_i p_o)}\right]^{\frac{1}{p_o}}I_{[1,+\infty)}(p_o).\end{aligned} \tag{7.7}$$

Simulamos um conjunto de dados com função de ligação *splogit*. Os ajustes dos modelos, considerando também esta função de ligação, apresentaram grandes variações nos resultados das estimativas dos parâmetros e, não raramente, não convergiam para um valor finito. Portanto, a inclusão de mais um parâmetro no modelo (p_o), e ainda de potência, causou grandes desvios nas estimativas clássicas dos demais parâmetros de interesse. Sendo a função de ligação *splogit* uma alternativa interessante sob o ponto de vista teórico, visto que, uma única função de ligação possa ser utilizada para o ajuste de dados discretos, não tendo com isto a necessidade de testar quatro ou mais funções de ligação para dados de natureza binária, temos como sugestão de continuidade de pesquisa analisar a família de ligação logito potência simétrica numa abordagem Bayesiana, como também estudar as demais famílias de ligação potência simétrica para modelagem de dados discretos.

Consideramos também obter as distribuições binomial negativa e Poisson de ordem k correlacionadas a partir da distribuição geométrica de ordem k correlacionada. Ambas possuem aplicações relevantes, como em estudos de prevalência, de incidência e de reincidência na área médica.

REFERÊNCIAS

ABDOU, H. A. Genetic programming for credit scoring: The case of egyptian public sector banks. **Expert Systems with Applications**, Elsevier, v. 36, n. 9, p. 11402–11417, 2009. Citado na página 94.

AKAIKE, H. A new look at the statistical model identification. **IEEE transactions on automatic control**, Ieee, v. 19, n. 6, p. 716–723, 1974. Citado nas páginas 50 e 85.

AKI, S.; HIRANO, K. Discrete distributions related to succession events in a two-state markov chain. In: VSP INTERNATIONAL SCIENCE PUBLISHERS. **Statistical Sciences and Data Analysis; Proceedings of the Third Pacific Area Statistical Conference**. [S.l.], 1993. p. 467–474. Citado nas páginas 9, 11, 24, 26, 33, 35 e 36.

____. Distributions of numbers of failures and successes until the first consecutive k successes. **Annals of the Institute of Statistical Mathematics**, Springer, v. 46, n. 1, p. 193–202, 1994. Citado na página 26.

BAESENS, B.; EGMONT-PETERSEN, M.; CASTELO, R.; VANTHIENEN, J. Learning bayesian network classifiers for credit scoring using markov chain monte carlo search. In: IEEE. **Pattern Recognition, 2002. Proceedings. 16th International Conference on**. [S.l.], 2002. v. 3, p. 49–52. Citado na página 94.

BENSIC, M.; SARLIJA, N.; ZEKIC-SUSAC, M. Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. **Intelligent Systems in Accounting, Finance and Management**, Wiley Online Library, v. 13, n. 3, p. 133–150, 2005. Citado na página 94.

BLACK, F.; SCHOLES, M. The pricing of options and corporate liabilities. **Journal of political economy**, The University of Chicago Press, v. 81, n. 3, p. 637–654, 1973. Citado na página 94.

BOLFARINE, H.; SANDOVAL, M. **Introdução à inferência estatística**. [S.l.]: SBM, 2001. v. 2. Citado nas páginas 85 e 122.

BREIMAN, L.; FRIEDMAN, J.; STONE, C. J.; OLSHEN, R. A. **Classification and regression trees**. [S.l.]: CRC press, 1984. Citado na página 94.

BROOKS, S. P. Discussion on the paper by Spiegelhalter, Best, Carlin, and van der Linde (2002). **Journal of the Royal Statistical Society B**, v. 64, p. 616–618, 2002. Citado na página 72.

CARLIN, B. P.; LOUIS, T. A. **Bayes and Empirical Bayes Methods for Data Analysis**. second. Boca Raton: Chapman & Hall/CRC, 2001. Citado na página 72.

CASELLA, G.; GEORGE, E. I. Explaining the gibbs sampler. **The American Statistician**, Taylor & Francis, v. 46, n. 3, p. 167–174, 1992. Citado na página 71.

CONWAY, R. W.; MAXWELL, W. L. A queuing model with state dependent service rates. **Journal of Industrial Engineering**, v. 12, n. 2, p. 132–136, 1962. Citado na página 25.

COOK, D. R. Detection of influential observation in linear regression. **Technometrics**, Taylor & Francis Group, v. 19, n. 1, p. 15–18, 1977. Citado nas páginas 53, 74 e 87.

COOK, R. D. Assessment of local influence. **Journal of the Royal Statistical Society, Series B**, v. 48, p. 133–169, 1986. Citado nas páginas 53, 54 e 88.

COOK, R. D.; PEÑA, D.; WEISBERG, S. The likelihood displacement: a unifying principle for influence measures. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 17, n. 3, p. 623–640, 1988. Citado nas páginas 53 e 87.

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. **Modelos lineares Generalizados: minicurso para o 12º SEAGRO e a 52ª reunião anual da RBRAS**. Santa Maria, RS: RBRAS-UFSM, 2007. Citado nas páginas 28 e 30.

COX, D.; OAKES, D. **Analysis of Survival Data**. London: Chapman and Hall, 1984. Citado na página 94.

DEMÉTRIO, C. G. B. **Modelos lineares generalizados em experimentação agronômica**. [S.l.]: USP/ESALQ, 2001. Citado na página 51.

DEY, D. K.; BIRMIWAL, L. R. Robust bayesian analysis using divergence measures. **Statistics & Probability Letters**, Elsevier, v. 20, n. 4, p. 287–294, 1994. Citado na página 75.

DINIZ, C.; LOUZADA, F. Métodos estatısticos para análise de dados de crédito. In: **6th Brazilian Conference on Statistical Modeling in Insurance and Finance, Maresias-SP**. [S.l.: s.n.], 2013. Citado na página 94.

DINIZ, C. A. R.; TUTIA, M. H.; LEITE, J. G. Bayesian analysis of a correlated binomial model. **Brazilian Journal of Probability and Statistics**, JSTOR, p. 68–77, 2010. Citado na página 26.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996. Citado nas páginas 30, 52, 74 e 87.

DURAND, D. *et al.* Risk elements in consumer instalment financing. **NBER Books**, National Bureau of Economic Research, Inc, 1941. Citado na página 94.

FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of eugenics**, Wiley Online Library, v. 7, n. 2, p. 179–188, 1936. Citado na página 94.

GABAI, H. Generalized fibonacci k-sequences. **Fibonacci Quart**, Citeseer, v. 8, p. 31–38, 1970. Citado na página 26.

GELFAND, A. E.; DEY, D. K.; CHANG, H. Model determination using predictive distributions with implementation via sampling-based methods. In: **Bayesian Statistics 4 (Peñíscola, 1991)**. New York: Oxford University Press, 1992. p. 147–167. Citado na página 72.

HAND, D. J.; HENLEY, W. E. Statistical classification methods in consumer credit scoring: a review. **Journal of the Royal Statistical Society: Series A (Statistics in Society)**, Wiley Online Library, v. 160, n. 3, p. 523–541, 1997. Citado na página 94.

HIRANO, K. Some properties of the distributions of order k. **Fibonacci Numbers and Their Applications**, Reidel Dordrecht, v. 42, p. p45–61, 1986. Citado na página 26.

HUANG, C.-L.; CHEN, M.-C.; WANG, C.-J. Credit scoring with a data mining approach based on support vector machines. **Expert systems with applications**, Elsevier, v. 33, n. 4, p. 847–856, 2007. Citado na página 94.

JIANG, X.; DEY, D. K.; PRUNIER, R.; WILSON, A. M. A new class of flexible link functions with application to species co-occurrence in cape floristic region. **The Annals of Applied Statistics**, Institute of Mathematical Statistics, v. 7, n. 4, p. 2180–2204, 2013. Citado na página 113.

KADANE, J. B. *et al.* Sums of possibly associated bernoulli variables: The conway–maxwell-binomial distribution. **Bayesian Analysis**, International Society for Bayesian Analysis, v. 11, n. 2, p. 403–420, 2016. Citado na página 25.

KOLEV, N.; MINKOVA, L.; NEYTCHEV, P. Inflated-parameter family of generalized power series distributions and their application in analysis of over dispersed insurance data. **Actuarial Research Clearing House**, Society Of Actuaries, v. 2, p. 295–320, 2000. Citado nas páginas 9, 11, 23, 24, 25, 26, 33, 34, 36, 37, 39, 40 e 47.

LEITE, J. G.; SINGER, J. M. **Métodos assintóticos em estatística: fundamentos e aplicações**. [S.l.]: Universidade de Sao Paulo, 1990. Citado nas páginas 51, 85 e 121.

LIMA, E. L. **Curso de análise, Volume 1**. [S.l.: s.n.], 2004. Citado na página 29.

LUCEÑO, A. A family of partially correlated poisson models for overdispersion. **Computational statistics & data analysis**, Elsevier, v. 20, n. 5, p. 511–520, 1995. Citado nas páginas 25, 26 e 37.

LUCEÑO, A.; CEBALLOS, F. D. Describing extra-binomial variation with partially correlated models. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 24, n. 6, p. 1637–1653, 1995. Citado nas páginas 25 e 26.

MARKOWITZ, H. Portfolio selection. **The journal of finance**, Wiley Online Library, v. 7, n. 1, p. 77–91, 1952. Citado na página 94.

MCCULLAGH, P.; NELDER, J. A. **Generalized Linear Models**. 2nd. ed. London, UK: Chapman & Hall, 1989. Citado nas páginas 24, 28 e 48.

OCHI, Y.; PRENTICE, R. L. Likelihood inference in a correlated probit regression model. **Biometrika**, Oxford University Press, v. 71, n. 3, p. 531–543, 1984. Citado na página 26.

PAULA, G. A. **Modelos de regressão: com apoio computacional**. [S.l.]: IME-USP São Paulo, 2004. Citado na página 30.

PENG, F.; DEY, D. Bayesian analysis of outlier problems using divergence measures. **Canadian Journal of Statistics**, Wiley Online Library, v. 23, n. 2, p. 199–213, 1995. Citado na página 74.

PERRODO, P. Improving scoring techniques in banking applications. **Applied stochastic models and data analysis**, Wiley Online Library, v. 9, n. 3, p. 231–244, 1993. Citado na página 94.

PHILIPPOU, A. The negative binomial distribution of order k and some of its properties. **Biometrical Journal**, Wiley Online Library, v. 26, n. 7, p. 789–794, 1984. Citado na página 26.

PHILIPPOU, A. N.; GEORGHIOU, C.; PHILIPPOU, G. N. A generalized geometric distribution and some of its properties. **Statistics & Probability Letters**, North-Holland, v. 1, n. 4, p. 171–175, 1983. Citado nas páginas 26 e 35.

PHILIPPOU, A. N.; MAKRI, F. S. Longest success runs and fibonacci-type polynomials. **The Fibonacci Quarterly**, Citeseer, v. 23, n. 4, p. 338–346, 1985. Citado na página 26.

____. Successes, runs and longest runs. **Statistics & Probability Letters**, North-Holland, v. 4, n. 4, p. 211–215, 1986. Citado na página 26.

PHILIPPOU, A. N.; MUWAFI, A. A. Waiting for the kth consecutive success and the fibonacci sequence of order k. **The Fibonacci Quarterly**, Citeseer, v. 20, p. 28–32, 1980. Citado nas páginas 9, 11, 24, 25, 26, 33, 34, 40 e 83.

PIRES, R. M.; DINIZ, C. A. R. Correlated binomial regression models. **Computational Statistics & Data Analysis**, Elsevier, v. 56, n. 8, p. 2513–2525, 2012. Citado na página 26.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2018. Disponível em: <https://www.R-project.org/>. Citado nas páginas 55 e 77.

SCHWARZ, G. *et al.* Estimating the dimension of a model. **The annals of statistics**, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. Citado nas páginas 50 e 85.

SHMUELI, G.; MINKA, T. P.; KADANE, J. B.; BORLE, S.; BOATWRIGHT, P. A useful distribution for fitting discrete data: revival of the conway–maxwell–poisson distribution. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 54, n. 1, p. 127–142, 2005. Citado na página 25.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. V. D. Bayesian measures of model complexity and fit. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 64, n. 4, p. 583–639, 2002. Citado na página 72.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. The deviance information criterion: 12 years on. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 76, n. 3, p. 485–493, 2014. Citado na página 72.

TONG, E. N.; MUES, C.; THOMAS, L. C. Mixture cure models in credit scoring: If and when borrowers default. **European Journal of Operational Research**, Elsevier, v. 218, n. 1, p. 132–139, 2012. Citado na página 94.

WANG, Y.; WANG, S.; LAI, K. K. A new fuzzy support vector machine to evaluate credit risk. **IEEE Transactions on Fuzzy Systems**, IEEE, v. 13, n. 6, p. 820–831, 2005. Citado na página 94.

WILLIAMS, D. A. Extra-binomial variation in logistic linear models. **Applied statistics**, JSTOR, p. 144–148, 1982. Citado na página 26.

YIQI, B. **Estimação e diagnóstico na disribuição Weibull-Binomial-Negativa em análise de sobrevivência**. Tese (Doutorado) — Universidade de São Paulo, 2012. Citado na página 75.

APÊNDICE

A

CONDIÇÕES DE REGULARIDADE

Teorema A.0.1. Sejam $Y_1, Y_2, \ldots, Y_m$ variáveis aleatórias independentes com função de probabilidade $f(y;\boldsymbol{\beta})$, $\boldsymbol{\beta} \in \boldsymbol{B} \subset \mathbb{R}^{R+1}$ satisfazendo as seguintes condições de regularidade:

1. Para $r,s = 0,\ldots,R$, $\frac{\partial}{\partial\beta_r} f(y;\boldsymbol{\beta})$ e $\frac{\partial^2}{\partial\beta_r\partial\beta_s} f(y;\boldsymbol{\beta})$ existem em quase toda parte e são tais que $|\frac{\partial}{\partial\beta_r} f(y;\boldsymbol{\beta})| \leq H_r(y)$ e $|\frac{\partial^2}{\partial\beta_r\partial\beta_s} f(y;\boldsymbol{\beta})| \leq G_{rs}(y)$ onde $\int_{\mathbb{R}} H_r(y)dy < \infty$ e $\int_{\mathbb{R}} G_{rs}(y)dy < \infty$;

2. Para $r,s = 0,\ldots,R$, $\frac{\partial}{\partial\beta_r} \log f(y;\boldsymbol{\beta})$ e $\frac{\partial^2}{\partial\beta_r\partial\beta_s} \log f(y;\boldsymbol{\beta})$ existem em quase toda parte e são tais que:

   a) a matriz de informação esperada de Fisher,

   $$\boldsymbol{I}(\boldsymbol{\beta}) = E[\frac{\partial}{\partial\boldsymbol{\beta}} \log f(Y_i;\boldsymbol{\beta})]^T \frac{\partial}{\partial\boldsymbol{\beta}} \log f(Y_i;\boldsymbol{\beta})],$$

   onde $\frac{\partial}{\partial\boldsymbol{\beta}} \log f(Y_i;\boldsymbol{\beta}) = [\frac{\partial}{\partial\beta_0} \log f(Y_i;\boldsymbol{\beta}) \ldots \frac{\partial}{\partial\beta_R} \log f(Y_i;\boldsymbol{\beta})]$, é finita e positiva definida.

   b) $E_{\boldsymbol{\beta}}\,[\,\sup\|\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \log f(Y_i;\boldsymbol{\beta}+h) - \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \log f(Y_i;\boldsymbol{\beta})\|\,] = \psi_\delta$, onde $\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \log f(Y_i;\boldsymbol{\beta}) = [\frac{\partial^2}{\partial\beta_r\partial\beta_s} \log f(Y_i;\boldsymbol{\beta})]_{rs}$, converge para zero com $\delta \to 0$.

Então o estimador de máxima verossimilhança, EMV, de $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, é tal que:

$$\sqrt{m}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}^{-1}(\boldsymbol{\beta}))$$

Pelo teorema A.0.1, demonstrado em (LEITE; SINGER, 1990) (p.112), tem-se que, sob as condições de regularidade, $\sqrt{m}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_{R+1}(\boldsymbol{0}, \boldsymbol{I}^{-1}(\boldsymbol{\beta}))$; isto é, a distribuição assintótica de $\sqrt{m}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ é Normal (R+1)-variada, em que $R+1$ é a quantidade de parâmetros no modelo. A matriz de informação esperada de Fisher, $\boldsymbol{I}(\boldsymbol{\beta})$, pode ser aproximada pela matriz de informação observada de Fisher, $\boldsymbol{J}(\boldsymbol{\beta})$. Dentre estas condições de regularidade tem-se, basicamente, que o suporte $S(\boldsymbol{y}) = \{\boldsymbol{y}, f(\boldsymbol{y}|\boldsymbol{\beta}) > \boldsymbol{0}\}$ deve ser independente de $\boldsymbol{\beta}$ e que a troca das ordens das

operações de derivação e de integracão sob a distribuicão da variável aleatória $\boldsymbol{Y}$ seja possível (BOLFARINE; SANDOVAL, 2001).

APÊNDICE

B

ALGORITMOS PARA O MRGCK CLÁSSICO

Algoritmo 1 – Gerador de m variáveis aletórias y_k do MRGCk.

Entrada: $\boldsymbol{\beta}$; ρ; $\boldsymbol{x}$; k e m
Saída: $\boldsymbol{y}_k = \boldsymbol{y} + k$
1: Calcular $g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta}) = \frac{\exp \boldsymbol{x}^T\boldsymbol{\beta}}{1+\exp \boldsymbol{x}^T\boldsymbol{\beta}}$ (logito)
2: Calcular $\boldsymbol{\pi} = g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta})$ e as probabilidades de transição da cadeia p_{00}, p_{01}, p_{10} e p_{11} conforme (2.13); (2.14); (2.15) e (2.16),
3: Calular $P(Y_{ki} = k) : P_{0i} = (1-\pi_i)p_{01i}p_{11i}^{k-1} + \pi_i p_{11i}^{k}$.
4: Gerar uma v.a. u_i de $U_i \sim U(0,1)$ e inicializar $y_i = -1$.
5: Gerar y_i:
se $u_i < P_{0i}$ **então** $y_i = 0$
senão
  $j = 1$; $aux = P(Y_k = j+k)$ (função que calcula a probabilidade em (2.12)); $P_i = P_{0i} + aux$
  enquanto $y_i = -1$ **faça**
    se $u_i < P_i$ **então** $y_i = j$
    senão
      $j = j+1$, $aux = P(Y_{ki} = j+k)$ e $P_i = P_i + aux$
    fim se
  fim enquanto
fim se
6: Repetir os passos 3 a 6 até m vezes ($i = 1, 2 \ldots m$)

Algoritmo 2 – Estimador de máxima verossimilhança

Entrada: $\boldsymbol{y}_k$; $\boldsymbol{x}$, k e função de verossimilhança.
Saída: $\hat{\boldsymbol{\beta}}, \hat{\rho}$ e matriz hessiana
1: Entrar com a verossimilhança em função de $\boldsymbol{\beta}$ e ρ.
2: Iniciar o processo numérico de maximização da verossimilhança pelas funções *maxLik* ou *optim* do *software R* atribuindo "chutes"iniciais para os valores dos parâmetros

As medidas de influência global são obtidas por cálculos utilizando as saídas do algoritmo 2. A função *optim* do R para maximização da função de verossimilhança nos fornece a

Algoritmo 3 – Resíduo quantílico aleatorizado

Entrada: m, $\boldsymbol{y}_k$; $\boldsymbol{x}$, k, $\hat{\boldsymbol{\beta}}$ e $\hat{\rho}$ obtidos pelo algoritmo 2
Saída: $\boldsymbol{r}_q$
1: Calcular $\hat{\boldsymbol{\pi}} = g^{-1}(\boldsymbol{x}^T\hat{\boldsymbol{\beta}})$
2: Obter as distribuições acumuladas de Y_{ki}:
para $i = 1$ to m **faça**
  $valor = y_{ki} - k$;
  se $valor > 0$ **então** inicializar $j = 0$ e $F = 0$;
    enquanto j<=valor-1 **faça** $pro = P(Y_{ki} = j+k)$, $F = F + pro$ e $j = j+1$
    fim enquanto
    obter $a_i = F$, $b_i = F + P(Y_{ki} = valor + k|\hat{\pi}_i, \hat{\rho})$ e gerar u_i de $U_i \sim U(a,b)$
  senão gerar u_i de $U_i \sim U(0, \hat{\pi}^k)$
  fim se
fim para
3: Obter $\boldsymbol{r}_q = \Phi^{-1}(\boldsymbol{u})$

Algoritmo 4 – Influência local

Entrada: m, $\boldsymbol{y}_k$; $\boldsymbol{x}$, k, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}; \hat{\rho})$ e $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$ obtidos pelo algoritmo 2, função de máxima verossimilhança $\mathscr{L}(\hat{\boldsymbol{\theta}})$ e função de verossimilhança perturbada $\mathscr{L}(\boldsymbol{\theta}; \boldsymbol{\gamma})$.
Saída: $|\boldsymbol{d}_{max}|$
1: Obter as derivadas parciais de segunda ordem $\boldsymbol{\Gamma} = \partial^2\mathscr{L}(\boldsymbol{\theta}; \boldsymbol{\gamma})/\partial\boldsymbol{\theta}'\partial\boldsymbol{\gamma}$
2: Calcular $\boldsymbol{A} = \boldsymbol{\Gamma}^T\boldsymbol{J}(\hat{\boldsymbol{\theta}})$
3: Obter $|\boldsymbol{d}_{max}|$ autovetor correspondente ao máximo valor absoluto dos autovalores de $\boldsymbol{A}$

matriz hessiana e então calculamos as distâncias de Cook. As estimativas e função de verossimilhança com os dados da amostra observada excluindo uma observação também são obtidas pelo algoritmo 2.


ICMC
USP
SÃO CARLOS