Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Modelos de Regressão Binomial Correlacionada

Rubiane Maria Pires

São Carlos
2012

Rubiane Maria Pires

Modelos de Regressão Binomial Correlacionada

Tese apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar como parte dos requisitos para obtenção do título de doutor em estatística.

Orientador: Prof. Dr. Carlos Alberto Ribeiro Diniz

São Carlos
2012

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária/UFSCar

P667mr

Pires, Rubiane Maria.
Modelos de regressão binomial correlacionada / Rubiane Maria Pires. -- São Carlos : UFSCar, 2012.
138 f.

Tese (Doutorado) -- Universidade Federal de São Carlos, 2012.

1. Estatística. 2. Análise de regressão. 3. Dados aumentados. 4. Análise de diagnósticos. 5. Distribuição binomial generalizada. I. Título.

CDD: 519.5 (20ª)

Rubiane Maria Pires

Modelos de Regressão Binomial Correlacionado

Tese apresentada à Universidade Federal de São Carlos, como parte dos requisitos para obtenção do título de Doutora em Estatística.

Aprovada em 18 de maio de 2012.

BANCA EXAMINADORA

Presidente _____
Prof. Dr. Carlos Alberto Ribeiro Diniz (DEs-UFSCar / Orientador)

1º Examinador _____
Prof. Dr. Carlos Alberto de Bragança Pereira (IME-USP)

2º Examinador _____
Prof. Dr. Enrico Antonio Colosimo (UFMG)

3º Examinador _____
Prof. Dr. Francisco Louzada Neto (ICMC-USP)

4º Examinador _____
Prof. Dr. Victor Hugo Lachos Dávila (UNICAMP)

Agradecimentos

Agradeço a minha mãe e demais familiares, pela formação de caráter em boas maneiras de conduzir minha vida.

Ao meu grande amor e companheiro Leandro Souza, pela amizade e incentivo, compartilhados nas pacientes horas de trabalhos, incentivando-me a seguir em frente.

A todos os meus amigos que sempre estiveram carinhosamente presentes contribuindo com críticas, sugestões e paciente tolerância.

Ao meu orientador, Carlos Diniz, os mais sinceros agradecimentos pela orientação segura, incentivadora e acolhedora, na elaboração e condução do trabalho.

Aos professores Enrico Colosimo e José Galvão Leite, membros da banca do exame de qualificação, pelas sugestões feitas.

Aos professores do Departamento de Estatística da Universidade Federal de São Carlos que me abriram as portas e me ofereceram o ambiente acolhedor e sadio para que eu pudesse realizar o meu doutorado.

Finalmente, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio concedido para este trabalho.

Resumo

Nesta tese é proposta uma classe de modelos de regressão binomial correlacionada baseados na distribuição binomial generalizada, proposta por Luceño (1995) e Luceño & Ceballos (1995). A estrutura de regressão é modelada usando diferentes funções de ligação e a relação de dependência entre os ensaios de Bernoulli é modelada usando diferentes estruturas de correlação. Uma estratégia de dados aumentados é utilizada para contornar a complexidade da função de verossimilhança. As abordagens clássica e Bayesiana são utilizadas no processo de ajuste dos modelos propostos. Análise de diagnóstico é desenvolvida com o objetivo de verificar as suposições iniciais do modelo e identificar a presença de *outliers* e/ou observações influentes. Estudos de simulação e aplicação em dados reais ilustram as metodologias. Propomos também uma nova classe de modelos de regressão binomial correlacionada, denominados modelos de regressão binomial correlacionada aditivo estrutural normal, que envolvem a presença de uma covariável com erro de medida. No processo de estimação para esta nova classe, dados aumentados e aproximação de integral são utilizadas para contornar a complexidade da função de verossimilhança.

Abstract

In this thesis, a class of correlated binomial regression models is proposed. The model is based on the generalized binomial distribution proposed by Luceño (1995) and Luceño & Ceballos (1995). The regression structure is modeled by using four different link functions and the dependence between the Bernoulli trials is modeled by using three different correlation structures. A data augmentation scheme is used in order to overcome the complexity of the mixture likelihood. Frequentist and Bayesian approaches are used in the model fitting process. A diagnostics analysis is provided in order to check the underlying model assumptions and to identify the presence of outliers and/or influential observations. Simulation studies are presented to illustrate the performance of the developed methodology. A real data set is analyzed by using the proposed models. Also the correlated binomial regression models is extended to include measurement error in a predictor. This new class of models is called additive normal structure correlated binomial regression models. The inference process also includes a data augmentation scheme to overcome the complexity of the mixture likelihood.

Sumário

1 Introdução **1**

2 Modelos de regressão binomial correlacionada (MRBC) **6**
2.1 Uma classe de modelos de regressão binomial correlacionada 6
2.2 Função de verosimilhança com dados aumentados 8

3 MRBC: Metodologia Clássica **11**
3.1 Estimação via algoritmo EM 11
3.2 Intervalos de confiança 12
3.2.1 Intervalos de confiança assintóticos 12
3.2.2 Intervalos de confiança *bootstrap* 15
3.2.3 Intervalos de confiança perfilados 16
3.3 Teste de hipótese 16
3.3.1 Teste de hipótese para os parâmetros de regressão 16
3.3.2 Teste de hipótese para o parâmetro da estrutura de correlação . . . 17
3.4 Diagnósticos 18
3.4.1 Resíduos 19
3.4.2 Diagnóstico de influencia global 20
3.5 Critérios de seleção de modelo 21
3.6 Estudos de simulação 22
3.6.1 Propriedade frequentista dos estimadores de máxima verossimilhança 26

4 MRBC: Metodologia Bayesiana **30**
4.1 Distribuição preditiva a *posteriori* 31
4.2 Densidade preditiva condicional ordinária 32
4.3 Diagnósticos 34
4.3.1 Resíduos Bayesianos 34
4.3.2 Diagnóstico de influencia Bayesiano 37
4.4 Critérios de seleção de modelo 38
4.5 Estudos de simulação 39
4.5.1 Sensibilidade em relação a distribuição a *priori* 42
4.5.2 Propriedade frequentista dos estimadores Bayesianos 43

5 Modelo de regressão beta-binomial (MRBB): Diagnóstico Bayesiano 46
5.1 Modelos de regressão beta-binomial . . . 47
5.2 Abordagem Bayesiana . . . 48
5.3 Diagnósticos . . . 49
5.4 Estudos de simulação . . . 52
5.5 MRBC e MRBB: Análise de dados reais . . . 57
5.5.1 Metodologia clássica: aplicação em planos de saúde . . . 57
5.5.2 Metodologia Bayesiana: aplicação em finanças . . . 61

6 Modelos de regressão binomial correlacionada aditivo estrutural normal (MRBCAEN) 67
6.1 Uma classe de modelos de regressão binomial correlacionada aditivo estrutural normal . . . 67
6.1.1 Aproximação da função de verosimilhança . . . 69
6.2 Função de verosimilhança com dados aumentados . . . 70
6.2.1 Aproximação da função de verosimilhança com dados aumentados . 71
6.3 Estimação via algoritmo EM . . . 72
6.4 Intervalos de confiança . . . 73
6.4.1 Intervalos de confiança assintóticos . . . 73
6.4.2 Intervalos de confiança *bootstrap* . . . 75
6.4.3 Intervalos de confiança perfilados . . . 76
6.5 Teste de hipótese . . . 76
6.6 Diagnósticos . . . 76
6.6.1 Resíduos . . . 77
6.6.2 Diagnóstico de influencia global . . . 78
6.7 Critérios de seleção de modelo . . . 79
6.8 Estudos de simulação . . . 80
6.8.1 Propriedade frequentista dos estimadores de máxima verossimilhança 84
6.9 Análise de dados reais . . . 88

7 Considerações finais e propostas futuras 91

A Condições de regularidade 97

B Método de Laplace 98

C Programas para o MRBC: Abordagem Clássica 99
C.1 Estimadores de máxima verossimilhança . . . 99
C.2 Intervalos de confiança . . . 101
C.3 Resíduos . . . 110
C.4 Influência global . . . 112
C.5 Seleção de modelos . . . 117

D Programas para o MRBC: Abordagem Bayesiana **119**
D.1 Gibbs com passo de Metropolis 119
D.2 CPO e valores preditos via CPO 122
D.3 Resíduos Bayesianos 123
D.4 Divergência de K-L 127

E Programas para o MRBCAEN **130**
E.1 Estimadores de máxima verossimilhança 130
E.2 Intervalos de confiança 132

Capítulo 1

Introdução

É comum, em situações práticas, a presença de conjuntos de dados envolvendo frequências de eventos, $Y_1, Y_2, \ldots, Y_m$, que ocorrem respectivamente em m clusters independentes, com a frequência Y_i baseada na soma de variáveis aleatórias dependentes de Bernoulli W_{ij}, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n_i$, $Y_i = \sum_{j=1}^{n_i} W_{ij}$, em que W_{ij} é um indicador do status do evento de interesse do j-ésimo indivíduo no i-ésimo cluster e n_i é o número de indivíduos no i-ésimo cluster. Em algumas situações, para cada cluster, estão disponíveis os valores de um conjunto de k covariáveis, $x_{i1}, x_{i2}, \ldots, x_{ik}$ e, para cada indivíduo dentro do cluster, o valor de q covariáveis $r_{i11}, \ldots, r_{i1n_i}, r_{i21}, \ldots, r_{i2n_i}, r_{iq1}, \ldots, r_{iqn_i}$. O interesse é modelar o comportamento da variável resposta, Y_i, como uma função de k covariáveis, $x_{i1}, x_{i2}, \ldots, x_{ik}$. Para este fim, propomos uma classe de modelos de regressão binomial correlacionada.

Kupper & Haseman (1978) desenvolveram o modelo binomial correlacionada. Essa distribuição foi obtida por meio de uma correção do modelo binomial convencional por um método proposto por Bahadur (1961) para inserir a dependência entre as variáveis de Bernoulli. Paul (1985, 1987) propôs uma distribuição binomial com três parâmetros que generaliza as distribuições binomial, beta-binomial e binomial correlacionada de Kupper & Haseman (1978). Ng (1989) propôs a distribuição binomial modificada, na qual a distribuição binomial convencional é modificada sequencialmente e a distribuição resultante torna-se mais dispersa (indicando correlação positiva entre as variáveis de Bernoulli), ou menos dispersa (indicando correlação negativa entre as variáveis de Bernoulli), que a distribuição binomial convencional. Fu & Sproule (1995) derivaram uma distribuição binomial com quatro parâmetros, permitindo que os ensaios de Bernoulli assumam valores α ou β, com $\alpha, \beta \in \mathbb{R}$ e $\alpha < \beta$, em vez dos valores usuais 0 ou 1. Yu & Zelterman (2002) desenvolveram uma nova distribuição discreta que descreve o comportamento da soma de variáveis aleatórias de Bernoulli dependentes. Tsai *et al.* (2003) apresentam um modelo que estuda a taxa global de erro ao testar hipóteses múltiplas. Este modelo envolve a distribuição da soma de ensaios de Bernoulli dependentes e esta distribuição é aproximada utilizando uma estrutura beta-binomial. Ao invés de utilizar o modelo beta-binomial, Gupta & Tao (2010) derivaram a distribuição exata da soma de variáveis aleatórias de Bernoulli dependentes e não identicamente distribuída e forneceram uma

aplicação. Luceño (1995) e Luceño & Ceballos (1995) propuseram uma distribuição binomial generalizada, que é discutida em detalhes em Diniz *et al.* (2010) e Salinas & Kolev (2003).

Existem alguns modelos de regressão na literatura que podem ser utilizados para encontrar a probabilidade de sucesso. Por exemplo, se assumimos que as variáveis aleatórias de Bernoulli dentro do cluster têm correlação comum e que o parâmetro de correlação é comum em todos os clusters, um modelo de regressão linear logístico (Williams, 1982) pode ser aplicado. Um modelo de regressão probito para modelar observações binárias equicorrelacionadas é descrito em Ochi & Prentice (1984). Regressão para dados binomial superdispersos é apresentado em Prentice (1986), que estende a distribuição beta-binomial permitindo correlação negativa entre variáveis binárias dentro do cluster, e propõe modelos de regressão para a taxa da variável resposta binária e para a correlação entre duas variáveis binárias. Prentice (1988) propôs métodos de regressão para análise de dados binários correlacionados quando cada observação binária pode ter sua própria covariável. Lindsey & Altham (1998) discutem e comparam três diferentes generalizações da distribuição binomial, beta-binomial (Skellam, 1948), binomial dupla (Efron, 1986; Lindsey, 1995) e binomial multiplicativa (Altham, 1978), que tratam de subdispersão ou superdispersão. Para cada distribuição, a probabilidade e o parâmetro de dispersão podem variar simultaneamente no que diz respeito a um conjunto de covariáveis de acordo com duas equações de regressão separadas.

Nesta tese, a distribuição binomial generalizada (Luceño, 1995; Luceño & Ceballos, 1995) é utilizada para construir o modelo de regressão binomial correlacionada, no qual a estrutura de regressão é modelada utilizando uma das quatro funções de ligação (logito, complementar log-log, log-log e probito) capazes de conectar a probabilidade de sucesso, p_i, a uma função de interesse das covariáveis associadas aos clusters. A dependência entre indivíduos de um mesmo cluster é modelada usando estruturas de correlação (Jennrich & Schluchter, 1986; Cressie, 1993; Zimmerman & Harville, 1991; Russell, 1996), levando em conta as covariáveis dos indivíduos dentro do cluster. O modelo de regressão binomial correlacionada permite apenas a presença de correlação positiva entre os indivíduos. Luceño & Ceballos (1995) e Kolev & Paiva (2008) apresentam uma motivação biológica e uma justificativa teórica, respectivamente, para esta restrição. A complexidade da função de verossimilhança, devido a presença de produtos de somas, é superada com uma estratégia de dados aumentados (Diebolt & Robert, 1994; Tanner & Wong, 1987).

Dois métodos de estimação, clássico e Bayesiano, são considerados para o modelo de regressão binomial correlacionada. Os dois próximos parágrafos descrevem em detalhes os pontos principais discutidos em cada método.

Na abordagem clássica, os estimadores de máxima verossimilhança e intervalos de confiança assintóticos são calculados para os $(k + 2)$ parâmetros presentes no modelo de regressão binomial correlacionada, levando em consideração a função de verossimilhança com dados observados e com dados aumentados. Além destes, também são construídos intervalos de confiança baseados em reamostragem e na log-verossimilhança perfilada. Testes de hipóteses, via teste da razão de verossimilhança e teste de Wald, são desen-

volvidos para subvetores dos parâmetros de regressão do modelo. Um teste de razão de verossimilhança é derivado para avaliar a plausibilidade da presença do parâmetro da estrutura de correlação. A análise de diagnóstico envolvendo análise de resíduos e medidas de influência global é construída. Os resíduos baseados nos valores preditos e os resíduos *deviance* são sugeridos para checar suposições do modelo. A distância de Cook generalizada e a distância da verossimilhança são consideradas como medidas de influência global. Dois critérios de seleção, AIC (Akaike, 1974) e BIC (Schwarz, 1978), são utilizados para selecionar variações do modelo proposto. Um estudo de simulação é realizado para ilustrar o processo de estimação e o desempenho das medidas de diagnóstico propostas. Além disso, um segundo estudo de simulação é apresentado para analisar as propriedades frequentistas dos estimadores de máxima verossimilhança e intervalos de confiança propostos neste trabalho.

Na bordagem Bayesiana, determinamos a distribuição a *posteriori* marginal dos $(k+2)$ parâmetros por meio de amostras das distribuições condicionais completas. A metodologia utilizada considera a função de verossimilhança aumentada, por meio da introdução de variáveis latentes, e distribuições a *priori* vagas para os parâmetros do modelo. Dando continuidade à construção da metodologia, definimos a distribuição preditiva a *posteriori*, a densidade preditiva condicional ordinária (CPO) e sugerimos um algoritmo para obter numericamente o valor predito para uma observação futura e para uma observação presente no conjunto de dados, respectivamente. Desenvolvemos uma análise de diagnóstico Bayesiano envolvendo análise de resíduos e medidas de influência local. Três tipos de resíduos são sugeridos, os resíduos baseado nos valores preditos pela densidade preditiva condicional ordinária (Cho *et al.*, 2009), os resíduos que dependem da distribuição a *posteriori* dos parâmetros do modelo (Albert & Chib, 1995) e os resíduos *deviance* Bayesianos (Spiegelhalter *et al.*, 2002). A calibração da divergência de Kullback-Leibler (Cho *et al.*, 2009; Cook & Weisberg, 1982) é utilizada como medida de influência local. O critério de seleção de modelo adotado é o critério de informação *deviance*, DIC, (Spiegelhalter *et al.*, 2002). Um estudo de simulação é realizado para ilustrar o processo inferencial, a sensibilidade na inferência dos parâmetros com diferentes distribuições a *priori* vagas e o desempenho das medidas de diagnóstico Bayesiano proposto. Um outro estudo de simulação é apresentado com as propriedades frequentistas dos estimadores Bayesianos utilizando a metodologia desenvolvida.

O objetivo principal desta tese é propor um modelo alternativo que permita ajustar dados binomiais com dependência entre os eventos de Bernoulli, quando estão disponíveis covariáveis dos clusters e dos indivíduos dentro dos clusters. No entanto, é usual a comparação de uma nova proposta com alguma outra frequentemente adotada na literatura. Desta forma, a escolha natural é o modelo beta-binomial, para o qual apresentamos uma análise de diagnóstico mais minuciosa, envolvendo análise de resíduos e medidas de influência para detecção de *outliers* e/ou observações influentes considerando uma abordagem Bayesiana. Dois conjuntos de dados reais são analisados utilizando a metodologia proposta para o modelo de regressão binomial correlacionada e os resultados são comparados com os fornecidos pelo modelo de regressão beta-binomial.

Na construção do modelo de regressão binomial correlacionada é suposto que as covariáveis disponíveis são observadas sem erro, ou que o erro existente é desprezível. Porém, pode ser de interesse do analista considerar uma covariável medida com erro no ajuste do modelo proposto. Neste cenário, não seria aconselhável conduzir a análise considerando o modelo inicialmente proposto, cujo ajuste poderia apresentar um vício na estimativa do parâmetro correspondente à covariável medida com erro, induzindo a inferências incorretas.

Motivados por esta necessidade, uma extensão do modelo de regressão binomial correlacionada é proposto considerando um cenário especifico. Ou seja, é construído um modelo de regressão binomial correlacionada aditivo estrutural normal, em que a covariável não observada segue uma distribuição normal com média μ e variância σ^2, e a covariável de fato observada pode ser expressa de forma aditiva com o erro de medida, que por sua vez segue uma distribuição normal com média zero e variância σ^2.

A construção do modelo de regressão binomial correlacionada aditivo estrutural normal também está baseada na distribuição binomial generalizada (Luceño, 1995; Luceño & Ceballos, 1995). Nesta modelagem é considerada apenas a função de ligação logito para conectar o parâmetro da probabilidade de sucesso a uma função de interesse das covariáveis dos clusters. A dependência entre indivíduos de um mesmo cluster é modelada usando estruturas de correlação (Jennrich & Schluchter, 1986; Cressie, 1993; Zimmerman & Harville, 1991; Russell, 1996), levando em conta as covariáveis dos indivíduos dentro do cluster. A complexidade da função de verossimilhança, devido a presença de produtos de somas e a integração da variável não observada, é superada com uma estratégia de dados aumentados (Diebolt & Robert, 1994; Tanner & Wong, 1987) e com uma aproximação de Laplace (Bernardo & Smith, 2000; Tanner, 1996), respectivamente. Uma metodologia clássica é desenvolvida para o modelo de regressão binomial correlacionada aditivo estrutural normal de forma similar à apresentada para o modelo de regressão binomial correlacionada.

Organização dos capítulos

Além deste primeiro capítulo, envolvendo a introdução do trabalho e a descrição sumarizada das metodologias desenvolvidas para os modelos propostos, este trabalho está organizado em mais seis capítulos descritos a seguir.

No Capítulo 2 apresentamos o desenvolvimento do modelo de regressão binomial correlacionada, no qual é inserida uma estrutura de regressão com as covariáveis dos clusters e uma estrutura de correlação com as covariáveis dos indivíduos. A função de verossimilhança é construída por meio da inserção de uma variável latente, tornando-a identificável.

A abordagem clássica é construída no Capítulo 3, iniciando com o processo de estimação via algoritmo EM. Apresentamos, também, os intervalos de confiança assintóticos, os intervalos de confiança perfilados e os intervalos de confiança *bootstrap*, testes de hipóteses para subvetores dos parâmetros de regressão do modelo e para o parâmetro da estrutura de correlação, uma análise de diagnóstico envolvendo análise de resíduos e medidas de influência global e critérios de seleção de modelos. Por fim apresentamos um

estudo de simulação ilustrando a metodologia clássica proposta.

No Capítulo 4 desenvolvemos a abordagem Bayesiana para o modelo, incluindo uma análise de diagnóstico com análise de resíduos e medidas de diagnóstico de influência local e um critério de seleção de modelos. Um estudo de simulação ilustrando a abordagem Bayesiana também é apresentado.

Consideramos, no Capítulo 5, uma análise de diagnóstico para o modelo de regressão beta-binomial, envolvendo análise de resíduos e medidas de influência para detecção de *outliers* e/ou observações influentes via abordagem Bayesiana. Os resultados obtidos no ajuste do modelo beta-binomial em dois conjuntos de dados reais são comparados aos resultados obtidos pelo modelo proposto nesta tese.

No Capítulo 6 uma extensão do modelo de regressão binomial correlacionada, modelo de regressão binomial correlacionada aditivo estrutural normal, é desenvolvida. Neste caso, é inserida uma estrutura de regressão com as covariáveis dos clusters, sendo uma delas não observada diretamente, e uma estrutura de correlação com as covariáveis dos indivíduos. A abordagem clássica para o modelo de regressão binomial correlacionada aditivo estrutural normal é construída neste capítulo, iniciando com o processo de estimação via algoritmo EM. Apresentamos também, neste capítulo, os intervalos de confiança assintóticos, baseados na função de verossimilhança aumentada e via reamostragem, uma análise de diagnóstico, envolvendo uma análise de resíduos e medidas de influência global, e critérios de seleção de modelos. Por fim, um estudo de simulação, ilustrando a metodologia clássica proposta e uma análise de dados reais, é apresentada.

No Capítulo 7 apresentamos as considerações finais e as propostas futuras.

Capítulo 2

Modelos de regressão binomial correlacionada (MRBC)

A distribuição binomial generalizada derivada por Luceño (1995), $BC(n, p, \rho)$, é obtida supondo que a variável Y, o número de sucessos em n ensaios de Bernoulli, é a soma de respostas binárias equicorrelacionadas com probabilidade de sucesso constante p e coeficiente de correlação comum ρ. Formalmente, seja $Y = \sum_{j=1}^{n} W_j$, em que $E(W_j) = p$, $\text{Var}(W_j) = p(1-p)$, $j = 1, \ldots, n$, e $\text{Corr}(W_s, W_t) = \rho$, para todo s e t, $s \neq t$. O caso em que $\rho > 0$ implica em $P(W_s = 1|W_t = 1) > P(W_s = 1)$, para todo s e t, $s \neq t$. Uma forma de expressar a distribuição de probabilidade de Y é por meio da mistura de duas variáveis: uma seguindo uma distribuição binomial, $B(n, p)$, com probabilidade da mistura $(1-\rho)$, e outra seguindo uma distribuição Bernoulli modificada, $BernM(p)$, assumindo valores zero ou n (Fu & Sproule, 1995), ao invés dos convencionais zero ou um, com probabilidade da mistura ρ. A distribuição de probabilidade de Y, dados n, p e ρ, é dada por

$$P(Y = y|y, n, p, \rho) = \binom{n}{y} p^y (1-p)^{n-y}(1-\rho) I_{A_1}(y) + p^{\frac{y}{n}}(1-p)^{\frac{n-y}{n}} \rho I_{A_2}(y), \qquad (2.1)$$

em que $A_1 = \{0, 1, \ldots, n\}$, $A_2 = \{0, n\}$, $y = 0, 1, \ldots, n$, $n \in \mathbb{N} - \{0\}$, $0 < p < 1$ e $0 \leq \rho \leq 1$. A média e variância deste modelo são $E(Y) = np$ e $\text{Var}(Y) = p(1-p)\{n + \rho\, n(n-1)\}$, que acomoda a variação extra-binomial quando $\rho \neq 0$. Note que o modelo $CB(n, p, \rho)$ é equivalente ao modelo binomial se $\rho = 0$.

Os modelos de regressão binomial correlacionada são construídos assumindo uma estrutura de regressão na distribuição binomial correlacionada (2.1), por meio de uma das quatro diferentes funções de ligação (logito, complementar log-log, log-log e probito).

2.1 Uma classe de modelos de regressão binomial correlacionada

Suponha que $Y_1, Y_2, \ldots, Y_m$ são variáveis aleatórias independentes tal que Y_i segue uma distribuição binomial correlacionada $CB(n_i, p_i, \rho_i)$. A variável resposta Y_i assume valores $0, 1, \ldots, n_i$, correspondendo ao número de indivíduos com o evento de interesse dentro do

i-ésimo cluster, isto é, $Y_i = \sum_{j=1}^{n_i} W_{ij}$, $i = 1 \ldots, m$, com $E(W_{ij}) = p_i$, $\text{Var}(W_{ij}) = p_i(1-p_i)$ e $\text{Corr}(W_{is}, W_{it}) = \rho_i$, para todo s e t, $s \neq t$. A média e a variância das variáveis resposta Y_i são dadas por $n_i p_i$ e $p_i(1-p_i)\{n_i + \rho_i\ n_i(n_i-1)\}$, respectivamente.

É conveniente, como a construção da função de verossimilhança para dados binários presente em McCullagh & Nelder (1989), considerarmos, inicialmente, a função de verossimilhança como uma função de vetores m-dimensionais $\boldsymbol{p} = (p_1, p_2, \ldots, p_m)^\top$ e $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_m)^\top$. Posteriormente, a função de verossimilhança é reescrita em função dos coeficientes $\beta_0, \ldots, \beta_k$, associados às covariáveis dos clusters, e o coeficiente γ, associado à estrutura de correlação. Assumindo que $y_1, y_2, \ldots, y_m$ são realizações de $Y_1, Y_2, \ldots, Y_m$, respectivamente, e usando (2.1), a função de verossimilhança pode ser escrita como

$$\mathcal{L}(\boldsymbol{p}, \boldsymbol{\rho}; m, \boldsymbol{n}, \boldsymbol{y}) = \prod_{i=1}^{m} \left\{ \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i - y_i}(1-\rho_i) + p_i^{\frac{y_i}{n_i}}(1-p_i)^{\frac{(n_i-y_i)}{n_i}} \rho_i I_{A_{2_i}}(y_i) \right\}, \quad (2.2)$$

em que $A_{2_i} = \{0, n_i\}$, $y_i = 0, ..., n_i$, $n_i \in \mathbb{N} - \{0\}$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)^\top$, $\boldsymbol{n} = (n_1, n_2, \ldots, n_m)^\top$, $0 < p_i < 1$ e $0 \leq \rho_i \leq 1$, com $i = 1, ..., m$.

Tabela 2.1: Funções de ligação para modelar p_i.

Função de ligação	$g^{-1}(\eta_i)$
Logito	$\exp\{\eta_i\} \,/\, [1 + \exp\{\eta_i\}]$
Complementar log-log	$1 - \exp\{-\exp\{\eta_i\}\}$
Log-log	$\exp\{-\exp\{-\eta_i\}\}$
Probito	$\Phi(\eta_i)$

Como nosso maior interesse é definir o modelo de regressão binomial correlacionada, é útil modelar a probabilidade de sucesso, p_i, e o parâmetro de correlação, ρ_i. Os p_i são modelados utilizando covariáveis comuns dos clusters, estas covariáveis são conectadas à probabilidade de sucesso do cluster por meio de funções de ligação que retornam valores no intervalo (0,1). As funções de ligação $g^{-1}(\eta_i)$ (logito, complementar log-log, log-log e probito) consideradas neste trabalho estão especificadas na Tabela 2.1, em que $\Phi(.)$ é a função de distribuição acumulada da distribuição Normal padrão; $\eta_i = \sum_{r=0}^{k} \beta_r x_{ir}$; os coeficientes $\beta_0, \beta_1, \ldots, \beta_k$ são parâmetros desconhecidos; $x_{i0} = 1$, para todo i e $x_{1i}, x_{2i}, \ldots, x_{ik}$ são os valores das k covariáveis para o i-ésimo cluster.

A estrutura de correlação é modelada considerando uma função das covariáveis específicas dos indivíduos que são capazes de relacionar a dependência entre eles em relação ao evento de interesse. De uma forma geral, a estrutura de correlação pode ser escrita como

$$\rho_i = h(v(\boldsymbol{r}_i), \gamma), \quad (2.3)$$

em que $h(v(\boldsymbol{r}_i), \gamma)$, a correlação entre quaisquer dois indivíduos dentro do i-ésimo cluster, é uma função não linear apropriada, monotônica e duas vezes diferenciável; $v(\boldsymbol{r}_i)$ representa o valor de uma função das covariáveis dos indivíduos, assumindo valores positivos; $\boldsymbol{r}_i = (r_{i11}, \ldots, r_{i1n_i}, r_{i21}, \ldots, r_{i2n_i}, r_{iq1}, \ldots, r_{iqn_i})^\top$, com r_{ilj} representando os valores da l-ésima

covariável para o j-ésimo indivíduo dentro do i-ésimo cluster, $i = 1, \ldots, m$, $l = 1, \ldots, q$ e $j = 1, \ldots, n_i$; γ é o parâmetro que determina a taxa de decaimento da correlação como uma função de $v(\boldsymbol{r}_i)$ (Sherman, 2011). Usando a idéia de estrutura de correlação espacial, possíveis escolhas da função $v(\boldsymbol{r}_i)$, podem ser, por exemplo, funções contínuas de soma de distâncias entre posições de vetores ou entre outros vetores avaliados os quais são capazes de relacionar a dependência entre os indivíduos dentro do cluster. Assim, candidatos para $v(\boldsymbol{r}_i)$, usando apenas duas covariáveis, r_{i1} e r_{i2}, podem ser a distância Euclidiana, definida como $\sqrt{\sum_{l=1,2}\sum_s\sum_{s<t}(r_{ils}-r_{ilt})^2}$, a distância Manhattan, definida como $\sum_{l=1,2}\sum_s\sum_{s<t}|r_{ils}-r_{ilt}|$, a distância máxima, definida como $\max_{s,t}|r_{i1s}-r_{i1t}|$, ou a distância mínima dada por $\min_{s,t}|r_{i2s}-r_{i2t}|$, com $s,t = 1, \ldots, n_i$. A Tabela 2.2 apresenta, para o modelo de regressão binomial correlacionada, diferentes opções para $h(v(\boldsymbol{r}_i), \gamma)$ (Cressie, 1993; Sherman, 2011).

Tabela 2.2: Algumas alternativas para h.

Estrutura de correlação	$h(v(\boldsymbol{r}_i), \gamma)$
Exponencial$^{(a)}$	$\exp\{-\gamma v(\boldsymbol{r}_i)\}$
Gaussiana$^{(a)}$	$\exp\{-[\gamma v(\boldsymbol{r}_i)]^2\}$
AR Contínua$^{(b)}$	$\gamma^{v(\boldsymbol{r}_i)}$

Nos casos (a): $\gamma > 0$ e em (b): $\gamma \in (0, 1)$.

Sejam $\boldsymbol{n} = (n_1, n_2, \ldots, n_m)^\top$, $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m)^\top$, $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ik})^\top$ e $\boldsymbol{r} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m)^\top$, $\boldsymbol{r}_i = (r_{i11}, \ldots, r_{i1n_i}, r_{i21}, \ldots, r_{i2n_i}, r_{iq1}, \ldots, r_{iqn_i})^\top$. Usando $g^{-1}(\eta_i)$ e (2.3), a função de verossimilhança (2.2) do vetor de parâmetros $\boldsymbol{\theta} = (\gamma, \beta_0, \ldots, \beta_k)^\top$, condicionado aos dados observados, $\mathcal{D} = (m, \boldsymbol{n}, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{r})^\top$, se expressa como

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) &= \prod_{i=1}^{m}\left\{\binom{n_i}{y_i} g^{-1}(\eta_i)^{y_i}(1-g^{-1}(\eta_i))^{n_i-y_i}\,(1-h(v(\boldsymbol{r}_i),\gamma))\right. \\ &\quad \left. + g^{-1}(\eta_i)^{\frac{y_i}{n_i}}(1-g^{-1}(\eta_i))^{\frac{n_i-y_i}{n_i}}\, h(v(\boldsymbol{r}_i),\gamma) I_{A_{2i}}(y_i)\right\}. \qquad (2.4)\end{aligned}$$

Modelos envolvendo misturas de distribuições apresentam problema de identificabilidade, decorrente da falta de informação da origem da resposta observada que pode ser proveniente de qualquer subpopulação, tornando não bem definido os métodos usuais de estimação (Titterington *et al.*, 1985). Entretanto, uma estratégia de dados aumentados (Tanner & Wong, 1987; Diebolt & Robert, 1994) é construída para este problema com o objetivo de tornar a função de verossimilhança identificável. A função de verossimilhança com dados aumentados (Tanner & Wong, 1987; Diebolt & Robert, 1994) para este modelo é apresentada na próxima seção.

2.2 Função de verosimilhança com dados aumentados

Para a construção do modelo de regressão binomial correlacionada, supomos que y_i é uma realização da distribuição binomial ou da distribuição Bernoulli modificada. Porém,

no processo de observação dos dados não somos capazes de identificar de qual distribuição y_i é proveniente, caracterizando uma observação incompleta. Com a intenção de suprir esta informação faltante, introduzimos uma variável latente Z_i, $i = 1, \ldots, m$, que indica de qual componente do modelo $CB(n_i, p_i, \rho_i)$ a observação y_i, $i = 1, \ldots, m$, é originária, ou seja,

$$Z_i = \begin{cases} 1, & \text{se a observação } y_i \text{ resulta de } MBern(p_i) \\ 0, & \text{se a observação } y_i \text{ resulta de } B(n_i, p_i) \end{cases}.$$

A probabilidade de sucesso da variável aleatória Z_i, condicionada a observação y_i, pode ser escrita como

$$\begin{aligned} \tau_i &= P(Z_i = 1|Y_i = y_i, n_i, p_i, \rho_i) = \frac{P(Y_i = y_i|Z_i = 1)P(Z_i = 1)}{P(Y_i = y_i)} \\ &= \frac{\rho_i p_i^{\frac{y_i}{n_i}}(1-p_i)^{\frac{n_i-y_i}{n_i}} I_{A_{2_i}}(y_i)}{\rho_i p_i^{\frac{y_i}{n_i}}(1-p_i)^{\frac{n_i-y_i}{n_i}} I_{A_{2_i}}(y_i) + (1-\rho_i)\binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i}}, \end{aligned} \tag{2.5}$$

em que $A_{2_i} = \{0, n_i\}$, $y_i = 0, ..., n_i$, $0 < p_i < 1$, $0 \leq \rho_i \leq 1$, $i = 1, ..., m$.

Assim,

$$\begin{aligned} &P(Z_i = z_i|Y_i = y_i, n_i, p_i, \rho_i) = \frac{\tau_i^{z_i}(1-\tau_i)^{1-z_i}}{P(Y_i = y_i)} \\ &= \frac{\left(\rho_i p_i^{\frac{y_i}{n_i}}(1-p_i)^{\frac{n_i-y_i}{n_i}} I_{A_{2_i}}(y_i)\right)^{z_i} \left((1-\rho_i)\binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i}\right)^{1-z_i}}{\rho_i p_i^{\frac{y_i}{n_i}}(1-p_i)^{\frac{n_i-y_i}{n_i}} I_{A_{2_i}}(y_i) + (1-\rho_i)\binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i}}, i = 1, \ldots, m. \end{aligned}$$

A função de verossimilhança completa com a presença da variável latente $\boldsymbol{z} = (z_1, z_2, \ldots, z_m)^\top$, é, então, definida como

$$\mathcal{L}(\boldsymbol{p}, \boldsymbol{\rho}; \boldsymbol{n}, \boldsymbol{y}, \boldsymbol{z}) = \left(\prod_{i=1}^{m} P(y_i|n_i, p_i, \rho_i)\right)\left(\prod_{i=1}^{m} P(z_i|y_i, n_i, p_i, \rho_i)\right)$$

$$= \prod_{i=1}^{m}\left\{\binom{n_i}{y_i}^{(1-z_i)} p_i^{\frac{y_i}{n_i}(z_i+n_i-n_i z_i)}(1-p_i)^{(n_i-y_i)\left(\frac{z_i}{n_i}+1-z_i\right)}\rho_i^{z_i}(1-\rho_i)^{(1-z_i)} I_{A_{2_i}}(y_i)^{z_i}\right\}. \tag{2.6}$$

A função de verossimilhança para $\boldsymbol{\theta} = (\gamma, \beta_0, \ldots, \beta_k)^\top$, condicionado aos dados completos $\mathcal{D}^* = (m, \boldsymbol{n}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{r})^\top$, é dada por

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}^*) &= \prod_{i=1}^{m}\left\{\binom{n_i}{y_i}^{(1-z_i)} g^{-1}(\eta_i)^{\frac{y_i}{n_i}(z_i+n_i-n_i z_i)}(1-g^{-1}(\eta_i))^{(n_i-y_i)\left(\frac{z_i}{n_i}+1-z_i\right)}\right. \\ &\quad \left. \times\ h(v(\boldsymbol{r}_i), \gamma)^{z_i}\left(1-h(v(\boldsymbol{r}_i), \gamma)\right)^{(1-z_i)} I_{A_{2_i}}(y_i)^{z_i}\right\}. \end{aligned} \tag{2.7}$$

A função de verossimilhança (2.7), baseada nos dados latentes, possibilita a utilização de um processo inferencial bem definido.

Usando a parametrização $\phi = \log(\gamma)$, para a estrutura de correlação exponencial ou Gaussiana, ou $\phi = \log(\gamma/(1-\gamma))$, para a estrutura de correlação AR contínua, o parâmetro γ pode ser estimado sem restrição. Assim, a função de log-verossimilhança para $\boldsymbol{\theta}^* = (\phi, \beta_0, \ldots, \beta_k)^\top$, condicionado aos dados completos $\mathcal{D}^* = (m, \boldsymbol{n}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{r})^\top$, é dada por

$\ell(\boldsymbol{\theta}^*; \mathcal{D}^*) \propto$

$$\sum_{i=1}^{m} \left\{ \frac{y_i}{n_i} \left(z_i + n_i - n_i z_i\right) \log(g^{-1}(\eta_i)) + (n_i - y_i) \left(\frac{z_i}{n_i} + 1 - z_i \right) \right.$$

$$\left. \times \log(1 - g^{-1}(\eta_i)) + z_i \log(h^*(v(\boldsymbol{r}_i), \phi)) + (1 - z_i) \log\left(1 - h^*(v(\boldsymbol{r}_i), \phi)\right) \right\}, \quad (2.8)$$

no qual $h^*(v(\boldsymbol{r}_i), \phi)$ é similar a função $h(v(\boldsymbol{r}_i), \gamma)$, apresentada na Tabela 2.2, considerando a parametrização $\gamma = \exp\{\phi\}/(1 + \exp\{\phi\})$ ou $\gamma = \exp\{\phi\}$.

Um procedimento clássico para inferência dos parâmetros presente no modelo de regressão binomial correlacionada é apresentado no Capítulo 3 e uma abordagem Bayesiana no Capítulo 4.

Capítulo 3

MRBC: Metodologia Clássica

Neste capítulo apresentamos uma abordagem clássica para o modelo de regressão binomial correlacionada. Os estimadores de máxima verossimilhança dos parâmetros são determinados via algoritmo EM. Intervalos de confiança são construídos para os $(k+2)$ parâmetros do modelo. Testes de hipóteses são desenvolvidos para testar a plausibilidade de possíveis valores para o vetor e para subvetores dos coeficientes de regressão do modelo e para testar a plausibilidade do parâmetro presente na estrutura de correlação. Uma análise de diagnóstico é desenvolvida envolvendo análise de resíduos e medidas de influência global. É apresentado ainda dois critérios de seleção de modelos. A metodologia desenvolvida é ilustrada através de conjuntos de dados simulados. No Capítulo 5.5 a análise envolve um conjunto de dados reais.

3.1 Estimação via algoritmo EM

Uma metodologia usual para a estimação dos parâmetros com a presença de dados latentes é o algoritmo EM (Tanner, 1996). Este método consiste de duas etapas, a esperança, passo E, e a maximização, passo M. O algoritmo determina no passo E o valor esperado da função de log-verossimilhança com dados completos, $\ell(\boldsymbol{\theta}^*|\mathcal{D}^*)$, considerando o conjunto de dados observado $\mathcal{D}$ e um valor inicial para o vetor de parâmetros $\boldsymbol{\theta}^*$, $\boldsymbol{\theta}^{(0)}$.

Como a variável latente Z_i segue uma distribuição de Bernoulli com parâmetro τ_i, então $E(Z_i) = \tau_i$, $i = 1\ldots, m$, expresso em (2.5). Assim, o passo E é dado por

$$E\left(\ell(\boldsymbol{\theta}^*;\mathcal{D}^*)|\boldsymbol{\theta}^{(0)},\mathcal{D}\right) \propto$$

$$\sum_{i=1}^{m}\left\{\frac{y_i}{n_i}\left(\tau_i + n_i - n_i\tau_i\right)\log\left(g^{-1}\left(\sum_{r=0}^{k}\beta_r x_{ir}\right)\right) + (1-\tau_i)\log\left(1-h^*(v(\boldsymbol{r}_i),\phi)\right)\right.$$

$$\left.+(n_i-y_i)\left(\frac{\tau_i}{n_i}+1-\tau_i\right)\log\left(1-g^{-1}\left(\sum_{r=0}^{k}\beta_r x_{ir}\right)\right) + \tau_i\log(h^*(v(\boldsymbol{r}_i),\phi))\right\}, \quad (3.1)$$

com $\ell(\boldsymbol{\theta}^*|\mathcal{D}^*)$ dada em (2.8), $p_i = g^{-1}\left(\sum_{r=0}^{k}\beta_r^{(0)}x_{ir}\right)$ e $\rho_i = h^*(v(\boldsymbol{r}_i),\phi^{(0)})$.

A função resultante do passo E depende do vetor de parâmetros, $\boldsymbol{\theta}^*$, dos dados observados, $\mathcal{D}$, e do vetor de variáveis latentes, $\boldsymbol{\tau} = (\tau_1^{(0)},\ldots,\tau_m^{(0)})^\top$. No passo M é feita a

maximização direta desta função em relação ao vetor de parâmetros $\boldsymbol{\theta}^*$, fornecendo, assim, uma atualização de $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\theta}^{(1)}$, e, consequentemente, uma atualização do vetor $\boldsymbol{\tau}^{(0)}$, $\boldsymbol{\tau}^{(1)}$. Os estimadores dos parâmetros são obtidos iterativamente, repetindo o passo E e o passo M, até a convergência dos parâmetros.

No Apêndice C.1 apresentamos a implementação usando o programa R (R Development Core Team, 2011) para obter os estimadores de máxima verossimilhança dos parâmetros.

3.2 Intervalos de confiança

Nesta seção, consideramos quatro formas de construção dos intervalos de confiança, duas baseadas nos intervalos de confiança assintóticos dos estimadores de máxima verossimilhança, uma por meio da função de verossimilhança com os dados observados e outra por meio da função de verossimilhança com os dados aumentados, uma baseada em reamostragem e uma baseada na função de log-verossimilhança perfilada.

No Apêndice C.2 apresentamos a implementação em R (R Development Core Team, 2011) dos intervalos de confiança considerados para os parâmetros do modelo de regressão binomial correlaciona.

3.2.1 Intervalos de confiança assintóticos

Nesta seção, os intervalos de confiança assintóticos dos estimadores de máxima verossimilhança são calculados para parâmetros presentes no modelo de regressão binomial correlacionada, levando em consideração a função de verossimilhança dos dados observados, expressa em (2.4), e a metodologia proposta por Louis (1982), baseada na função de verossimilhança com dados aumentados, expressa em (2.7).

Sob as condições de regularidade (ver Apêndice A), a distribuição assintótica para $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ é uma distribuição Normal multivariada $N_{k+2}(\boldsymbol{\theta}, I(\boldsymbol{\theta})^{-1})$ (Cox & Hinkley, 1979), em que $I(\boldsymbol{\theta})$ é a matriz de informação, que pode ser aproximada pela matriz de informação observada de Fisher, $J(\boldsymbol{\theta})$. Para o modelo de regressão binomial correlacionada, a matriz J, com dimensão $(k+2)\times(k+2)$, pode ser escrita como

$$J(\hat{\boldsymbol{\theta}}) = -\left.\frac{\partial^2 \ell(\boldsymbol{\theta};\mathcal{D})}{\partial\boldsymbol{\theta}\boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \tag{3.2}$$

cujos elementos são dados por

$$\frac{\partial^2 \ell(\boldsymbol{\theta};\mathcal{D})}{\partial\gamma^2} = \sum_{i=1}^{m}\left\{\frac{\partial^2 h(v(\boldsymbol{r}_i),\gamma)}{\partial\gamma^2}\frac{[-A+B]}{[C+D]} - \left(\frac{\partial h(v(\boldsymbol{r}_i),\gamma)}{\partial\gamma}\right)^2\frac{[-A+B]^2}{[C+D]^2}\right\},$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta};\mathcal{D})}{\partial\beta_r\partial\gamma} = \sum_{i=1}^{m}\Bigg\{\Bigg[\frac{\partial h(v(\boldsymbol{r}_i),\gamma)}{\partial\gamma}\Bigg[A\Bigg[-y_i(g^{-1}(\eta_i))^{-1}+(n_i-y_i)(1-g^{-1}(\eta_i))^{-1}\Bigg]$$
$$+B\Bigg[\frac{y_i}{n_i}(g^{-1}(\eta_i))^{-1}-\frac{(n_i-y_i)}{n_i}(1-g^{-1}(\eta_i))^{-1}\Bigg]\Bigg]\frac{\partial g^{-1}(\eta_i)}{\partial\beta_r}\Bigg][C+D]^{-1}$$
$$-[C+D]^{-2}\Bigg[\frac{\partial g^{-1}(\eta_i)}{\partial\beta_r}\Bigg[C\Bigg[y_i(g^{-1}(\eta_i))^{-1}-(n_i-y_i)(1-g^{-1}(\eta_i))^{-1}\Bigg]$$
$$+D\Bigg[\frac{y_i}{n_i}(g^{-1}(\eta_i))^{-1}-\frac{(n_i-y_i)}{n_i}(1-g^{-1}(\eta_i))^{-1}\Bigg]\Bigg]\Bigg[\frac{\partial h(v(\boldsymbol{r}_i),\gamma)}{\partial\gamma}[-A+B]\Bigg]\Bigg]\Bigg\}$$

e

$$\frac{\partial^2 \ell(\boldsymbol{\theta};\mathcal{D})}{\partial\beta_r\partial\beta_s} = \sum_{i=1}^{m}\Bigg\{\Bigg[\frac{\partial g^{-1}(\eta_i)}{\partial\beta_r}\frac{\partial g^{-1}(\eta_i)}{\partial\beta_s}\Bigg[C\Bigg[y_i^2(g^{-1}(\eta_i))^{-2}-y_i(g^{-1}(\eta_i))^{-2}$$
$$-2y_i(n_i-y_i)(g^{-1}(\eta_i))^{-1}(1-g^{-1}(\eta_i))^{-1}+(n_i-y_i)^2(1-g^{-1}(\eta_i))^{-2}$$
$$-(n_i-y_i)(1-g^{-1}(\eta_i))^{-2}\Bigg]+D\Bigg[\frac{y_i^2}{n_i^2}(g^{-1}(\eta_i))^{-2}-\frac{y_i}{n_i}(g^{-1}(\eta_i))^{-2}$$
$$-2\frac{y_i}{n_i^2}(n_i-y_i)(g^{-1}(\eta_i))^{-1}(1-g^{-1}(\eta_i))^{-1}+\frac{(n_i-y_i)^2}{n_i^2}(1-g^{-1}(\eta_i))^{-2}$$
$$-\frac{(n_i-y_i)}{n_i}(1-g^{-1}(\eta_i))^{-2}\Bigg]\Bigg]+\frac{\partial^2 g^{-1}(\eta_i)}{\partial\beta_r\beta_s}\Bigg[C\Bigg[y_i(g^{-1}(\eta_i))^{-1}-(n_i-y_i)$$
$$\times(1-g^{-1}(\eta_i))^{-1}\Bigg]+D\Bigg[\frac{y_i}{n_i}(g^{-1}(\eta_i))^{-1}-\frac{(n_i-y_i)}{n_i}(1-g^{-1}(\eta_i))^{-1}\Bigg]\Bigg]\Bigg]$$
$$\times[C+D]^{-1}-\Bigg[\frac{\partial g^{-1}(\eta_i)}{\partial\beta_r}\Bigg[C\Bigg[y_i(g^{-1}(\eta_i))^{-1}-(n_i-y_i)(1-g^{-1}(\eta_i))^{-1}\Bigg]$$
$$+D\Bigg[\frac{y_i}{n_i}(g^{-1}(\eta_i))^{-1}-\frac{(n_i-y_i)}{n_i}(1-g^{-1}(\eta_i))^{-1}\Bigg]\Bigg]^2\frac{\partial g^{-1}(\eta_i)}{\partial\beta_s}\Bigg][C+D]^{-2}\Bigg\},$$

$r,s=0,\ldots,k$, em que A, B, C e D são descritos na Tabela 3.1.

Tabela 3.1: Funções auxiliares no cálculo da matriz de informação de Fisher observada.

$A=\binom{n_i}{y_i}(g^{-1}(\eta_i))^{y_i}(1-g^{-1}(\eta_i))^{n_i-y_i}$
$B=(g^{-1}(\eta_i))^{\frac{y_i}{n_i}}(1-g^{-1}(\eta_i))^{\frac{n_i-y_i}{n_i}}I_{A_{2i}}(y_i)$
$C=\binom{n_i}{y_i}(g^{-1}(\eta_i))^{y_i}(1-g^{-1}(\eta_i))^{n_i-y_i}(1-h(v(\boldsymbol{r}_i),\gamma))$
$D=(g^{-1}(\eta_i))^{\frac{y_i}{n_i}}(1-g^{-1}(\eta_i))^{\frac{n_i-y_i}{n_i}}h(v(\boldsymbol{r}_i),\gamma)I_{A_{2i}}(y_i)$

As derivadas específicas para cada função de ligação e estrutura de correlação são apresentadas na Tabela 3.2 e na Tabela 3.3, respectivamente.

Outra forma de obter a matriz de informação de Fisher observada, utilizando a função de verossimilhança com dados completos (Louis, 1982), é considerando a matriz

$$J_a(\hat{\boldsymbol{\theta}}) = -E\left[\frac{\partial^2 \ell(\boldsymbol{\theta};\mathcal{D}^*)}{\partial\boldsymbol{\theta}\boldsymbol{\theta}^\top}\middle|\mathcal{D}\right]\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} - E\left[\frac{\partial \ell(\boldsymbol{\theta};\mathcal{D}^*)}{\partial\boldsymbol{\theta}}\frac{\partial \ell(\boldsymbol{\theta};\mathcal{D}^*)}{\partial\boldsymbol{\theta}}^\top\middle|\mathcal{D}\right]\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},$$

que pode ser aproximada via Monte Carlo por

$$J_{\hat{a}}(\hat{\boldsymbol{\theta}}) = -\frac{1}{Q}\sum_{q=1}^{Q}\frac{\partial^2 \ell(\boldsymbol{\theta};\mathcal{D}_q^*)}{\partial\boldsymbol{\theta}\boldsymbol{\theta}^\top}\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} - \frac{1}{Q}\sum_{q=1}^{Q}\frac{\partial \ell(\boldsymbol{\theta};\mathcal{D}_q^*)}{\partial\boldsymbol{\theta}}\frac{\partial \ell(\boldsymbol{\theta};\mathcal{D}_q^*)}{\partial\boldsymbol{\theta}}^\top\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \tag{3.3}$$

sendo $\mathcal{D}_q^* = (\mathcal{D}, \boldsymbol{z}_q)^\top$, $q = 1,\ldots,Q$, com $\boldsymbol{z}_q = (z_{1q},\ldots,z_{mq})^\top$ e $z_{iq} \sim Ber(\tau_i)$, $i = 1,\ldots,m$, com τ_i obtido em (2.5) considerando $p_i = g^{-1}(\sum_{r=0}^{k}\beta_r x_{ir})$ e $\rho_i = h(v(\boldsymbol{r}_i),\gamma)$, e

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\theta};\mathcal{D}_q^*)}{\partial\beta_r} &= \sum_{i=1}^{m}\Bigg\{\frac{y_i}{n_i}(z_{iq}+n_i-n_i z_{iq})\left[g^{-1}(\eta_{ir})\right]^{-1}\frac{\partial g^{-1}(\eta_{ir})}{\partial\beta_r} - (n_i - z_{iq})\left[\frac{z_{iq}}{n_i}+1-z_{iq}\right]\\ &\quad\times\left[1-g^{-1}(\eta_{ir})\right]^{-1}\frac{\partial g^{-1}(\eta_{ir})}{\partial\beta_r}\Bigg\},\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\theta};\mathcal{D}_q^*)}{\partial\beta_r\partial\beta_s} &= \sum_{i=1}^{m}\Bigg\{\left[\left[g^{-1}(\eta_{ir})\right]^{-1}\frac{\partial^2 g^{-1}(\eta_{ir})}{\partial\beta_r\partial\beta_s} - \left[g^{-1}(\eta_{ir})\right]^{-2}\frac{\partial g^{-1}(\eta_{ir})}{\partial\beta_r}\frac{\partial g^{-1}(\eta_{ir})}{\partial\beta_s}\right]\\ &\quad\times\frac{y_i}{n_i}(z_{iq}+n_i-n_i z_{iq}) - \Bigg[\left[1-g^{-1}(\eta_{ir})\right]^{-1}\frac{\partial^2 g^{-1}(\eta_{ir})}{\partial\beta_r\partial\beta_s} + \left[1-g^{-1}(\eta_{ir})\right]^{-2}\\ &\quad\times\frac{\partial g^{-1}(\eta_{ir})}{\partial\beta_r}\frac{\partial g^{-1}(\eta_{ir})}{\partial\beta_s}\Bigg](n_i-y_i)\left(\frac{z_{iq}}{n_i}+1-z_{iq}\right)\Bigg\},\end{aligned}$$

com $r, s = 0,\ldots,k$,

$$\frac{\partial \ell(\boldsymbol{\theta};\mathcal{D}_q^*)}{\partial\gamma} = \sum_{i=1}^{m}\left\{\frac{\partial h(v(\boldsymbol{r}_i),\gamma)}{\partial\gamma}\left[\frac{z_{iq}}{h(v(\boldsymbol{r}_i),\gamma)} - \frac{1-z_{iq}}{1-h(v(\boldsymbol{r}_i),\gamma)}\right]\right\}$$

e

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\theta};\mathcal{D}_q^*)}{\partial\gamma^2} &= \sum_{i=1}^{m}\Bigg\{\frac{\partial^2 h(v(\boldsymbol{r}_i),\gamma)}{\partial\gamma^2}\left[\frac{z_{iq}}{h(v(\boldsymbol{r}_i),\gamma)} - \frac{1-z_{iq}}{1-h(v(\boldsymbol{r}_i),\gamma)}\right]\\ &\quad+\left[\frac{\partial h(v(\boldsymbol{r}_i),\gamma)}{\partial\gamma}\right]^2\left[-\frac{z_{iq}}{h(v(\boldsymbol{r}_i),\gamma)^2} - \frac{1-z_{iq}}{(1-h(v(\boldsymbol{r}_i),\gamma))^2}\right]\Bigg\}.\end{aligned}$$

Devido a ortogonalidade entre os parâmetros, as derivadas $\dfrac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D}^*)}{\partial\boldsymbol{\beta}\partial\gamma}$ são todas iguais a zero. As derivadas específicas para cada função de ligação e estrutura de correlação são apresentadas nas Tabelas 3.2 e 3.3, respectivamente.

O intervalo de confiança assintótico, com nível de confiança $100\times(1-\alpha)\%$, para o r-ésimo componente do vetor de parâmetros $\boldsymbol{\theta}$, θ_r, $r = 1,\ldots,k+2$, pode ser calculado utilizando

$$\hat{\theta}_r \pm \mathcal{Z}_{\alpha/2}\sqrt{J^{-1}_{(r)}(\hat{\boldsymbol{\theta}})}\ , \tag{3.4}$$

em que $\mathcal{Z}_{\alpha/2}$ é o valor do $(\alpha/2)$-ésimo quantil superior da distribuição Normal padrão e $J^{-1}_{(r)}(\hat{\boldsymbol{\theta}})$ é o r-ésimo elemento da diagonal principal da inversa de $J(\hat{\boldsymbol{\theta}})$, ou $J_{\hat{a}}(\hat{\boldsymbol{\theta}})$, que corresponde ao estimador da variância do estimador de interesse.

Tabela 3.2: Derivadas, em relação aos parâmetros $\boldsymbol{\beta}$, para cada função de ligação em estudo.

Ligação	$\frac{\partial g^{-1}(\eta_i)}{\partial \beta_r}$	$\frac{\partial^2 g^{-1}(\eta_i)}{\partial \beta_r \partial \beta_s}$
Logito	$x_{ir}\exp\{\eta_i\}\left[1+\exp\{\eta_i\}\right]^{-2}$	$-x_{ir}x_{is}\exp\{\eta_i\}\left[\exp\{\eta_i\}-1\right]\left[1+\exp\{\eta_i\}\right]^{-3}$
C. log-log	$x_{ir}\exp\{\eta_i-\exp\{\eta_i\}\}$	$x_{ir}x_{is}\left[\exp\{\eta_i-\exp\{\eta_i\}\}-\exp\{2\eta_i-\exp\{\eta_i\}\}\right]$
Log-log	$x_{ir}\exp\{-\eta_i-\exp\{-\eta_i\}\}$	$-x_{ir}x_{is}\left[\exp\{-\eta_i-\exp\{-\eta_i\}\}-\exp\{-2\eta_i-\exp\{-\eta_i\}\}\right]$
Probito	$x_{ir}\phi(\eta_i)$	$-(2\pi)^{-\frac{1}{2}}x_{ir}x_{is}\eta_i\exp\{-(\eta_i)^2/2\}$

Em que $\phi(.)$ é a densidade da distribuição Normal padrão no ponto.

Tabela 3.3: Derivadas da estrutura de correlação, em relação ao parâmetro γ.

Estrutura	Exponencial	Gaussiana	AR contínua
$\frac{\partial h(v(\boldsymbol{r}_i),\gamma)}{\partial\gamma}$	$-v(\boldsymbol{r}_i)\exp\{-\gamma v(\boldsymbol{r}_i)\}$	$-2\gamma\left[v(\boldsymbol{r}_i)\right]^2\exp\{-\left[\gamma v(\boldsymbol{r}_i)\right]^2\}$	$v(\boldsymbol{r}_i)\lambda^{v(\boldsymbol{r}_i)}\lambda^{-1}$
$\frac{\partial^2 h(v(\boldsymbol{r}_i),\gamma)}{\partial\gamma^2}$	$\left[v(\boldsymbol{r}_i)\right]^2\exp\{-\gamma v(\boldsymbol{r}_i)\}$	$\frac{2\left[v(\boldsymbol{r}_i)\right]^2\exp\{-\left[\gamma v(\boldsymbol{r}_i)\right]^2\}}{\left[2(\gamma v(\boldsymbol{r}_i))^2-1\right]^{-1}}$	$\gamma^{v(\boldsymbol{r}_i)-2}v(\boldsymbol{r}_i)\left[v(\boldsymbol{r}_i)-1\right]$

A aproximação à distribuição normal para os estimadores de máxima verossimilhança dos parâmetros do modelo é válida assintoticamente, ou seja, quando a amostra é grande. Portanto, a construção destes intervalos de confiança podem não apresentar resultados acurados para amostras pequenas e moderadas. Neste sentido, uma duas formas de construção dos intervalos de confiança não-paramétricos são consideradas nas seções seguintes. A principal vantagem destes métodos é a sua capacidade de derivar intervalos que independem de estimativas da variância.

3.2.2 Intervalos de confiança *bootstrap*

O método *bootstrap* (Efron, 1979) é uma técnica de reamostragem utilizada para aproximar a distribuição teórica de uma variável aleatória por sua distribuição empírica. No processo de reamostragem pode ser considerada uma especificação paramétrica, em que novos conjuntos de dados são gerados, ou uma não-paramétrica, em que conjuntos de dados são construídos com base nos dados observados. Nesta seção, descrevemos como obter as estimativas intervalares empíricas via *bootstrap* não-paramétrico.

Os intervalos de confiança *bootstrap* não-paramétrico são obtidos simulando B amostras com reposição de tamanho m dos dados originais, $\mathcal{D}^{*(1)}$, $\mathcal{D}^{*(2)}, \ldots, \mathcal{D}^{*(B)}$. Para cada

reamostra, $\mathcal{D}^{*(b)}, b = 1, \ldots, B$, os estimadores de máxima verossimilhança dos parâmetros do modelo são obtidos conforme a Seção 3.1. Os intervalos com nível de confiança $100 \times (1 - \alpha)\%$, para cada um dos parâmetros, são construídos calculando os quantis $(1 - \alpha/2)$ e $(\alpha/2)$ dos respectivos B estimadores de máxima verossimilhança.

3.2.3 Intervalos de confiança perfilados

Nesta seção, intervalos de plausibilidade para os parâmetros do modelo de regressão binomial correlacionada são construídos com base na distribuição teórica da estatística de razão de log-verossimilhança perfilada (Kalbfleisch, 1985; Pawitan, 2001).

A função de verossimilhança perfilada para o r-ésimo componente do vetor de parâmetros $\boldsymbol{\theta}$, θ_r, $r = 1, \ldots, k + 2$, é definida como

$$\ell_r(\theta_r) = \max_{\theta_r}\{\ell(\hat{\boldsymbol{\theta}}_{(-r)}; \mathcal{D})\}, \tag{3.5}$$

sendo $\ell(\hat{\boldsymbol{\theta}}_{(-r)}; \mathcal{D}) = \log\{\mathcal{L}(\hat{\boldsymbol{\theta}}_{(-r)}; \mathcal{D})\}$ a função de log-verossimilhança, apresentada em (2.4), considerando $\hat{\boldsymbol{\theta}}_{(-r)}$ o vetor de parâmetros $\boldsymbol{\theta}$ assumindo $\hat{\boldsymbol{\theta}}$ exceto na r-ésima posição. A estatística de razão de log-verossimilhança para θ_r é dada por

$$W_r^* = 2\{\ell_r(\hat{\theta}_r) - \ell_r(\theta_r)\},$$

em que $\hat{\theta}_r$ é o estimador de máxima verossimilhança de θ_r e W_r^* segue uma distribuição χ^2_1. Ao utilizar a distribuição teórica, um intervalo de plausibilidade aproximado de $100(1 - \alpha)\%$ para θ_r pode ser obtido por

$$l_r(\theta_r) \geq \ell_r(\hat{\theta}_r) - \frac{1}{2}\chi^2_{1,(1-\alpha)}, \tag{3.6}$$

em que $\chi^2_{1,(1-\alpha)}$ é o quantil $(1 - \alpha)$ da distribuição χ^2_1.

Para obtermos a solução em (3.6) consideramos um intervalo de possíveis valores para θ_r, uma idéia é utilizar valores em torno de $\hat{\theta}_r$. Calculamos para cada um destes valores a função em (3.5) e, posteriormente, determinamos os valores de θ_r que satisfazem a inequação (3.6).

3.3 Teste de hipótese

Nesta seção, iremos definir o teste de hipótese para os parâmetros de regressão e o teste de hipótese para o parâmetro de correlação presente na estrutura de correlação.

3.3.1 Teste de hipótese para os parâmetros de regressão

Se o interesse reside em testar a plausibilidade de valores específicos para um subvetor de parâmetros $\boldsymbol{\beta}^*$, de $\boldsymbol{\beta}$, podemos considerar o teste de razão de verossimilhança. Para o modelo de regressão binomial correlacionada, este subvetor pode ser definido da seguinte forma:

1. Ajuste o modelo como descrito na Seção 3.1 e na Seção 3.2.

2. Utilize o teste de Wald, como descrito na Subseção 3.3.1, para cada parâmetro em $\boldsymbol{\beta}$ e selecione o subconjunto de parâmetros para os quais o teste foi não significativo.

Para testar as hipóteses $H_0 : \boldsymbol{\beta}^* = 0$ contra $H_1 : \boldsymbol{\beta}^* \neq 0$, utilize o teste de razão de verossimilhança, considerando a estatística

$$LR_\beta = -2\left\{\ell(\tilde{\boldsymbol{\theta}}; \mathcal{D}) - \ell(\hat{\boldsymbol{\theta}}; \mathcal{D})\right\}, \tag{3.7}$$

com $\ell(\boldsymbol{\theta}; \mathcal{D}) = \log\{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})\}$, e $\ell(\tilde{\boldsymbol{\theta}}; \mathcal{D})$ e $\ell(\hat{\boldsymbol{\theta}}; \mathcal{D})$ as funções de verossimilhança avaliadas respectivamente no estimador de máxima verossimilhança sob restrição, $\tilde{\boldsymbol{\theta}}$, e no estimador de máxima verossimilhança usual, $\hat{\boldsymbol{\theta}}$. A estatística de razão de verossimilhança LR_β segue, assintoticamente, uma distribuição χ^2 com o número de graus de liberdade igual ao número de restrições em H_0. A hipótese H_0 é rejeitada para grandes valores da estatística de teste LR_β. Se a hipótese H_0 não é rejeitada, subvetores de $\boldsymbol{\beta}^*$ devem ser testados.

Teste de Wald

Seja $\hat{\beta}_r$ um estimador de máxima verossimilhança de β_r, correspondente a r-ésima covariável e $J^{-1}_{(r)}(\hat{\boldsymbol{\theta}})$ o estimador da variância de $\hat{\beta}_r$. O teste de Wald para testar $H_0 : \beta_r = 0$ contra $H_1 : \beta_r \neq 0$, $r = 0, \ldots, k$, é dado por

$$W_r = \frac{\hat{\beta}_r{}^2}{J^{-1}_{(r)}(\hat{\boldsymbol{\theta}})}, \quad r = 0, \ldots, k, \tag{3.8}$$

em que $J^{-1}_{(r)}(\hat{\boldsymbol{\theta}})$ é o r-ésimo elemento da diagonal principal da inversa da matriz $J(\hat{\boldsymbol{\theta}})$ avaliada no estimador de máxima verossimilhança do vetor de parâmetros $\boldsymbol{\theta}$, definido em (3.2). A estatística W_r segue uma distribuição Normal padrão. A hipótese H_0 é rejeitada para pequenos e grandes valores da estatística W_r.

3.3.2 Teste de hipótese para o parâmetro da estrutura de correlação

Ao ajustar um modelo de regressão binomial correlacionada consideramos a suposição de correlação comum, ρ_i, entre os eventos de Bernoulli dentro do i-ésimo cluster. Porém, tal suposição pode não ser satisfeita, ou seja, um ajuste do modelo de regressão binomial poderia ter sido considerado. Uma forma de verificar esta suposição de dependência é testar a plausibilidade do valor do parâmetro de correlação. Na construção do modelo de regressão binomial correlacionada, ρ_i é modelado por meio de uma estrutura de correlação que envolve o parâmetro γ. Assim, testar a suposição de correlação comum, ρ_i, entre os eventos de Bernoulli dentro do i-ésimo cluster é análogo a testar a plausibilidade do valor do parâmetro γ.

Para testar as hipóteses $H_0 : \gamma = 0$ contra $H_1 : \gamma > 0$, para as estruturas de correlação exponencial e Gaussiana, ou $H_0 : \gamma = 1$ contra $H_1 : \gamma < 1$, para a estrutura de correlação AR contínua, iremos considerar o teste de razão de verossimilhança, por meio da estatística

$$LR_\gamma = -2\left\{\ell^B(\hat{\boldsymbol{\beta}};\mathcal{D}) - \ell(\hat{\boldsymbol{\theta}};\mathcal{D})\right\}, \tag{3.9}$$

sendo

$\ell^B(\hat{\boldsymbol{\beta}};\mathcal{D}) =$

$$\sum_{i=1}^{m}\left\{\log\binom{n_i}{y_i} + y_i\log\left(g^{-1}\left(\sum_{r=0}^{k}\hat{\beta}_r x_{ir}\right)\right) + (n_i - y_i)\log\left(1 - g^{-1}\left(\sum_{r=0}^{k}\hat{\beta}_r x_{ir}\right)\right)\right\},$$

a função de log-verossimilhança do modelo de regressão binomial avaliado nos respectivos estimadores de máxima verossimilhança fornecidos por este modelo; $\ell(\hat{\boldsymbol{\theta}};\mathcal{D}) = \log\{\mathcal{L}(\hat{\boldsymbol{\theta}};\mathcal{D})\}$, a função de log-verossimilhança avaliada no vetor de estimadores de máxima verossimilhança, $\hat{\boldsymbol{\theta}}$. A estatística de razão de verossimilhança LR_γ segue, assintoticamente, uma distribuição χ^2_1. A hipótese H_0 é rejeitada para grandes valores da estatística de teste LR_γ.

3.4 Diagnósticos

Entre as suposições impostas na construção do modelo de regressão binomial correlacionada, podemos ressaltar: (i) independência entre as variáveis resposta, (ii) as variáveis resposta seguem uma distribuição binomial correlacionada, $BC(n_i, p_i, \rho_i)$, (iii) correlação positiva entre as variáveis de Bernoulli dentro do cluster, $\rho_i > 0$, (iv) função de ligação e (v) estrutura de correlação. Nesta seção, uma análise de diagnóstico é proposta para a verificação dos pressupostos (i)-(v), e para a identificação de *outliers* e/ou observações influentes.

O pressuposto de independência pode ser verificado por meio dos gráficos dos resíduos versus a ordem das observações, quando a mesma está disponível. Dois diferentes tipos de resíduos são construídos para verificar o ajuste do modelo, bem como a presença de *outliers*. Na análise de resíduos pode ser verificada a adequação das variáveis resposta à distribuição binomial correlacionada, $BC(n_i, p_i, \rho_i)$, a função de ligação e a estrutura de correlação adotadas. Para verificar se $\rho_i > 0$, podemos observar a plausibilidade do valor do parâmetro da estrutura de correlação, γ, por meio dos intervalos de confiança obtidos na etapa de estimação e por meio do teste de hipótese definido na Seção 3.3.2. Para os casos em que esta suposição não for satisfeita, isto é, $\gamma = 1$ ou $\gamma = 0$, o modelo de regressão binomial usual pode ser considerado.

A detecção de observações influentes é feita utilizando a distância de Cook generalizada e a distância da verossimilhança (Zhu *et al.*, 2001) envolvendo observações deletadas (Cook & Weisberg, 1982).

3.4.1 Resíduos

Após o ajuste do modelo, é de interesse verificar a proximidade dos valores obtidos pelo ajuste em relação aos valores observados no conjunto de dados. Nesta seção, dois tipos de resíduos são construídos para o modelo de regressão binomial correlacionada. Um resíduo que depende do valor esperado de Y_i e um outro resíduo baseado na *deviance*.

Nos casos simulados, ambos os resíduos se mostraram eficientes na detecção de observações influentes (casos perturbados). No caso de dados não perturbados, os resíduos baseados nos valores preditos são mais sensíveis às características das observações, enquanto que os resíduos *deviance* são mais robustos. Ou melhor, resíduos *deviance* só identificam um caso como *outlier* quando o mesmo se destaca muito das demais observações.

Três gráficos, baseados nos resíduos padronizados, podem ser utilizados para verificar as suposições iniciais e identificar má especificação do modelo. Os gráficos (i) resíduo contra a ordem das observações, quando a mesma está disponível, para identificação de dependência temporal das observações; (ii) resíduo contra função das covariáveis, para verificar a necessidade de inserir outras funções de covariáveis, além da linear, na parte sistemática do modelo; (ii) resíduo contra valores preditos, onde espera-se que o conjunto de resíduos esteja próximo de zero. Se isso não ocorrer, ou seja, se há um conjunto de pontos longe de zero, tem-se uma indicação de que o modelo está mal ajustado aos dados.

No Apêndice C.3 apresentamos a implementação usando o programa R (R Development Core Team, 2011) dos resíduos padronizados propostos neste trabalho.

Resíduos baseados nos valores preditos

O resíduo padronizado baseado no valor predito, para o modelo de regressão binomial correlacionada, é definido como

$$r_i^{spp} = \frac{y_i - n_i\hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)\{n_i + \hat{\rho}_i n_i(n_i-1)\}}}, \quad i = 1, \ldots, m, \tag{3.10}$$

em que o par (y_i, n_i) é a informação da i-ésima observação, $\hat{p}_i = g^{-1}(\sum_{r=0}^{k} \hat{\beta}_r x_{ir})$ e $\hat{\rho}_i = h(v(\boldsymbol{r}_i), \hat{\gamma})$, com $\hat{\gamma}$ e $\hat{\beta}_r, r = 0, \ldots, k$, a estimativa de máxima verossimilhança de γ e $\beta_r, r = 0, \ldots, k$, respectivamente.

Resíduos *deviance*

O resíduo *deviance* padronizado para o modelo de regressão binomial correlacionada é definido por

$$r_i^{spd} = \frac{\text{sinal }(y_i - n_i\hat{p}_i)\sqrt{2\ell(y_i; \mathcal{D}_{(i)}, \hat{\gamma}) - 2\ell(\hat{\boldsymbol{\beta}}; \mathcal{D}_{(i)}, \hat{\gamma})}}{\sqrt{\hat{p}_i(1-\hat{p}_i)\{n_i + \hat{\rho}_i n_i(n_i-1)\}}}, \quad i = 1, \ldots, m, \tag{3.11}$$

sendo

$$\begin{aligned}\ell(y_i;\mathcal{D}_{(i)},\hat{\gamma}) &= \log\left\{\binom{n_i}{y_i}\left(\frac{y_i}{n_i}\right)^{y_i}\left(1-\frac{y_i}{n_i}\right)^{n_i-y_i}(1-h(v(\boldsymbol{r}_i),\hat{\gamma}))\right.\\ &\quad\left.+\left(\frac{y_i}{n_i}\right)^{\frac{y_i}{n_i}}\left(1-\frac{y_i}{n_i}\right)^{\frac{n_i-y_i}{n_i}}h(v(\boldsymbol{r}_i),\hat{\gamma})I_{A_{2i}}(y_i)\right\},\end{aligned}$$

a função de log-verossimilhança saturada, considerando para a estimação do parâmetro p_i a proporção y_i/n_i do i-ésimo cluster e o valor do parâmetro da estrutura de correlação, γ, é substituído pelo valor da estimativa de máxima verossimilhança, $\hat{\gamma}$; $\ell(\hat{\boldsymbol{\beta}};\mathcal{D}_{(i)},\hat{\gamma})$ a função de log-verossimilhança avaliada nos estimadores de máxima verossimilhança com $\hat{p_i} = g^{-1}(\sum_{r=0}^{k}\hat{\beta}_r x_{ir})$ e $\hat{\rho}_i = h(v(\boldsymbol{r}_i),\hat{\gamma})$ com $\hat{\gamma}$ e $\hat{\beta}_r, r = 0,\ldots,k$, a estimativa de máxima verossimilhança de γ e $\beta_r, r = 0,\ldots,k$, respectivamente; e $\mathcal{D}_{(i)} = (n_i, y_i, \boldsymbol{x}_i, \boldsymbol{r}_i)^\top$ é a informação disponível da i-ésima observação.

3.4.2 Diagnóstico de influencia global

Nesta seção, apresentamos duas métricas para avaliar a influência global das observações nas estimativas dos parâmetros do modelo de regressão binomial correlacionada: a distância de Cook generalizada e a distância da verossimilhança (Zhu *et al.*, 2001). Essas metodologias são eficazes quando existe apenas um *outlier* (She & Owen, 2011). She & Owen (2011) sugerem um método alternativo na presença de múltiplos *outliers*. Entretanto, esta ferramenta não foi adaptada para o modelo proposto.

No Apêndice C.4 apresentamos a implementação usando o programa R (R Development Core Team, 2011) destes duas métricas de avaliação de influência global.

Distância de Cook Generalizada

A distância de Cook generalizada (Zhu *et al.*, 2001) pode ser utilizada para quantificar o impacto da i-ésima observação no estimador de máxima verossimilhança de $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, sendo obtida por meio de

$$C_i = \left(\hat{\boldsymbol{\theta}}_{(-i)} - \hat{\boldsymbol{\theta}}\right)^\top J(\hat{\boldsymbol{\theta}})\left(\hat{\boldsymbol{\theta}}_{(-i)} - \hat{\boldsymbol{\theta}}\right),$$

em que $\hat{\boldsymbol{\theta}}_{(-i)}$ é o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ baseado na função de verossimilhança com dados aumentados, $\mathcal{L}(\boldsymbol{\theta};\mathcal{D}^*)$, apresentada em (2.7), com a i-ésima observação $(n_i, y_i, \boldsymbol{x}_i, \boldsymbol{r}_i)^\top$ deletada e $J(\hat{\boldsymbol{\theta}})$ é a matriz de informação de Fisher observada, dada em (3.2). Quando o número de clusters, m, é grande, Cook & Weisberg (1982) sugerem a seguinte aproximação para $\hat{\boldsymbol{\theta}}_{(-i)}$:

$$\hat{\boldsymbol{\theta}}_{(-i)} = \hat{\boldsymbol{\theta}} + J(\hat{\boldsymbol{\theta}})^{-1}U(\hat{\boldsymbol{\theta}}_{(-i)}), \tag{3.12}$$

em que

$$U(\hat{\boldsymbol{\theta}}_{(-i)}) = \left.\frac{\partial\ell(\boldsymbol{\theta};\mathcal{D}_{(-i)})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{(-i)}}.$$

O vetor de escores, $U(\boldsymbol{\theta}_{(-i)})$, com a i-ésima observação deletada, tem dimensão $(k+2)$ e os termos são dados por:

$$\frac{\partial \ell(\boldsymbol{\theta}; \mathcal{D}_{(-i)})}{\partial \gamma} = \sum_{i=1}^{m-1} \left\{ \frac{\partial h(v(\boldsymbol{r}_i), \gamma)}{\partial \gamma} \frac{[-A+B]}{[C+D]} \right\}$$

e

$$\frac{\partial \ell(\boldsymbol{\theta}; \mathcal{D}_{(-i)})}{\partial \beta_r} = \sum_{i=1}^{m-1} \left\{ \frac{\partial g^{-1}(\eta_i)}{\partial \beta_r} \left[C \left[y_i (g^{-1}(\eta_i))^{-1} - (n_i - y_i)(1 - g^{-1}(\eta_i))^{-1} \right] \right. \right.$$
$$\left. \left. + D \left[\frac{y_i}{n_i} (g^{-1}(\eta_i))^{-1} - \frac{(n_i - y_i)}{n_i} (1 - g^{-1}(\eta_i))^{-1} \right] \right] [C+D]^{-1} \right\},$$

com $r, s = 0, \ldots, k$, e A, B, C e D descritos na Tabela 3.1. As derivadas presentes nestas funções são mostradas em detalhes nas Tabelas 3.2 e 3.3.

Usando a aproximação apresentada em (3.12), a distância de Cook generalizada, $C_{i_{app}}$, é reescrita como

$$C_{i_{app}} = U(\hat{\boldsymbol{\theta}}_{(-i)})^\top J(\hat{\boldsymbol{\theta}}) U(\hat{\boldsymbol{\theta}}_{(-i)}).$$

Distância da Verossimilhança

A distância da verossimilhança também pode ser utilizada para quantificar a diferença entre $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_{(-i)}$. Essa métrica é dada por

$$LD_i = 2 \left\{ \ell(\hat{\boldsymbol{\theta}}; \mathcal{D}) - \ell(\hat{\boldsymbol{\theta}}_{(-i)}; \mathcal{D}) \right\}, \tag{3.13}$$

na qual $\ell(\hat{\boldsymbol{\theta}}; \mathcal{D})$ e $\ell(\hat{\boldsymbol{\theta}}_{(-i)}; \mathcal{D})$ são a função de log-verossimilhança avaliada no estimador de máxima verossimilhança, $\hat{\boldsymbol{\theta}}$, e no estimador de máxima verossimilhança com a i-ésima observação $(n_i, y_i, \boldsymbol{x}_i, \boldsymbol{r}_i)^\top$ deletada, respectivamente.

A i-ésima observação é considerada como influente se o valor da distância de Cook generalizada ou o valor da distância da verossimilhança é grande. Estes valores podem ser comparados com os valores críticos da distribuição χ^2_{k+2}. É importante ressaltar que a utilização deste critério pode impedir a detecção de observações influentes quando de fato existem. O quantil da distribuição χ^2_k, para k de moderado a grande podem produzir valores que sejam grandes o suficiente para inviabilizar a detecção.

Nos casos em que uma maior variabilidade nos valores obtidos por estas métricas ocorre, sugerimos que um caso com valor moderado não deve ser considerado como indicativo de influência. Porém, em situações em que a maior parte das observações apresenta valores das métricas muito próximas de zero, uma alteração moderada pode ser estudada como uma possível observação influente.

3.5 Critérios de seleção de modelo

É possível ajustarmos um conjunto potencial de modelos de regressão binomial correlacionada utilizando o mesmo conjunto de dados. Isto ocorre quando, por exemplo,

consideramos diferentes subconjuntos de covariáveis ou diferentes estruturas de correlação e/ou diferentes funções de ligação no processo de análise. Nestes casos, é conveniente o uso de um critério de seleção de modelos.

O critério de informação Bayesiano (BIC) e o critério de informação de Akaike (AIC) podem ser utilizados para selecionar o melhor modelo de regressão binomial correlacionada. O BIC (Schwarz, 1978) é calculado pela expressão

$$BIC = -2\ell(\hat{\boldsymbol{\theta}}; \mathcal{D}) + (k+2)\log(m), \tag{3.14}$$

em que $\ell(\hat{\boldsymbol{\theta}}; \mathcal{D})$ é o valor da função de log-verossimilhança avaliada no estimador usual de máxima verossimilhança, $(k+2)$ é o número de parâmetros no modelo e m é o número de clusters. O AIC (Akaike, 1974) é calculado pela expressão

$$AIC = -2\ell(\hat{\boldsymbol{\theta}}; \mathcal{D}) + (k+2). \tag{3.15}$$

Os menores valores do BIC e do AIC indicam o melhor modelo.

No Apêndice C.5 apresentamos a implementação usando o programa R (R Development Core Team, 2011) destes dois critérios de seleção de modelos.

3.6 Estudos de simulação

Nesta seção consideramos um estudo de simulação com apenas uma amostra para ilustrar o processo de estimação e o desempenho das medidas de diagnóstico propostas para o modelo de regressão binomial correlacionada. Um outro estudo com 1000 amostras é realizado para analisar as propriedades frequentistas dos estimadores de máxima verossimilhança e seus intervalos de confiança propostos.

Dados simulados

A geração da amostra envolve $m = 100$ clusters com as variáveis resposta, Y_i, $i = 1, \ldots, 100$, seguindo uma distribuição $BC(n_i, p_i, \rho_i)$, com os n_i gerados de uma distribuição B(45, 0.5). O parâmetro da estrutura de correlação utilizado na simulação é $\gamma = 0,2$ e os $v(\boldsymbol{r}_i)$ assumem valores de uma distribuição U(0,1). Duas covariáveis são consideradas, x_{i1} e x_{i2}. Os valores das covariáveis x_{i1} provêm de uma distribuição U(0,2) e os valores da covariável x_{i2} provêm de uma distribuição N(0,4). Os valores dos coeficientes de regressão são $\beta_0 = 2$, $\beta_1 = -2$ e $\beta_2 = -2$. São considerados neste estudo a função de ligação logito e a estrutura de correlação exponencial.

Estimação

As quatro funções de ligação, logito, complementar log-log, log-log e probito, foram consideradas na análise. Como esperávamos, os valores obtidos pelos critérios de seleção de modelos, AIC e BIC, confirma a função de ligação logito como o melhor ajuste, conforme pode ser visto na Tabela 3.4.

Tabela 3.4: Valores obtidos pelos critérios de seleção de modelo ajustando o conjunto de dados simulado.

Critérios	Logito	Complementar log-log	Log-log	Probito
AIC	125,913	127,016	138,554	127,177
BIC	154,754	155,857	167,395	156,018

As estimativas de máxima verossimilhança e os intervalos de confiança assintóticos para os parâmetros do modelo com ligação logito são mostradas na Tabela 3.5, para o intervalo de confiança baseado na função de verossimilhança com dados aumentados foram consideradas 3000 iterações para a aproximação de Monte Carlo, para o intervalo de confiança *bootstrap* foram considerados 100 reamostras e para obter os intervalos de confiança perfilados foram considerados 5000 valores em torno do estimador de máxima verossimilhança. Note que os intervalos de confiança assintóticos contêm o verdadeiro valor dos parâmetros. É importante ressaltar que o intervalo de confiança do parâmetro da estrutura de correlação, γ, não contêm o valor zero, corroborando com a necessidade de ajuste de um modelo binomial correlacionada.

Tabela 3.5: Estimativas de máxima verossimilhança e intervalo de confiança de 95% para γ, β_0, β_1 e β_2 para a ligação logito.

	Valor real	EMV	IC 95% via $\mathcal{L}(\boldsymbol{\theta};\mathcal{D})$	IC 95% via $\mathcal{L}(\boldsymbol{\theta};\mathcal{D}^*)$ 95%	IC 95% via *bootstrap*	IC 95% via perfilada
γ	0,20	0,237	(0,055 ; 0,419)	(0,050 ; 0,424)	(0,096 ; 0,452)	(0,101 ; 0,463)
β_0	2,00	2,205	(1,414 ; 2,995)	(1,410 ; 2,999)	(1,624 ; 3,669)	(1,832 ; 2,581)
β_1	-2,00	-2,158	(-3,074 ; -1,243)	(-3,073 ; -1,244)	(-3,647 ; -1,392)	(-2,624 ; -1,690)
β_2	-2,00	-2,114	(-2,588 ; -1,640)	(-2,597 ; -1,630)	(-2,728 ; 1,695)	(-2,493 ; -1,776)

Análise de resíduos

Os gráficos de resíduos contra valores preditos são apresentados para os dois tipos de resíduos desenvolvidos neste trabalho (ver Seção 3.4.1). Os gráficos dos resíduos contra valores preditos, apresentados na Figura 3.1a e 3.1b, verificam a coerência do modelo proposto. Observe que ambos os gráficos não indicam observações aberrantes, principalmente os resíduos *deviance* padronizados (Figura 3.1b), que apresentam todos os valores próximos de zero. Não identificamos a necessidade de inserir outras funções das covariáveis no modelo, assim, os gráficos envolvendo as funções das covariáveis contra os resíduos, sugeridos na Seção 3.4.1, foram omitidos.

Diagnósticos de influência

Nessa seção examinamos o desempenho das medidas de diagnósticos de influência, a distância de Cook generalizada e a distância da verossimilhança. Para isto, as observações

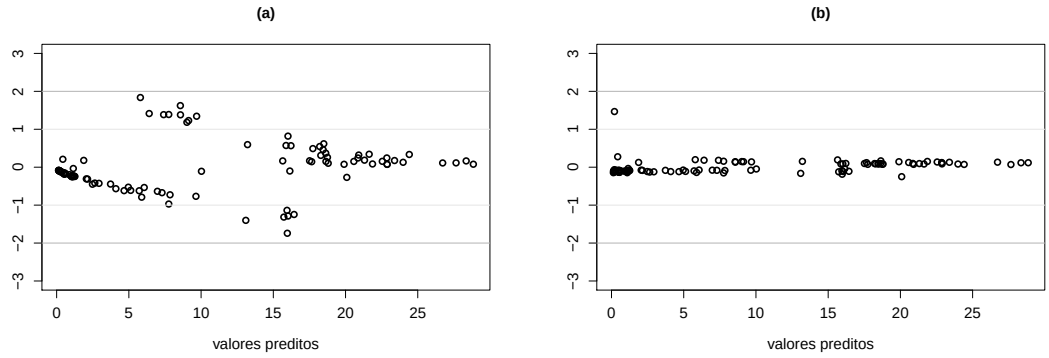

Figura 3.1: (a) resíduos padronizados versus valores preditos, (b) resíduos *deviance* padronizados versus valores preditos.

25 e 48 foram perturbadas, criando assim observações influentes no conjunto de dados. As perturbações foram feitas na covariável x_{i2}, na forma $x_{i2} = x_{i2} + 8sd(x_2)$, $i = 25$ e 48, em que $sd(x_2)$ corresponde ao desvio padrão da covariável x_2. Combinações de presença e ausência destas observações perturbadas foram usadas para formar novos conjuntos de dados. Estes novos conjuntos de dados foram utilizados na obtenção de estimativas dos parâmetros do modelo. A Tabela 3.6 apresenta os estimadores de máxima verossimilhança e a mudança relativa da estimativas dos parâmetros em relação aos dados originais.

Tabela 3.6: Estimador de máxima verossimilhança e mudança relativa em relação aos dados originais simulados.

	Parâmetros							
Casos	γ		β_0		β_1		β_2	
perturbados	EMV	%	EMV	%	EMV	%	EMV	%
{Nenhum}	0,24	-	2,21	-	-2,16		-2,11	-
Caso {25}	0,19	-20,80	1,48	-33,03	-1,35	-37,50	-1,52	-27,96
Caso {48}	0,19	-20,80	1,64	-25,79	-1,55	-28,24	-1,50	-28,91
Casos {25} e {48}	0,17	-29,17	1,04	-52,94	-0,82	-62,04	-1,02	-51,66

Como mencionado na Seção 3.4.2, valores altos da distância de Cook generalizada e da distância da verossimilhança evidenciam a presença de pontos influentes no conjunto de dados. Uma ilustração de um dos casos analisados na Tabela 3.6, perturbação dos casos 25 e 48, é apresentado na Figura 3.2. A Figura 3.2a mostra as distâncias de Cook generalizada, a Figura 3.2b mostra as distâncias de Cook generalizada envolvendo apenas os parâmetros da função de ligação, a Figura 3.2c mostra as distâncias de Cook generalizada envolvendo apenas o parâmetro da estrutura de correlação e finalmente é apresentado na Figura 3.2d as distâncias da verossimilhança. Conforme esperávamos, a distância de Cook generalizada, que envolve apenas o parâmetro da estrutura de correlação, Figura 3.2c, não evidenciou impacto com as perturbações, uma vez que as mesmas ocorreram na covariável x_{i2} que está relacionada a função de ligação. Observamos também que a distância de Cook generalizada e distância da verossimilhança, Figura 3.2a e Figura 3.2d,

apresentam valores sensivelmente mais elevados para as observações 25 e 48.

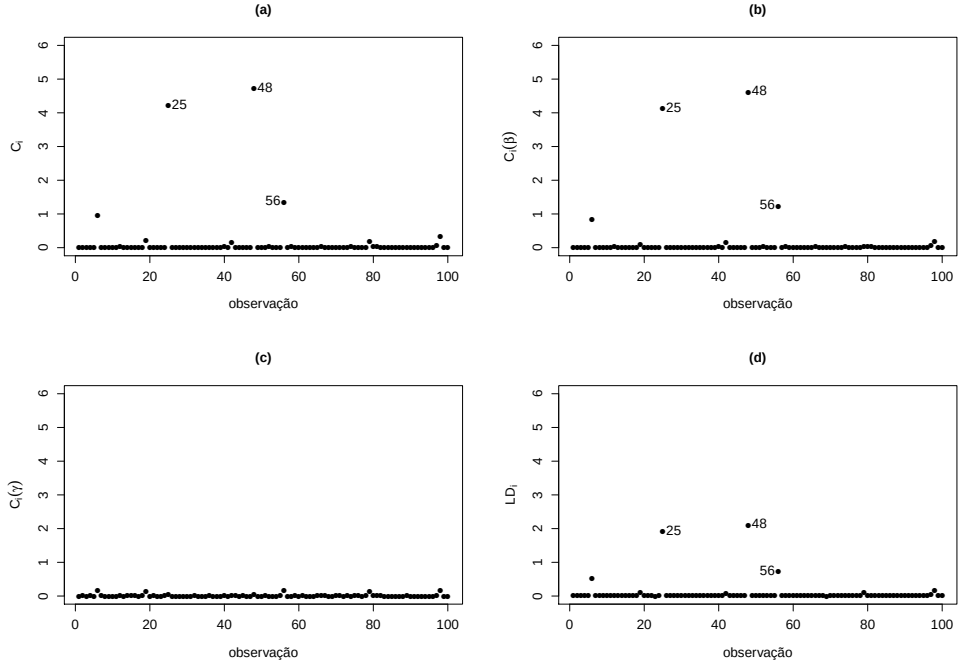

Figura 3.2: (a) distâncias de Cook generalizada, (b) distâncias de Cook generalizada para os parâmetros da função de ligação, (c) distâncias de Cook generalizada para o parâmetro da estrutura de correlação e (d) distâncias da verossimilhança.

Modelo usual

Nesta seção ajustamos o modelo de regressão usual para os dados simulado na Seção 3.6 utilizando a função de ligação logito. Ou seja, ajustamos um modelo supondo a não existência de correlação entre os eventos de Bernoulli. Na Tabela 3.7 apresentamos as estimativas de máxima verossimilhança dos parâmetros do ajuste via modelo de regressão binomial correlacionada, obtidos na Seção 3.6, e do ajuste via modelo de regressão binomial usual. Observe que os valores estimados no ajuste do modelo de regressão binomial usual forneceram valores muito diferentes dos fixados para a simulação. Ou seja, fica evidente que desconsiderar a dependência entre os eventos de Bernoulli podem levar a conclusões erradas.

Na Figura 3.3a e na Figura 3.3b apresentam os valores das probabilidades de sucesso geradas contra as probabilidades de sucesso estimadas pelo modelo de regressão binomial usual e pelo modelo de regressão binomial correlacionada, respectivamente. Os valores destas probabilidades estimadas no modelo usual são quase sempre subestimadas, enquanto que no modelo de regressão binomial correlacionada são muito próximas das verdadeiras. Além das evidências gráficas que o modelo de regressão binomial correlacionada apresenta melhor ajuste, realizamos o teste qui-quadrado comparando os valores preditos pelos modelos ajustados contra os valores das variáveis resposta observados e obtivemos para o modelo regressão binomial correlacionada o valor-p de 0,862 e para o modelo de

Tabela 3.7: Estimativas de máxima verossimilhança para γ, β_0, β_1 e β_2 via modelo de regressão binomial correlacionada e via modelo de regressão binomial usual.

Parâmetros	Valor verdadeiro	Binomial Correlacionada	Binomial usual
γ	0,20	0,237	-
β_0	2,00	2,205	-0,269
β_1	-2,00	-2,158	-0,823
β_2	-2,00	-2,114	-0,662

- indica que o modelo não envolve o parâmetro em questão.

regressão binomial o valor-p de 0,051, confirmando que os valores ajustados pelo modelo de regressão binomial correlacionada são mais coerentes com os valores observados nas variáveis resposta.

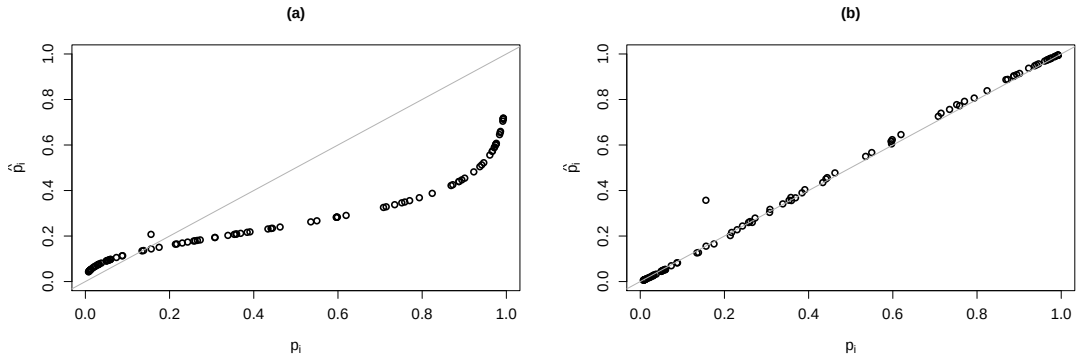

Figura 3.3: (a) probabilidade de sucesso real versus probabilidade de sucesso estimada pelo modelo usual, (b) probabilidade de sucesso real versus probabilidade de sucesso estimada pelo MRBC.

3.6.1 Propriedade frequentista dos estimadores de máxima verossimilhança

As estimativas das probabilidades de cobertura dos intervalos assintóticos, dos intervalos *bootstrap* e dos intervalos perfilados foram construídas para o nível de confiança fixado em 95%. A determinação das estimativas das probabilidades de cobertura foram obtidas calculando a proporção de intervalos que continham o verdadeiro valor dos parâmetros fixados na geração dos dados, baseada em um processo de simulação similar ao descrito na Seção 3.6, exceto pelas estrutura de correlação e função de ligação utilizadas. O cálculo da proporção está baseado na simulação de 1.000 amostras com $m = 500$ clusters, sendo que para obtenção dos intervalos assintóticos foram consideradas 500 iterações para aproximação de Monte Carlo, para os intervalos perfilados foram considerados 500 valores em torno dos estimadores de máxima verossimilhança. Para os intervalos de confiança *bootstrap* foram consideradas apenas 300 amostras simuladas e para cada amostra simulada

100 reamostragens. Nesta parte da análise utilizamos quatro funções de ligação, logito, complementar log-log, log-log e probito, e as três estruturas de correlação, exponencial, Gaussiana e AR contínua.

As estimativas das probabilidades de cobertura para os intervalos de confiança com variância estimada via função de verossimilhança com dados observados são denotados por PC_u e para os intervalos de confiança com variância estimada via função de verossimilhança com dados aumentados são denotados por PC_a. Os resultados apresentados na Tabela 3.8 indicam que as estas estimativas (PC_u e PC_a) estão entre 92% e 98%, considerando uma cobertura nominal de 95%, para os quatro parâmetros em todos os cenários. Ambos as formas de obtenção das estimativas da variância do estimador de máxima verossimilhança fornecem resultados muito similares mostrando que os processos são equivalentes.

Para os intervalos de confiança perfilados observamos que as estimativas das probabilidades de cobertura (PC_p) para o parâmetro γ estão entre 92% e 97%, considerando uma cobertura nominal de 95%, exceto para a função de ligação complementar log-log e probito com a estrutura de correlação Gaussiana que forneceram 83% e 86%, respectivamente. Porém, as estimativas das probabilidades de cobertura para os parâmetros de regressão, β_0, β_1 e β_2, ficaram sensivelmente abaixo da cobertura nominal, pois apresentaram valores entre 33% e 92%.

No caso dos intervalos de confiança *bootstrap* as estimativas das probabilidades de cobertura (PC_b) estão entre 87% e 98%, considerando a cobertura nominal de 95%. Porém, apenas 100 reamostras são utilizadas no processo.

Na Tabela 3.9 apresentamos a média dos erros quadráticos médios, considerando a variância estimada via função de verossimilhança com dados observados (EQM_u) e com a variância estimada via função de verossimilhança com dados aumentados (EQM_a), e média dos vícios dos estimadores de máxima verossimilhança considerando as quatro funções de ligação e as três estruturas de correlação. Para determinar estes valores foram geradas 1.000 amostras para cada cenário. Em cada amostra foi calculado o estimador de máxima verossimilhança, o EQM_u, o EQM_a e o vício. Posteriormente, obtivemos a média destes 1.000 EQMs e vícios, para cada parâmetro. Os resultados mostram valores próximos de zero.

Tabela 3.8: Estimativas das probabilidades de cobertura dos intervalos assintóticos (PC_u e PC_a), dos intervalos via *bootstrap* (PC_b) e dos intervalos via log-verossimilhança perfilada (PC_p) com confiança de 95% para diferentes funções de ligação e estruturas de correlação.

		Estrutura de correlação											
		Exponencial				Gaussiana				AR Contínua			
	Ligação	PC_u	PC_a	PC_b	PC_p	PC_u	PC_a	PC_b	PC_p	PC_u	PC_a	PC_b	PC_p
	Logito	0,94	0,94	0,94	0,58	0,96	0,96	0,95	0,62	0,95	0,95	0,90	0,58
	C. log-log	0,94	0,94	0,91	0,48	0,94	0,94	0,93	0,52	0,94	0,95	0,98	0,51
	Log-log	0,95	0,95	0,90	0,39	0,94	0,94	0,95	0,33	0,95	0,95	0,94	0,34
β_0	Probito	0,94	0,94	0,94	0,52	0,93	0,93	0,94	0,47	0,96	0,96	0,94	0,56
	Logito	0,94	0,94	0,95	0,62	0,96	0,95	0,94	0,66	0,94	0,95	0,90	0,63
	C. log-log	0,95	0,95	0,91	0,58	0,94	0,94	0,92	0,62	0,94	0,94	0,96	0,60
	Log-log	0,95	0,95	0,92	0,52	0,94	0,94	0,95	0,50	0,96	0,96	0,94	0,47
β_1	Probito	0,95	0,95	0,94	0,61	0,94	0,94	0,93	0,56	0,96	0,96	0,93	0,61
	Logito	0,95	0,96	0,93	0,89	0,97	0,97	0,93	0,92	0,94	0,95	0,94	0,87
	C. log-log	0,93	0,94	0,91	0,78	0,92	0,92	0,91	0,77	0,95	0,96	0,95	0,78
	Log-log	0,94	0,94	0,92	0,57	0,94	0,94	0,92	0,55	0,94	0,94	0,91	0,58
β_2	Probito	0,94	0,94	0,93	0,82	0,92	0,92	0,92	0,79	0,94	0,95	0,93	0,79
	Logito	0,94	0,94	0,90	0,94	0,97	0,97	0,87	0,92	0,94	0,95	0,94	0,95
	C. log-log	0,93	0,94	0,94	0,95	0,97	0,98	0,93	0,83	0,93	0,94	0,90	0,94
	Log-log	0,96	0,96	0,96	0,96	0,98	0,98	0,91	0,97	0,94	0,94	0,95	0,94
γ	Probito	0,94	0,95	0,93	0,95	0,92	0,92	0,93	0,86	0,95	0,95	0,92	0,96

Probabilidade de cobertura nominal de 95%.

Tabela 3.9: Média dos erros quadráticos médios (EQM_u e EQM_a) e média dos vícios dos estimadores de máxima verossimilhança para diferentes funções de ligação e estruturas de correlação.

		Estrutura de correlação								
		Exponencial			Gaussiana			AR Contínua		
	Ligação	Vício	EQM_u	EQM_a	Vício	EQM_u	EQM_a	Vício	EQM_u	EQM_a
	Logito	0,001	0,226	0,227	0,004	0,348	0,349	0,007	0,104	0,105
	C. log-log	0,014	0,168	0,170	-0,040	0,248	0,249	0,006	0,077	0,078
	Log-log	0,003	0,183	0,183	-0,008	0,296	0,296	0,001	0,086	0,087
β_0	Probito	0,007	0,158	0,164	-0,055	0,263	0,271	-0,000	0,069	0,070
	Logito	-0,000	0,391	0,392	0,041	0,614	0,614	-0,008	0,170	0,170
	C. log-log	-0,017	0,249	0,250	0,068	0,380	0,380	-0,006	0,111	0,112
	Log-log	0,001	0,220	0,220	0,014	0,354	0,354	-0,002	0,102	0,102
β_1	Probito	-0,005	0,230	0,236	0,074	0,402	0,412	0,003	0,100	0,101
	Logito	-0,004	0,120	0,122	-0,005	0,176	0,177	-0,006	0,058	0,059
	C. log-log	-0,011	0,111	0,113	-0,023	0,177	0,177	-0,003	0,052	0,053
	Log-log	-0,011	0,158	0,158	0,001	0,248	0,248	-0,001	0,074	0,074
β_2	Probito	-0,009	0,118	0,126	0,033	0,188	0,197	-0,002	0,051	0,052
	Logito	0,000	0,035	0,036	0,006	0,047	0,047	-0,001	0,027	0,028
	C. log-log	0,000	0,036	0,037	-0,015	0,048	0,048	0,001	0,028	0,028
	Log-log	0,001	0,131	0,031	0,011	0,041	0,041	-0,000	0,022	0,022
γ	Probito	0,001	0,035	0,037	0,015	0,047	0,048	0,001	0,027	0,028

Capítulo 4

MRBC: Metodologia Bayesiana

Neste capítulo desenvolvemos uma abordagem Bayesiana para o modelo de regressão binomial correlacionada. O método envolve uma estratégia de dados aumentado e algoritmos MCMC são considerados para obter as estimativas a *posteriori* para os parâmetros.

Utilizando a parametrização $\gamma = \exp\{\phi\}$, para estrutura de correlação exponencial ou Gaussiana, ou $\gamma = \exp\{\phi\}/(1 + \exp\{\phi\})$, para estrutura de correlação AR contínua, a abordagem Bayesiana assume $n_i, i = 1, \ldots, m$, conhecido e independência a *priori* entre os parâmetros $\beta_0, \beta_1, \ldots, \beta_k$ e ϕ. Uma distribuição a *priori* para o vetor de parâmetros $\boldsymbol{\theta}^* = (\phi, \beta_0, \ldots, \beta_r)^\top$ pode ser dada por uma distribuição Normal multivariada $(k+2)$-dimensional, com vetor de médias zero e matriz de variância-covariância $\boldsymbol{\Sigma} = \text{diag}\{\lambda_1, \ldots, \lambda_{k+2}\}$, com hiperparâmetros λ_r, $r = 1, \ldots, k+2$ conhecidos.

A distribuição a *posteriori* conjunta dos parâmetros é obtida combinando a função de verossimilhança (2.8) e a distribuição a *priori* de $\boldsymbol{\theta}^*$. Assim, a distribuição a *posteriori* conjunta é dada por

$$\pi(\boldsymbol{\theta}^*|\mathcal{D}^*) \quad \propto \quad \mathcal{L}(\boldsymbol{\theta}^*; \mathcal{D}^*)\ \pi(\boldsymbol{\theta}^*). \tag{4.1}$$

Esta distribuição não é tratável analiticamente. Assim, a inferência Bayesiana pode ser conduzida considerando métodos MCMC tal como o algoritmo Gibbs com passos de Metropolis (Robert & Casella, 2004). As densidades condicionais completas para os parâmetros ϕ e β_r, necessárias para implementar o algoritmo de Metropolis, são dadas por

$$\pi(\phi|\lambda_1, \mathcal{D}^*) \propto \prod_{i=1}^{m} \left\{h^*(v(\boldsymbol{r}_i), \phi)^{z_i}(1 - h^*(v(\boldsymbol{r}_i), \phi))^{1-z_i}\right\} \exp\left\{-\frac{\phi^2}{2\lambda_1}\right\}$$

e

$$\pi(\beta_r|\boldsymbol{\beta}_{(-r)}, \lambda_{r+1}, \mathcal{D}^*) \propto$$

$$\prod_{i=1}^{m} \left\{g^{-1}(\eta_i)^{\frac{y_i}{n_i}(z_i+n_i-n_iz_i)}(1 - g^{-1}(\eta_i))^{(n_i-y_i)\left(\frac{z_i}{n_i}+1-z_i\right)}\right\} \exp\left\{-\frac{\beta_r^2}{2\lambda_{r+1}}\right\}.$$

Pelo fato das densidades condicionais completas não terem forma fechada o algoritmo de Metropolis-Hastings pode ser utilizado para gerar amostras da distribuição a *posteriori* conjunta dos parâmetros por meio das densidades condicionais.

Para gerar amostras de $\boldsymbol{\theta}^*$ na primeira iteração da cadeia, considere o algoritmo:

1. Gere valores iniciais para $\boldsymbol{\theta}^*$, $\boldsymbol{\theta}^{(0)} = (\phi^{(0)}, \beta_0^{(0)}, \ldots, \beta_k^{(0)})^\top$.
2. Gere $\boldsymbol{z}^{(0)} = (z_1^{(0)}, \ldots, z_m^{(0)})^\top$, em que $z_i^{(0)} \sim Bern(\tau_i^{(0)})$, τ_i dada por (2.5);
3. Obtenha os dados completos $\mathcal{D}^{*(0)} = (\mathcal{D}, \boldsymbol{z}^{(0)})^\top$.
4. Gere uma amostra candidata $\boldsymbol{\theta}^{(c)}$ de $N_{k+2}(\boldsymbol{\theta}^{(0)}, \xi I_{k+2})$, em que ξ é um valor que deve ser escolhido tal que a taxa de aceitação seja razoável e I_{k+2} é uma matriz identidade de ordem $(k+2)$.
5. Gere número aleatórios $u_r, r = 1, \ldots, k+2$, de uma distribuição uniforme no intervalo $(0, 1)$.
6. Para cada candidato $\phi^{(c)}$, e cada candidato $\beta_r^{(c)}$, $r = 1, \ldots, k+2$, ambos em $\boldsymbol{\theta}^{(c)}$, calcule as razões

$$R_{_{MH1}} = \frac{\pi(\phi^{(c)}|\lambda_1, \mathcal{D}^{*(0)})}{\pi(\phi^{(0)}|\lambda_1, \mathcal{D}^{*(0)})} \text{ e } R_{_{MH2}} = \frac{\pi(\beta_r^{(c)}|\boldsymbol{\beta}_{(-r)}^{(0)}, \lambda_{r+1}, \mathcal{D}^{*(0)})}{\pi(\beta_r^{(0)}|\boldsymbol{\beta}_{(-r)}^{(0)}, \lambda_{r+1}, \mathcal{D}^{*(0)})}.$$

7. Se $u_1 \leq R_{_{MH1}}$, então aceita o ponto candidato $\phi^{(1)} = \phi^{(c)}$, caso contrário $\phi^{(1)} = \phi^{(0)}$ e se $u_{r+1} \leq R_{_{MH2}}$, então aceita o ponto candidato $\beta_r^{(1)} = \beta_r^{(c)}$, caso contrário $\beta_r^{(1)} = \beta_r^{(0)}$.
8. Com $\boldsymbol{\theta}^{(0)}$ substituindo $\boldsymbol{\theta}^{(1)}$, repita o processo para atualizar $\boldsymbol{\theta}^*$.

No Apêndice D.1 apresentamos a implementação usando o programa R (R Development Core Team, 2011) para obter as cadeias de Gibbs com passos de Metropolis e dados aumentados fazendo uso do passeio aleatório.

No algoritmo descrito acima os candidatos são gerados via passeio aleatório, porém, para algumas covariáveis dos clusters, a utilização do algoritmo Gibbs com passos de Metropolis via passeio aleatório pode demandar maior tempo computacional na obtenção da convergência. Nestes casos, podemos considerar candidatos de uma cadeia independente da forma $\boldsymbol{\theta}^{(c)} \sim N_{k+2}(\hat{\boldsymbol{\theta}}, \xi I_{k+2})$, em que $\hat{\boldsymbol{\theta}}$ é o estimador de máxima verossimilhança de $\boldsymbol{\theta}^*$, os quais podem ser obtidos conforme descrito na Seção 3.1, e ξ é um valor que garante o envelopamento da condicional completa.

4.1 Distribuição preditiva a *posteriori*

A distribuição de uma observação futura, $\tilde{y}$, condicionada a $\mathcal{D}$ é dada pela distribuição preditiva a *posteriori* (Gelman *et al.*, 2003). A distribuição preditiva a *posteriori* para o vetor de parâmetros $\boldsymbol{\theta}^* = (\phi, \beta_0, \ldots, \beta_k)^\top$, condicionado a $(n, \boldsymbol{x}, \boldsymbol{r})^\top$, informações associadas

a observação futura, com $\boldsymbol{x} = (x_1, \ldots, x_k)^\top$ e $\boldsymbol{r} = (r_{11}, \ldots, r_{1n}, r_{21}, \ldots, r_{2n}, r_{q1}, \ldots, r_{qn})^\top$, é definida como segue

$$\pi(\tilde{y}|\mathcal{D}) = \int_\Theta \pi(\tilde{y}|\mathcal{D}, \boldsymbol{\theta}^*)\pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*\right) d\boldsymbol{\theta}, \tag{4.2}$$

com $\pi(\tilde{y}|\mathcal{D}, \boldsymbol{\theta}^*) = \binom{n}{\tilde{y}} p^{\tilde{y}}(1-p)^{n-\tilde{y}}(1-\rho) + p^{\frac{\tilde{y}}{n}}(1-p)^{\frac{(n-\tilde{y})}{n}}\rho I_{A_2}(\tilde{y})$,

em que $\rho = h^*(v(\boldsymbol{r}), \phi)$ e $p = g^{-1}(\sum_{r=0}^{k} \beta_r x_r)$.

Considerando os estimadores de Bayes para os parâmetros dos coeficientes de regressão, $\beta_r, r = 1, \ldots, k$, e para o parâmetro da estrutura de correlação, ϕ, podemos predizer o valor de $\tilde{y}$ usando (4.2). Para isto, usamos uma estimativa de Monte Carlo da densidade $\pi(\tilde{y}|\mathcal{D}, \boldsymbol{\theta}^*)$ com amostras MCMC da distribuição a *posteriori* $\pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*\right)$ (Chen *et al.*, 2000). Entretanto, para realizar este método, é necessário primeiro fixar um valor para $\tilde{y}$ em $\{0, \ldots, n\}$. O seguinte algoritmo fornece o valor predito para $\tilde{y}$.

(a) Gere uma amostra $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q$ de tamanho Q de $\pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*\right)$. Para cada $\boldsymbol{\theta}_q = (\phi_q, \beta_{0q}, \ldots, \beta_{kq})^\top$, $q = 1, \ldots, Q$, calcule os valores de $\hat{p}_q$ e $\hat{\rho}_q$, sendo

$$\hat{p}_q = g^{-1}\left(\sum_{r=0}^{k} \hat{\beta}_{rq} x_r\right) \quad \text{e} \quad \hat{\rho}_q = h^*(v(\boldsymbol{r}), \hat{\phi}_q).$$

(b) Para cada $\tilde{y}$ em $\{0, \ldots, n\}$, obtenha a estimativa de Monte Carlo para $\pi(\tilde{y}|\mathcal{D})$, dado por

$$\hat{\pi}(\tilde{y}|\mathcal{D}) = \frac{1}{Q}\sum_{q=1}^{Q} \binom{n}{\tilde{y}} \hat{p}_q^{\tilde{y}}(1-\hat{p}_q)^{n-\tilde{y}}(1-\hat{\rho}_q) + \hat{p}_q^{\frac{\tilde{y}}{n}}(1-\hat{p}_q)^{\frac{(n-\tilde{y})}{n}}\hat{\rho}_q I_{A_2}(\tilde{y}).$$

(c) O valor de $\tilde{y}$ em $\{0, \ldots, n\}$ que maximiza $\hat{\pi}(\tilde{y}|\mathcal{D})$ é o valor predito para a observação futura, condicionado a observação de $(n, \boldsymbol{x}, \boldsymbol{r})^\top$.

4.2 Densidade preditiva condicional ordinária

Uma outra forma de avaliar a qualidade do ajuste de um modelo, é verificando a capacidade que o mesmo tem de capturar as respostas observadas. Para isto, comparamos os valores das respostas preditas pelo modelo ajustado via densidade preditiva condicional ordinária (CPO) (Cho *et al.*, 2009) com os valores de fato observados. A densidade preditiva condicional ordinária para a i-ésima observação, condicionada a $\mathcal{D}_{(-i)}$, é definida como

$$CPO_i \;=\; \int_\Theta \pi(y_i|\mathcal{D}, \boldsymbol{\theta}^*)\pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*_{(-i)}\right) d\boldsymbol{\theta}, \tag{4.3}$$

que pode ser reescrita como

$$\begin{aligned} CPO_i &= \int_\Theta \pi(y_i|\mathcal{D},\boldsymbol{\theta}^*)\frac{\mathcal{L}(\boldsymbol{\theta}^*;\mathcal{D}^*_{(-i)})\pi(\boldsymbol{\theta}^*)}{\int_\Theta \mathcal{L}(\boldsymbol{\theta}^*;\mathcal{D}^*_{(-i)})\pi(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^*}d\boldsymbol{\theta}^* \\ &= \int_\Theta \frac{\mathcal{L}(\boldsymbol{\theta}^*;\mathcal{D}^*)\pi(\boldsymbol{\theta}^*)}{\int_\Theta \frac{\pi(y_i|\mathcal{D},\boldsymbol{\theta}^*)}{\pi(y_i|\mathcal{D},\boldsymbol{\theta}^*)}\mathcal{L}(\boldsymbol{\theta}^*;\mathcal{D}^*_{(-i)})\pi(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^*}d\boldsymbol{\theta}^* \\ &= \int_\Theta \frac{\mathcal{L}(\boldsymbol{\theta}^*;\mathcal{D}^*)\pi(\boldsymbol{\theta}^*)}{\int_\Theta \frac{1}{\pi(y_i|\mathcal{D},\boldsymbol{\theta}^*)}\mathcal{L}(\boldsymbol{\theta}^*;\mathcal{D}^*)\pi(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^*}d\boldsymbol{\theta}^* \end{aligned}$$

$$\begin{aligned} &= \frac{1}{\int_\Theta \frac{1}{\pi(y_i|\mathcal{D},\boldsymbol{\theta}^*)}\mathcal{L}(\boldsymbol{\theta}^*;\mathcal{D}^*)\pi(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^*}\int_\Theta \mathcal{L}(\boldsymbol{\theta}^*;\mathcal{D}^*)\pi(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^* \\ &= \frac{1}{\int_\Theta \frac{1}{\pi(y_i|\mathcal{D},\boldsymbol{\theta}^*)}\mathcal{L}(\boldsymbol{\theta}^*;\mathcal{D}^*)\pi(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^*}\frac{1}{\frac{1}{\int_\Theta \mathcal{L}(\boldsymbol{\theta}^*;\mathcal{D}^*)\pi(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^*}} \\ &= \frac{1}{\int_\Theta \frac{1}{\pi(y_i|\mathcal{D},\boldsymbol{\theta}^*)}\frac{\mathcal{L}(\boldsymbol{\theta}^*;\mathcal{D}^*)\pi(\boldsymbol{\theta}^*)}{\int_\Theta \mathcal{L}(\boldsymbol{\theta}^*;\mathcal{D}^*)\pi(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^*}} \\ &= \frac{1}{\int_\Theta \frac{1}{\pi(y_i|\mathcal{D},\boldsymbol{\theta}^*)}\pi(\boldsymbol{\theta}^*|\mathcal{D}^*)}, \end{aligned} \tag{4.4}$$

$i = 1, \ldots, m$, com

$$\pi(y_i|\mathcal{D},\boldsymbol{\theta}^*) = \binom{n_i}{y_i}p_i^{y_i}(1-p_i)^{n_i-y_i}(1-\rho_i) + p_i^{\frac{y_i}{n_i}}(1-p_i)^{\frac{(n_i-y_i)}{n_i}}\rho I_{A_{2_i}}(y_i), \tag{4.5}$$

em que $\rho_i = h^*(v(\boldsymbol{r}_i),\phi)$ e $p_i = g^{-1}(\sum_{r=0}^k \beta_r x_{r_i})$.

Considerando os estimadores de Bayes para os parâmetros dos coeficientes de regressão, $\beta_r, r = 1, \ldots, k$, e para o parâmetro da estrutura de correlação, ϕ, podemos predizer o valor de y_i usando (4.3). Para isto, usamos uma estimativa de Monte Carlo da densidade $\pi(y_i|\mathcal{D},\boldsymbol{\theta}^*)$ com amostras MCMC da distribuição a *posteriori* $\pi(\boldsymbol{\theta}^*|\mathcal{D}^*)$ (Chen *et al.*, 2000). Entretanto, para realizar este método, é necessário primeiro fixar um valor para y_i em $\{0, \ldots, n_i\}$. O seguinte algoritmo fornece o valor predito para y_i.

(a) Gere uma amostra $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q$ de tamanho Q de $\pi(\boldsymbol{\theta}^*|\mathcal{D}^*)$. Para cada $\boldsymbol{\theta}_q = (\phi_q, \beta_{0q}, \ldots, \beta_{kq})^\top$, $q = 1, \ldots, Q$, calcule os valores de $\hat{p}_q$ e $\hat{\rho}_q$, sendo

$$\hat{p}_{iq} = g^{-1}\left(\sum_{r=0}^k \hat{\beta}_{rq}x_{ir}\right) \quad \text{e} \quad \hat{\rho}_{iq} = h^*(v(\boldsymbol{r}_i),\hat{\phi}_q).$$

(b) Para cada y_i em $\{0, \ldots, n_i\}$, obtenha a estimativa de Monte Carlo para CPO_i, dado por

$$\widehat{CPO_i} = \left\{ \frac{1}{Q} \sum_{q=1}^{Q} \left(\binom{n_i}{y_i} \hat{p}_{iq}^{\,y_i} (1-\hat{p}_{iq})^{n_i - y_i} (1-\hat{\rho}_{iq}) + \hat{p}_{iq}^{\,\frac{y_i}{n_i}} (1-\hat{p}_{iq})^{\frac{(n_i-y_i)}{n_i}} \hat{\rho}_{iq} I_{A_{2_i}}(y_i) \right)^{-1} \right\}^{-1}.$$

(c) O valor de y_i em $\{0, \ldots, n_i\}$ que maximiza $\widehat{CPO_i}$ é o valor predito para a variável resposta y_i.

No Apêndice D.2 apresentamos a implementação usando o programa R (R Development Core Team, 2011) para calcular o valor da CPO para cada cluster e também obter numericamente o valor predito para variável resposta de cada cluster.

4.3 Diagnósticos

Entre as suposições impostas na construção do modelo de regressão binomial correlacionada, podemos ressaltar: (i) independência entre as variáveis resposta, (ii) as variáveis resposta seguem uma distribuição binomial correlacionada, $BC(n_i, p_i, \rho_i)$, (iii) correlação positiva entre as variáveis de Bernoulli dentro do cluster, $\rho_i > 0$, (iv) função de ligação e (v) estrutura de correlação. Nesta seção, três tipos de resíduos Bayesianos e uma medida de influência local via perspectiva Bayesiana são propostos para a verificação dos pressupostos (i)-(v) e para identificar *outliers* e/ou observações influentes. Para verificar as suposições que as variáveis resposta seguem a distribuição binomial correlacionada $BC(n_i, p_i, \rho_i)$ e que $\rho_i > 0$ podemos observar a significância do parâmetro da estrutura de correlação, γ, por meio dos intervalos de credibilidade inter-quantil e HPD (Chen & Shao, 1999) obtidos no processo inferencial. Para os casos em que esta suposição não for satisfeita, isto é, $\gamma = 0$ ou $\gamma = 1$, o modelo de regressão binomial usual pode ser considerado. A detecção de observações influentes é feita utilizando a divergência de Kullback-Leibler (Cho *et al.*, 2009) envolvendo observações deletadas (Cho *et al.*, 2009).

4.3.1 Resíduos Bayesianos

Após o ajuste do modelo, é de interesse verificar a proximidade dos valores obtidos pelo ajuste em relação aos valores observados no conjunto de dados. Nesta seção, três tipos de resíduos são construídos para o modelo de regressão binomial correlacionada. Um resíduo que depende dos valores preditos via densidade preditiva condicional ordinária (Cho *et al.*, 2009), um outro resíduo que depende da distribuição a *posteriori* dos parâmetros envolvidos no modelo, baseado no resíduo proposto por Albert & Chib (1995) e um resíduo baseado na *deviance* Bayesiana (Spiegelhalter *et al.*, 2002). Nos casos simulados, os resíduos se mostraram igualmente eficientes na detecção de observações influentes (casos perturbados).

Três gráficos, baseados nos resíduos padronizados, podem ser utilizados para verificar as suposições iniciais e identificar má especificação do modelo. Os gráficos (i) resíduo contra a ordem das observações, quando a mesma está disponível, para identificação de dependência temporal das observações; (ii) resíduo contra função das covariáveis, para verificar a necessidade de inserir outras funções de covariáveis, além da linear, na parte sistemática do modelo; (ii) resíduo contra valores preditos para verificar o ajuste global do modelo.

No Apêndice D.3 apresentamos a implementação usando o programa R (R Development Core Team, 2011) dos resíduos Bayesianos padronizados.

Resíduos baseado na densidade preditiva condicional ordinária

O resíduo baseado na densidade preditiva condicional ordinária para a i-ésima observação, r_i^{pp}, é calculado como $r_i^{pp} = y_i - \tilde{y}_i$, em que y_i é a i-ésima resposta observada e $\tilde{y}_i$ é a moda da distribuição preditiva condicional ordinária condicionada aos valores das covariáveis, $(x_{i0}, \ldots, x_{ik}, r_{i11}, \ldots, r_{i1n_i}, r_{i21}, \ldots, r_{i2n_i}, r_{iq1}, \ldots, r_{iqn_i})^\top$, em que x_{ir} é o valor da r-ésima covariável dentro do i-ésimo cluster, $i = 1, \ldots, m$ e $r = 0, \ldots, k$, e r_{ilj} é o valor da l-ésima covariável para o j-ésimo indivíduo dentro do i-ésimo cluster, $i = 1, \ldots, m$, $l = 1, \ldots, q$ e $j = 1, \ldots, n_i$. O resíduo padronizado, r_i^{spp}, baseado na densidade preditiva condicional ordinária para o modelo de regressão binomial correlacionada é definido como

$$r_i^{spp} = \frac{r_i^{pp}}{\sqrt{\hat{p}_i(1-\hat{p}_i)\{n_i + \hat{\rho}_i n_i(n_i-1)\}}}, \qquad i = 1, \ldots, m, \tag{4.6}$$

com $\hat{p}_i$ e $\hat{\rho}_i$ estimadores de Bayes de p_i e ρ_i, respectivamente.

É esperado que o conjunto de resíduos esteja próximo de zero. Se isso não ocorrer, isto é, se existir um conjunto de pontos afastado de zero temos um indicativo de que o modelo está mal ajustado aos dados.

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

O resíduo padronizado baseado na distribuição a *posteriori* dos parâmetros para o modelo de regressão binomial correlacionada é definido pela transformação

$$R_{iq}^{spd} = \frac{y_i - E(Y_i|\mathcal{D}, \boldsymbol{\theta}_q)}{\sqrt{\mathrm{Var}(Y_i|\mathcal{D}, \boldsymbol{\theta}_q)}} = \frac{y_i - n_i p_{iq}}{\sqrt{p_{iq}(1-p_{iq})(n_i + \rho_{iq} n_i(n_i-1))}}, \tag{4.7}$$

com $i = 1, \ldots, m$ e $q = 1, \ldots, Q$; em que p_{iq} e ρ_{iq} são não observados e n_i e y_i são informações da amostra para a i-ésima observação. O novo índice q presente em R_{iq}^{spd} se faz necessário para justificar, para cada i, uma amostra de tamanho Q destes resíduos. Como a relação (4.7) é uma função do vetor de parâmetros, $\boldsymbol{\theta}_q = (\phi_q, \beta_{0q}, \ldots, \beta_{kq})^\top$, ela carrega toda a incerteza refletida nos parâmetros a *posteriori*, refletindo na distribuição a *posteriori* dos resíduos, R_i^{spd}, (Albert & Chib, 1995). A distribuição a *posteriori* dos resíduos pode ser resumida usando amostras de um amostrador MCMC. Isto é, ao gerar uma amostra $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q$ de tamanho Q de $\pi(\boldsymbol{\theta}^*|\mathcal{D}^*)$, obtemos para o par (y_i, n_i), $i =$

$1, \ldots, m$, uma amostra de tamanho Q dos resíduos em (4.7). É esperado que a média dos resíduos esteja centrada em zero. Se isso não ocorrer, ou se a amostra dos resíduos apresentar uma alta variabilidade, a observação pode ser considerada um *outlier*.

Resíduos baseado na *deviance* Bayesiana

O resíduo *deviance* Bayesiano para a i-ésima observação, é uma composição da dificuldade de estimar $\boldsymbol{\theta}$ na presença de y_i, p_{D_i}, e da contribuição de cada observação para o total global da *deviance*, $D_i(\hat{\boldsymbol{\theta}})$. O resíduo *deviance* Bayesiano é definido para o modelo de regressão binomial correlacionada por

$$r_i^{db} = \text{sinal}(y_i - n_i\hat{p}_i)\sqrt{p_{D_i} + D_i(\hat{\boldsymbol{\theta}})}, \qquad i = 1, \ldots, m, \tag{4.8}$$

sendo o valor de p_{D_i} para a i-ésima observação, aproximado via método de Monte Carlo por

$$p_{D_i} = -\frac{2}{Q}\sum_{q=1}^{Q}\left\{\log\frac{\pi(\boldsymbol{\theta}_q|\mathcal{D}^*_{(i)})}{\pi(\boldsymbol{\theta}_q)}\right\} + 2\log\frac{\pi(\hat{\boldsymbol{\theta}}|\mathcal{D}^*_{(i)})}{\pi(\hat{\boldsymbol{\theta}})}, \tag{4.9}$$

em que $\boldsymbol{\theta}_q = (\gamma_q, \beta_{0q}, \ldots, \beta_{kq})^\top$ representa uma amostra de tamanho Q gerada da distribuição a *posteriori* $\pi(\boldsymbol{\theta}^*|\mathcal{D}^*)$ e $\hat{\boldsymbol{\theta}}$ é o estimador de Bayes de $\boldsymbol{\theta}$; $\pi(\boldsymbol{\theta}_q|\mathcal{D}^*_{(i)})$ e $\pi(\hat{\boldsymbol{\theta}}|\mathcal{D}^*_{(i)})$ são a distribuição a *posteriori* aplicada em $\boldsymbol{\theta}_q$ e em $\hat{\boldsymbol{\theta}}$, respectivamente, e $\pi(\boldsymbol{\theta}_q)$ e $\pi(\hat{\boldsymbol{\theta}})$ são a distribuição a *priori* aplicada em $\boldsymbol{\theta}_q$ e $\hat{\boldsymbol{\theta}}$, respectivamente. O valor de D_i para a i-ésima observação é dado por

$$D_i(\hat{\boldsymbol{\theta}}) = -2\log\left\{\binom{n_i}{y_i}\hat{p}_i^{y_i}(1-\hat{p}_i)^{n_i-y_i}(1-\hat{\rho}_i) + \hat{p}_i^{\frac{y_i}{n_i}}(1-\hat{p}_i)^{\frac{(n_i-y_i)}{n_i}}\hat{\rho}_i I_{A_{2_i}}(y_i)\right\}. \tag{4.10}$$

Assim, o resíduo *deviance* Bayesiano padronizado para a i-ésima observação pode ser obtido através da expressão

$$r_i^{sdb} = \frac{r_i^{db}}{\sqrt{\hat{p}_i(1-\hat{p}_i)\{n_i + \hat{\rho}_i n_i(n_i-1)\}}}, \qquad i = 1, \ldots, m, \tag{4.11}$$

sendo $\hat{p}_i = g^{-1}(\sum_{r=0}^{k}\hat{\beta}_r x_{ir})$ e $\hat{\rho}_i = h(v(\boldsymbol{r}_i), \hat{\gamma}))$.

Ao plotar os valores dos resíduos *deviance*, r_i^{db}, contra os valores de p_{D_i} para um determinado modelo ajustado, considerando marcações de curvas da forma $x^2 + y = c$, temos os pontos situados ao longo de tal parábola como indicativo da quantia de contribuição $\text{DIC}_i = c$ da observação no valor global do DIC (Spiegelhalter *et al.*, 2002). A Seção 4.5 e 5.5.2 apresentam uma ilustração deste gráfico para diferentes funções de ligação do modelo de regressão binomial correlacionada.

Espera-se que o conjunto de resíduos *deviance* padronizados esteja próximo de zero. Se isso não ocorrer, ou seja, se há um conjunto de pontos longe de zero, é uma indicação de que o modelo está mal ajustado aos dados.

4.3.2 Diagnóstico de influencia Bayesiano

Para avaliar a sensibilidade na estimação dos parâmetros, é considerado um diagnóstico Bayesiano de influência via deleção de caso (Cook & Weisberg, 1982) baseado na divergência de Kullback-Leibler (K-L) (Cho *et al.*, 2009), o qual é calculado utilizando a densidade preditiva condicional ordinária definida para o modelo de regressão binomial correlacionada em (4.3). A divergência de K-L entre a distribuição a *posteriori* com os dados completos, $\mathcal{D}^*$, e a distribuição a *posteriori* com a i-ésima observação deletada, $\mathcal{D}^*_{(-i)}$, é definida por

$$K\left(\pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*\right), \pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*_{(-i)}\right)\right) = \int_\Theta \pi(y_i|\mathcal{D}, \boldsymbol{\theta}^*) \log\left(\frac{\pi(\boldsymbol{\theta}^*|\mathcal{D}^*)}{\pi(\boldsymbol{\theta}^*|\mathcal{D}^*_{(-i)})}\right) d\boldsymbol{\theta}^*. \quad (4.12)$$

Primeiramente considere,

$$\frac{\pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*\right)}{\pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*_{(-i)}\right)} = \frac{\mathcal{L}(\boldsymbol{\theta}^*; \mathcal{D}^*) \int_\Theta \mathcal{L}(\boldsymbol{\theta}^*; \mathcal{D}^*_{(-i)})\pi(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^*}{\mathcal{L}(\boldsymbol{\theta}^*; \mathcal{D}^*_{(-i)}) \int_\Theta \mathcal{L}(\boldsymbol{\theta}^*; \mathcal{D}^*)\pi(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^*}$$

$$= \pi(y_i|\mathcal{D}, \boldsymbol{\theta}^*) \frac{\int_\Theta \mathcal{L}(\boldsymbol{\theta}^*; \mathcal{D}^*_{(-i)})\pi(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^*}{\int_\Theta \pi(y_i|\mathcal{D}, \boldsymbol{\theta}^*)\mathcal{L}(\boldsymbol{\theta}^*; \mathcal{D}^*_{(-i)})\pi(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^*}$$

$$= \pi(y_i|\mathcal{D}, \boldsymbol{\theta}^*) \frac{\int_\Theta \mathcal{L}(\boldsymbol{\theta}^*; \mathcal{D}^*_{(-i)})\pi(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^*}{\int_\Theta \pi(y_i|\mathcal{D}, \boldsymbol{\theta}^*)\left(\pi(\boldsymbol{\theta}^*|\mathcal{D}^*_{(-i)}) \int_\Theta \mathcal{L}(\boldsymbol{\theta}^*; \mathcal{D}^*_{(-i)})\pi(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^*\right) d\boldsymbol{\theta}^*}$$

$$= \pi(y_i|\mathcal{D}, \boldsymbol{\theta}^*) \frac{\int_\Theta \mathcal{L}(\boldsymbol{\theta}^*; \mathcal{D}^*_{(-i)})\pi(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^*}{\int_\Theta \mathcal{L}(\boldsymbol{\theta}^*; \mathcal{D}^*_{(-i)})\pi(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^* \int_\Theta \pi(y_i|\mathcal{D}, \boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*|\mathcal{D}^*_{(-i)})d\boldsymbol{\theta}^*}$$

$$= \pi(y_i|\mathcal{D}, \boldsymbol{\theta}^*) \frac{1}{\int_\Theta \pi(y_i|\mathcal{D}, \boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*|\mathcal{D}^*_{(-i)})d\boldsymbol{\theta}^*}. \quad (4.13)$$

Aplicando a função log em ambos os lados de (4.13), obtemos

$$\log\left(\frac{\pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*\right)}{\pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*_{(-i)}\right)}\right) = \log(\pi(y_i|\mathcal{D}, \boldsymbol{\theta}^*)) - \log\left(\int_\Theta \pi(y_i|\mathcal{D}, \boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*|\mathcal{D}^*_{(-i)})d\boldsymbol{\theta}^*\right). \quad (4.14)$$

Substituindo (4.14) e (4.3) em (4.12), a divergência de K-L pode ser escrita como

$$\begin{aligned} K\left(\pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*\right), \pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*_{(-i)}\right)\right) &= \int_\Theta \log\left(\pi(y_i|\mathcal{D}, \boldsymbol{\theta}^*)\right) \pi(\boldsymbol{\theta}^*|\mathcal{D}^*)d\boldsymbol{\theta}^* \\ &\quad - \int_\Theta \log(CPO_i)\pi(\boldsymbol{\theta}^*|\mathcal{D})d\boldsymbol{\theta}^* \\ &= E_{\boldsymbol{\theta}}\left[\, \log\{\pi(y_i|\mathcal{D}, \boldsymbol{\theta}^*)\}|\mathcal{D}\, \right] - \log\left(CPO_i\right), \quad (4.15) \end{aligned}$$

em que $\pi(y_i|\mathcal{D},\boldsymbol{\theta}^*)$ é dada em (4.5).

A estimação da divergência de K-L pode ser conduzida utilizando o método de Monte Carlo com amostras MCMC da distribuição a *posteriori* $\pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*\right)$. Então, se $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_Q$ é uma amostra de tamanho Q de $\pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*\right)$, a estimação de Monte Carlo para (4.15) é dada por

$$\hat{K}\left(\pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*\right),\pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*_{(-i)}\right)\right)=\frac{1}{Q}\sum_{q=1}^{Q}\log[\pi(y_i|\mathcal{D},\boldsymbol{\theta}_q)]+\log\left\{\frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{\pi(y_i|\mathcal{D},\boldsymbol{\theta}_q)}\right\}. \quad (4.16)$$

Grandes valores da divergência de K-L evidenciam a presença de pontos influentes no conjunto de dados. Para verificar se a i-ésima observação é um ponto influente ou não, é determinada uma calibração da divergência de K-L, como mostrado em Cho *et al.* (2009) e Peng & Dey (1995), dada por

$$p_i^*=0,5\left[1+\sqrt{1-\exp\left\{-2\hat{K}\left(\pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*\right),\pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*_{(-i)}\right)\right)\right\}}\right].$$

Se o valor de p_i^* é muito maior que 0,5, o i-ésimo caso é considerado influente.

No Apêndice D.4 apresentamos a implementação usando o programa R (R Development Core Team, 2011) dos valores da divergência de K-L para cada cluster.

4.4 Critérios de seleção de modelo

É possível ajustarmos um conjunto potencial de modelos de regressão binomial correlacionada utilizando o mesmo conjunto de dados. Isto ocorre quando, por exemplo, consideramos diferentes subconjuntos de covariáveis ou diferentes estruturas de correlação e/ou diferentes funções de ligação no processo de análise. Nestes casos é conveniente o uso de um critério de seleção de modelos.

O critério de seleção Bayesiano (DIC) (Spiegelhalter *et al.*, 2002) para o modelo de regressão binomial correlacionada é determinado por

$$\widehat{DIC}=2\bar{D}-\hat{D},$$

em que

$$\bar{D}=\frac{1}{Q}\sum_{q=1}^{Q}D(\boldsymbol{\theta}_q)\quad\text{e}\quad\hat{D}=D\left(\frac{1}{Q}\sum_{q=1}^{Q}\beta_{0q},\ \ldots,\frac{1}{Q}\sum_{q=1}^{Q}\beta_{kq},\ \frac{1}{Q}\sum_{q=1}^{Q}\gamma_q\right),$$

com $D(\boldsymbol{\theta}_q)=-2\sum_{i=1}^{m}\log\left(\pi(y_i|\mathcal{D},\boldsymbol{\theta}_q)\right)$, e $\boldsymbol{\theta}_q=(\phi_q,\beta_{0q},\ldots,\beta_{kq})^\top$ uma amostra gerada da distribuição a *posteriori* $\pi\left(\boldsymbol{\theta}^*|\mathcal{D}^*\right)$. A função $\pi(y_i|\mathcal{D},\boldsymbol{\theta}_q)$ é como definida em (4.5). Para calcular $\hat{D}$, gere Q amostras MCMC $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_q$ e determine $\frac{1}{Q}\sum_{q=1}^{Q}\gamma_q$, $\frac{1}{Q}\sum_{q=1}^{Q}\beta_{0q},\ldots,$ $\frac{1}{Q}\sum_{q=1}^{Q}\beta_{kq}$. Esses valores são usados diretamente em $-2\sum_{i=1}^{m}\log\left(\pi(y_i|\mathcal{D},\boldsymbol{\theta})\right)$. O modelo com melhor ajuste é aquele que apresenta o menor valor de DIC.

A função apresentada no Apêndice D.3 fornece o valor do DIC, a implementação é feita usando o programa R (R Development Core Team, 2011).

4.5 Estudos de simulação

Nesta seção, consideramos um estudo de simulação com apenas uma amostra para ilustrar o processo inferencial, a sensibilidade na inferência dos parâmetros com diferentes distribuições a *priori* e o desempenho das medidas de diagnóstico propostas neste trabalho. Além disto, um outro estudo de simulação com 1000 amostras foi realizado para analisar as propriedades frequentistas dos estimadores Bayesianos.

Dados simulados

A geração da amostra envolve $m = 100$ clusters com as variáveis resposta, Y_i, $i = 1, \ldots, 100$, seguindo uma distribuição $BC(n_i, p_i, \rho_i)$, com os n_i gerados de uma distribuição B(45; 0,5). O parâmetro da estrutura de correlação utilizado na simulação é $\gamma = 0,2$ e os $v(\boldsymbol{r}_i)$ assumem valores de uma distribuição U(0,1). Duas covariáveis são consideradas, x_{i1} e x_{i2}. Os valores das covariáveis x_{i1} são provenientes de uma distribuição U(0,2) e os valores da covariável x_{i2} são provenientes de uma distribuição N(0,4). Os valores dos coeficientes de regressão são $\beta_0 = 2$, $\beta_1 = -2$ e $\beta_3 = -2$. São considerados neste estudo a função de ligação logito e a estrutura de correlação exponencial. A distribuição a *priori* para o parâmetro β_r é N(0,λ_r), $r = 0, 1, 2$ e para o parâmetro ϕ é N(0,λ_3), com $\lambda_r = 10.000$, $r = 0, 1, 2, 3$.

Estimação

As quatro funções de ligação, logito, complementar log-log, log-log e probito, foram consideradas na análise. Como esperávamos, os valores obtidos pelo critério DIC de seleção de modelos confirma a função de ligação logito como o melhor ajuste, conforme pode ser visto na Tabela 4.1.

Tabela 4.1: Valores obtidos pelo critério DIC ajustando o modelo de regressão binomial correlacionada com diferentes funções de ligação.

Critério	Logito	Complementar log-log	Log-log	Probito
DIC	112,597	119,840	129,030	115,141

Os resumos da distribuição a *posteriori* para o modelo com ligação logito são mostrados na Tabela 4.2. Note que ambos os intervalos de credibilidade inter-quantil e HPD contêm o verdadeiro valor dos parâmetros. É importante ressaltar que os intervalos de credibilidade do parâmetro da estrutura de correlação, γ, não contêm o valor zero, corroborando com a necessidade de ajuste de um modelo binomial correlacionada. Ressaltamos também que, ao ajustar estes dados considerando o modelo de regressão binomial usual, obtivemos o valor do DIC igual a 881,275. Ambos os modelos usual e binomial correlacionada, foram ajustados com a função de ligação logito, evidenciando que o modelo de regressão binomial correlacionada é mais adequado a estes dados. O diagnóstico de convergência dos parâmetros foi verificado via CODA, por meio do teste de Geweke que apresentou valores entre (-1,5; 1,5).

Tabela 4.2: Resumos das densidades marginais a *posteriori* de γ, β_0, β_1 e β_2 para a função de ligação logito.

	Valor verdadeiro	Média a *posteriori*	Mediana a *posteriori*	Intervalo inter-quantil credibilidade 95%	Intervalo HPD credibilidade 95%
γ	0,20	0,158	0,151	(0,065 ; 0,294)	(0,047 ; 0,269)
β_0	2,00	2,995	2,930	(1,953 ; 4,260)	(1,991 ; 4,269)
β_1	-2,00	-2,496	-2,462	(-3,669 ; -1,531)	(-3,519 ; -1,455)
β_2	-2,00	-2,305	-2,266	(-2,955 ; -1,777)	(-2,911 ; -1,761)

Análise de resíduos

Como mencionado na Seção 4.3.1, Spiegelhalter *et al.* (2002) sugerem analisar o gráfico dos resíduos *deviance* Bayesianos contra a alavancagem das observações, p_{D_i}, considerando marcações que permitem avaliar a contribuição de cada observação no valor global do DIC. Na Figura 4.1 são apresentados estes gráficos para os ajustes do modelo de regressão binomial correlacionada com as quatro funções de ligação: Figura 4.1a logito, Figura 4.1b complementar log-log, Figura 4.1c log-log e Figura 4.1d probito. As contribuições das observações são muito semelhantes para todas as funções de ligação, principalmente para o modelo ajustado com as ligações logito e probito. A Figura 4.1e mostra este gráfico para o ajuste do modelo de regressão binomial usual (MRB) com a função de ligação logito, justificando o alto valor do DIC para este ajuste em relação ao modelo proposto.

Dois gráficos de resíduos padronizados contra valores preditos, apresentados na Figura 4.3a e 4.3c, e um gráfico do resíduo padronizado contra o valor esperado, apresentado na Figura 4.3b, verificam o ajuste global do modelo, mostrando coerência e ausência de *outliers*. Não identificamos a necessidade de inserir outras funções das covariáveis no modelo, assim, os gráficos envolvendo as funções das covariáveis contra os resíduos, sugeridos na Seção 4.3.1, foram omitidos. A média das amostras geradas das distribuições a *posteriori* dos resíduos padronizados são mostradas na Figura 4.2, e seus respectivos intervalos HPD versus valores esperados de $E(Y_i|\mathcal{D}, \boldsymbol{\theta})$ são mostrados na Figura 4.3b. As amplitudes dos boxplots, apresentados nessa figura, são pequenas devido a baixa variabilidade da distribuição. Os pontos que apresentam maiores dispersões para as distribuições dos resíduos são os pontos que requerem uma maior investigação.

Diagnósticos de influência

Nessa seção examinamos o desempenho das medidas de diagnósticos de influência, a divergência de K-L e a calibração. Para isto, as observações 6 e 42 foram perturbadas, criando assim observações influentes no conjunto de dados. As perturbações foram feitas na covariável x_{i2}, na forma $x_{i2} = x_{i2} + 8sd(x_2)$, $i = 6$ e 42, em que $sd(x_2)$ corresponde ao desvio padrão da covariável x_2. Combinações de presença e ausência destas observações perturbadas foram usadas para formar novos conjuntos de dados. Estes novos conjuntos de dados foram utilizados na obtenção de estimativas dos parâmetros do modelo. A Tabela

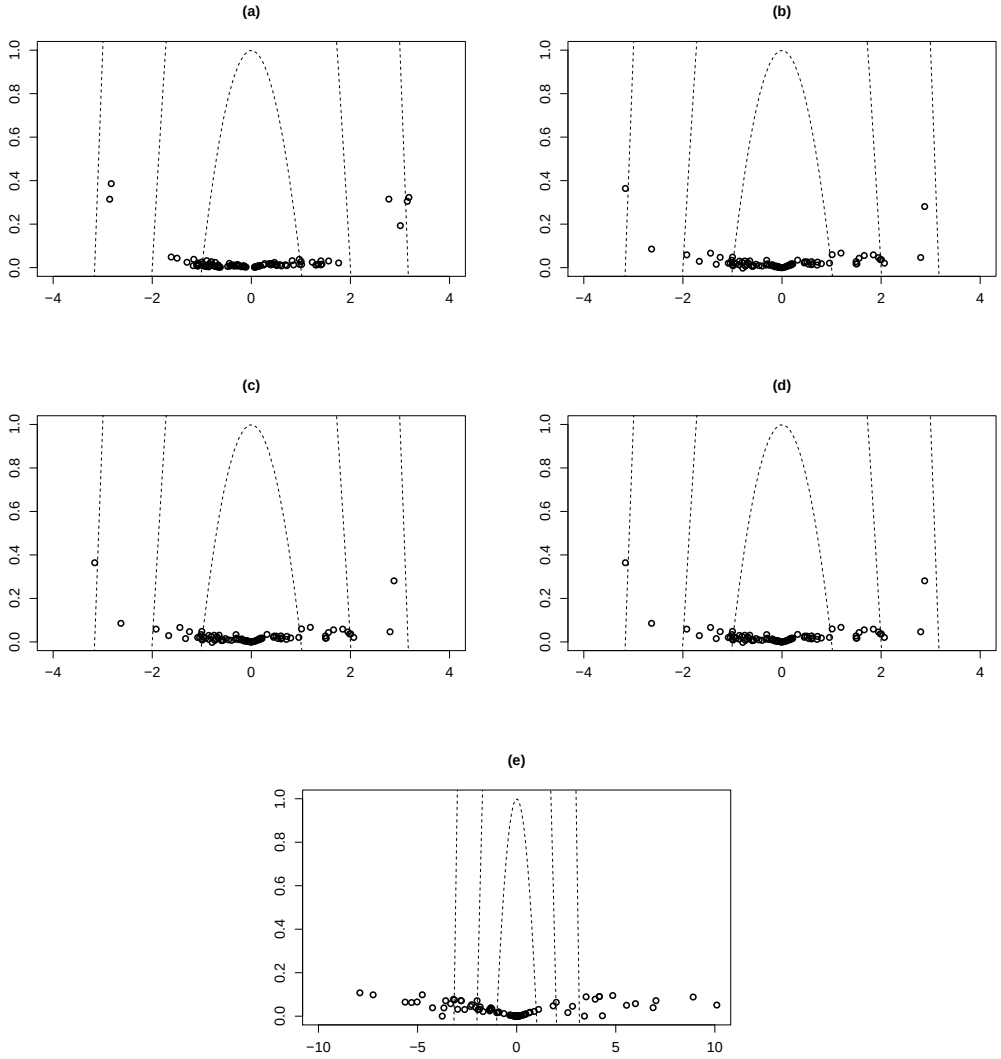

Figura 4.1: Resíduos *deviance* versus alavancagem Bayesiana, considerando $c = 1, 4$ e 10, (a) MRBC: ligação logito, (b) MRBC: ligação complementar log-log, (c) MRBC: ligação log-log, (d) MRBC: ligação probito, (e) MRB: ligação logito.

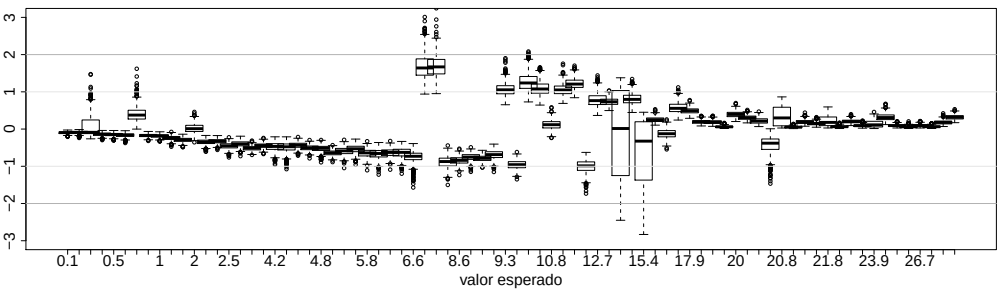

Figura 4.2: Boxplots das amostras da distribuição a *posteriori* dos resíduos para cada observação versus o valor esperado de $E(Y_i|\mathcal{D}, \boldsymbol{\theta})$.

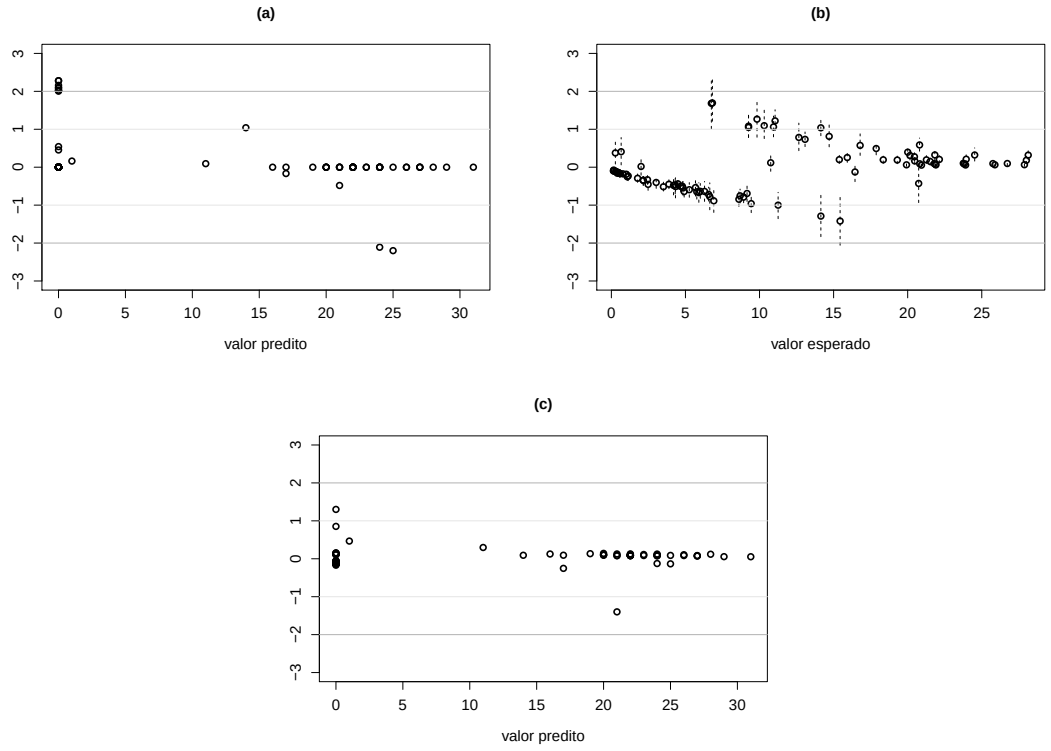

Figura 4.3: (a) resíduos padronizados via densidade preditiva condicional ordinária versus valores preditos, (b) médias e intervalos HPD (95%) das amostras dos resíduos padronizados via distribuição a *posteriori* versus valores esperados, (c) resíduos *deviance* padronizados versus valores preditos.

4.3 apresenta a média da distribuição a *posteriori* e a mudança relativa da estimativas dos parâmetros em relação aos dados originais.

Como mencionado na Seção 4.3.2, valores altos da divergência de K-L evidenciam a presença de pontos influentes no conjunto de dados e valores altos da calibração, próximas de um, confirmam esta presença. A Tabela 4.4 apresenta o valor da divergência de K-L e da calibração para os casos analisados na Tabela 4.3. Por exemplo, ao analisarmos o conjunto de dados sem perturbações, obtemos DKL = 0,007 e DKL = 0,009 para os casos 6 e 42, respectivamente. Porém, ao analisarmos o conjunto com os casos 6 e 42 perturbados, obtemos DKL = 1,654 e DKL = 2,778 para os casos 6 e 42, respectivamente. Além disso, o valor da calibração na ausência de perturbação é mais próxima de 0,5, mas na presença de caso perturbados no conjunto de dados esta métrica se aproxima de um, confirmando a influência da observação da estimativa dos parâmetros.

4.5.1 Sensibilidade em relação a distribuição a *priori*

Para avaliar a sensibilidade do modelo na escolha da distribuição a *priori*, conduzimos um estudo usando dois diferentes tipos de distribuição a *priori*, $N(0, 10^4)$ e Constante, para cada um dos parâmetros $\beta_r, r = 0, 1, 2$, e duas diferentes *prioris*, $LN(0, 10^4)$ e $\mathrm{G}(10^{-4}, 10^{-4})$, para γ e Constante para $\log(\gamma)$. O conjunto de dados simulado corre-

Tabela 4.3: Média a *posteriori* e mudança relativa em relação aos dados originais para os dados simulados.

Casos	γ		β_0		β_1		β_2	
perturbados	Média	%	Média	%	Média	%	Média	%
{Nenhum}	0,16	-	3,00	-	-2,50	-	-2,31	-
Caso {6}	0,20	25,00	1,84	-38,67	-1,62	-35,20	-1,69	-26,84
Caso {42}	0,16	-	1,73	-42,33	-1,50	-40,00	-1,61	-30,30
Casos {6} e {42}	0,21	31,25	1,47	-51,00	-1,49	-40,40	-1,24	-46,32

Tabela 4.4: Valores obtidos para divergência de K-L e calibração para situações de perturbação dos dados.

	Casos analisados			
Casos	{6}		{42}	
perturbados	DKL	Calibração	DKL	Calibração
{Nenhum}	0,007	0,558	0,010	0,570
Caso {6}	2,939	0,999	0,009	0,565
Caso {42}	0,008	0,562	2,891	0,999
Casos {6} e {42}	1,654	0,991	2,778	0,999

sponde ao mesmo analisado na Seção 4.5. Os resumos a *posteriori*, estimativas pontuais (médias e medianas) e os intervalos inter-quantil e HPD com 95% de credibilidade, para os parâmetros $\beta_0, \beta_1, \beta_2$, para ambas as distribuições a *priori*, são extremamente similares. Os resumos a *posteriori* são quase os mesmos, usando as *prioris* $LN(0, 10^4)$ e Constante, para o parâmetro γ e $\log(\gamma)$, respectivamente. Então, apenas os resultados usando a *prior* $N(0, 10^4)$ para os parâmetros $\beta_r, r = 0, 1, 2$, e os resultados usando as *priors* $LN(0, 10^4)$ e $\mathrm{G}(10^{-4}, 10^{-4})$ para o parâmetro γ são mostrados na Tabela 4.5.

Tabela 4.5: Resumos das distribuições marginais a *posteriori* para os parâmetros β_0, β_1, β_2 e γ.

	Distribuição a *priori*	Média a *posteriori*	Mediana a *posteriori*	Intervalo inter-quantil credibilidade 95%	Intervalo HPD credibilidade 95%
β_0	$N(0, 10^4)$	2,995	2,930	(1,953 ; 4,260)	(1,991 ; 4,269)
β_1	$N(0, 10^4)$	-2,496	-2,462	(-3,669 ; -1,531)	(-3,519 ; -1,455)
β_2	$N(0, 10^4)$	-2,305	-2,266	(-2,955 ; -1,777)	(-2,911 ; -1,761)
γ	$LN(0, 10^4)$	0,158	0,151	(0,065 ; 0,294)	(0,047 ; 0,269)
γ	$\mathrm{G}(10^{-4}, 10^{-4})$	0,178	0,173	(0,077 ; 0,310)	(0,070 ; 0,299)

4.5.2 Propriedade frequentista dos estimadores Bayesianos

As estimativas das probabilidades de cobertura são baseadas em 1.000 simulações com $m = 500$ clusters e os cenários são simulados como descrito na Seção 4.5, exceto pelas estrutura de correlação e função de ligação utilizadas. As quatro funções de ligação, logito,

complementar log-log, log-log e probito, e as três estruturas de correlação, exponencial, Gaussiana e AR contínua, são usadas nesta análise. A distribuição a *priori* $N(0, 10^4)$ é considerada para cada um dos parâmetros β_0, β_1, β_2 e ϕ. Os resultados apresentados na Tabela 4.6 indicam que as estimativas das probabilidades de cobertura estão entre 92% e 96%, considerando uma probabilidade de cobertura nominal de 95%, para os quatro parâmetros em todos os cenários. Para a estrutura de correlação exponencial, as estimativas das probabilidades de cobertura dos intervalos de credibilidade inter-quantil e HPD para β_0, β_1 e β_2, usando a função de ligação logito, são 94% ou 95%, entretanto para γ, essas estimativas são 94% e 93% respectivamente. Resultados similares para β_0, β_1 e β_2 são encontrados para a estrutura de correlação Gaussiana. Entretanto, para γ, as estimativas das probabilidades de cobertura estão entre 91% e 93%. Para a estrutura de correlação AR contínua as estimativas das probabilidades de cobertura para γ estão entre 94% e 95%.

Tabela 4.6: Probabilidades de cobertura estimadas para os intervalos de credibilidade inter-quantil e HPD para diferentes funções de ligação e estruturas de correlação.

		Estrutura de Correlação					
		Exponencial		Gaussiana		AR Contínua	
	Função de Ligação	Inter-quantil	HPD	Inter-quantil	HPD	Inter-quantil	HPD
	Logito	0,95	0,94	0,95	0,95	0,92	0,93
	C. log-log	0,94	0,95	0,94	0,94	0,94	0,93
	Log-log	0,94	0,94	0,93	0,93	0,93	0,93
β_0	Probito	0,93	0,93	0,95	0,95	0,94	0,94
	Logito	0,94	0,95	0,95	0,95	0,94	0,93
	C. log-log	0,94	0,94	0,94	0,94	0,93	0,93
	Log-log	0,95	0,94	0,93	0,93	0,94	0,94
β_1	Probito	0,94	0,94	0,96	0,95	0,94	0,93
	Logito	0,95	0,95	0,94	0,94	0,94	0,93
	C. log-log	0,94	0,93	0,94	0,94	0,95	0,94
	Log-log	0,93	0,93	0,93	0,93	0,92	0,92
β_2	Probito	0,93	0,94	0,95	0,95	0,94	0,93
	Logito	0,94	0,93	0,93	0,93	0,94	0,94
	C. log-log	0,94	0,94	0,93	0,91	0,95	0,95
	Log-log	0,96	0,96	0,93	0,92	0,94	0,94
γ	Probito	0,95	0,94	0,92	0,92	0,94	0,95

Probabilidade de cobertura nominal de 95%.

A Tabela 4.7 apresenta a média dos erros quadráticos médios e vícios das medianas a *posteriori* considerando as quatro funções de ligação e as três estruturas de correlação. Para determinar estes valores, geramos 1.000 amostras MCMC das distribuições a *posteriori* para cada parâmetro. Em cada amostra MCMC, são calculados a mediana a *posteriori*, o erro quadrático médio e o vício. Por fim, as médias dos 1,000 EQMs e vícios, para cada

parâmetro são determinadas. Os resultados mostram que os valores estão próximos de zero. Resultados similares são obtidos para a média a *posteriori*. A estrutura de correlação AR contínua apresenta os melhores resultados globais para erro quadrático médio e vício, e a estrutura de correlação Gaussiana apresenta os piores resultados globais para erro quadrático médio e vício.

Tabela 4.7: Médias dos erro quadráticos médio (EQM) e vícios da mediana a *posteriori* para diferentes funções de ligação e estruturas de correlação.

		Estrutura de Correlação					
		Exponencial		Gaussiana		AR Contínua	
	Função de ligação	EQM	Vício	EQM	Vício	EQM	Vício
	Logito	0,058	0,006	0,133	-0,006	0,015	-0,002
	C. log-log	0,034	0,013	0,077	0,011	0,008	0,003
	Log-log	0,080	0,020	0,189	0,029	0,020	0,000
β_0	Probito	0,023	0,004	0,059	0,022	0,006	0,001
	Logito	0,056	-0,006	0,122	0,008	0,014	0,003
	C. log-log	0,035	-0,011	0,079	-0,009	0,008	-0,003
	Log-log	0,067	-0,020	0,160	-0,031	0,016	-0,002
β_1	Probito	0,023	-0,006	0,067	-0,012	0,007	-0,002
	Logito	0,021	-0,014	0,066	-0,012	0,005	-0,001
	C. log-log	0,018	-0,012	0,041	-0,018	0,004	-0,003
	Log-log	0,050	-0,021	0,112	-0,035	0,012	-0,003
β_2	Probito	0,012	-0,007	0,049	-0,018	0,004	-0,002
	Logito	0,002	-0,001	0,003	-0,008	0,001	-0,001
	C. log-log	0,002	-0,001	0,003	-0,008	0,001	0,000
	Log-log	0,002	0,000	0,003	-0,009	0,001	0,001
γ	Probito	0,002	-0,000	0,003	-0,007	0,001	0,001

Capítulo 5

Modelo de regressão beta-binomial (MRBB): Diagnóstico Bayesiano

Para dados do tipo binomial com eventos de Bernoulli dependentes não é aconselhável modelar a probabilidade de sucesso ajustando um modelo de regressão binomial usual, pois a suposição de independência entre as variáveis de Bernoulli não é satisfeita. Como uma alternativa, é proposta nesta tese uma classe de modelos de regressão binomial correlacionada. Um idéia natural é comparar o modelo proposto com uma classe de modelos que também permitam ajustar dados do tipo binomial com eventos de Bernoulli dependentes. Observamos na literatura que conjuntos de dados desta natureza são frequentemente modelados usando regressão beta-binomial. Este modelo é baseado na distribuição beta-binomial (Skellam, 1948), obtida pela mistura da distribuição binomial e da distribuição beta, ou melhor, se $Y_i|p_i \sim B(n_i, p_i)$ com $p_i \sim \text{Beta}(\alpha_1, \alpha_2)$, então a distribuição marginal de Y_i resulta na distribuição beta-binomial com parâmetros α_1 e α_2. Outras escolhas para a distribuição de p_i estão presentes na literatura, por exemplo, a distribuição retangular contínua (Horsnell, 1957), a distribuição triangular (Horsnell, 1957), a distribuição binomial multiplicativa (Altham, 1978), a distribuição logística-normal (Williams, 1982), a distribuição probito-normal (Ochi & Prentice, 1984) e a distribuição binomial dupla (Efron, 1986; Lindsey, 1995).

Prentice (1986), Lindsey & Altham (1998) e Kahn & Raftery (1996) apresentaram uma metodologia clássica para o modelo de regressão beta-binomial e Kahn & Raftery (1996) apresentaram uma abordagem Bayesiana. Nestes artigos, a modelagem da probabilidade de sucesso é feita por meio da função de ligação logito. Prentice (1986) estendeu a distribuição beta-binomial permitindo também a correlação negativa entre os eventos de Bernoulli.

Na construção dos modelos de regressão beta-binomial, em geral, é considerada a função de ligação logito ao modelar a probabilidade de sucesso (Prentice, 1986; Kahn & Raftery, 1996; Lindsey & Altham, 1998), e ao modelar a correlação entre os eventos de Bernoulli (Prentice, 1986; Lindsey & Altham, 1998). Neste capítulo, a probabilidade de sucesso correspondente ao evento de interesse para um determinado cluster é modelada usando uma entre quatro diferentes funções de ligação (logito, complementar log-log, log-

log e probito) envolvendo cováriaveis dos clusters. E a dependência entre os indivíduos presentes no mesmo cluster é modelada usando três diferentes estruturas de correlação (exponencial, AR contínua e Gaussiana) (Jennrich & Schluchter, 1986; Zimmerman & Harville, 1991; Cressie, 1993; Russell, 1996; Sherman, 2011) as quais envolvem covariáveis dos indivíduos dentro do cluster. As análises, considerando o modelo de regressão beta-binomial, geralmente limitam-se a apresentação dos estimadores dos parâmetros e critérios para seleção do modelo com melhor ajuste. O objetivo deste capítulo é explorar uma análise de diagnóstico mais minuciosa, envolvendo análise de resíduos e medidas de influência para detecção de *outliers* e/ou observações influentes para o modelo de regressão beta-binomial via abordagem Bayesiana. Para a construção da análise de diagnóstico, é considerada uma parametrização específica do modelo de regressão beta-binomial (Prentice, 1986) apresentada na Seção 5.1. Neste capítulo também é apresentado a análise de dois conjuntos de dados reais via modelo de regressão binomial correlacionada e modelo de regressão beta-binomial.

5.1 Modelos de regressão beta-binomial

Suponha que a soma de variáveis aleatórias de Bernoulli dependentes, Y, segue uma distribuição condicional $Y|p \sim B(n,p)$. Assumindo que p segue uma distribuição conjugada Beta, $Beta(\alpha_1,\alpha_2)$, $\alpha_1 > 0$ e $\alpha_2 > 0$. Então, a distribuição marginal de Y segue uma distribuição beta-binomial com parâmetros α_1 e α_2, e pode ser expressa como

$$P(Y=y|n,\alpha_1,\alpha_2) = \binom{n}{y}\frac{B(y+\alpha_1, n-y+\alpha_2)}{B(\alpha_1,\alpha_2)}, \tag{5.1}$$

where $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx = \Gamma(a)\Gamma(b)/\Gamma(a+b)$.

Prentice (1986) propôs uma parametrização em (5.1) considerando a probabilidade de sucesso e a correlação positiva no i-ésimo cluster como sendo $p_i = \alpha_1/(\alpha_1+\alpha_2)$ e $\rho_i = 1/(\alpha_1+\alpha_2+1)$, respectivamente. Usando esta parametrização a distribuição beta-binomial pode ser escrita como

$$P(Y=y|n,p,\zeta) = \binom{n}{y}\prod_{j=0}^{y-1}(p+\zeta j)\prod_{j=0}^{n-y-1}((1-p)+\zeta j)\left[\prod_{j=0}^{n-1}(1+\zeta j)\right]^{-1}, \tag{5.2}$$

no qual $\prod_{j=0}^{x} c_j = 0$, para qualquer $x < 0$, $y = 0,1,\ldots,n$, $n \in \mathbb{N} - \{0\}$, $0 < p < 1$ e $0 < \rho < 1$. A média e variância deste modelo são $E(Y) = np$ e $\mathrm{Var}(Y) = np(1-p)(1+(n-1)\rho)$, respectivamente.

Seja $y_1, y_2, \ldots, y_m$ um conjunto de valores observados de $Y_1, Y_2, \ldots, Y_m$, respectivamente, com $Y_i \sim BB(p_i(1-\rho_i)\rho_i^{-1}, (p_i-1)(\rho_i-1)\rho_i^{-1})$. A função de verossimilhança é dada por

$$\mathcal{L}(\boldsymbol{p},\boldsymbol{\zeta};m,\boldsymbol{n},\boldsymbol{y}) = \prod_{i=1}^{m}\left\{\binom{n_i}{y_i}\prod_{j=0}^{y_i-1}(p_i+\zeta_i j)\prod_{j=0}^{n_i-y_i-1}((1-p_i)+\zeta_i j)\left[\prod_{j=0}^{n_i-1}(1+\zeta_i j)\right]^{-1}\right\} \tag{5.3}$$

Prentice (1986), Kahn & Raftery (1996) e Lindsey & Altham (1998) sugerem modelar a probabilidade de sucesso utilizando covariáveis dos clusters via função de ligação logito. Neste trabalho, outras três funções de ligação também são consideradas para modelar p_i. As funções de ligação logito, complementar log-log, log-log e probito são denotadas por $g^{-1}(\eta_i)$ e estão especificadas na Tabela 2.1. Considerando disponíveis, para o i-ésimo cluster, os valores do conjunto de k covariáveis, $x_{i1}, x_{i2}, \ldots, x_{ik}$; $\eta_i = \sum_{r=0}^{k} \beta_r x_{ir}$; os coeficientes $\beta_0, \beta_1, \ldots, \beta_k$ são parâmetros de regressão desconhecidos a serem estimados e $x_{i0} = 1$, para todo i.

Uma alternativa às abordagens apresentadas em Prentice (1986) e em Lindsey & Altham (1998), as quais fazem uso da função de ligação logito, é modelar o parâmetro de correlação do i-ésimo cluster ρ_i por meio de uma estrutura de correlação na forma (2.3), discutida na construção do modelo de regressão binomial correlacionada (ver Capítulo 2). Usando a parametrização $\phi = \log(\gamma)$, para estrutura de correlação exponencial ou Gaussiana, ou $\phi = \log(\gamma/(1-\gamma))$, para estrutura de correlação AR contínua, o parâmetro da estrutura de correlação γ pode ser estimado sem restrição.

Prentice (1986) apresentou uma extensão do modelo de regressão beta-binomial que permite modelar, além da correlação positiva, também a correlação negativa entre os eventos de Bernoulli adaptando a função de ligação logito. Para utilizar esta extensão na modelagem aqui apresentada é necessário considerar em (2.3) $\rho_i = 2h(v(\boldsymbol{r}_i), \gamma) - 1$, assim $-1 < \rho_i < 1$, porém esta abordagem não será discutida neste trabalho.

Seja $\mathcal{D}$ os dados observados, usando $g^{-1}(\eta_i)$ e (2.3), a função de verossimilhança (5.3) pode ser expressa como uma função dos parâmetros $\boldsymbol{\theta}^* = (\phi, \beta_0, \beta_1, \ldots, \beta_k)^\top$. Então,
$\mathcal{L}(\boldsymbol{\theta}^*; \mathcal{D}) =$

$$\prod_{i=1}^{m} \left\{ \binom{n_i}{y_i} \prod_{j=0}^{y_i-1} \left(g^{-1}(\eta_i) + \frac{h^*(v(\boldsymbol{r}_i), \phi) j}{1 + h^*(v(\boldsymbol{r}_i), \phi)} \right) \right.$$
$$\left. \times \prod_{j=0}^{n_i-y_i-1} \left((1 - g^{-1}(\eta_i)) + \frac{h^*(v(\boldsymbol{r}_i), \phi) j}{1 + h^*(v(\boldsymbol{r}_i), \phi)} \right) \left[\prod_{j=0}^{n_i-1} \left(1 + \frac{h^*(v(\boldsymbol{r}_i), \phi) j}{1 + h^*(v(\boldsymbol{r}_i), \phi)} \right) \right]^{-1} \right\} \quad (5.4)$$

no qual $\prod_{j=0}^{x} c_j = 0$, para qualquer $x < 0$, $y_i = 0, 1, \ldots, n_i$; $n_i \in \mathbb{N} - \{0\}$; $h^*(v(\boldsymbol{r}_i), \phi)$ é uma função similar a $h(v(\boldsymbol{r}_i), \gamma)$, apresentada na Tabela 2.2, considerando a parametrização necessária.

5.2 Abordagem Bayesiana

Considere uma distribuição a *priori*, $\pi(\boldsymbol{\theta}^*)$, para $\boldsymbol{\theta}^* = (\phi, \beta_0, \beta_1, \ldots, \beta_k)^\top$, $\boldsymbol{\theta}^* \sim N_{k+2}(\mathbf{0}, \Sigma)$, $\Sigma = \text{diag}\{\lambda_1, \ldots, \lambda_{k+2}\}$ com hiperparâmetros conhecidos λ_r, $r = 1, \ldots, k+2$, e $n_i, i = 1, \ldots, m$, conhecido, a distribuição a *posteriori* conjunta para $\boldsymbol{\theta}^*$ é dada por

$$\pi(\boldsymbol{\theta}^* | \mathcal{D}) \quad \propto \quad \mathcal{L}(\boldsymbol{\theta}^*; \mathcal{D}) \; \pi(\boldsymbol{\theta}^*). \quad (5.5)$$

A distribuição a *posteriori*, $\pi(\boldsymbol{\theta}^* | \mathcal{D})$, é tratável via perspectiva Bayesiana. A inferência Bayesiana pode ser conduzida considerando métodos MCMC tal como o algoritmo de

Gibbs com passos de Metropolis. As densidades condicionais completas para os parâmetros ϕ e β_r são dadas por

$$\pi(\phi|\boldsymbol{\beta},\lambda_1,\mathcal{D}) \;\propto\; \exp\left\{-\frac{\phi^2}{2\lambda_1}\right\}\prod_{i=1}^{m}\left\{\prod_{j=0}^{y_i-1}\left(g^{-1}(\eta_i)+\frac{h^*(v(\boldsymbol{r}_i),\phi)j}{1+h^*(v(\boldsymbol{r}_i),\phi)}\right)\right.$$

$$\left.\times\prod_{j=0}^{n_i-y_i-1}\left((1-g^{-1}(\eta_i))+\frac{h^*(v(\boldsymbol{r}_i),\phi)j}{1+h^*(v(\boldsymbol{r}_i),\phi)}\right)\left[\prod_{j=0}^{n_i-1}\left(1+\frac{h^*(v(\boldsymbol{r}_i),\phi)j}{1+h^*(v(\boldsymbol{r}_i),\phi)}\right)\right]^{-1}\right\}$$

e

$$\pi(\beta_r|\boldsymbol{\beta}_{(-r)},\phi,\lambda_{r+1},\mathcal{D}) \;\propto\; \exp\left\{-\frac{\beta_r^2}{2\lambda_{r+1}}\right\}$$

$$\times\prod_{i=1}^{m}\left\{\prod_{j=0}^{y_i-1}\left(g^{-1}(\eta_i)+\frac{h^*(v(\boldsymbol{r}_i),\phi)j}{1+h^*(v(\boldsymbol{r}_i),\phi)}\right)\prod_{j=0}^{n_i-y_i-1}\left((1-g^{-1}(\eta_i))+\frac{h^*(v(\boldsymbol{r}_i),\phi)j}{1+h^*(v(\boldsymbol{r}_i),\phi)}\right)\right\}.$$

Densidade preditiva condicional ordinária

A densidade preditiva condicional ordinária para a i-ésima observação, condicionada a $\mathcal{D}_{(-i)}$, considerando o modelo de regressão beta-binomial, é definida como

$$CPO_i \;=\; \int_\Theta \pi(y_i|\mathcal{D},\boldsymbol{\theta}^*)\pi\left(\boldsymbol{\theta}^*|\mathcal{D}_{(-i)}\right)d\boldsymbol{\theta} = \left\{\int_\Theta \frac{1}{\pi(y_i|\mathcal{D},\boldsymbol{\theta}^*)}\pi\left(\boldsymbol{\theta}^*|\mathcal{D}\right)d\boldsymbol{\theta}\right\}^{-1}, \quad (5.6)$$

$i=1,\ldots,m$, com

$$\pi(y_i|\mathcal{D},\boldsymbol{\theta}^*) = \binom{n_i}{\tilde{y_i}}\prod_{j=0}^{\tilde{y_i}-1}(p_i+\zeta_i j)\prod_{j=0}^{n_i-\tilde{y_i}-1}((1-p_i)+\zeta_i j)\left[\prod_{j=0}^{n_i-1}(1+\zeta_i j)\right]^{-1}, \quad (5.7)$$

em que $\zeta_i = h^*(v(\boldsymbol{r}_i),\phi)/(1+h^*(v(\boldsymbol{r}_i),\phi))$ e $p_i = g^{-1}(\sum_{r=0}^{k}\beta_r x_r)$.

Considerando os estimadores de Bayes para os parâmetros dos coeficientes de regressão, $\beta_r, r=1,\ldots,k$, e para o parâmetro da estrutura de correlação, ϕ, podemos predizer o valor de y_i usando (5.6). Para isso, uma estimação de Monte Carlo da densidade $\pi(\tilde{y_i}|\mathcal{D},\boldsymbol{\theta}^*)$ pode ser obtida por meio de amostras MCMC da distribuição a *posteriori* $\pi\left(\boldsymbol{\theta}^*|\mathcal{D}\right)$. A aplicação do algoritmo apresentado na Seção 4.2 considerando (5.7) fornece numericamente a moda da densidade preditiva condicional ordinária como uma predição para $\tilde{y_i}$.

5.3 Diagnósticos

Entre as suposições feitas na construção do modelo de regressão beta-binomial podemos destacar: (i) independência entre as variáveis resposta, (ii) as variáveis resposta seguem uma distribuição beta-binomial $BB(p_i(1-\rho_i)\rho_i^{-1},(p_i-1)(\rho_i-1)\rho_i^{-1})$, (iii) função de ligação e (iv) estrutura de correlação. Nesta seção, três diferentes tipos de resíduos

Bayesianos e uma medida de influência local considerando uma perspectiva Bayesiana são propostas para checar as suposições do modelo e para identificar a presença de *outliers* e/ou observações influentes. O pressuposto de independência é verificado por meio dos gráficos dos resíduos versus a ordem das observações, quando a mesma está disponível. Para checar a suposição que as variáveis resposta seguem distribuição beta-binomial $BB(p_i(1-\rho_i)\rho_i^{-1}, (p_i-1)(\rho_i-1)\rho_i^{-1})$ é necessário verificar a adequação do modelo aos dados por meio da análise de resíduos. Além disso, nesta seção apresentamos um conveniente critério para seleção de modelo.

Resíduos baseado na densidade preditiva condicional ordinária

O resíduo baseado na densidade preditiva condicional ordinária (Cho *et al.*, 2009) para a i-ésima observação, r_i^{pp}, é calculado como $r_i^{pp} = y_i - \tilde{y}_i$, em que y_i é a i-ésima resposta observada e $\tilde{y}_i$ é a moda da distribuição preditiva condicional ordinária condicionado aos valores das covariáveis, $(x_{i0}, \ldots, x_{ik}, r_{i11}, \ldots, r_{i1n_i}, r_{i21}, \ldots, r_{i2n_i}, r_{iq1}, \ldots, r_{iqn_i})^\top$, em que x_{ir} é o valor da r-ésima covariável dentro do i-ésimo cluster, $i = 1, \ldots, m$ e $r = 0, \ldots, k$, e r_{ilj} é o valor da l-ésima covariável para o j-ésimo indivíduo dentro do i-ésimo cluster, $i = 1, \ldots, m$, $l = 1, \ldots, q$ e $j = 1, \ldots, n_i$. O resíduo padronizado, r_i^{spp}, baseado na densidade preditiva condicional ordinária para o modelo de regressão beta-binomial é definido como

$$r_i^{spp} = \frac{r_i^{pp}}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)\{1+(n_i-1)\hat{\rho}_i\}}}, \qquad i = 1, \ldots, m,$$

com $\hat{p}_i$ e $\hat{\rho}_i$ estimadores de Bayes de p_i e ρ_i, respectivamente.

É esperado que o conjunto de resíduos esteja próximo de zero. Se isso não ocorrer é um indicativo de que o modelo está mal ajustado aos dados.

Resíduos baseado na distribuição a *posteriori* dos parâmetros do modelo

O resíduo padronizado baseado na distribuição a *posteriori* dos parâmetros para o modelo de regressão beta-binomial é definido pela transformação

$$R_{iq}^{spd} = \frac{y_i - E(Y_i|\mathcal{D}, \boldsymbol{\theta}_q)}{\sqrt{\mathrm{Var}(Y_i|\mathcal{D}, \boldsymbol{\theta}_q)}} = \frac{y_i - n_i p_{iq}}{\sqrt{n_i p_{iq}(1-p_{iq})\{1+(n_i-1)\rho_{iq}\}}}, \tag{5.8}$$

com $i = 1, \ldots, m$ e $q = 1, \ldots, Q$; em que p_{iq} e ρ_{iq} são não observados e n_i e y_i são informações da amostra para a i-ésima observação. O novo índice q presente em R_{iq}^{spd} se faz necessário para justificar, para cada i, uma amostra de tamanho Q destes resíduos. Como a relação (5.8) é uma função do vetor de parâmetros, $\boldsymbol{\theta}_q$, esta relação carrega toda a incerteza refletida nos parâmetros a *posteriori*, refletindo na distribuição a *posteriori* dos resíduos, R_i^{spd}. A distribuição a *posteriori* dos resíduos pode ser resumida usando amostras de um amostrador MCMC. Isto é, ao gerar uma amostra $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q$ de tamanho Q de $\pi(\boldsymbol{\theta}^*|\mathcal{D})$, obtemos para o par (y_i, n_i), $i = 1, \ldots, m$, uma amostra de tamanho Q dos resíduos em (5.8). É esperado que a média dos resíduos esteja centrada em zero. Se isso não ocorrer, ou se a amostra dos resíduos apresentar uma alta variabilidade, a observação pode ser considerada um *outlier*.

Resíduos *deviance* Bayesiano

O resíduo *deviance* Bayesiano para a i-ésima observação é uma composição da dificuldade de estimar $\boldsymbol{\theta}$ na presença de y_i, p_{D_i}, e da contribuição de cada observação para o total global da *deviance*, $D_i(\hat{\boldsymbol{\theta}})$. O resíduo *deviance* Bayesiano é definido para o modelo de regressão beta-binomial por

$$r_i^{db} = \text{sinal}(y_i - n_i\hat{p}_i)\sqrt{p_{D_i} + D_i(\hat{\boldsymbol{\theta}})}, \qquad i = 1, \ldots, m, \tag{5.9}$$

sendo o valor de p_{D_i} para a i-ésima observação aproximado via método de Monte Carlo por

$$p_{D_i} = -\frac{2}{Q}\sum_{q=1}^{Q}\left\{\log\frac{\pi(\boldsymbol{\theta}_q|\mathcal{D}_{(i)})}{\pi(\boldsymbol{\theta}_q)}\right\} + 2\log\frac{\pi(\hat{\boldsymbol{\theta}}|\mathcal{D}_{(i)})}{\pi(\hat{\boldsymbol{\theta}})}, \tag{5.10}$$

em que $\boldsymbol{\theta}_q = (\gamma_q, \beta_{0q}, \ldots, \beta_{kq})^\top$ representa uma amostra de tamanho Q gerada da distribuição a *posteriori* $\pi(\boldsymbol{\theta}^*|\mathcal{D})$, $\hat{\boldsymbol{\theta}}$ o estimador de Bayes de $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}_q|\mathcal{D}_{(i)})$ e $\pi(\hat{\boldsymbol{\theta}}|\mathcal{D}_{(i)})$ são a distribuição a *posteriori* aplicada em $\boldsymbol{\theta}_q$ e em $\hat{\boldsymbol{\theta}}$, respectivamente, e $\pi(\boldsymbol{\theta}_q)$ e $\pi(\hat{\boldsymbol{\theta}})$ são a distribuição a *priori* aplicada em $\boldsymbol{\theta}_q$ e $\hat{\boldsymbol{\theta}}$, respectivamente. O valor de D_i para a i-ésima observação é dado por

$$D_i(\hat{\boldsymbol{\theta}}) = -2\log\left\{\binom{n_i}{y_i}\prod_{j=0}^{y_i-1}\left(\hat{p}_i + \hat{\zeta}_i j\right)\prod_{j=0}^{n_i-y_i-1}\left((1-\hat{p}_i) + \hat{\zeta}_i j\right)\left[\prod_{j=0}^{n_i-1}\left(1+\hat{\zeta}_i j\right)\right]^{-1}\right\}, \tag{5.11}$$

com $\hat{\zeta}_i = h(v(\boldsymbol{r}_i), \hat{\gamma})/(1 + h(v(\boldsymbol{r}_i), \hat{\gamma}))$ e $\hat{p}_i = g^{-1}(\sum_{r=0}^{k}\hat{\beta}_r x_{ir})$.

Assim, o resíduo *deviance* Bayesiano padronizado para a i-ésima observação pode ser obtido através da expressão

$$r_i^{sdb} = \frac{r_i^{db}}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)\{1 + (n_i-1)\hat{\rho}_i\}}}, \qquad i = 1, \ldots, m, \tag{5.12}$$

sendo $\hat{p}_i = g^{-1}(\sum_{r=0}^{k}\hat{\beta}_r x_{ir})$ e $\hat{\rho}_i = h(v(\boldsymbol{r}_i), \hat{\gamma})$.

Ao plotar os valores dos resíduos *deviance*, r_i^{db}, contra os valores de p_{D_i} para um determinado modelo ajustado, considerando marcações de curvas da forma $x^2 + y = c$, temos os pontos situados ao longo de tal parábola como indicativo da quantia de contribuição $\text{DIC}_i = c$ da observação no valor global do DIC (Spiegelhalter *et al.*, 2002).

Espera-se que o conjunto de resíduos *deviance* padronizados esteja próximo de zero. Se isso não ocorrer é uma indicação de que o modelo está mal ajustado aos dados.

Diagnóstico de influência

O diagnóstico Bayesiano de influência do modelo de regressão beta-binomial é similar ao apresentado para o MRBC na Seção 4.3.2.

Critério de seleção de modelo

O critério de seleção Bayesiano (DIC) para o modelo de regressão beta-binomial é similar ao apresentado para o MRBC na Seção 4.4.

5.4 Estudos de simulação

Nesta seção é considerado um estudo de simulação para ilustrar o processo inferencial e o desempenho das medidas de diagnóstico propostas neste trabalho. O estudo de simulação envolve duas etapas. Na primeira, os dados simulados na Seção 4.5 são analisados via modelo de regressão beta-binomial. Na segunda etapa, analisamos um conjunto de dados simulado do modelo de regressão beta-binomial, consideramos na análise também o modelo de regressão binomial correlacionada.

Dados simulados via MRBC

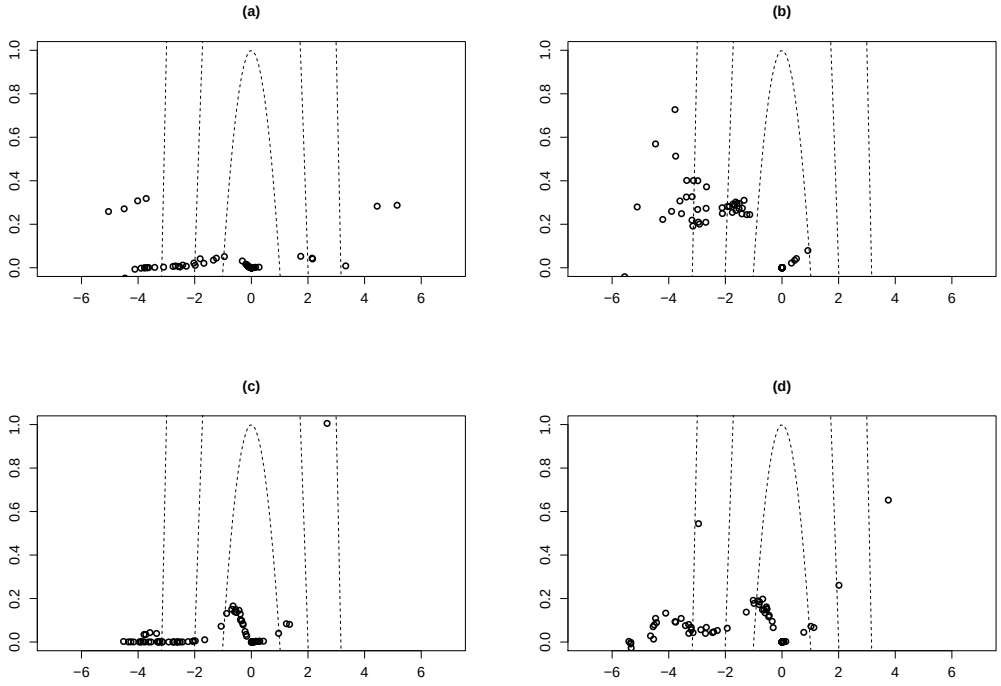

Figura 5.1: MRBB: Resíduos *deviance* versus alavancagem Bayesiana, considerando $c = 1, 4$ e 10, (a) ligação logito, (b) ligação complementar log-log, (c) ligação log-log, (d) ligação probito.

Os dados simulados na Seção 4.5, provenientes do modelo de regressão binomial correlacionada, são ajustados via modelo de regressão beta-binomial considerando as quatro funções de ligação, logito, complementar log-log, log-log e probito com a estrutura de correlação exponencial. As mesmas distribuições a *priori* são utilizadas. Observe que os valores obtidos pelo critério DIC de seleção de modelos, mostrados na Tabela 5.1, confirmam a função de ligação logito e o modelo MRBC como o melhor ajuste.

As contribuições de cada observação no valor global do DIC para o modelo de regressão binomial correlacionada são mostrados na Figura 4.1, e para o modelo de regressão beta-binomial na Figura 5.1. As mesmas marcações exibidas na Figura 4.1 são fixadas, a maior

Tabela 5.1: Valores obtidos pelo critério DIC ajustando o modelo de regressão binomial correlacionada (MRBC) e o modelo de regressão beta-binomial (MRBB) com diferentes funções de ligação.

Critério DIC	Logito	Complementar log-log	Log-log	Probito
MRBC	112,597	119,840	129,030	115,141
MRBB	589,949	790,120	461,225	711,129

amplitude dos resíduos *deviance* Bayesianos para os ajustes do MRBB justificam maiores valores do critério DIC de seleção de modelos.

Dados simulados via MRBB

A geração da amostra envolve $m = 100$ clusters com as variáveis resposta, $Y_i \sim BB(p_i(1-\rho_i)\rho_i^{-1}, (p_i-1)(\rho_i-1)\rho_i^{-1})$, $i = 1, \ldots, 100$, com os n_i gerados de uma distribuição B(45; 0,5). O parâmetro da estrutura de correlação utilizado na simulação é $\gamma = 0,2$ e os $v(\boldsymbol{r}_i)$ assumem valores de uma distribuição U(0,1). Duas covariáveis são consideradas, x_{i1} e x_{i2}. Os valores das covariáveis x_{i1} são provenientes de uma distribuição U(0,2) e os valores da covariável x_{i2} são provenientes de uma distribuição N(0,4). Os valores dos coeficientes de regressão são $\beta_0 = 2$, $\beta_1 = -2$ e $\beta_3 = -2$. São considerados neste estudo a função de ligação log-log e a estrutura de correlação exponencial. A distribuição a *priori* para o parâmetro β_r é N(0,10^4), $r = 0, 1, 2$ e para o parâmetro ϕ é N(0,10^4).

Estimação

O modelo de regressão binomial correlacionada e o modelo de regressão beta-binomial são ajustados com as quatro funções de ligação, logito, complementar log-log, log-log e probito. Como esperávamos, os valores obtidos pelo critério DIC de seleção de modelos confirma o MRBB com a função de ligação log-log como o melhor ajuste, conforme pode ser visto na Tabela 5.2.

Tabela 5.2: Valores obtidos pelo critério DIC ajustando o modelo de regressão binomial correlacionada (MRBC) e o modelo de regressão beta-binomial (MRBB) com diferentes funções de ligação.

Critério DIC	Logito	Complementar log-log	Log-log	Probito
MRBC	317,800	332,284	322,149	319,505
MRBB	189,927	191,853	186,190	189,507

Os resumos da distribuição a *posteriori* para o modelo com ligação log-log são mostrados na Tabela 5.3. Note que ambos, intervalo de credibilidade inter-quantil e HPD, contêm o verdadeiro valor dos parâmetros. O diagnóstico de convergência dos parâmetros foi verificado via CODA, por meio do teste de Geweke o qual apresentou valores entre (-0,5; 0,5).

Tabela 5.3: Resumos das densidades marginais a *posteriori* de γ, β_0, β_1 e β_2 para a função de ligação log-log.

	Valor verdadeiro	Média a *posteriori*	Mediana a *posteriori*	Intervalo inter-quantil credibilidade 95%	Intervalo HPD credibilidade 95%
γ	0,20	0,325	0,317	(0,164 ; 0,545)	(0,165 ; 0,548)
β_0	2,00	2,317	2,321	(2,080 ; 2,598)	(2,070 ; 2,553)
β_1	-2,00	-2,223	-2,221	(-2,479 ; -1,996)	(-2,428 ; -1,945)
β_2	-2,00	-2,145	-2,146	(-2,371 ; -1,928)	(-2,323 ; -1,889)

Análise de resíduos

Na Figura 5.2 e na Figura 5.3 são apresentadas marcações que permitem avaliar a contribuição de cada observação no valor global do DIC para o modelo de regressão beta-binomial e para o modelo de regressão binomial correlacionada, respectivamente, para diferentes funções de ligação. Observe que a contribuição das observações nos ajustes do modelo de regressão beta-binomial apresentam-se mais concentradas em torno de zero, justificando os menores valores do critério DIC.

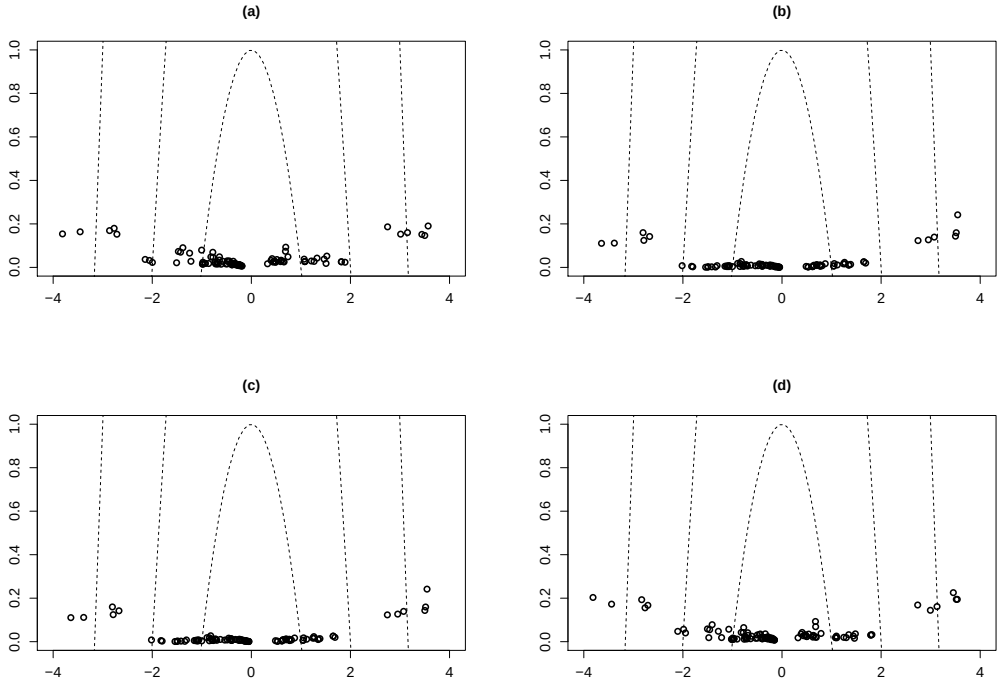

Figura 5.2: MRBB: resíduos *deviance* versus alavancagem Bayesiana, considerando $c = 1, 4$ e 10, (a) ligação logito, (b) ligação complementar log-log, (c) ligação log-log, (d) ligação probito.

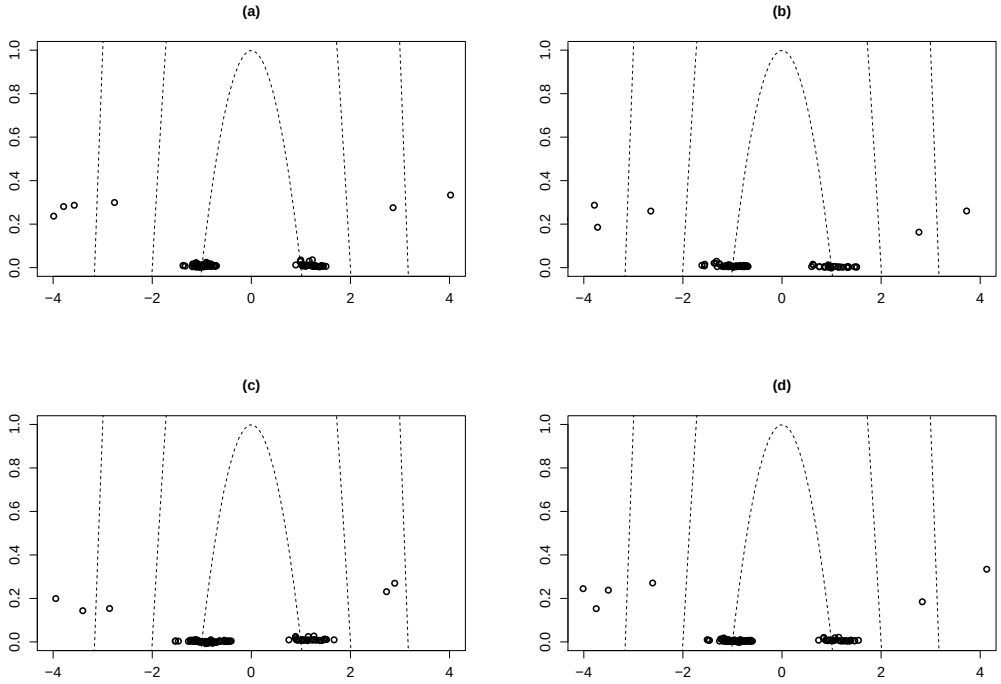

Figura 5.3: MRBC: resíduos *deviance* versus alavancagem Bayesiana, considerando $c = 1, 4$ e 10, (a) ligação logito, (b) ligação complementar log-log, (c) ligação log-log, (d) ligação probito.

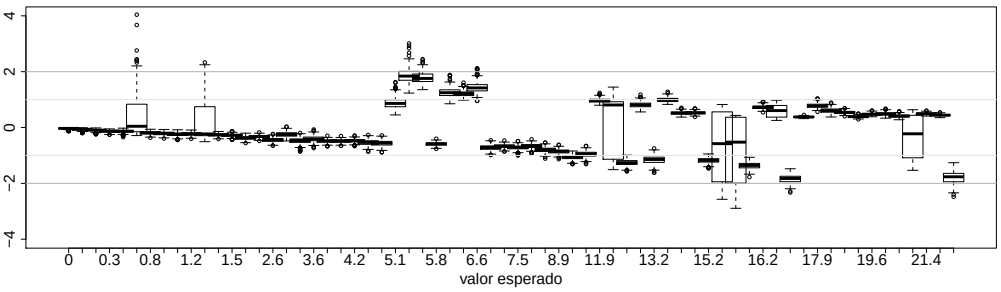

Figura 5.4: Boxplots das amostras da distribuição a *posteriori* dos resíduos para cada observação versus o valor esperado de $E(Y_i|\mathcal{D}, \boldsymbol{\theta})$.

Gráficos dos resíduos padronizados contra valores preditos, apresentados na Figura 5.5a e 5.5c, e um gráfico do resíduo padronizado contra o valor esperado, apresentado na Figura 5.5b, são utilizados para verificarmos o ajuste global do modelo. A média das amostras geradas das distribuições a *posteriori* dos resíduos padronizados, mostradas na Figura 5.4, e seus respectivos intervalos HPD versus valores esperados de $E(Y_i|\mathcal{D}, \boldsymbol{\theta})$ são mostrados na Figura 5.5b. Os resíduos baseados na distribuição a *posteriori* dos parâmetros e os resíduos baseados no *deviance* Bayesiano mostram que o ajuste dos dados está adequado e não evidenciam *outliers*.

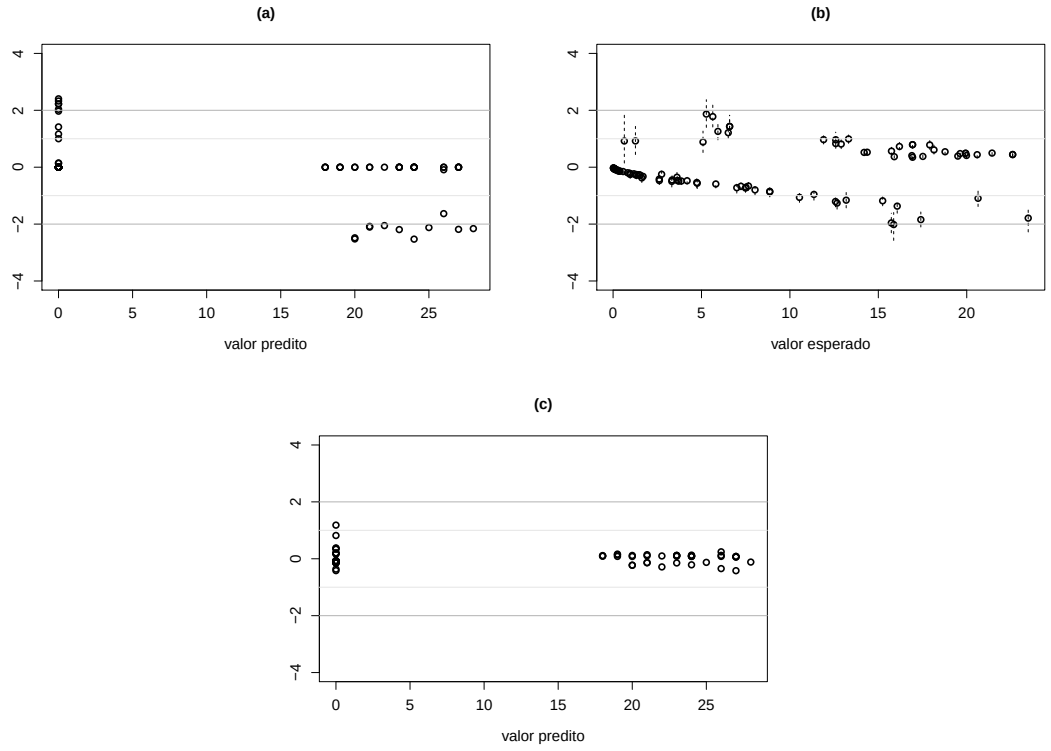

Figura 5.5: (a) resíduos padronizados via densidade preditiva condicional ordinária versus valores preditos, (b) médias e intervalos HPD (95%) das amostras dos resíduos padronizados via distribuição a *posteriori* versus valores esperados, (c) resíduos *deviance* padronizados versus valores preditos.

A partir dos resultados mostrados nos gráficos observamos que os resíduos baseados na densidade preditiva condicional ordinária indicam um grupo de pontos, para valores preditos entre 20 e 30, mais afastados de zero. Uma possível justificativa para este afastamento é a forma bimodal que a densidade preditiva condicional ordinária da variável aleatória $\tilde{Y}_i$ pode assumir. Esta forma bimodal da distribuição prejudica a predição desta variável aleatória em alguns casos. Isto nos permite dizer que estes resíduos não podem ser utilizados, de forma única, para avaliar a qualidade do ajuste de um modelo de regressão beta-binomial. É importante ressaltar que a densidade preditiva condicional ordinária para os modelos de regressão propostos nesta tese, em todos os casos analisados, apresentaram forma unimodal.

Diagnósticos de influência

Nessa seção examinamos o desempenho das medidas de diagnósticos de influência, a divergência de K-L e a calibração. Para isto, a observação 91 foi perturbada, criando assim um caso influente no conjunto de dados. A perturbação foi feita na covariável x_{i2}, na forma $x_{i2} = x_{i2} + 4sd(x_2)$, $i = 91$, em que $sd(x_2)$ corresponde ao desvio padrão da covariável x_2. Duas combinações de presença e ausência desta observação perturbada foram usadas para formar novos conjuntos de dados. Estes novos conjuntos de dados foram utilizados

na obtenção de estimativas dos parâmetros do modelo. A Tabela 5.4 apresenta a média da distribuição *posteriori* e a mudança relativa da estimativas dos parâmetros em relação aos dados originais.

Tabela 5.4: Média a *posteriori* e mudança relativa em relação aos dados originais para os dados simulados.

Casos	γ		β_0		β_1		β_2	
perturbados	Média	%	Média	%	Média	%	Média	%
{Nenhum}	0,33	-	2,32	-	-2,22	-	-2,15	-
Caso {91}	0,22	-33,33	0,37	-84,05	-0,69	-68,92	-0,83	-61,39

A Figura 5.6 mostra os valores da divergência de K-L com dos dados simulados e com a presença da perturbação na observação 91, para os quais foram obtidos valores da calibração de 0,821 e 0,998, respectivamente.

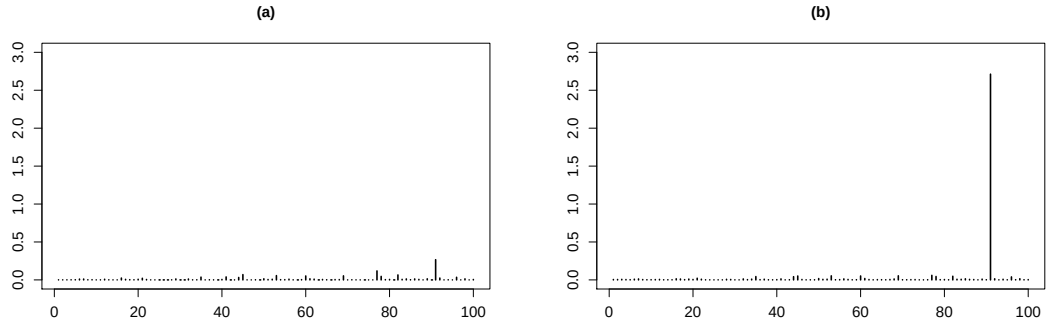

Figura 5.6: (a) DKL: dados sem perturbação, (b) DKL: dados com o caso 91 perturbado.

5.5 MRBC e MRBB: Análise de dados reais

Dois conjuntos de dados reais são analisados usando o modelo de regressão binomial correlacionada e o modelo de regressão beta-binomial: o primeiro, analisado sob a perspectiva clássica, trata de uma aplicação na área de renovação de contratos de planos de saúde coletivos empresariais, onde o interesse é determinar a probabilidade do grupo de funcionários da empresa utilizar serviços de alto custo oferecidos pela operadora de saúde e a dependência entre os funcionários da empresa com relação ao evento de interesse; o segundo, analisado sob enfoque Bayesiano, trata de aplicação em finanças, cujo interesse é determinar a probabilidade de *default* de uma corporação (empresa que possui outras empresas) e a relação de dependência entre as empresas de um mesmo grupo.

5.5.1 Metodologia clássica: aplicação em planos de saúde

Considere um conjunto de dados reais de uma operadora de plano de saúde no Brasil contendo uma carteira de empresas (clusters) (o conjunto de dados está disponível em

http://www.ufscar.br/~des/docente/carlos/Dados/Dados2.txt). Para cada funcionário (indivíduo) dentro da empresa, é observada a ocorrência ou não de serviços de alto custos, tais como cirurgia oncológica, prótese, quimioterapia e hemodiálise. Os dados disponíveis da i-ésima empresa, $i = 1, 2, \ldots, 160$, com n_i funcionários, consiste de $W_{i1}, W_{i2}, \ldots, W_{in_i}$, cada um assumindo valor zero ou um, dependendo do status do funcionário (0 = não ocorrência; 1 = ocorrência). A variável resposta para a i-ésima empresa, $Y_i = \sum_{j=1}^{n_i} W_{ij}$, assume valores em $\{0, 1, \ldots, n_i\}$ de acordo com o número de funcionários que utilizaram serviços de saúde de alto custo.

A estrutura de dependência entre as variáveis de Bernoulli dentro da empresa pode ser explicada pelo fato dos funcionários estarem expostos ao mesmo ambiente. Algumas covariáveis estão disponíveis no conjunto de dados, tais como número médio de consultas por funcionário, custo médio dos exames, ocorrência de procedimentos cirúrgicos, número de terapias, número de procedimentos de urgência e emergência, número de dias entre o início de vigência dos serviços do plano de saúde e a primeira utilização do serviço de alto custo para cada funcionário e informações específicas das empresas (tamanho, número de funcionários, atividade de atuação). O principal interesse da análise é ajustar um modelo de regressão que possa ser utilizado para determinar a probabilidade de ocorrência de serviços de alto custo na empresa. Uma decisão sobre a renovação ou não do plano de saúde pode ser tomada com base na magnitude dessa probabilidade.

Duas covariáveis dos clusters são consideradas na análise: *número médio de consultas por funcionário*, x_{i1}, e *custo médio padronizado dos exames*, x_{i2}. A covariável *número de dias entre o início de vigência dos serviços do plano de saúde e a primeira utilização de algum serviço de saúde entre o s-ésimo e o t-ésimo funcionário*, r_{i1}, é utilizada para explicar a dependência entre as variáveis de Bernoulli dentro da empresa. De fato, consideramos a variável $\min_{s,t} |r_{i1s} - r_{i1t}|$, o mínimo de dias entre a utilização entre o s-ésimo e t-ésimo funcionário, que assume valores entre zero, se ambos os funcionários usam serviços no mesmo dia, e 365, se não houve uso do plano pelos funcionários durante a vigência do contrato. Esta variável é padronizada no intervalo [0,1] pela transformação $\min_{s,t} |r_{i1s} - r_{i1t}|/365$. É intuitivo supor que quanto maior a diferença entre os tempos de utilização do plano menor a relação entre o uso do serviço. Por esta razão, consideramos na análise a estrutura de correlação AR contínua, dada por

$$h(v(\boldsymbol{r}_i), \gamma) = \gamma^{\frac{\min_{s,t} |r_{i1s} - r_{i1t}|}{365}},$$

com $i = 1, \ldots, m$ e $s, t = 1, \ldots, n_i$.

O modelo de regressão binomial correlacionada e o modelo de regressão binomial são ajustados aos dados considerando as quatro funções de ligação discutidas neste trabalho. Os resultados obtidos pelos critérios de seleção de modelo, AIC e BIC, são apresentados na Tabela 5.5 e identificam o modelo de regressão binomial correlacionada com a ligação complementar log-log e logito, com valores dos critérios muito próximos, como os melhores ajustes.

Tabela 5.5: Valores obtidos pelos critérios de seleção ajustando o modelo de regressão binomial correlacionada (MRBC) e o modelo de regressão binomial (MRB) com diferentes funções de ligação.

Modelo	Critérios	Logito	Complementar log-log	Log-log	Probito
	AIC	397,861	397,811	398,390	398,143
MRBC	*BIC*	430,462	430,412	430,992	430,744
	AIC	538,471	538,495	551,946	538,242
MRB	*BIC*	547,696	547,720	564,172	547,468

As estimativas de máxima verossimilhança e intervalos de confiança assintóticos, *bootstrap* e perfilado para os parâmetros do modelo de regressão binomial correlacionada, considerando a função de ligação complementar log-log, são mostradas na Tabela 5.6. É importante observar que os intervalos de confiança para o parâmetro da estrutura de correlação, γ, não contêm o valor zero, confirmando a necessidade do ajuste de um modelo que incorpore correlação positiva entre os eventos de Bernoulli.

Tabela 5.6: Estimativas de máxima verossimilhança para γ, β_0, β_1 e β_2, para os dados de planos de saúde.

Parâmetros	γ	β_0	β_1	β_2
EMV	0,223	-3,833	0,206	0,322
IC_u 95%	(0,121 ; 0,325)	(-4,411 ; -3,254)	(0,032 ; 0,381)	(0,161 ; 0,484)
IC_a 95%	(0,122 ; 0,325)	(-4,410 ; -3,255)	(0,033 ; 0,380)	(0,161 ; 0,484)
IC_b 95%	(0,132 ; 0,323)	(-4,508 ; -3,133)	(-0,008 ; 0,401)	(0,153 ; 0,496)
IC_p 95%	(0,134 ; 0,332)	(-3,978 ; -3,690)	(0,163 ; 0,250)	(0,164 ; 0,477)

IC_u: intervalo de confiança assintótico usando (3.2), IC_a: intervalo de confiança assintótico usando (3.3), IC_u: intervalo de confiança *bootstrap* e IC_p: intervalo de confiança perfilado.

A suposição de independência e a presença de *outliers* podem ser verificadas examinando o gráfico dos resíduos contra a ordem, se a ordem estiver disponível. Os resíduos padronizados via valores preditos e o resíduo *deviance* padronizado são apresentados nas Figuras 5.7a e 5.7b. A especificação do modelo e a presença de *outliers* são observados ao examinar os resíduos contra os valores preditos. Ambos os gráficos indicam uma boa especificação do modelo.

Para identificar observações influentes no conjunto as distâncias de Cook generalizada e distâncias da verossimilhança foram obtidas e são apresentadas na Figura 5.8. Os maiores valores obtidos para a distância de Cook generalizada e para a distância da verossimilhança ocorrem para o caso 85, com valores 1,147 e 0,557 respectivamente.

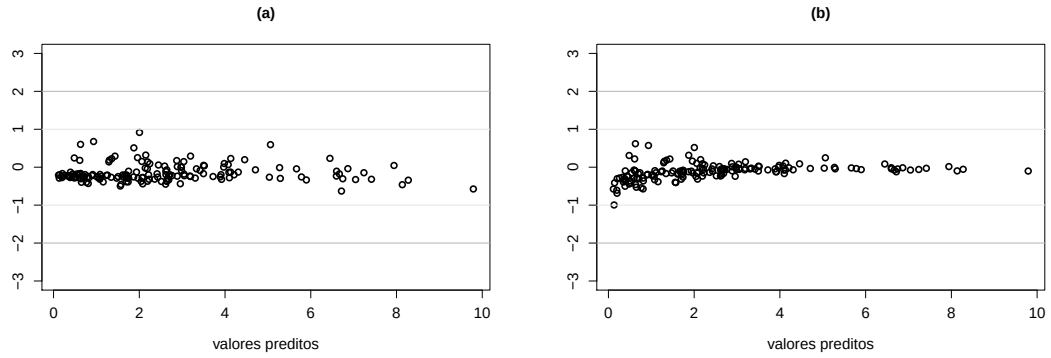

Figura 5.7: (a) resíduos padronizados via valores preditos versus valores preditos, (b) resíduos *deviance* padronizados versus valores preditos.

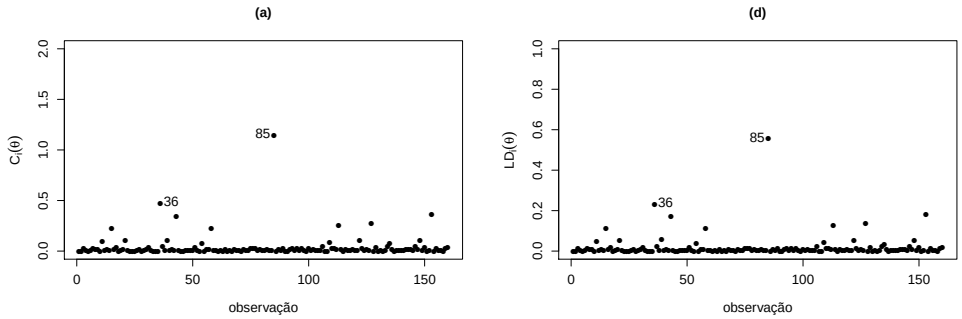

Figura 5.8: (a) distâncias de Cook generalizada e (b) distâncias da verossimilhança.

A avaliação do impacto causado pelo casos 36 e 85 na estimação dos parâmetros é mostrada na Tabela 5.7, que apresenta os estimadores de máxima verossimilhança e a mudança relativa deste estimador com respeito ao valor obtido omitindo cada caso.

Tabela 5.7: Estimadores de máxima verossimilhança e mudança relativa em relação aos dados originais para os dados de saúde ajustados via MRBC.

	Parâmetros							
Casos	γ		β_0		β_1		β_2	
omitidos	EMV	%	EMV	%	EMV	%	EMV	%
{Nenhum}	0,22	-	-3,83	-	0,21	-	0,32	-
Caso {36}	0,22	-	-4,03	5,30	0,26	19,23	0,32	-
Caso {85}	0,23	4,35	-3,57	-6,90	0,12	-75,00	0,24	-33,33

Para reforçar a necessidade de usar o modelo proposto neste conjunto de dados, ajustamos o modelo de regressão binomial usual obteve AIC igual a 538,495 e BIC igual a 547,720, o modelo de regressão beta-binomial, com AIC igual a 438,546 e BIC igual a 462,997, enquanto o modelo de regressão binomial correlacionada, com AIC igual a 397,811 e BIC igual a 430,412. Os modelos de regressão binomial correlacionada e beta-binomial foram ajustados usando a função de ligação complementar log-log e estrutura de

correlação AR contínua. A análise descrita nesta seção indica que o modelo alternativo oferece um bom ajuste para esse conjunto de dados.

A tomada de decisão em relação a renovação dos contratos, baseada na análise conduzida nesta seção, estabelece que a probabilidade de utilização de um serviço de alto custo por uma empresa é dada por

$$\hat{p_i} = 1 - \exp\left\{-\exp\{-3,833 + 0,206x_{i1} + 0,322x_{i2}\}\right\},$$

sendo x_{i1}: *número médio de consultas por funcionário*, e x_{2i}: *custo médio padronizado dos exames*. E a correlação entre quaisquer duas empresas dentro da corporação por

$$\hat{\rho_i} = 0,223^{\frac{\min_{s,t}|r_{i1s} - r_{i1t}|}{365}},$$

sendo r_{i1} *o mínimo de dias entre a utilização entre o s-ésimo e t-ésimo funcionário.*

5.5.2 Metodologia Bayesiana: aplicação em finanças

Nesta seção, um conjunto de dados reais de um escritório de crédito no Brasil contendo uma carteira de $m = 148$ corporações (clusters) que possuem controle acionário total ou parcial com, pelo menos, duas empresas, é analisada usando o modelo proposto (o conjunto de dados está disponível em em `http://www.ufscar.br/~des/docente/carlos/Dados/Dados1.txt`). Para cada empresa dentro da corporação, é observada a condição de *default*, isto é, se a empresa tem ou não cumprido as suas obrigações legais, no período de janeiro a dezembro de 2009, de acordo com o contrato da dívida. Os dados disponíveis na i-ésima corporação, $i = 1, 2, \ldots, 148$, com n_i empresas, consistem de $W_{i1}, W_{i2}, \ldots, W_{in_i}$, cada uma assumindo valor zero ou um, dependendo do status de *default* da empresa (0 = Não *Default*; 1 = *Default*). A variável resposta, $Y_i = \sum_{j=1}^{n_i} W_{ij}$, assume valores em $\{0, 1, \ldots, n_i\}$ de acordo com o número de empresas em *default* dentro da i-ésima corporação. Um total de dez covariáveis foram inicialmente disponibilizadas no conjunto de dados, mas apenas três covariáveis são consideradas na análise. Covariáveis com um excesso de dados faltantes, valores suspeitos e com o mesmo valor em todos os casos foram retiradas do conjunto de dados. As três covariáveis das corporações são *o número de sócios pessoa física na corporação*, x_{i1}; *o porte da corporação* x_{2i} e *o número de protestos da corporação com qualquer atraso no último ano* x_{i3}. A covariável específica, r_{i1}, considerada na análise, é *a receita das empresas.*

A presença de correlação entre duas empresas dentro da corporação é possível devido a diferentes fatos, tais como, empresas compartilhando o mesmo conselho administrativo, compartilhando as mesmas prioridades econômicas da corporação e as empresas enfrentando as fraquezas inerentes da corporação, o que limitaria sua capacidade de superar uma crise financeira. Além disso, é intuitivo supor que, quanto maior a diferença de receita entre duas empresas dentro da corporação, menor a dependência entre elas, isto é, empresas com maiores receitas não dependem de empresas com receitas menores, enquanto que empresas com receitas similares teriam uma maior dependência entre si, a fim de manter as suas posições de mercado, serviços comerciais, vendas e informações.

Por estas razões, a estrutura de correlação exponencial parece ser uma boa opção. Essa estrutura de correlação é dada por

$$h(v(\boldsymbol{r}_i), \gamma) = \exp(-\gamma \max_{s,t} |r_{i1s} - r_{i1t}| / \max_{i} \max_{s,t} |r_{i1s} - r_{i1t}|),$$

com $s, t = 1, \ldots, n_i$ e $i = 1, \ldots, m$. Observe que, se todas as empresas em uma corporação têm as mesmas receitas, elas estarão em *default* em bloco, ou nenhuma delas em *default* (correlação perfeita), o que poderia ser um caso limite.

O interesse principal na análise é ajustar um modelo de regressão que possa ser utilizado para determinar a probabilidade de inadimplência de uma nova corporação. Assumindo que uma corporação está em *default* se pelo menos uma das empresas está em *default*. Como é do nosso conhecimento, a abordagem mais comum para modelagem deste tipo de dados é através do modelo de regressão logística. O modelo de regressão binomial correlacionada é usado como uma nova abordagem na análise com o objetivo de proporcionar inferências mais realistas.

As mesmas *prioris* vagas usadas no estudo de simulação são utilizadas aqui. É considerada a *priori* $N(0, 10^4)$ para cada um dos parâmetros $\beta_0, \beta_1, \beta_2, \beta_3$ e $\log(\gamma)$. O modelo de regressão binomial correlacionada e o modelo de regressão binomial foram ajustados aos dados usando as quatro funções de ligação. Os resultados obtidos pelo critério de seleção DIC são mostrados na Tabela 5.8. Como pode ser observado pelos resultados presentes na Tabela 5.8, o modelo de regressão binomial correlacionada com função de ligação logito é identificado como a melhor escolha.

Tabela 5.8: Valores obtidos pelo critério DIC ajustando o modelo de regressão binomial correlacionada (MRBC) e o modelo de regressão binomial (MRB) com diferentes funções de ligação.

Modelo	Logito	Complementar log-log	Log-log	Probito
MRBC	100,442	101,981	103,406	102,006
MRB	214,956	214,746	220,857	220,025

Tabela 5.9: Resumos das densidades marginais a *posteriori* de γ, β_0, β_1, β_2 e β_3, para os dados de inadimplência.

Parâmetros	Média a *posteriori*	Mediana a *posteriori*	Intervalo de Credibilidade 95%	Intervalo HPD 95%
γ	1,811	1,789	(1,370 ; 2,435)	(1,359 ; 2,425)
β_0	-3,158	-3,157	(-4,331 ; -2,079)	(-4,133 ; -1,952)
β_1	-2,533	-2,533	(-3,439 ; -1,639)	(-3,465 ; -1,723)
β_2	0,361	0,359	(0,182 ; 0,588)	(0,181 ; 0,583)
β_3	0,422	0,407	(0,044 ; 0,866)	(0,044 ; 0,866)

Os resumos da distribuição a *posteriori* para o modelo com ligação logito são mostrados na Tabela 5.9. É importante ressaltar que tanto o intervalo de credibilidade inter-quantil

como o intervalo de credibilidade HPD do parâmetro da estrutura de correlação, γ, não contêm o valor zero, corroborando com a necessidade do ajuste de um modelo que incorpore correlação positiva entre os eventos de Bernoulli. O diagnóstico de convergência dos parâmetros foi verificado via CODA, por meio do teste de Geweke, que apresentou valores entre (-1,2; 1,2).

Na Figura 5.9 são apresentados os gráficos dos resíduos *deviance* Bayesianos para os ajustes do modelo de regressão binomial correlacionada com as quatro funções de ligação: Figura 5.9a logito, Figura 5.9b complementar log-log, Figura 5.9c log-log e Figura 5.9d probito. As contribuições das observações são muito semelhantes para todas as funções de ligação.

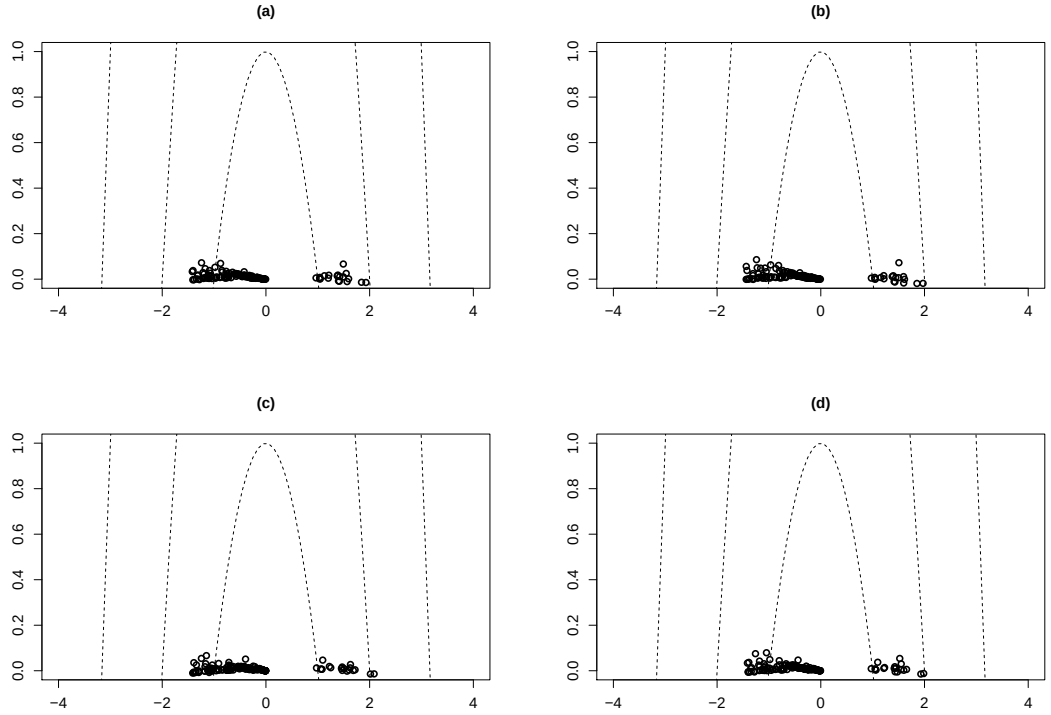

Figura 5.9: MRBC: Resíduos *deviance* versus alavancagem Bayesiana, considerando $c = 1, 4, 10$, (a) ligação logito, (b) ligação complementar log-log, (c)ligação log-log, (d) ligação probito.

A suposição de independência e a presença de *outliers* podem ser verificadas examinando o gráfico dos resíduos padronizados contra a ordem, se a ordem estiver disponível. A especificação do modelo e a presença de *outliers* são observadas ao examinar os resíduos padronizados contra os valores preditos obtidos pela densidade preditiva condicional ordinária. Esses gráficos são apresentados na Figura 5.10a, na Figura 5.10b e na Figura 5.10c, para os três tipos de resíduos Bayesianos. Na Figura 5.10a, os resíduos são baseados nos valores preditos via densidade preditiva condicional ordinária, como descrito em (4.6). Na Figura 5.10b, os resíduos são baseados na distribuição a *posteriori* dos parâmetros do modelo. As médias e intervalos de credibilidade HPD, com 95% de credibilidade, para

as amostras das distribuições dos resíduos a *posteriori* contra os valores esperados são apresentados na Figura 5.10b. Os boxplots das amostras MCMC das distribuições a *posteriori* dos resíduos padronizados, para cada valor esperado, $E(Y_i|\mathcal{D}, \boldsymbol{\theta})$, são apresentados na Figura 5.11. As amplitudes destes boxplots são pequenas devido à baixa variação dos pontos em cada amostra. Casos com dispersões elevadas requerem mais investigações. Na Figura 5.10c os resíduos *deviance* Bayesianos evidenciam a observação 113 como um possível *outlier*.

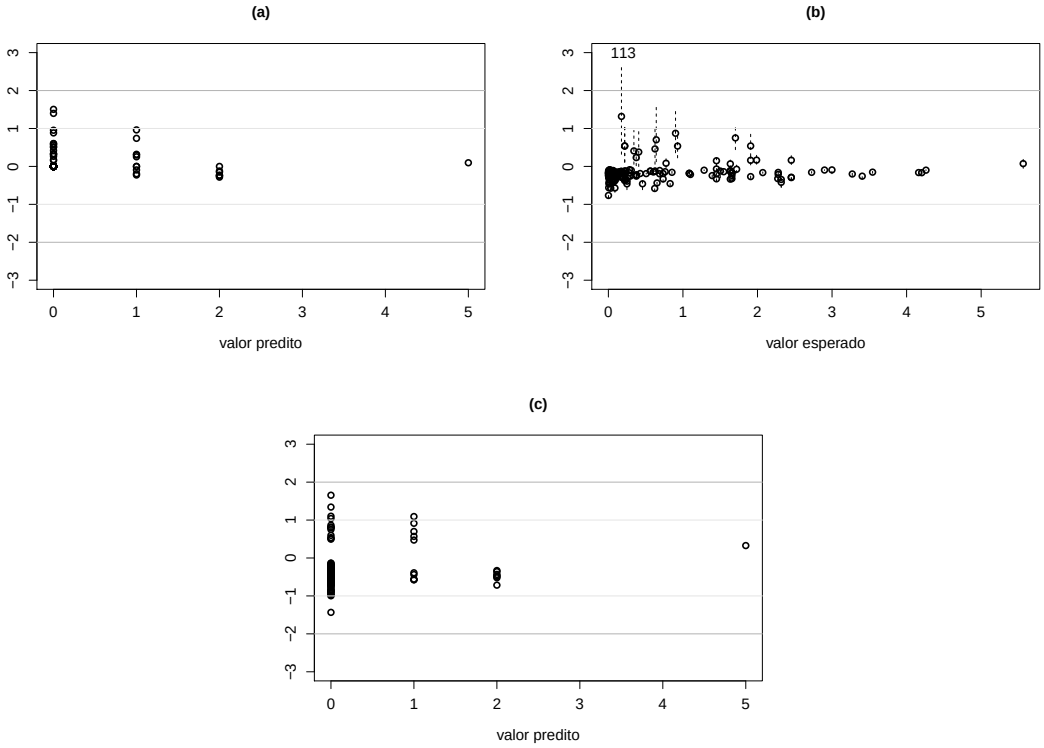

Figura 5.10: MRBC: (a) resíduos padronizados via densidade preditiva condicional ordinária versus valores preditos, (b) resíduos padronizados via distribuição a *posteriori* dos parâmetros do modelo versus valores esperados, (c) resíduos *deviance* padronizados versus valores preditos.

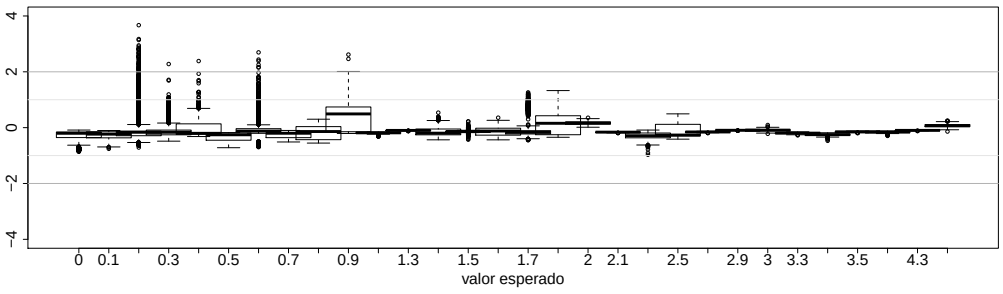

Figura 5.11: MRBC: boxplots das amostras da distribuição a *posteriori* dos resíduos para cada observação versus os valores esperados de $E(Y_i|\mathcal{D}, \boldsymbol{\theta})$.

A avaliação do impacto causado pelo caso 113 na estimação dos parâmetros é mostrada na Tabela 5.10. Os dados não apresentaram nenhum valor de calibração maior que 0,61, indicando que, apesar do caso 113 ter sido indicado como um *outlier*, a divergência de K-L não classifica este ponto como uma observação influente no processo de estimação. A Tabela 5.10 apresenta a média a *posteriori* e a mudança relativa deste valor médio com respeito a média obtida omitindo o caso 113. A calibração desta observação é igual a 0,56.

Tabela 5.10: Média a *posteriori* e mudança relativa em relação aos dados originais para os dados de inadimplência ajustados via MRBC.

	Parâmetros									
Casos	γ		β_0		β_1		β_2		β_3	
omitidos	Média	%	Média	%	Média	%	Média	%	Média	%
{Nenhum}	1,81	-	-3,16	-	-2,53	-	0,36	-	0,42	-
Caso {113}	1,80	-0,55	-3,19	0,94	-2,67	-5,53	0,36	-	0,46	9,52

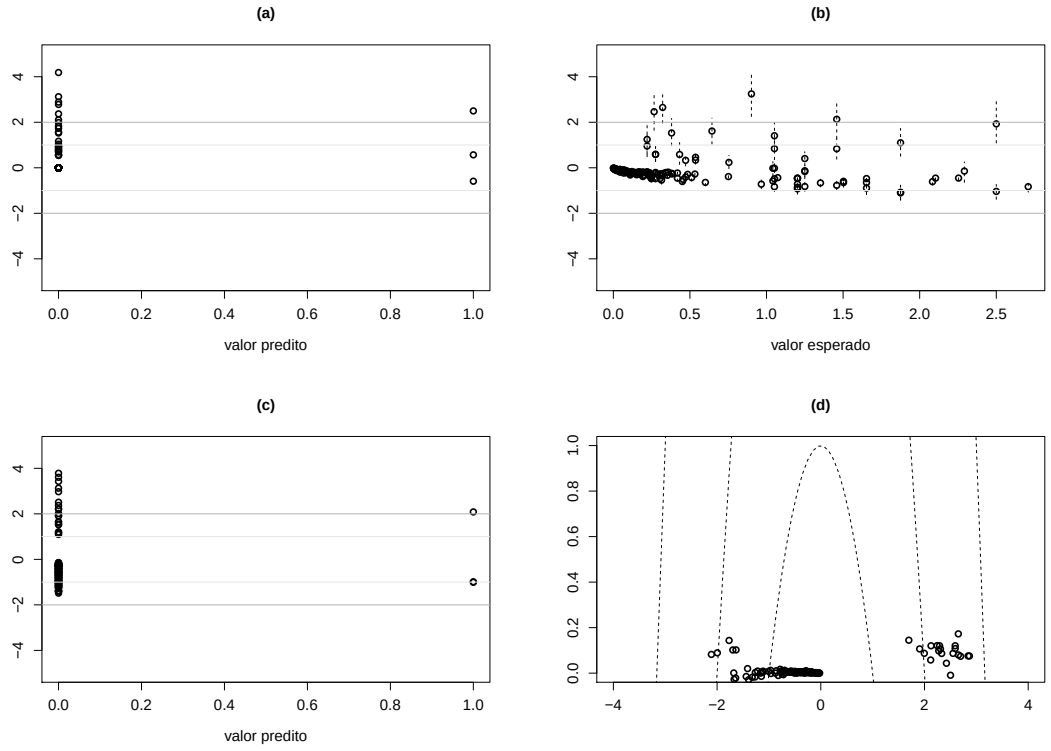

Figura 5.12: MRBB: (a) resíduos padronizados via densidade preditiva condicional ordinária versus valores preditos, (b) resíduos padronizados via distribuição a *posteriori* dos parâmetros do modelo versus valores esperados, (c) resíduos *deviance* padronizados versus valores preditos, (d) resíduos *deviance* versus alavancagem Bayesiana.

Para reforçar a necessidade de usar o modelo proposto neste conjunto de dados, ajustamos o modelo de regressão binomial usual que apresenta valor do DIC de 214,956, e o

modelo de regressão beta-binomial com valor do DIC de 195,179. Os modelos de regressão binomial correlacionada e beta-binomial foram ajustados usando a função de ligação logito e a estrutura de correlação exponencial.

Além do valor do DIC em seu favor, a análise mostrada nesta seção indica que o modelo alternativo proporciona um ajuste muito bom para este conjunto de dados. As Figuras 5.12a, 5.12b, 5.12c e 5.12d são similares às Figuras 5.10a, 5.10b, 5.10c, 5.9d e 5.11 para o modelo de regressão beta-binomial. Todos estes gráficos mostram valores de resíduos grandes, indicando que o modelo de regressão beta-binomial não se ajusta bem aos dados.

Ao considerar o conjunto de dados completo, a tomada de decisão baseada na análise conduzida pelo modelo de regressão binomial correlacionada nesta seção estabelece a probabilidade de *default* na i-ésima corporação como sendo

$$\hat{p_i} = \frac{\exp\{-3,158 - 2,533x_{i1} + 0,361x_{i2} + 0,422x_{i3}\}}{1 + \exp\{-3,158 - 2,533x_{i1} + 0,361x_{i2} + 0,422x_{i3}\}},$$

sendo x_{i1}: *o número de sócios pessoa física na corporação*, x_{2i}: *o porte da corporação* e x_{i3}: *o número de protestos da corporação com qualquer atraso no último ano*. E a correlação entre quaisquer duas empresas dentro da corporação por

$$\hat{\rho_i} = \exp\{-1,811 \max_{s,t} |r_{i1s} - r_{i1t}| / (\max_{i} \max_{s,t} |r_{i1s} - r_{i1t}|)\},$$

sendo r_{i1} *a receita das empresas.*

Capítulo 6

Modelos de regressão binomial correlacionada aditivo estrutural normal (MRBCAEN)

Ao analisar o modelo de regressão binomial correlacionada, proposto no Capítulo 2, algumas extensões podem ser feitas, como, por exemplo, inserir uma covariável com erro de medida. Uma motivação utilizando o conjunto de dados de planos de saúde, analisado na Seção 5.5.1, é considerar uma covariável, tal como o *índice de predisposição a uma determinada doença*, construído para os funcionários da empresa. Este índice seria uma observação medida com erro da verdadeira predisposição à doença de cada funcionário, possível apenas por meio de uma análise genética, que seria economicamente inviável e portanto quase sempre não disponível. Na presença de uma ou mais variáveis medidas com erro o modelo de regressão binomial correlacionada não é aconselhável, uma vez que é imposta, na modelagem, a suposição que as covariáveis não apresentem erro ou apresentem erro desprezível. Se uma análise é conduzida via MRBC na presença de uma variável com erro de medida, um vício na estimativa do parâmetro correspondente pode ocorrer, induzindo à inferências incorretas.

A proposta deste capítulo é estender o MRBC, permitindo inserir na modelagem uma variável medida com erro, considerando um processo aditivo.

O modelo de regressão binomial correlacionada aditivo estrutural normal está baseado na distribuição binomial generalizada.

6.1 Uma classe de modelos de regressão binomial correlacionada aditivo estrutural normal

Considerando que um grupo de covariáveis comuns aos clusters esteja disponível, a probabilidade de sucesso, p_i, pode ser modelada por meio de covariáveis do i-ésimo cluster. Dentre estas covariáveis está presente u_i, que não é observada diretamente, ou seja, apenas uma função de u_i medida com erro é, de fato, conhecida. Suponha, agora, que este valor

observado possa ser expresso como sendo $w_i = u_i + \epsilon_i$, uma realização da função de variáveis aleatórias $W_i = U_i + \epsilon_i$ com $U_i \sim N(\mu, \sigma^2)$ e $\epsilon_i \sim N(0, b\sigma^2)$, com $b > 0$ uma constante conhecida que indica a proporção de erro na variância observada em relação a variância da covariável sem contaminação, U_i e Y_i independentes de ϵ_i. Assim, $W_i \sim N(\mu, (1+b)\sigma^2)$.

Como em modelos lineares generalizados, funções de ligação podem ser utilizadas para conectar p_i com uma função linear das covariáveis η_i, da forma

$$\eta_i = \sum_{r=0}^{k} \beta_r x_{ir} + \delta u_i, \tag{6.1}$$

em que os coeficientes $\delta, \beta_0, \beta_1, \ldots, \beta_k$ são parâmetros desconhecidos; $x_{i0} = 1$, para todo i e $x_{1i}, x_{2i}, \ldots, x_{ik}$ são os valores das k covariáveis para o i-ésimo cluster e u_i é uma covariável não observada. Consideramos neste capítulo, apenas a função de ligação logito, porém as demais funções de ligação apresentadas na Tabela 2.1 podem ser utilizadas.

Para modelar a correlação, ρ_i, dos indivíduos no i-ésimo cluster, considere uma função de covariáveis destes indivíduos capazes de informar sobre a dependência dentro do cluster em relação ao evento de interesse. Uma estrutura de correlação pode ser escrita como em (2.3), discutida na construção do modelo de regressão binomial correlacionada (ver Capítulo 2).

Para determinar a função de verossimilhança do modelo de regressão binomial correlacionada aditivo estrutural normal, é necessário obter a distribuição conjunta, (Y_i, W_i), que pode ser obtida por

$$\begin{aligned} F_{Y_i,W_i}(y_i, w_i) &= \int_{-\infty}^{+\infty} F_{Y_i,W_i|U_i}(y_i, w_i|u_i) F_{U_i}(u_i) du_i \\ &= \int_{-\infty}^{+\infty} F_{Y_i|U_i}(y_i|u_i) F_{W_i|Y_i,U_i}(w_i|y_i, u_i) F_{U_i}(u_i) du_i \\ &= \int_{-\infty}^{+\infty} F_{Y_i|U_i}(y_i|u_i) F_{\epsilon_i|Y_i,U_i} P(w_i - u_i|y_i, u_i) F_{U_i}(u_i) du_i. \end{aligned} \tag{6.2}$$

Sejam $Y_1|U_1, \ldots, Y_m|U_m$ uma sequência independente de variáveis aleatórias $CB(n_i, p_i, \rho_i)$, $i = 1, \ldots, m$; $\boldsymbol{n} = (n_1, n_2, \ldots, n_m)^\top$, $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m)^\top$, $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ik})^\top$, $\boldsymbol{w}_i = (w_1, \ldots, w_m)^\top$ e $\boldsymbol{r} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m)^\top$, $\boldsymbol{r}_i = (r_{i11}, \ldots, r_{i1n_i}, r_{i21}, \ldots, r_{i2n_i}, r_{iq1}, \ldots, r_{iqn_i})^\top$. Usando (6.1), (2.3) e (6.2), a função de verossimilhança do vetor de parâmetros $\boldsymbol{\theta} = (\gamma, \delta, \mu, \sigma^2, \beta_0, \ldots, \beta_k)^\top$, condicionado aos dados observados, $\mathcal{D} = (m, \boldsymbol{n}, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}, \boldsymbol{r})^\top$, para o modelo de regressão binomial correlacionada aditivo estrutural normal é dada por

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) &= \prod_{i=1}^{m} \Bigg\{ (2\pi\sigma^2 b^{\frac{1}{2}})^{-1} \exp\left\{ -\frac{1}{2\sigma^2} \left(\frac{w_i^2}{b} + \mu^2 - \left(\frac{(b\mu + w_i)^2}{b(1+b)} \right) \right) \right\} \\ &\quad \times \int_{-\infty}^{+\infty} \exp\left\{ -\frac{\frac{1+b}{b}}{2\sigma^2} \left(u_i - \left(\frac{b\mu + w_i}{1+b} \right) \right)^2 \right\} \\ &\quad \times \Bigg\{ \binom{n_i}{y_i} g^{-1}(\eta_i)^{y_i} (1 - g^{-1}(\eta_i))^{n_i - y_i} \left(1 - h(v(\boldsymbol{r}_i), \gamma)\right) \\ &\quad + g^{-1}(\eta_i)^{\frac{y_i}{n_i}} (1 - g^{-1}(\eta_i))^{\frac{n_i - y_i}{n_i}} h(v(\boldsymbol{r}_i), \gamma) I_{A_{2i}}(y_i) \Bigg\} du_i \Bigg\}. \end{aligned} \tag{6.3}$$

em que $y_i = 0, 1, \ldots, n_i$; $w_i, u_i, \mu \in \mathbb{R}$; $b, \sigma^2 > 0$; $n_i \in \mathbb{N} - \{0\}$; $A_{2_i} = \{0, n_i\}$; sendo $g^{-1}(\eta_i)$ a função de ligação logito, apresentada na Tabela 2.1, com η_i na forma (6.1), e $h(v(\boldsymbol{r}_i), \gamma)$ uma das estruturas de correlação sugeridas da Tabela 2.2.

Note a complexidade da função de verossimilhança (6.3). A função envolve uma integral, que não possui solução analítica, com produto de somas. Uma aproximação para a integral em (6.3) é apresentada na próxima seção.

6.1.1 Aproximação da função de verosimilhança

Devido as dificuldades observadas no cálculo analítico da integral em (6.3), o método de Laplace (Bernardo & Smith, 2000; Tanner, 1996) é utilizado para aproximar a integral

$$\begin{aligned} I_{\mathcal{D}} &= \int_{-\infty}^{+\infty} \exp\left\{-\frac{\frac{1+b}{b}}{2\sigma^2}\left(u_i - \left(\frac{b\mu + w_i}{1+b}\right)\right)^2\right\}\left\{\binom{n_i}{y_i} g^{-1}(\eta_i)^{y_i}(1 - g^{-1}(\eta_i))^{n_i - y_i}\right. \\ &\quad \left.\times (1 - h(v(\boldsymbol{r}_i), \gamma)) + g^{-1}(\eta_i)^{\frac{y_i}{n_i}}(1 - g^{-1}(\eta_i))^{\frac{n_i - y_i}{n_i}} h(v(\boldsymbol{r}_i), \gamma) I_{A_{2i}}(y_i)\right\} du_i. \end{aligned}$$

Considerando a formulação apresentada no Apêndice B, seja

$$h(u_i) = \frac{\frac{1+b}{b}}{2N\sigma^2}\left(u_i - \left(\frac{b\mu + w_i}{1+b}\right)\right)^2, \text{ com } N \to \infty,$$

cujo ponto de máximo é $\hat{u}_i = \dfrac{b\mu + w_i}{1+b}$. E $\left(\dfrac{d^2 h(u_i)}{du_i^2}\right)^{-1} = \left(\dfrac{\frac{1+b}{b}}{N\sigma^2}\right)^{-1}$. Assim, a integral $I_{\mathcal{D}}$ é aproximada por

$$\begin{aligned} \hat{I}_{\mathcal{D}} &= \left(2\pi\sigma^2\right)^{\frac{1}{2}}\left(\frac{1+b}{b}\right)^{-\frac{1}{2}}\left\{\binom{n_i}{y_i} g^{-1}(\eta_i^*)^{y_i}(1 - g^{-1}(\eta_i^*))^{n_i - y_i}\,(1 - h(v(\boldsymbol{r}_i), \gamma))\right. \\ &\quad \left. + g^{-1}(\eta_i^*)^{\frac{y_i}{n_i}}(1 - g^{-1}(\eta_i^*))^{\frac{n_i - y_i}{n_i}} h(v(\boldsymbol{r}_i), \gamma) I_{A_{2i}}(y_i)\right\}, \end{aligned}$$

com $\eta_i^* = \sum_{r=0}^{k} \beta_r x_{ir} + \delta(b\mu + w_i)/(1+b)$.

A função de verossimilhança $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$, utilizando $\hat{I}_{\mathcal{D}}$, é aproximada por

$$\begin{aligned} \mathcal{L}_a(\boldsymbol{\theta}; \mathcal{D}) &= \prod_{i=1}^{m}\left\{(2\pi\sigma^2(1+b))^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\frac{w_i^2}{b} + \mu^2 - \left(\frac{(b\mu + w_i)^2}{b(1+b)}\right)\right)\right\}\right. \\ &\quad \times\left\{\binom{n_i}{y_i} g^{-1}(\eta_i^*)^{y_i}(1 - g^{-1}(\eta_i^*))^{n_i - y_i}\,(1 - h(v(\boldsymbol{r}_i), \gamma))\right. \\ &\quad \left.\left. + g^{-1}(\eta_i^*)^{\frac{y_i}{n_i}}(1 - g^{-1}(\eta_i^*))^{\frac{n_i - y_i}{n_i}} h(v(\boldsymbol{r}_i), \gamma) I_{A_{2i}}(y_i)\right\}\right\}, \end{aligned} \tag{6.4}$$

com $\eta_i^* = \sum_{r=0}^{k} \beta_r x_{ir} + \delta(b\mu + w_i)/(1+b)$.

Modelos envolvendo misturas de distribuições apresentam problema de identificabilidade, decorrente da falta de informação da origem da resposta observada que pode ser

proveniente de qualquer subpopulação, tornando não bem definido os métodos usuais de estimação (Titterington *et al.*, 1985). Entretanto, na próxima seção, uma estratégia de dados aumentados (Tanner & Wong, 1987; Diebolt & Robert, 1994) é construída para este problema com o objetivo de tornar a função de verossimilhança (6.3) identificável.

6.2 Função de verosimilhança com dados aumentados

Ao observar as realizações da variável aleatória $Y_i|U_i$, não é possível identificar de qual termo da distribuição binomial correlacionada y_i é originária, uma vez que y_i pode ser uma realização de uma distribuição binomial ou de uma distribuição Bernoulli modificada. Para caracterizamos os dados completos introduzimos uma variável latente $Z_i|U_i$, $i = 1, \ldots, m$, que indica o componente do modelo $CB(n_i, p_i, \rho_i)$ da qual a observação y_i, $i = 1, \ldots, m$, é originária, ou seja,

$$Z_i|U_i = \begin{cases} 1, & \text{se a observação } y_i \text{ resulta de } MBern(p_i) \\ 0, & \text{se a observação } y_i \text{ resulta de } B(n_i, p_i) \end{cases}.$$

Para construirmos a função de verossimilhança com dados aumentados é necessário determinar a distribuição conjunta (Y_i, W_i, Z_i). A probabilidade de sucesso da variável aleatória, Z_i, condicionada a observação do vetor (y_i, u_i), $(Z_i|Y_i, U_i)$, é dada por

$$\begin{aligned} \tau_i &= P(Z_i = 1|Y_i = y_i, U_i = u_i) = \frac{P(Y_i = y_i, Z_i = 1|U_i = u_i)}{P(Y_i = y_i|U_i = u_i)} \\ &= \frac{P(Y_i = y_i|Z_i = 1, U_i = u_i)P(Z_i = 1|U_i = u_i)}{P(Y_i = y_i|U_i = u_i)} \\ &= \frac{\rho_i p_i^{\frac{y_i}{n_i}}(1-p_i)^{\frac{n_i-y_i}{n_i}} I_{A_{2_i}}(y_i)}{\rho_i p_i^{\frac{y_i}{n_i}}(1-p_i)^{\frac{n_i-y_i}{n_i}} I_{A_{2_i}}(y_i) + (1-\rho_i)\binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i}}. \end{aligned} \tag{6.5}$$

A distribuição condicional de $(Z_i|Y_i, U_i)$ é escrita como

$$\begin{aligned} P(Z_i = z_i|Y_i = y_i, U_i = u_i) &= \frac{\tau_i^{z_i}(1-\tau_i)^{1-z_i}}{P(Y_i = y_i|U_i = u_i)} \\ &= \frac{\left(\rho_i p_i^{\frac{y_i}{n_i}}(1-p_i)^{\frac{n_i-y_i}{n_i}} I_{A_{2_i}}(y_i)\right)^{z_i} \left((1-\rho_i)\binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i}\right)^{1-z_i}}{\rho_i p_i^{\frac{y_i}{n_i}}(1-p_i)^{\frac{n_i-y_i}{n_i}} I_{A_{2_i}}(y_i) + (1-\rho_i)\binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i}}, \end{aligned} \tag{6.6}$$

para $i = 1, \ldots, m$.

A distribuição conjunta (Y_i, W_i, Z_i) pode ser expressa na forma

$$\begin{aligned} F_{Y_i,W_i,Z_i}(y_i, z_i, w_i) &= \int_{-\infty}^{+\infty} F_{Y_i,W_i,Z_i|U_i}(y_i, z_i, w_i|u_i) F_{U_i}(u_i) du_i \\ &= \int_{-\infty}^{+\infty} P_{Z_i|Y_i,U_i}(z_i|y_i, u_i) F_{Y_i,W_i|U_i}(y_i, w_i|u_i) F_{U_i}(u_i) du_i. \end{aligned} \tag{6.7}$$

Utilizando (6.6) em (6.7), a função de verossimilhança para $\boldsymbol{\theta} = (\gamma, \delta, \mu, \sigma^2, \beta_0, \ldots, \beta_k)^\top$, condicionado aos dados completos $\mathcal{D}^* = (m, \boldsymbol{n}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{x}, \boldsymbol{r})^\top$, é escrita como

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}^*) &= \prod_{i=1}^{m} \Bigg\{ (2\pi\sigma^2 b^{\frac{1}{2}})^{-1} \binom{n_i}{y_i}^{(1-z_i)} h(v(\boldsymbol{r}_i), \gamma)^{z_i} (1 - h(v(\boldsymbol{r}_i), \gamma))^{(1-z_i)} \\
&\quad \times \exp\left\{ -\frac{1}{2\sigma^2} \left(\frac{w_i^2}{b} + \mu^2 - \left(\frac{(b\mu + w_i)^2}{b(1+b)} \right) \right) \right\} I_{A_{2_i}}(y_i)^{z_i} \\
&\quad \times \int_{-\infty}^{+\infty} \exp\left\{ -\frac{\frac{1+b}{b}}{2\sigma^2} \left(u_i - \left(\frac{b\mu + w_i}{1+b} \right) \right)^2 \right\} \\
&\quad \times g^{-1}(\eta_i)^{\frac{y_i}{n_i}(z_i + n_i - n_i z_i)} (1 - g^{-1}(\eta_i))^{(n_i - y_i)\left(\frac{z_i}{n_i} + 1 - z_i\right)} du_i, \Bigg\}, \qquad (6.8)
\end{aligned}$$

em que $y_i = 0, 1, \ldots, n_i$; $w_i, u_i, \mu \in \mathbb{R}$; $b, \sigma^2 > 0$; $n_i \in \mathbb{N} - \{0\}$; $A_{2_i} = \{0, n_i\}$; sendo $g^{-1}(\eta_i)$ a função de ligação logito apresentada na Tabela 2.1, com η_i na forma (6.1), e $h(v(\boldsymbol{r}_i), \gamma)$ uma das estruturas de correlação sugeridas da Tabela 2.2.

Observe que a função de verossimilhança com dados completos (6.8) também envolve uma integral, que não possui resolução analítica. Na próxima seção, é sugerida uma aproximação para esta integral.

6.2.1 Aproximação da função de verosimilhança com dados aumentados

Devido às dificuldades observadas no cálculo analítico da integral em (6.8), o método de Laplace (Bernardo & Smith, 2000; Tanner, 1996) é utilizado para aproximar a integral

$$\begin{aligned}
I_{\mathcal{D}^*} &= \int_{-\infty}^{+\infty} \exp\left\{ -\frac{\frac{1+b}{b}}{2\sigma^2} \left(u_i - \left(\frac{b\mu + w_i}{1+b} \right) \right)^2 \right\} \\
&\quad \times g^{-1}(\eta_i)^{\frac{y_i}{n_i}(z_i + n_i - n_i z_i)} (1 - g^{-1}(\eta_i))^{(n_i - y_i)\left(\frac{z_i}{n_i} + 1 - z_i\right)} du_i.
\end{aligned}$$

Considerando uma aproximação análoga a apresentada na Seção 6.1.1, a integral $I_{\mathcal{D}^*}$ é aproximada por

$$\hat{I}_{\mathcal{D}^*} = \left(2\pi\sigma^2\right)^{\frac{1}{2}} \left(\frac{1+b}{b} \right)^{-\frac{1}{2}} g^{-1}(\eta_i^*)^{\frac{y_i}{n_i}(z_i + n_i - n_i z_i)} (1 - g^{-1}(\eta_i^*))^{(n_i - y_i)\left(\frac{z_i}{n_i} + 1 - z_i\right)}, \qquad (6.9)$$

com $\eta_i^* = \sum_{r=0}^{k} \beta_r x_{ir} + \delta(b\mu + w_i)/(1+b)$.

Assim, a função de verossimilhança $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}^*)$, utilizando (6.9), é aproximada por

$$\begin{aligned}
\mathcal{L}_a(\boldsymbol{\theta}; \mathcal{D}^*) &= \prod_{i=1}^{m} \Bigg\{ (2\pi\sigma^2(1+b))^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \left(\frac{w_i^2}{b} + \mu^2 - \left(\frac{(b\mu + w_i)^2}{b(1+b)} \right) \right) \right\} \\
&\quad \times \binom{n_i}{y_i}^{(1-z_i)} I_{A_{2_i}}(y_i)^{z_i} h(v(\boldsymbol{r}_i), \gamma)^{z_i} (1 - h(v(\boldsymbol{r}_i), \gamma))^{(1-z_i)} \\
&\quad \times g^{-1}(\eta_i^*)^{\frac{y_i}{n_i}(z_i + n_i - n_i z_i)} (1 - g^{-1}(\eta_i^*))^{(n_i - y_i)\left(\frac{z_i}{n_i} + 1 - z_i\right)} \Bigg\}, \qquad (6.10)
\end{aligned}$$

com $\eta_i^* = \sum_{r=0}^{k} \beta_r x_{ir} + \delta(b\mu + w_i)/(1+b)$.

A função de verossimilhança aproximada (6.10), baseada nos dados latentes, possibilita a condução de um processo inferencial bem definido.

Usando a parametrização $\phi = \log(\gamma)$, para a estrutura de correlação exponencial ou Gaussiana, ou $\phi = \log(\gamma/(1-\gamma))$, para a estrutura de correlação AR contínua, o parâmetro γ pode ser estimado sem restrição. Considerando, ainda, a parametrização $\psi = \log(\sigma^2)$, o parâmetro σ^2 pode ser estimado sem restrição.

Assim, função de log-verossimilhança para $\boldsymbol{\theta^*} = (\phi, \delta, \mu, \psi, \beta_0, \ldots, \beta_k)^\top$, condicionado aos dados completos, $\mathcal{D}^* = (m, \boldsymbol{n}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{w}, \boldsymbol{r})^\top$, é dada por

$$\begin{aligned}\ell_a(\boldsymbol{\theta}^*; \mathcal{D}^*) \;\propto\; & \sum_{i=1}^{m} \left\{ -\frac{1}{2}\psi - \frac{1}{2\exp\{\psi\}} \left(\frac{w_i^2}{b} + \mu^2 - \left(\frac{(b\mu + w_i)^2}{b(1+b)} \right) \right) \right. \\ & + (1 - z_i) \log(1 - h^*(v(\boldsymbol{r}_i), \phi)) + (n_i - y_i) \left(\frac{z_i}{n_i} + 1 - z_i \right) \log(1 - g^{-1}(\eta_i^*)) \\ & \left. + z_i \log(h^*(v(\boldsymbol{r}_i), \phi)) + \frac{y_i}{n_i}(z_i + n_i - n_i z_i) \log(g^{-1}(\eta_i^*)) \right\}, \end{aligned} \tag{6.11}$$

na qual $h^*(v(\boldsymbol{r}_i), \phi)$ é similar a função $h(v(\boldsymbol{r}_i), \gamma)$, apresentada na Tabela 2.2, considerando a parametrização $\gamma = \exp\{\phi\}/(1 + \exp\{\phi\})$ ou $\gamma = \exp\{\phi\}$ e $\eta_i^* = \sum_{r=0}^{k} \beta_r x_{ir} + \delta(b\mu + w_i)/(1+b)$.

Um método inferencial é desenvolvido para obter os estimadores de máxima verossimilhança, seus respectivos intervalos de confiança assintóticos, intervalos de confiança baseados em reamostragem e intervalos de confiança baseado na função de log-verossimilhança perfilada para os $(k+5)$ parâmetros do modelo de regressão binomial correlacionada aditivo estrutural normal. O método está baseado em uma estratégia de dados aumentados e no algoritmo EM. Uma análise de diagnóstico é desenvolvida envolvendo análise de resíduos e medidas de influência global. É apresentado ainda dois critérios de seleção de modelos.

6.3 Estimação via algoritmo EM

Uma metodologia usual para estimação de parâmetros na presença de variáveis latente é o algoritmo EM (Tanner, 1996). Este método consiste em dois passos, a esperança, passo E, e maximização, passo M. No passo E é determinado o valor esperado da log-verossimilhança aproximada com os dados completos, $\ell_a(\boldsymbol{\theta}^*; \mathcal{D}^*)$, condicionado aos dados observados, $\mathcal{D}$, e um valor inicial para $\boldsymbol{\theta}^*$, $\boldsymbol{\theta}^{(0)}$.

Como a variável latente, $Z_i|U_i$, segue uma distribuição de Bernoulli com parâmetro τ_i, então $E(Z_i) = \tau_i$, $i = 1 \ldots, m$, expresso em (6.5). Assim, o passo E é dado por

$$E\left(\ell_a(\boldsymbol{\theta};\mathcal{D}^*)|\boldsymbol{\theta}^{(0)},\mathcal{D}\right) \propto$$

$$\sum_{i=1}^{m}\left\{-\frac{1}{2}\psi - \frac{1}{2\exp\{\psi\}}\left(\frac{w_i^2}{b}+\mu^2-\left(\frac{(b\mu+w_i)^2}{b(1+b)}\right)\right)\right.$$

$$+(1-\tau_i)\log(1-h^*(v(\boldsymbol{r}_i),\phi)) + (n_i-y_i)\left(\frac{\tau_i}{n_i}+1-\tau_i\right)\log(1-g^{-1}(\eta_i^*))$$

$$\left.+\tau_i\log(h^*(v(\boldsymbol{r}_i),\phi)) + \frac{y_i}{n_i}(\tau_i+n_i-n_i\tau_i)\log(g^{-1}(\eta_i^*))\right\}, \tag{6.12}$$

na qual $h^*(v(\boldsymbol{r}_i),\phi)$ é similar a função $h(v(\boldsymbol{r}_i),\gamma)$, apresentada na Tabela 2.2, considerando a parametrização $\gamma = \exp\{\phi\}/(1+\exp\{\phi\})$ ou $\gamma = \exp\{\phi\}$.

A função resultante do passo E depende do vetor de parâmetros, $\boldsymbol{\theta}^*$, dos dados observados, $\mathcal{D}$, e do vetor latente, $\boldsymbol{\tau} = (\tau_1^{(0)},\ldots,\tau_m^{(0)})^\top$. No passo M, a maximização direta desta função resultante é feita com respeito ao vetor de parâmetros $\boldsymbol{\theta}^*$, fornecendo uma atualização de $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\theta}^{(1)}$, e, portanto, uma atualização do vetor $\boldsymbol{\tau}^{(0)}$, $\boldsymbol{\tau}^{(1)}$. As estimativas dos parâmetros são obtidas iterativamente repetindo estes dois passos até que os critérios de convergência sejam atingidos.

No Apêndice E.1 apresentamos a implementação usando o programa R (R Development Core Team, 2011) para obter os estimadores de máxima verossimilhança dos parâmetros.

6.4 Intervalos de confiança

Nesta seção, os intervalos de confiança assintóticos dos estimadores de máxima verossimilhança são calculados para os $(k+5)$ parâmetros presentes no modelo de regressão binomial correlacionada aditivo estrutural normal por meio da metodologia proposta por Louis (1982), baseada na função de verossimilhança aproximada com dados aumentados, expressa em (6.10). Também são considerados os intervalos de confiança *bootstrap* e via log-verossimilhança perfilada.

No Apêndice E.2 apresentamos a implementação usando o programa R (R Development Core Team, 2011) para obter os intervalos de confiança discutidos nesta seção.

6.4.1 Intervalos de confiança assintóticos

Sob as condições de regularidade (ver Apêndice A), a distribuição assintótica para $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ é uma distribuição Normal multivariada $N_{k+5}(\boldsymbol{\theta}, I(\boldsymbol{\theta})^{-1})$, em que $I(\boldsymbol{\theta})$ é a matriz de informação (Cox & Hinkley, 1979), que pode ser aproximada pela matriz de informação observada de Fisher, $J_a(\boldsymbol{\theta})$. Para o modelo de regressão binomial correlacionada aditivo estrutural normal a matriz J_a, com dimensão $(k+5)\times(k+5)$, utilizando a função de verossimilhança aproximada com dados completos, é dada pela matriz

$$J_a(\hat{\boldsymbol{\theta}}) = -E\left[\frac{\partial^2\ell_a(\boldsymbol{\theta};\mathcal{D}^*)}{\partial\boldsymbol{\theta}\boldsymbol{\theta}^\top}\middle|\mathcal{D}\right]\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} - E\left[\frac{\partial\ell_a(\boldsymbol{\theta};\mathcal{D}^*)}{\partial\boldsymbol{\theta}}\frac{\partial\ell_a(\boldsymbol{\theta};\mathcal{D}^*)}{\partial\boldsymbol{\theta}}^\top\middle|\mathcal{D}\right]\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},$$

que pode ser aproximada via Monte Carlo por

$$J_{\hat{a}}(\hat{\boldsymbol{\theta}}) = -\frac{1}{Q}\sum_{q=1}^{Q}\frac{\partial^2\ell_a(\boldsymbol{\theta};\mathcal{D}_q^*)}{\partial\boldsymbol{\theta}\boldsymbol{\theta}^\top}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} - \frac{1}{Q}\sum_{q=1}^{Q}\frac{\partial\ell_a(\boldsymbol{\theta};\mathcal{D}_q^*)}{\partial\boldsymbol{\theta}}\frac{\partial\ell_a(\boldsymbol{\theta};\mathcal{D}_q^*)}{\partial\boldsymbol{\theta}}^\top\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \tag{6.13}$$

sendo $\mathcal{D}_q^* = (\mathcal{D}, \boldsymbol{z}_q)^\top$, $q = 1, \ldots, Q$, com $\boldsymbol{z}_q = (z_{1q}, \ldots, z_{mq})^\top$ e $z_{iq} \sim Ber(\tau_i)$, $i = 1, \ldots, m$, com τ_i obtido em (6.5) considerando $p_i = g^{-1}(\sum_{r=0}^{k}\beta_r x_{ir} + \delta(\mu + w_i)/2)$, $\rho_i = h(v(\boldsymbol{r}_i), \gamma)$, e

$$\begin{aligned}\frac{\partial\ell_a(\boldsymbol{\theta};\mathcal{D}^*)}{\partial\zeta_1} &= \sum_{i=1}^{m}\Bigg\{f_1(\zeta_1) + \frac{\partial g^{-1}(\eta_i^*)}{\partial\zeta_1}\Bigg[\frac{y_i}{n_i}(z_{iq} + n_i - n_i z_{iq})\left[g^{-1}(\eta_i^*)\right]^{-1} \\ &\quad -(n_i - z_{iq})\left(\frac{z_{iq}}{n_i} + 1 - z_{iq}\right)\left[1 - g^{-1}(\eta_i^*)\right]^{-1}\Bigg]\Bigg\},\end{aligned}$$

$$\begin{aligned}\frac{\partial^2\ell_a(\boldsymbol{\theta};\mathcal{D}^*)}{\partial\zeta_1\partial\zeta_2} &= \sum_{i=1}^{m}\Bigg\{f_2(\zeta_1,\zeta_2) + \left[\left[g^{-1}(\eta_i^*)\right]^{-1}\frac{\partial^2 g^{-1}(\eta_i^*)}{\partial\zeta_1\partial\zeta_2} - \left[g^{-1}(\eta_i^*)\right]^{-2}\frac{\partial g^{-1}(\eta_i^*)}{\partial\zeta_1}\frac{\partial g^{-1}(\eta_i^*)}{\partial\zeta_2}\right] \\ &\quad \times\frac{y_i}{n_i}(z_{iq} + n_i - n_i z_{iq}) - \Bigg[\left[1 - g^{-1}(\eta_i^*)\right]^{-1}\frac{\partial^2 g^{-1}(\eta_i^*)}{\partial\zeta_1\partial\zeta_2} + \left[1 - g^{-1}(\eta_i^*)\right]^{-2} \\ &\quad \times\frac{\partial g^{-1}(\eta_i^*)}{\partial\zeta_1}\frac{\partial g^{-1}(\eta_i^*)}{\partial\zeta_2}\Bigg](n_i - y_i)\left(\frac{z_{iq}}{n_i} + 1 - z_{iq}\right)\Bigg\},\end{aligned}$$

com ζ_1 e ζ_2 assumindo os parâmetros δ, μ, σ^2, β_r e β_s com $r, s = 0, \ldots, k$ e $f_1(\zeta_1)$ e $f_2(\zeta_1, \zeta_2)$ são apresentadas na Tabela 6.1,

$$\frac{\partial\ell(\boldsymbol{\theta};\mathcal{D}^*)}{\partial\gamma} = \sum_{i=1}^{m}\left\{\frac{\partial h(v(\boldsymbol{r}_i),\gamma)}{\partial\gamma}\left[\frac{z_{iq}}{h(v(\boldsymbol{r}_i),\gamma)} - \frac{1 - z_{iq}}{1 - h(v(\boldsymbol{r}_i),\gamma)}\right]\right\}$$

e

$$\begin{aligned}\frac{\partial^2\ell(\boldsymbol{\theta};\mathcal{D}^*)}{\partial\gamma^2} &= \sum_{i=1}^{m}\Bigg\{\frac{\partial^2 h(v(\boldsymbol{r}_i),\gamma)}{\partial\gamma^2}\left[\frac{z_{iq}}{h(v(\boldsymbol{r}_i),\gamma)} - \frac{1 - z_{iq}}{1 - h(v(\boldsymbol{r}_i),\gamma)}\right] \\ &\quad + \left[\frac{\partial h(v(\boldsymbol{r}_i),\gamma)}{\partial\gamma}\right]^2\left[-\frac{z_{iq}}{h(v(\boldsymbol{r}_i),\gamma)^2} - \frac{1 - z_{iq}}{(1 - h(v(\boldsymbol{r}_i),\gamma))^2}\right]\Bigg\}.\end{aligned}$$

Devido a ortogonalidade entre os parâmetros, as derivadas $\dfrac{\partial^2\ell_a(\boldsymbol{\theta}|\mathcal{D}^*)}{\partial\zeta_1\partial\gamma}$ são todas iguais a zero.

As derivadas de primeira ordem, em relação aos parâmetros presentes na função de ligação, δ, μ e β_r, para a função de ligação logito são dadas por

$$\frac{\partial g^{-1}(\eta_i^*)}{\partial\beta_r} = x_{ir}A_{log} \qquad \frac{\partial g^{-1}(\eta_i^*)}{\partial\delta} = \frac{b\mu + w_i}{1 + b}A_{log} \qquad \frac{\partial g^{-1}(\eta_i^*)}{\partial\mu} = \frac{b\delta}{1 + b}A_{log}$$

Tabela 6.1: Funções auxiliares no cálculo das derivadas da função de log-verossimilhança aumentada.

$f_1(\mu) = \dfrac{w_i - \mu}{\sigma^2(1+b)}$	$f_1(\sigma^2) = \dfrac{-\sigma^2(1+b) + w_i^2 - 2w_i\mu + \mu^2}{2(\sigma^2)^2(1+b)}$	
$f_2(\mu,\mu) = -\dfrac{1}{\sigma^2(1+b)}$	$f_2(\sigma^2,\sigma^2) = \dfrac{\sigma^2(1+b) - 2w_i^2 - 2\mu^2 + 4\mu w_i}{2(\sigma^2)^3(1+b)}$	$f_2(\mu,\sigma^2) = \dfrac{\mu - w_i}{(\sigma^2)^2(1+b)}$

Para as combinações envolvendo β_r, β_s e δ temos $f_1(.)$ e $f_2(.)$ iguais a zero.

As derivadas de segunda ordem, em relação aos parâmetros δ, μ, β_r e β_s $s, r = 0, \ldots, k$, para a função de ligação logito são dadas por

$$\begin{aligned}
\frac{\partial^2 g^{-1}(\eta_i^*)}{\partial\beta_r\partial\beta_s} &= x_{ir}x_{is}B_{log} & \frac{\partial^2 g^{-1}(\eta_i^*)}{\partial\beta_r\partial\delta} &= \frac{b\mu + w_i}{1+b}x_{ir}B_{log} \\
\frac{\partial^2 g^{-1}(\eta_i^*)}{\partial\beta_r\partial\mu} &= \frac{b\delta}{1+b}x_{ir}B_{log} & \frac{\partial^2 g^{-1}(\eta_i^*)}{\partial\delta\partial\delta} &= \frac{(b\mu + w_i)^2}{(1+b)^2}B_{log} \\
\frac{\partial^2 g^{-1}(\eta_i^*)}{\partial\delta\partial\mu} &= \frac{b}{(1+b)^2}\left[\delta(b\mu + w_i) - (1+b)C_{log}\right]B_{log} & \frac{\partial^2 g^{-1}(\eta_i^*)}{\partial\mu\partial\mu} &= \frac{b^2\delta^2}{(1+b)^2}B_{log}
\end{aligned}$$

As funções A_{log}, B_{log} e C_{log} presentes na derivadas de primeira e segunda ordem da função de ligação logito são apresentados na Tabela 6.2.

Tabela 6.2: Funções auxiliares no cálculo das derivadas, para a ligação logito.

$A_{log} = \exp\{\eta_i^*\}\left[1 + \exp\{\eta_i^*\}\right]^{-2}$
$B_{log} = -\exp\{\eta_i^*\}\left[\exp\{\eta_i^*\} - 1\right]\left[1 + \exp\{\eta_i^*\}\right]^{-3}$
$C_{log} = (\exp\{\eta_i^*\} + 1)(\exp\{\eta_i^*\} - 1)^{-1}$

As derivadas de primeira e segunda ordem em relação as estruturas de correlação são apresentadas na Tabela 3.3.

O intervalo de confiança assintótico, com nível de confiança de $100 \times (1-\alpha)\%$, para o r-ésimo componente do vetor de parâmetros $\boldsymbol{\theta}$, θ_r, $r = 1, \ldots, k+5$, pode ser calculado utilizando

$$\hat{\theta}_r \pm \mathcal{Z}_{\alpha/2}\sqrt{J^{-1}_{a(r)}(\hat{\boldsymbol{\theta}})}\ , \tag{6.14}$$

em que $\mathcal{Z}_{\alpha/2}$ é o valor do $(\alpha/2)$-ésimo quantil superior da distribuição Normal padrão e $J^{-1}_{a(r)}(\hat{\boldsymbol{\theta}})$ é o r-ésimo elemento da diagonal principal da inversa de $J_{\hat{a}}(\hat{\boldsymbol{\theta}})$, que corresponde ao estimador da variância do estimador de interesse.

6.4.2 Intervalos de confiança *bootstrap*

Os intervalos de confiança *bootstrap* não-paramétrico para os parâmetros do modelo de regressão binomial correlacionada aditivo estrutural normal são obtidos de forma análoga à descrição apresentada na Seção 3.2.2.

6.4.3 Intervalos de confiança perfilados

Os intervalos de plausibilidade para os parâmetros do modelo de regressão binomial correlacionada aditivo estrutural normal são construídos de forma análoga à descrita na Seção 3.2.3, considerando a função de verossimilhança apresentada em (6.4).

6.5 Teste de hipótese

Ao ajustar um modelo de regressão binomial correlacionada aditivo estrutural normal consideramos a suposição de significância do coeficiente da covariável medida com erro, δ. Porém, tal suposição pode não ser satisfeita, ou seja, um ajuste do modelo de regressão binomial correlacionada poderia ter sido considerado. Uma forma de verificar esta suposição é testar as hipóteses $H_0 : \delta = 0$ contra $H_1 : \delta \neq 0$ utilizando o teste de razão de verossimilhança, por meio da estatística

$$LR_\delta = -2\left\{\ell(\hat{\boldsymbol{\theta}}; \mathcal{D}) - \ell_a(\hat{\boldsymbol{\theta}}; \mathcal{D})\right\}, \tag{6.15}$$

sendo $\ell(\hat{\boldsymbol{\theta}}; \mathcal{D}) = \log\{\mathcal{L}(\hat{\boldsymbol{\theta}}; \mathcal{D})\}$ a função de log-verossimilhança do modelo de regressão binomial correlacionada (expressão (2.4)) avaliado nos respectivos estimadores de máxima verossimilhança fornecidos por este modelo; $\ell_a(\hat{\boldsymbol{\theta}}; \mathcal{D}) = \log\{\mathcal{L}_a(\hat{\boldsymbol{\theta}}; \mathcal{D})\}$, a função de log-verossimilhança avaliada no vetor de estimadores de máxima verossimilhança, $\hat{\boldsymbol{\theta}}$. A estatística de razão de verossimilhança LR_δ segue, assintoticamente, uma distribuição χ^2_3. A hipótese H_0 é rejeitada para grandes valores da estatística de teste LR_δ.

6.6 Diagnósticos

Entre as suposições impostas na construção do modelo de regressão binomial correlacionada podemos ressaltar: (i) a suposição de independência entre as variáveis resposta, (ii) a suposição que as variáveis resposta seguem uma distribuição binomial correlacionada, $BC(n_i, p_i, \rho_i)$, (iii) a suposição de correlação positiva entre as variáveis de Bernoulli dentro do cluster, $\rho_i > 0$, (iv) função de ligação, (v) estrutura de correlação, e (vi) a significância do parâmetro corresponde a covariável medida com erro, $\delta \neq 0$. Nesta seção uma análise de diagnóstico é proposta para a verificação dos pressupostos (i)-(vi), além de identificação de *outliers* e/ou observações influentes.

A suposição que $\rho_i > 0$ é verificada ao observar a significância do parâmetro da estrutura de correlação, γ, por meio dos intervalos de confiança obtidos na etapa de estimação. Para os casos em que esta suposição não for satisfeita, isto é, $\gamma = 1$ ou $\gamma = 0$, o modelo de regressão binomial usual pode ser considerado. Também pode ser verificado por meio dos intervalos de confiança a significância do parâmetro δ, se esta suposição não for satisfeita, isto é, $\delta = 0$, o modelo de regressão binomial correlacionada, proposto no Capítulo 2, deve ser considerado.

A detecção de observações influentes é feita utilizando a distância de Cook generalizada e a distância da verossimilhança (Zhu *et al.*, 2001) envolvendo observações deletadas (Cook & Weisberg, 1982).

6.6.1 Resíduos

Nesta seção, dois tipos de resíduos são construídos para o modelo de regressão binomial correlacionada aditivo estrutural normal. Um resíduo que depende do valor esperado de Y_i e um outro resíduo baseado na *deviance*.

Nos casos simulados, ambos os resíduos se mostraram eficientes na detecção de observações influentes (casos perturbados). No caso de dados não perturbados, os resíduos baseados nos valores preditos são mais sensíveis às características das observações, enquanto que os resíduos *deviance* são mais robustos. Ou melhor, resíduos *deviance* só identificam um caso como *outlier* quando o mesmo se destaca muito das demais observações.

Três gráficos, baseados nos resíduos padronizados, podem ser utilizados para verificar as suposições iniciais e identificar má especificação do modelo. Os gráficos (i) resíduo contra a ordem das observações, quando a mesma está disponível, para identificação de dependência temporal das observações; (ii) resíduo contra função das covariáveis, para verificar a necessidade de inserir outras funções de covariáveis, além da linear, na parte sistemática do modelo; (ii) resíduo contra valores preditos, onde espera-se que o conjunto de resíduos esteja próximo de zero. Se isso não ocorrer é uma indicação de que o modelo está mal ajustado aos dados.

Resíduos baseados nos valores preditos

O resíduo padronizado baseado no valor predito para o modelo de regressão binomial correlacionada aditivo estrutural normal é definido como

$$r_i^{spp} = \frac{y_i - n_i\hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)\{n_i + \hat{\rho}_i n_i(n_i-1)\}}}, \qquad i = 1, \ldots, m, \tag{6.16}$$

em que o par (y_i, n_i) são informações da i-ésima observação, $\hat{p}_i = g^{-1}(\sum_{r=0}^{k}\hat{\beta}_r x_{ir} + \hat{\delta}(b\hat{\mu} + w_i)/(1+b))$ e $\hat{\rho}_i = h(v(\boldsymbol{r}_i), \hat{\gamma})$, sendo $\hat{\delta}$, $\hat{\mu}$, $\hat{\sigma^2}$, $\hat{\beta}_r$, $r = 0, \ldots, k$, e $\hat{\gamma}$ os estimadores de máxima verossimilhança de δ, μ, σ^2, β_r, $r = 0, \ldots, k$, e γ, respectivamente.

Resíduos *deviance*

O resíduo *deviance* padronizado para o modelo de regressão binomial correlacionada aditivo estrutural normal é definido por

$$r_i^{spd} = \frac{\text{sinal }(y_i - n_i\hat{p}_i)\sqrt{2\ell_a(y_i; \mathcal{D}_{(i)}, \hat{\gamma}, \hat{\mu}, \hat{\sigma^2}) - 2\ell_a(\hat{\boldsymbol{\beta}}, \hat{\delta}; \mathcal{D}_{(i)}, \hat{\gamma}, \hat{\mu}, \hat{\sigma^2})}}{\sqrt{\hat{p}_i(1-\hat{p}_i)\{n_i + \hat{\rho}_i n_i(n_i-1)\}}}, \qquad i = 1, \ldots, m, \tag{6.17}$$

sendo

$\ell_a(y_i; \mathcal{D}_{(i)}, \hat{\gamma}, \hat{\mu}, \hat{\sigma^2}) =$

$$-\frac{1}{2}\log(2\pi\hat{\sigma}^2(1+b)) - \frac{1}{2\hat{\sigma}^2}\left(\frac{w_i^2}{b} + \hat{\mu}^2 - \left(\frac{(b\hat{\mu}+w_i)^2}{b(1+b)}\right)\right) + \log\left\{\binom{n_i}{y_i}\left(\frac{y_i}{n_i}\right)^{y_i}\right.$$

$$\left.\times\left(1-\frac{y_i}{n_i}\right)^{n_i-y_i}(1-h(v(\boldsymbol{r}_i),\hat{\gamma})) + \left(\frac{y_i}{n_i}\right)^{\frac{y_i}{n_i}}\left(1-\frac{y_i}{n_i}\right)^{\frac{n_i-y_i}{n_i}} h(v(\boldsymbol{r}_i),\hat{\gamma})I_{A_{2i}}(y_i)\right\},$$

a função de log-verossimilhança aproximada saturada, considerando para a estimação do parâmetro p_i a proporção y_i/n_i do i-ésimo cluster e os valores dos parâmetros γ, μ e σ^2 seus respectivos estimadores de máxima verossimilhança, $\hat{\gamma}$, $\hat{\mu}$ e $\hat{\sigma}^2$; $\ell_a(\hat{\boldsymbol{\beta}}, \hat{\delta}; \mathcal{D}_{(i)}, \hat{\gamma}, \hat{\mu}, \hat{\sigma^2})$ é a função de log-verossimilhança aproximada avaliada nos estimadores de máxima verossimilhança com $\hat{p}_i = g^{-1}(\sum_{r=0}^{k}\hat{\beta}_r x_{ir} + \hat{\delta}(b\hat{\mu}+w_i)/(1+b))$ e $\hat{\rho}_i = h(v(\boldsymbol{r}_i),\hat{\gamma})$, sendo $\hat{\delta}$, $\hat{\mu}$, $\hat{\sigma^2}$, $\hat{\beta}_r$, $r = 0,\ldots,k$, e $\hat{\gamma}$ os estimadores de máxima verossimilhança de δ, μ, σ^2, β_r, $r = 0,\ldots,k$, e γ, respectivamente; e $\mathcal{D}_{(i)} = (n_i, y_i, \boldsymbol{x}_i, w_i, \boldsymbol{r}_i)^\top$ é a informação disponível da i-ésima observação.

6.6.2 Diagnóstico de influencia global

Nesta seção, apresentamos duas métricas para avaliar a influência global das observações na estimativa dos parâmetros no modelo de regressão binomial correlacionada, a distância de Cook generalizada e a distância da verossimilhança (Zhu *et al.*, 2001).

Distância de Cook Generalizada

A distância de Cook generalizada, similar ao apresentado na Seção 3.4.2, é dada por

$$C_i = \left(\hat{\boldsymbol{\theta}}_{(-i)} - \hat{\boldsymbol{\theta}}\right)^\top J_a(\hat{\boldsymbol{\theta}})\left(\hat{\boldsymbol{\theta}}_{(-i)} - \hat{\boldsymbol{\theta}}\right)$$

em que $\hat{\boldsymbol{\theta}}_{(-i)}$ é o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ baseado em $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}^*)$, apresentada em (6.8), com a i-ésima observação $(n_i, y_i, \boldsymbol{x}_i, w_i, \boldsymbol{r}_i)^\top$ deletada e $J_a(\hat{\boldsymbol{\theta}})$ é a matriz de informação de Fisher observada dada em (6.13). Quando o número de clusters, m, é grande, Cook & Weisberg (1982) sugerem a seguinte aproximação para $\hat{\boldsymbol{\theta}}_{(-i)}$:

$$\hat{\boldsymbol{\theta}}_{(-i)} = \hat{\boldsymbol{\theta}} + J_a(\hat{\boldsymbol{\theta}})^{-1}U(\hat{\boldsymbol{\theta}}_{(-i)}), \tag{6.18}$$

em que

$$U(\hat{\boldsymbol{\theta}}_{(-i)}) = \left.\frac{\partial \ell_a(\boldsymbol{\theta}; \mathcal{D}_{(-i)})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{(-i)}}.$$

O vetor de escores, $U(\boldsymbol{\theta}_{(-i)})$, com a i-ésima observação deletada, tem dimensão $(k+5)$ e os termos são dados por:

$$\frac{\partial \ell_a(\boldsymbol{\theta}; \mathcal{D}_{(-i)})}{\partial \gamma} = \sum_{i=1}^{m-1}\left\{\frac{\partial h(v(\boldsymbol{r}_i),\gamma)}{\partial \gamma}\frac{[-A+B]}{[C+D]}\right\}$$

e

$$\frac{\partial \ell_a(\boldsymbol{\theta};\mathcal{D}_{(-i)})}{\partial \zeta_1} = \sum_{i=1}^{m-1}\Bigg\{ f_1(\zeta_1) + \frac{\partial g^{-1}(\eta_i^*)}{\partial \zeta_1}\Bigg[C\Bigg[y_i(g^{-1}(\eta_i^*))^{-1} - (n_i - y_i)(1-g^{-1}(\eta_i^*))^{-1}\Bigg] + D\Bigg[\frac{y_i}{n_i}(g^{-1}(\eta_i^*))^{-1} - \frac{(n_i-y_i)}{n_i}(1-g^{-1}(\eta_i^*))^{-1}\Bigg]\Bigg][C+D]^{-1}\Bigg\},$$

com ζ_1 assumindo os parâmetros δ, μ, σ^2 e β_r com $r, s = 0, \ldots, k$; $f_1(\zeta_1)$ conforme mostra a Tabela 6.1; A, B, C e D como descritos na Tabela 3.1. As derivadas presentes nestas funções são mostradas na Seção 6.4.

Usando a aproximação apresentada em (6.18), a distância de Cook generalizada, $C_{i_{app}}$, é reescrita como

$$C_{i_{app}} = U(\hat{\boldsymbol{\theta}}_{(-i)})^\top J_a(\hat{\boldsymbol{\theta}}) U(\hat{\boldsymbol{\theta}}_{(-i)}).$$

Distância da Verossimilhança

A distância da verossimilhança é dada por

$$LD_i = 2\left\{\ell_a(\hat{\boldsymbol{\theta}};\mathcal{D}) - \ell_a(\hat{\boldsymbol{\theta}}_{(-i)};\mathcal{D})\right\}, \tag{6.19}$$

na qual $\ell_a(\hat{\boldsymbol{\theta}};\mathcal{D})$ e $\ell_a(\hat{\boldsymbol{\theta}}_{(-i)};\mathcal{D})$ são a função de log-verossimilhança aproximada avaliada no estimador de máxima verossimilhança, $\hat{\boldsymbol{\theta}}$, e no estimador de máxima verossimilhança com a i-ésima observação $(n_i, y_i, \boldsymbol{x}_i, w_i, \boldsymbol{r}_i)^\top$ deletada, respectivamente.

A i-ésima observação é considerada como influente se o valor da distância de Cook generalizada ou o valor da distância da verossimilhança é grande. Estes valores podem ser comparados com os valores críticos da distribuição χ^2_{k+5}.

6.7 Critérios de seleção de modelo

O critério de informação Bayesiano (BIC) e o critério de informação de Akaike (AIC) podem ser utilizados para selecionar o melhor modelo de regressão binomial correlacionada. O BIC é calculado pela expressão

$$BIC = -2\ell_a(\hat{\boldsymbol{\theta}};\mathcal{D}) + (k+5)\log(m), \tag{6.20}$$

em que $\ell_a(\hat{\boldsymbol{\theta}};\mathcal{D})$ é o valor da função log-verossimilhança aproximada avaliado no estimador usual de máxima verossimilhança, $(k+5)$ é o número de parâmetros no modelo e m é o número de clusters. O AIC é calculado pela expressão

$$AIC = -2\ell_a(\hat{\boldsymbol{\theta}};\mathcal{D}) + (k+5). \tag{6.21}$$

6.8 Estudos de simulação

Nesta seção, consideramos um estudo de simulação com uma amostra para ilustrar o processo de estimação e o desempenho das medidas de diagnóstico propostas para o modelo de regressão binomial correlacionada aditivo estrutural normal. Um outro estudo com 300 amostras é realizado para analisar as propriedades frequentistas dos estimadores de máxima verossimilhança e seus intervalos de confiança.

Dados simulados

A geração da amostra envolve $m = 100$ clusters com as variáveis resposta, $Y_i|U_i$, $i = 1, \ldots, 100$, seguindo uma distribuição $BC(n_i, p_i, \rho_i)$, com n_i gerados de uma distribuição $\mathrm{U}_d(30, 50)$. O parâmetro da estrutura de correlação utilizado na simulação é $\gamma = 0,5$ e $v(\boldsymbol{r}_i)$ assumindo valores de uma distribuição U(0,2). Duas covariáveis são consideradas, x_{i1} e w_i, a covariável x_{i1} assumindo valores de uma distribuição N(0,1) e a covariável observada com erro $w_i = u_i + \epsilon_i$, com u_i assumindo valores de uma distribuição $N(1; 0,5)$ e ϵ_i assumindo valores de uma distribuição $N(0; 0,5)$, ou seja, consideramos $b = 1$, a variância do erro é igual a variância da covariável não observada. Os valores dos coeficientes de regressão são $\beta_0 = 1$, $\beta_1 = -1$ e $\delta = 1$. São considerados neste estudo a função de ligação logito e a estrutura de correlação exponencial.

Estimação

O modelo de regressão binomial correlacionada aditivo estrutural normal com função de ligação logito e estrutura de correlação exponencial é considerado na análise, com 3000 iterações para aproximação de Monte Carlo para construção dos intervalos de confiança assintóticos, $B = 100$ reamostras para construção dos intervalos de confiança *bootstrap* e $B = 5000$ valores em torno dos estimadores de máxima verossimilhança para construção dos intervalos de confiança perfilados.

Tabela 6.3: Estimativas de máxima verossimilhança e intervalos com confiança de 95% para os parâmetros γ, δ, μ, σ^2, β_0 e β_1.

Parâmetros	Valor real	EMV	IC 95% via $\mathcal{L}_a(\boldsymbol{\theta}; \mathcal{D}^*)$	IC 95% via *bootstrap*	IC 95% via perfilada
γ	0,50	0,595	(0,392 ; 0,797)	(0,463 ; 0,803)	(0,419 ; 0,821)
δ	1,00	1,032	(0,686 ; 1,378)	(0,455 ; 1,324)	(0,903 ; 1,168)
μ	1,00	1,047	(0,852 ; 1,242)	(0,885 ; 1,215)	(0,885 ; 1,212)
σ^2	0,50	0,495	(0,398 ; 0,592)	(0,366 ; 0,585)	(0,381 ; 0,662)
β_0	1,00	0,872	(0,471 ; 1,273)	(0,502 ; 1,391)	(0,723 ; 1,030)
β_1	-1,00	-1,017	(-1,197 ; -0,838)	(-1,214 ; -0,702)	(-1,176 ; -0,856)

As estimativas de máxima verossimilhança e os intervalos de confiança para os parâmetros do modelo ajustado são mostrados na Tabela 6.3. Note que, os intervalos de

confiança contêm o verdadeiro valor dos parâmetros. É importante ressaltar que os intervalos de confiança dos parâmetros da estrutura de correlação, γ, e do coeficiente de regressão da variável contaminada com erro, δ, não contêm o valor zero, corroborando com a necessidade de ajuste de um modelo binomial correlacionada aditivo estrutural normal.

Análise de resíduos

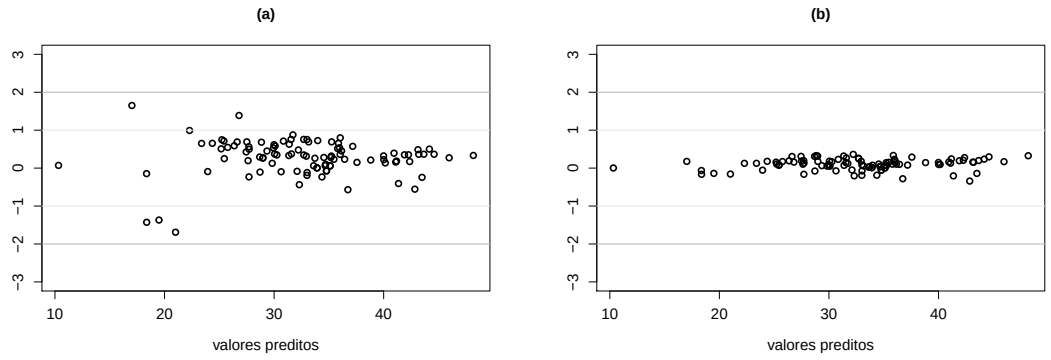

Figura 6.1: (a) resíduos padronizados versus valores preditos, (b) resíduos *deviance* padronizados versus valores preditos.

Os gráficos de resíduos contra valores preditos são apresentados, para cada tipo de resíduos, discutidos neste trabalho (ver Seção 6.6.1). Os gráficos dos resíduos contra valores preditos, apresentados nas Figuras 6.1a e 6.1b, verificam a coerência do modelo proposto e a ausência de *outliers*. Observe que ambos os resíduos padronizados não indicam observações aberrantes, principalmente os resíduos *deviance* padronizados, pois estão mais próximos de zero.

Diagnósticos de influência

Nessa seção, examinamos o desempenho das medidas de diagnósticos de influência, a distância de Cook generalizada e a distância da verossimilhança. Para isto, a observação 49 foi perturbada, criando, assim, uma observação influente no conjunto de dados. A perturbação foi feita na covariável w_i, na forma $w_{49} = w_{49} + 5sd(w)$, em que $sd(w)$ corresponde ao desvio padrão da covariável w_i. As duas combinações com a presença e ausência desta observação perturbada foram usadas para formar dois novos conjuntos de dados. Estes novos conjuntos de dados foram utilizados na obtenção de estimativas dos parâmetros do modelo.

A Tabela 6.4 apresenta os estimadores de máxima verossimilhança e a mudança relativa da estimativas dos parâmetros em relação aos dados originais. Observe os impactos ocorridos nos parâmetros δ e σ^2, que estão diretamente relacionados à covariável w.

Como mencionado na Seção 6.6.2, valores altos da distância de Cook generalizada e da distância da verossimilhança evidenciam a presença de pontos influentes no conjunto

Tabela 6.4: Estimador de máxima verossimilhança e mudança relativa em relação aos dados originais simulados.

	Casos perturbados		%
Parâmetros	{Nenhum}	Caso {49}	mudança
γ	0,595	0,577	-3,025
δ	1,032	0,573	-44,477
μ	1,047	1,097	4,776
σ^2	0,495	0,662	33,737
β_0	0,872	1,309	50,115
β_1	-1,017	-0,937	7,866

de dados. Uma ilustração do caso analisado na Tabela 6.4 é apresentado na Figura 6.2. A Figura 6.2a mostram as distâncias de Cook generalizada e a Figura 6.2b mostram as distâncias da verossimilhança para o conjunto de dados original. A Figura 6.2c mostram as distâncias de Cook generalizada e a Figura 6.2d mostram as distâncias da verossimilhança para o conjunto de dados considerando a perturbação na obervação 49. Observamos, nas Figuras 6.2c e 6.2d, valores sensivelmente mais elevados para a observação 49.

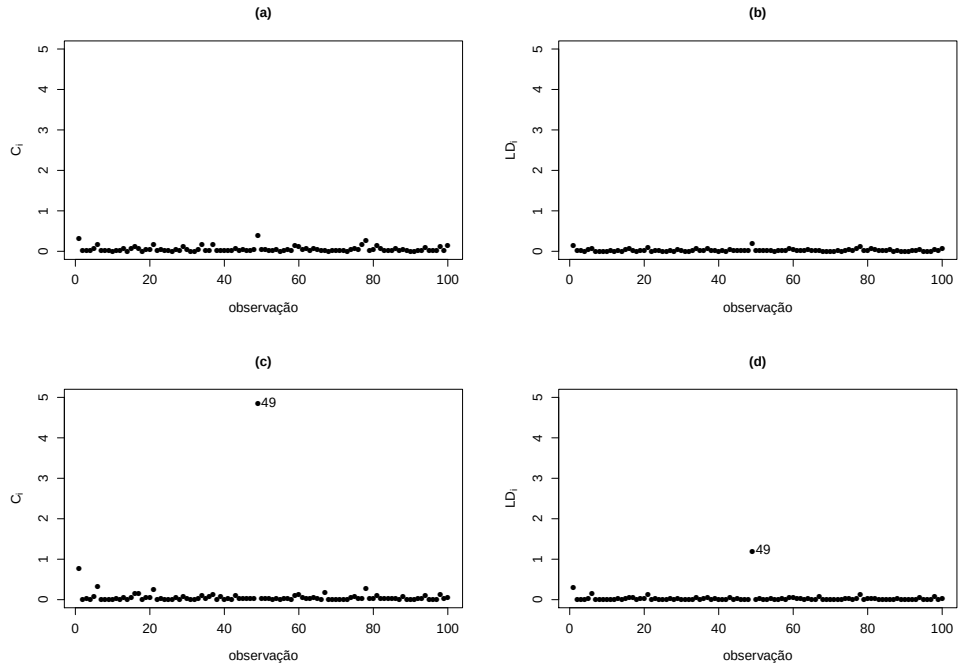

Figura 6.2: Dados originais simulados: (a) distâncias de Cook generalizada, (b) distâncias da verossimilhança; Dados com a observação 49 perturbada: (c) distâncias de Cook generalizada, (d) distâncias da verossimilhança.

Flutuação na variância do erro, σ^2, com diferentes valores de b

O modelo de regressão binomial correlacionada aditivo estrutural normal deve ser considerado ao inserir uma covariável com erro de medida na forma aditiva, como formulado no Capítulo 6, na modelagem da probabilidade de sucesso do cluster. Nesta seção, apresentamos para a particular amostra simulada, descrita na Seção 6.8, o impacto do aumento do erro de medida no estimador do coeficiente de regressão, δ, da variável contaminada. Para esta análise foram considerados quatro valores de σ^2, sendo estes 0,25, 0,50, 1,00 e 2,00, e quatro valores de b, sendo estes 0,10, 0,25, 0,50, 1,00. A análise é conduzida considerando o modelo correto, MRBCAEN, e o MRBC, proposto no Capítulo 2. Os resultados são apresentados na Tabela 6.5. Fica evidente nos casos analisados o vício na estimativa do parâmetro ao considerar o MRBC, apenas o modelo de regressão binomial correlacionada aditivo estrutural normal é capaz de captar o aumento do erro de medida.

Tabela 6.5: Estimador de máxima verossimilhança de δ via MRBCAEN e MRBC considerando diferentes valores para σ^2 e b.

$\text{Var}(\epsilon_i) = b\sigma^2$		Valor	MRBCAEN		MRBC	
b	σ^2	real	EMV	IC_a	EMV	IC_a
	0,25	1,00	0,902	(0,562 ; 1,243)	0,820	(0,507 ; 1,133)
	0,50	1,00	0,931	(0,687 ; 1,175)	0,838	(0,640 ; 1,037)
	1,00	1,00	0,937	(0,759 ; 1,115)	0,852	(0,691 ; 1,014)
0,10	2,00	1,00	0,931	(0,797 ; 1,064)	0,846	(0,724 ; 0,968)
	0,25	1,00	0,935	(0,570 ; 1,301)	0,748	(0,450 ; 1,047)
	0,50	1,00	0,954	(0,697 ; 1,211)	0,763	(0,553 ; 0,973)
	1,00	1,00	0,954	(0,771 ; 1,137)	0,763	(0,614 ; 0,912)
0,25	2,00	1,00	0,923	(0,792 ; 1,054)	0,738	(0,630 ; 0,847)
	0,25	1,00	0,977	(0,570 ; 1,384)	0,652	(0,372 ; 0,931)
	0,50	1,00	0,987	(0,704 ; 1,270)	0,658	(0,464 ; 0,853)
	1,00	1,00	0,984	(0,787 ; 1,181)	0,656	(0,520 ; 0,792)
0,50	2,00	1,00	0,930	(0,794 ; 1,067)	0,620	(0,524 ; 0,717)
	0,25	1,00	1,029	(0,543 ; 1,515)	0,515	(0,264 ; 0,765)
	0,50	1,00	1,032	(0,695 ; 1,369)	0,516	(0,342 ; 0,690)
	1,00	1,00	1,031	(0,799 ; 1,263)	0,515	(0,395 ; 0,635)
1,00	2,00	1,00	0,955	(0,798 ; 1,111)	0,477	(0,394 ; 0,560)

Com o intuito de apresentar uma análise mais detalhada em relação a propriedade frequentista dos intervalos assintóticos do estimador de máxima verossimilhança do MRBC, apresentados na Seção 3.2, na estimação do parâmetro δ, realizamos um estudo de simulação similar ao processo apresentado na Seção 3.6.1, no qual são utilizadas 300 amostras geradas de forma análoga a descrição apresentada na Seção 6.8, exceto por considerar diferentes valores para a variância do erro, diferentes proporções do erro de medida e diferentes estruturas de correlação. Os resultados da Tabela 6.6 mostram claramente que o

aumento da proporção do erro de medida (b) e do valor da variância do erro (σ^2) causam vício na estimativa do coeficiente de regressão correspondente a covariável medida com erro, δ, impactando nas estimativas das probabilidades de cobertura, evidenciando que o modelo usual tem dificuldade em estimar corretamente o coeficiente de regressão da variável medida com erro para valores da variância maior que 10% do valor da variância do processo sem erro.

Tabela 6.6: Estimativa da probabilidade de cobertura dos intervalos com confiança assintóticos de 95% (PC e PC_a) do estimador de máxima verossimilhança de δ via MRBC, com 3000 iterações para aproximação de Monte Carlo para construção dos intervalos de confiança assintóticos, para diferentes valores da variância do erro (σ^2), diferentes proporções de erro (b) e diferentes estruturas de correlação.

$\text{Var}(\epsilon_i) = b\sigma^2$		Exponencial		AR Contínua		Gaussiana	
b	σ^2	PC	PC_a	PC	PC_a	PC	PC_a
	0,25	0,9000	0,9000	0,8733	0,8733	0,8767	0,8767
	0,50	0,8400	0,8400	0,8067	0,8067	0,8467	0,8467
	1,00	0,7000	0,7000	0,6567	0,6567	0,7567	0,7567
0,10	2,00	0,4833	0,4833	0,5200	0,5200	0,6667	0,6667
	0,25	0,6467	0,6467	0,6467	0,6500	0,7433	0,7433
	0,50	0,4667	0,4667	0,4333	0,4333	0,6100	0,6100
	1,00	0,2567	0,2567	0,1767	0,1767	0,3867	0,3867
0,25	2,00	0,0900	0,0867	0,0800	0,0800	0,2200	0,2200
	0,25	0,3500	0,3467	0,2367	0,2367	0,5000	0,5000
	0,50	0,1433	0,1433	0,0633	0,0633	0,2300	0,2300
	1,00	0,0400	0,0400	0,0167	0,0167	0,1100	0,1100
0,50	2,00	0,0133	0,0133	0,0067	0,0067	0,0233	0,0233
	0,25	0,0567	0,0567	0,0233	0,0233	0,1400	0,1367
	0,50	0,0067	0,0067	0,0033	0,0033	0,0367	0,0367
	1,00	0,0033	0,0033	0,0000	0,0000	0,0100	0,0100
1,00	2,00	0,0000	0,0000	0,0000	0,0000	0,0067	0,0067

6.8.1 Propriedade frequentista dos estimadores de máxima verossimilhança

As estimativas das probabilidades de cobertura dos intervalos assintóticos, dos intervalos *bootstrap* e dos intervalos perfilados foram construídas para o nível de confiança fixado em 95%. A determinação das estimativas das probabilidades de cobertura foram obtidas calculando a proporção de intervalos que continham o verdadeiro valor de cada um dos parâmetros fixados na geração dos dados, baseada em um processo de simulação similar ao descrito na Seção 6.8, exceto pelas estrutura de correlação utilizadas, variância do erro e proporção de erro em relação a variância do processo não contaminado. O cálculo da proporção está baseado na simulação de 300 amostras com $m = 500$ clusters, sendo que

para obtenção dos intervalos assintóticos foram consideradas 500 iterações para aproximação de Monte Carlo, para os intervalos perfilados foram considerados 500 valores em torno dos estimadores de máxima verossimilhança e para os intervalos *bootstrap* foram consideradas 100 reamostragens da amostra original simulada. Nesta parte da análise utilizamos três estruturas de correlação, exponencial, Gaussiana e AR contínua, três valores diferentes para a variância do processo $\sigma^2 = 0,25$, $\sigma^2 = 0,5$ e $\sigma^2 = 1$, e dois valores para a proporção da variância do erro de medida em relação a variância do processo $b = 0,10$ e $b = 0,25$.

Tabela 6.7: Estimativa da probabilidade de cobertura dos intervalos com confiança de 95% (PC_a, PC_b, PC_p) considerando $b = 0,10$, para diferentes valores de σ^2 e diferentes estruturas de correlação.

	Estrutura de correlação exponencial								
	$\sigma^2 = 0,25$			$\sigma^2 = 0,50$			$\sigma^2 = 1,00$		
	PC_a	PC_b	PC_p	PC_a	PC_b	PC_p	PC_a	PC_b	PC_p
γ	0,91	0,94	0,95	0,93	0,90	0,95	0,91	0,94	0,95
δ	0,88	0,92	0,58	0,80	0,89	0,63	0,64	0,82	0,60
μ	0,72	0,70	0,74	0,49	0,47	0,68	0,10	0,09	0,25
σ^2	0,94	0,94	0,96	0,81	0,73	0,87	0,38	0,37	0,48
β_0	0,86	0,91	0,51	0,79	0,88	0,55	0,76	0,85	0,59
β_1	0,90	0,93	0,80	0,83	0,90	0,70	0,70	0,80	0,56
	Estrutura de correlação Gaussiana								
	$\sigma^2 = 0,25$			$\sigma^2 = 0,50$			$\sigma^2 = 1,00$		
	PC_a	PC_b	PC_p	PC_a	PC_b	PC_p	PC_a	PC_b	PC_p
γ	0,95	0,91	0,94	0,92	0,93	0,91	0,91	0,93	0,87
δ	0,82	0,88	0,50	0,80	0,89	0,60	0,65	0,89	0,59
μ	0,67	0,63	0,71	0,41	0,46	0,65	0,08	0,10	0,21
σ^2	0,94	0,92	0,93	0,79	0,84	0,89	0,46	0,35	0,54
β_0	0,83	0,87	0,43	0,79	0,85	0,50	0,72	0,89	0,56
β_1	0,91	0,90	0,76	0,88	0,92	0,74	0,63	0,89	0,48
	Estrutura de correlação AR contínua								
	$\sigma^2 = 0,25$			$\sigma^2 = 0,50$			$\sigma^2 = 1,00$		
	PC_a	PC_b	PC_p	PC_a	PC_b	PC_p	PC_a	PC_b	PC_p
γ	0,93	0,92	0,92	0,91	0,95	0,91	0,92	0,91	0,91
δ	0,88	0,93	0,58	0,84	0,88	0,69	0,79	0,78	0,74
μ	0,71	0,70	0,73	0,48	0,49	0,68	0,09	0,07	0,19
σ^2	0,95	0,95	0,91	0,84	0,81	0,90	0,39	0,38	0,50
β_0	0,88	0,90	0,58	0,88	0,85	0,62	0,81	0,82	0,68
β_1	0,93	0,94	0,84	0,89	0,93	0,81	0,81	0,84	0,68

Probabilidade de cobertura nominal de 95%.

Os resultados apresentados nas Tabelas 6.7 e 6.8 indicam que as estimativas das probabilidades de cobertura para os intervalos utilizados neste trabalho estão mais próximos da cobertura nominal ao considerar a variância menor, ou seja, $\sigma^2 = 0,25$, exceto para o parâmetro μ, para todas as estruturas de correlação. Esta proximidade da cobertura nominal é sensivelmente diminuída para os parâmetros presentes na função de ligação, β_0, β_1, δ, μ e σ^2 com o aumento da variância do processo ($\sigma^2 = 0,50$ e $\sigma^2 = 1,00$) e do erro de medida consequentemente, pois a variância do erro é dada por $b\sigma^2$, conforme já discutido anteriormente.

Tabela 6.8: Estimativa da probabilidade de cobertura dos intervalos com confiança de 95% (PC_a, PC_b, PC_p) considerando $b = 0,25$, para diferentes valores de σ^2 e diferentes estruturas de correlação.

	Estrutura de correlação exponencial								
	$\sigma^2 = 0,25$			$\sigma^2 = 0,50$			$\sigma^2 = 1,00$		
	PC_a	PC_b	PC_p	PC_a	PC_b	PC_p	PC_a	PC_b	PC_p
γ	0,95	0,91	0,96	0,89	0,89	0,95	0,89	0,87	0,93
δ	0,85	0,89	0,56	0,79	0,89	0,57	0,56	0,78	0,50
μ	0,72	0,69	0,89	0,50	0,55	0,62	0,12	0,15	0,22
σ^2	0,92	0,93	0,91	0,89	0,85	0,93	0,49	0,52	0,58
β_0	0,84	0,89	0,45	0,86	0,91	0,59	0,83	0,91	0,58
β_1	0,88	0,91	0,76	0,82	0,89	0,65	0,61	0,79	0,43
	Estrutura de correlação Gaussiana								
	$\sigma^2 = 0,25$			$\sigma^2 = 0,50$			$\sigma^2 = 1,00$		
	PC_a	PC_b	PC_p	PC_a	PC_b	PC_p	PC_a	PC_b	PC_p
γ	0,95	0,93	0,94	0,93	0,90	0,92	0,90	0,93	0,89
δ	0,88	0,88	0,51	0,77	0,90	0,53	0,51	0,85	0,46
μ	0,66	0,71	0,79	0,52	0,52	0,70	0,12	0,12	0,25
σ^2	0,93	0,93	0,95	0,85	0,80	0,89	0,53	0,46	0,61
β_0	0,88	0,87	0,44	0,81	0,90	0,51	0,77	0,84	0,58
β_1	0,90	0,95	0,78	0,80	0,93	0,64	0,58	0,83	0,42
	Estrutura de correlação AR contínua								
	$\sigma^2 = 0,25$			$\sigma^2 = 0,50$			$\sigma^2 = 1,00$		
	PC_a	PC_b	PC_p	PC_a	PC_b	PC_p	PC_a	PC_b	PC_p
γ	0,95	0,92	0,95	0,96	0,89	0,93	0,91	0,88	0,92
δ	0,91	0,91	0,54	0,87	0,87	0,69	0,71	0,75	0,65
μ	0,71	0,66	0,87	0,56	0,47	0,70	0,16	0,12	0,24
σ^2	0,97	0,93	0,96	0,85	0,84	0,90	0,54	0,47	0,63
β_0	0,89	0,91	0,56	0,90	0,88	0,61	0,82	0,86	0,69
β_1	0,88	0,93	0,78	0,84	0,88	0,72	0,75	0,70	0,59

Probabilidade de cobertura nominal de 95%.

Tabela 6.9: Média dos erros quadráticos médios (EQM_a) e média dos vícios dos estimadores de máxima verossimilhança considerando $b = 0,10$, para diferentes valores da variância do erro e diferentes estruturas de correlação.

	Estrutura de correlação exponencial					
	$\sigma^2 = 0,25$		$\sigma^2 = 0,50$		$\sigma^2 = 1,00$	
	EQM_a	Vício	EQM_a	Vício	EQM_a	Vício
γ	0,0570	-0,0145	0,0562	-0,0153	0,0560	-0,0197
δ	0,1577	-0,0787	0,1145	-0,0856	0,0885	-0,0941
μ	0,0249	-0,0328	0,0375	-0,0652	0,0662	-0,1428
σ^2	0,0158	-0,0038	0,0316	-0,0279	0,0737	-0,1240
β_0	0,1576	0,0915	0,1129	0,0843	0,0862	0,0765
β_1	0,0775	0,0176	0,0771	0,0389	0,0839	0,0845
	Estrutura de correlação Gaussiana					
	$\sigma^2 = 0,25$		$\sigma^2 = 0,50$		$\sigma^2 = 1,00$	
	EQM_a	Vício	EQM_a	Vício	EQM_a	Vício
γ	0,0347	0,0116	0,0348	0,0117	0,0346	0,0185
δ	0,1408	-0,0802	0,0998	-0,0740	0,0780	-0,0922
μ	0,0251	-0,0351	0,0377	-0,0672	0,0674	-0,1467
σ^2	0,0159	-0,0028	0,0316	-0,0295	0,0725	-0,1157
β_0	0,1393	0,0829	0,0967	0,0706	0,0764	0,0723
β_1	0,0678	0,0253	0,0676	0,0414	0,0753	0,0889
	Estrutura de correlação AR contínua					
	$\sigma^2 = 0,25$		$\sigma^2 = 0,50$		$\sigma^2 = 1,00$	
	EQM_a	Vício	EQM_a	Vício	EQM_a	Vício
γ	0.0467	-0,0059	0.0469	-0,0124	0,0463	-0,0066
δ	0.2353	-0,0687	0.1708	-0,0909	0,1230	-0,0896
μ	0.0250	-0,0340	0.0375	-0,0649	0,0670	-0,1455
σ^2	0.0160	-0,0020	0.0314	-0,0279	0,0739	-0,1225
β_0	0.2317	0,0892	0.1662	0,0983	0,1219	0,0758
β_1	0.1142	0,0167	0.1164	0,0401	0,1158	0,0817

Nas Tabelas 6.9 e 6.10 apresentamos a média dos erros quadráticos médios via variância estimada assintoticamente (EQM_a) e média dos vícios dos estimadores de máxima verossimilhança, considerando a função de ligação logito, três estruturas de correlação, três situações diferentes para a variância do erro de medida e para os valores da proporção do erro $b = 0,10$ e $b = 0,25$. Para determinar estes valores foram geradas 300 amostras para cada cenário. Em cada amostra foi calculado o estimador de máxima verossimilhança, o EQM_a e o vício. Posteriormente, obtivemos a média destes 300 EQM_as e vícios, para cada parâmetro. Os resultados mostram que os vícios médios e os erros quadráticos médios médios para os parâmetros envolvidos na função de ligação, β_0, β_1, δ, μ e σ^2 apresentam-se maiores que o parâmetro γ, presente na estrutura de correlação. Isso se deve a influência direta do erro de medida nos estimadores de máxima verossimilhança destes parâmetros.

Tabela 6.10: Média dos erros quadráticos médios (EQM_a) e média dos vícios dos estimadores de máxima verossimilhança considerando $b = 0,25$, para diferentes valores da variância do erro e diferentes estruturas de correlação.

	Estrutura de correlação exponencial					
	$\sigma^2 = 0,25$		$\sigma^2 = 0,50$		$\sigma^2 = 1,00$	
	EQM_a	Vício	EQM_a	Vício	EQM_a	Vício
γ	0,0565	-0,0129	0,0561	-0,0245	0,0554	-0,0283
δ	0,1701	-0,0916	0,1214	-0,0825	0,0949	-0,1156
μ	0,0266	-0,0334	0,0404	-0,0690	0,0707	-0,1460
σ^2	0,0159	-0,0024	0,0315	-0,0211	0,0716	-0,1101
β_0	0,1689	0,0980	0,1130	0,0664	0,0836	0,0594
β_1	0,0766	0,0214	0,0801	0,0517	0,0864	0,1054
	Estrutura de correlação Gaussiana					
	$\sigma^2 = 0,25$		$\sigma^2 = 0,50$		$\sigma^2 = 1,00$	
	EQM_a	Vício	EQM_a	Vício	EQM_a	Vício
γ	0,0346	0,0080	0,0346	0,0124	0,0345	0,0212
δ	0,1443	-0,0732	0,1087	-0,0861	0,0865	-0,1174
μ	0,0270	-0,0393	0,0398	-0,0654	0,0709	-0,1459
σ^2	0,0159	-0,0035	0,0317	-0,0244	0,0707	-0,1051
β_0	0,1422	0,0757	0,1026	0,0673	0,0754	0,0563
β_1	0,0677	0,0189	0,0695	0,0524	0,0783	0,1015
	Estrutura de correlação AR contínua					
	$\sigma^2 = 0,25$		$\sigma^2 = 0,50$		$\sigma^2 = 1,00$	
	EQM_a	Vício	EQM_a	Vício	EQM_a	Vício
γ	0,0461	-0,0047	0,0462	-0,0090	0,0462	-0,0167
δ	0,2409	-0,0984	0,1720	-0,0730	0,1389	-0,1137
μ	0,0268	-0,0347	0,0398	-0,0637	0,0697	-0,1415
σ^2	0,0160	-0,0005	0,0314	-0,0245	0,0707	-0,1054
β_0	0,2362	0,0991	0,1702	0,0606	0,1279	0,0733
β_1	0,1163	0,0303	0,1162	0,0600	0,1229	0,1034

6.9 Análise de dados reais

Nesta seção iremos ajustar o conjunto de dados de planos de saúde, apresentado na Seção 5.5.1, considerando o modelo de regressão binomial correlacionada aditivo estrutural normal com função de ligação logito e estrutura de correlação AR contínua, supondo que a covariável custo médio padronizado dos exames, x_{i2} foi medida com erro.

As estimativas de máxima verossimilhança e intervalos de confiança para os parâmetros do modelo de regressão binomial correlacionada aditivo estrutural normal com função de ligação logito e estrutura de correlação AR contínua são mostradas na Tabela 6.11, para o intervalo de confiança baseado na função de verossimilhança com dados aumentados foram consideradas 3000 iterações para a aproximação de Monte Carlo, para o intervalo

Tabela 6.11: Estimativas de máxima verossimilhança e intervalos de confiança de 95% para γ, δ, μ, σ^2, β_0 e β_1.

Parâmetros	EMV	IC 95% via $\mathcal{L}_a(\boldsymbol{\theta}; \mathcal{D}^*)$	IC 95% via *bootstrap*	IC 95% via perfilada
γ	0,223	(0,119 ; 0,327)	(0,117 ; 0,334)	(0,136 ; 0,334)
δ	0,659	(0,330 ; 0,988)	(0,235 ; 0,991)	(0,354 ; 0,996)
μ	0,000	(-0,155 ; 0,154)	(-0,165 ; 0,180)	(-0,144 ; 0,176)
σ^2	0,497	(0,388 ; 0,606)	(0,348 ; 0,649)	(0,407 ; 0,638)
β_0	-3,821	(-4,408 ; -3,235)	(-4,535 ; -2,983)	(-3,966 ; -3,645)
β_1	0,210	(0,033 ; 0,387)	(-0,060 ; 0,410)	(0,194 ; 0,258)

de confiança *bootstrap* foram considerados 100 reamostras, para obter os intervalos de confiança perfilados foram considerados 5000 valores em torno do estimador de máxima verossimilhança e adotamos $b = 1$. É importante observar que os intervalos de confiança para o parâmetro da estrutura de correlação, γ, e coeficiente de regressão, δ da covariável contaminada não contêm o valor zero, confirmando a necessidade do ajuste de um modelo binomial correlacionada aditivo estrutural normal.

Os resíduos padronizados baseado nos valores preditos e os resíduos *deviance* padronizados são apresentados na Figura 6.3a e na Figura 6.3b. A especificação do modelo e a presença de *outliers* são observados ao examinar os resíduos contra os valores preditos. Ambos os gráficos indicam uma boa especificação do modelo.

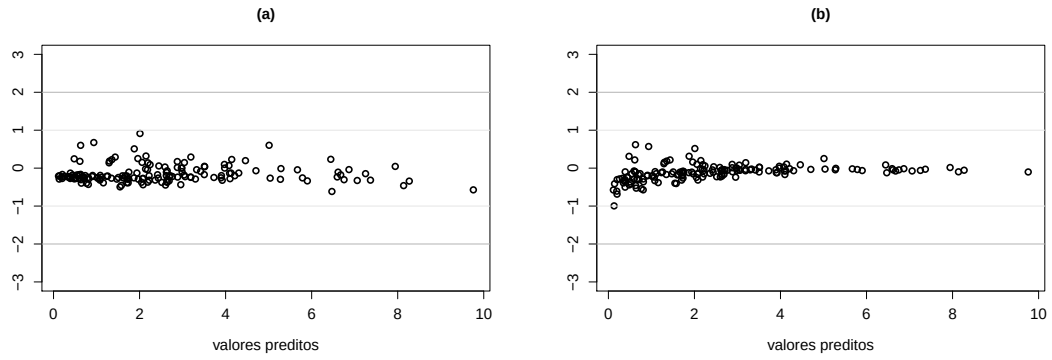

Figura 6.3: (a) resíduos padronizados via valores preditos versus valores preditos, (b) resíduos *deviance* padronizados versus valores preditos.

Para identificar observações influentes no conjunto de dados, as distâncias de Cook generalizada e distâncias da verossimilhança foram obtidas e são apresentadas na Figura 6.4. Os valores obtidos indicam que o maior valor da distância de Cook e da distância da verossimilhança são 1,045 e 0,475, respectivamente, para o caso 85.

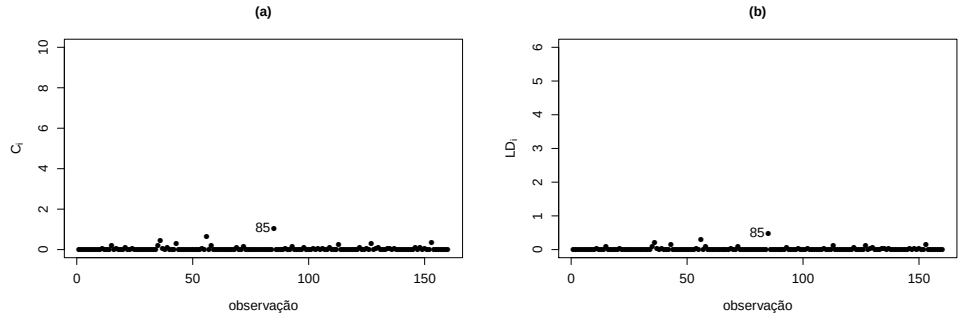

Figura 6.4: (a) distâncias de Cook generalizada; (b) distâncias da verossimilhança.

Capítulo 7

Considerações finais e propostas futuras

Considerações finais

Para modelar dados cujas respostas são frequências de eventos que ocorreram em clusters independentes com presença de covariáveis, geralmente utilizamos o modelo de regressão binomial usual. Porém, uma suposição relevante no ajuste do modelo de regressão binomial usual é a independência entre os ensaios de Bernoulli. Para situações em que a independência não é admissível, como presentes nos conjuntos de dados reais de inadimplência e de planos de saúde, propomos o modelo de regressão binomial correlacionada. Esta estrutura de regressão, além de modelar a probabilidade de sucesso de um evento de interesse para um determinado cluster por meio de covariáveis, permite inserir uma estrutura de correlação para modelar a relação de dependência existente entre os eventos de Bernoulli dentro dos clusters.

Neste trabalho, desenvolvemos o modelo de regressão binomial correlacionada e usamos abordagens clássica e Bayesiana no processo de ajuste do modelo. Propomos uma análise de diagnóstico envolvendo resíduos e medidas de influência. Estendemos o modelo de regressão binomial correlacionada para o caso em que uma covariável medida com erro é inserida na modelagem. O modelo de regressão binomial correlacionada aditivo estrutural normal é desenvolvido usando a função de ligação logito e três estruturas de correlação. O procedimento clássico foi utilizado no processo de ajuste do modelo.

Propostas futuras

Para o modelo de regressão binomial correlacionada, pretendemos desenvolver uma medida de detecção de multicolinearidade e um método para seleção de variáveis. A presença de relação entre as covariáveis dentro de um mesmo cluster é natural, a preocupação reside em detectar se está dependência afeta os coeficientes de regressão. Para isso temos interesse em apresentar uma medida de detecção de multicolinearidade Bayesiana baseada em Leamer (1973). O problema de multicolinearidade em modelos Bayesianos com incerteza a *priori* consiste em identificar se a colinearidade das covariáveis influenciam na média da distribuição a *posteriori* dos parâmetros. No desenvolvimento do modelo de regressão binomial correlacionada consideramos um número pequeno de covariáveis para

ajustar o modelo. Em uma situação prática, porém, podemos nos deparar com um número de covariáveis extremamente grande fazendo necessário determinar uma metodologia específica para seleção de variáveis para o modelo de regressão binomial correlacionada.

Um particular cenário foi desenvolvido para o modelo de regressão binomial correlacionada aditivo estrutural normal. Outras propostas relacionadas a esta modelagem envolvem considerar outras distribuições, além da distribuição normal, para a covariável não observada e para o erro, como, por exemplo, distribuições assimétricas ou distribuições que assumem apenas valores positivos. Uma abordagem para outras formas do erro, como a multiplicativa pode ser considerada. Todo o desenvolvimento do modelo aborda apenas uma função de ligação entre as covariáveis dos clusters e a probabilidade de sucesso, a ligação logito, todavia outras funções, como a complementar log-log, log-log e probito, podem ser utilizadas. É interessante apresentar uma abordagem Bayesiana para todas as situações, mencionadas inclusive para a apresentada nesta tese.

Referências

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

Albert, J. H. & Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika*, **82**(4), 747–759.

Altham, P. M. E. (1978). Two generalizations of the binomial distribution. *Journal of the Royal Statistical Society. Series C*, **27**(2), 162–167.

Bahadur, R. R. (1961). *A representation of the joint distribution of responses to n dichotomous items*. Studies Item Analysis and Prediction. Solomon, Stanford University Press.

Bernardo, J. M. & Smith, A. F. M. (2000). *Bayesian Theory*. Wiley, Chichester.

Chen, M. H. & Shao, Q. M. (1999). Monte carlo estimation of bayesian credible and hpd intervals. *Journal of Computational and Graphical Satatistics*, **8**(1), 69–92.

Chen, M. H., Shao, Q. M. & Ibrahim, J. (2000). *Monte Carlo Methods in Bayesian Computation*. Statistics. Springer, New York.

Cho, H., Ibrahim, J. G., Sinha, D. & Zhu, H. (2009). Bayesian case in influence diagnostics for survival models. *Biometrics*, **65**(1), 116–124.

Cook, R. & Weisberg, S. (1982). *Residuals and influence in regression*. Monographs on statistics and applied probability. Chapman and Hall, London.

Cox, D. & Hinkley, D. (1979). *Theoretical Statistics*. Chapman and Hall.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley-Interscience, New York.

Diebolt, J. & Robert, C. P. (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B*, **56**(2), 363–375.

Diniz, C. A. R., Tutia, M. H. & Leite, J. G. (2010). Bayesian analysis of a correlated binomial model. *Brazilian Journal of Probability and Statistics*, **24**(1), 68–77.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**(1), 1–26.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, **81**(395), 709–721.

Fu, J. & Sproule, R. (1995). A generalization of the binomial distribution. *Communications in Statistics - Theory and Methods*, **24**(10), 2645–2658.

Gelman, A., Carlin, J., Stern, H. & Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman and Hall/CRC Texts in Statistical Science, London, second edition.

Gupta, R. C. & Tao, H. (2010). A generalized correlated binomial distribution with application in multiple testing problems. *Metrika*, **71**(1), 59–77.

Horsnell, G. (1957). Economical acceptance sampling schemes. *Journal of the Royal Statistical Society. Series A*, **120**(2), 148–201.

Jennrich, R. I. & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, **42**(4), 805–820.

Kahn, M. J. & Raftery, A. E. (1996). Discharge rates of medicare stroke patients to skilled nursing facilities: Bayesian logistic regression with unobserved heterogeneity. *Journal of the American Statistical Association*, **91**(433), 29–41.

Kalbfleisch, J. G. (1985). *Probability and Statistical Inference*, volume 1 of *Springer texts in statistics*. Springer-Verlag, New York, second edition.

Kolev, N. & Paiva, D. (2008). Random sums of exchangeable variables and actuarial applications. *Insurance: Mathematics and Economics*, **42**(1), 147–153.

Kupper, L. L. & Haseman, J. K. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics*, **34**(1), 69–76.

Leamer, E. E. (1973). Multicollinearity: A bayesian interpretation. *The Review of Economics and Statistics*, **55**(3), 371–380.

Lindsey, J. K. (1995). *Modelling frequency and count data*. Oxford Statistical Science. Oxford University.

Lindsey, J. K. & Altham, P. M. E. (1998). Analysis of the human sex ratio by using overdispersion models. *Journal of the Royal Statistical Society. Series C*, **1**(47), 149–157.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B*, **44**(2), 226–233.

Luceño, A. (1995). A family of partially correlated poisson models for overdispersion. *Computational Statistics and Data Analysis*, **20**(5), 511–520.

Luceño, A. & Ceballos, F. (1995). Describing extra-binomial variation with partially correlated models. *Communications in Statistics - Theory and Methods*, **24**(6), 1637–1653.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London, second edition.

Ng, T.-H. (1989). A new class of modified binomial distributions with applications to certain toxicological experiments. *Communications in Statistics - Theory and Methods*, **18**(9), 3477–3492.

Ochi, Y. & Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika*, **71**(3), 531–543.

Paul, S. R. (1985). A three-parameter generalization of the binomial distribution. *Communications in Statistics - Theory and Methods*, **14**(6), 1497–1506.

Paul, S. R. (1987). On the beta-correlated binomial (bcb) distribution - a three-parameter generalization of the binomial distribution. *Communications in Statistics - Theory and Methods*, **16**(5), 1473–1478.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford Science Publications. Clarendon Press.

Peng, F. & Dey, D. (1995). Bayesian analysis of outlier problems using divergence measures. *The Canadian Journal of Statistics*, **23**(2), 199–213.

Prentice, R. L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association*, **81**(394), 321–327.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**(4), 1033–1048.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Robert, C. & Casella, G. (2004). *Monte Carlo statistical methods*. Springer texts in statistics. Springer.

Russell, D. W. (1996). Heterogeneous variance: Covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*, **1**(2), 205–230.

Salinas, D. & Kolev, N. (2003). *Soma de variáveis aleatórias equicorrelacionadas e aplicações em análise de risco e séries temporais discretas*. Instituto de Matemática e Estatística da Universidade de São Paulo.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**(2), 461–464.

She, Y. & Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, **106**(494), 626–639.

Sherman, M. (2011). *Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties*. Wiley Series in Probability and Statistics. John Wiley and Sons.

Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society. Series B*, **10**(2), 257–261.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B*, **64**(4), 583–639.

Tanner, M. A. (1996). *Tools For Statistical Inference: Methods For The Exploration Of Posterior Distributions And Likelihood Functions*. Springer Series in Statistics. Springer, New York.

Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**(398), 528–540.

Titterington, D. M., Smith, A. F. M. & Makov, U. (1985). *Statistical analysis of finite mixture distributions*. Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley.

Tsai, C., Hsueh, H. & Chen, J. J. (2003). Estimation of false discovery rates in multiple testing: Application to gene microarray data. *Biometrics*, **59**(4), 1071–1081.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Journal of the Royal Statistical Society. Series C*, **31**(2), 144–148.

Yu, C. & Zelterman, D. (2002). Sums of dependent bernoulli random variables and disease clustering. *Statistics and Probability Letters*, **57**(4), 363–373.

Zhu, H., Lee, S.-Y., Wei, B.-C. & Zhou, J. (2001). Case-deletion measures for models with incomplete data. *Biometrika*, **88**(3), 727–737.

Zimmerman, D. L. & Harville, D. A. (1991). A random field approach to the analysis of field-plot experiments and other spatial experiments. *Biometrics*, **47**(1), 223–239.

Apêndice A

Condições de regularidade

Condições de regularidade para utilizar a distribuição assintótica dos estimadores de máxima verossimilhança:

a) A distribuição de probabilidade do modelo binomial correlacionada e $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$ devem ser duas vezes diferenciável em relação a $\boldsymbol{\theta}$.

b) $\sum_{y_1=1}^{n_1} \dots \sum_{y_m=1}^{n_m} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) = 1$.

c) Os limites das somas em b) não devem depender dos parâmetros.

d) Diferenciação sob o sinal de integração é permitida.

Apêndice B

Método de Laplace

Bernardo & Smith (2000) discutem que o método de Laplace é um algoritmo de aproximação numérica atraente. Aproximar uma integral utilizando o método de Laplace consiste em utilizar uma expansão de Taylor de segunda ordem no integrando (Bernardo & Smith, 2000). Para a aplicação do método de Laplace (Tanner, 1996) são necessárias algumas suposições. Sejam N um escalar ($N \to \infty$) e $-h(u_i)$ uma função suave, duas vezes diferenciável, que possui apenas um ponto de máximo $\hat{u}_i$ no intervalo de integração $[a, b]$, com $-\infty \leq a < b \leq +\infty$.

Pelo método de Laplace a integral

$$I(N) = \int_a^b \phi(u_i) e^{-Nh(u_i)} du_i,$$

com $\phi(u_i)$ uma função de u_i, é aproximada pela função

$$\hat{I}(N) \approx \phi(\hat{u}_i) \sqrt{\frac{2\pi}{N}} \; |\beta|^{\frac{1}{2}} \; e^{-Nh(\hat{u}_i)}, \text{ com } \beta = \left(\frac{d^2 h(u_i)}{du_i^2} \right)^{-1}.$$

Na integral $I(N)$, se N é grande, a contribuição do integrando está fundamentalmente em torno de $\hat{u}_i$. Então, o aproxima por uma expansão de Taylor de segunda ordem em torno de $\hat{u}_i$. Neste caso, Tanner (1996) também menciona que $I(N) = \hat{I}(N) + \{1 + O(1/N)\}$.

Apêndice C

Programas para o MRBC: Abordagem Clássica

Neste capítulo, apresentamos os principais programas computacionais utilizados no MRBC via abordagem clássica. As funções foram implementadas em R (R Development Core Team, 2011) para obter os estimadores em máxima verossimilhança, os intervalos de confiança, os resíduos padronizados, as métricas para detecção de observações influentes e os critérios de seleção de modelo, discutidos no Capítulo 3.

C.1 Estimadores de máxima verossimilhança

Função para obter os estimadores de máxima verossimilhança, $\hat{\boldsymbol{\theta}}$, do vetor de parâmetros $\boldsymbol{\theta}$, usando o algoritmo EM, conforme descrito na Seção 3.1.

```
EMV.MRBC <- function(dados, fun.p, fun.rho, chutes, erro = 1e-08){

BETAS = as.matrix(chutes)
y <- dados[,1]
n <- dados[,2]
v <- dados[,3]
X <- dados[,4:ncol(dados)]

vero <- function(beta){
preditor = X%*%(as.matrix(beta[-1]))
if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }
if(fun.rho==1) { rho = exp(-exp(beta[1])*v) }
if(fun.rho==2) { rho = (exp(beta[1])/(1+exp(beta[1])))^v }
if(fun.rho==3) { rho = exp(-(exp(beta[1])*v)^2) }
ll = sum ( log(p^((y/n)*(z+n-n*z))) + log((1-p)^((n-y)*((z/n)+1-z)))
return(ll)
}
```

```
dif <- 0.1

while(dif > erro) {

preditor = X%*%BETAS[-1]
if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }
if(fun.rho==1) { rho = exp(-exp(BETAS[1])*v) }
if(fun.rho==2) { rho = (exp(BETAS[1])/(1+exp(BETAS[1])))^v }
if(fun.rho==3) { rho = exp(-(exp(BETAS[1])*v)^2) }

v1 <- rho*p^(y/n)*(1-p)^((n-y)/n)
v2 <- (1-rho)*choose(n,y)*p^y*(1-p)^(n-y)
z <-  v1 /(v1+v2)
z[!((y==0)|(y==n))]=0

res <- optim(BETAS, vero, method="BFGS", hessian=T, control=list(fnscale=-1))

if(fun.rho==1) { dif.gama <- abs(exp(BETAS[1]) - exp(res$par[1])) }
if(fun.rho==2) { dif.gama <- abs(exp(BETAS[1])/
                          (1+exp(BETAS[1])) - exp(res$par[1])/(1+exp(res$par[1]))) }
if(fun.rho==3) { dif.gama <- abs(exp(BETAS[1]) - exp(res$par[1])) }
dif.betas <- abs(BETAS[-1] - res$par[-1])

dif <- max(dif.gama, dif.betas)
BETAS <- res$par

if(fun.rho==1) { EMV <- c(exp(BETAS[1]), BETAS[-1]) }
if(fun.rho==2) { EMV <- c(exp(BETAS[1])/(1+exp(BETAS[1])), BETAS[-1]) }
if(fun.rho==3) { EMV <- c(exp(BETAS[1]), BETAS[-1]) }

EMV <- matrix(EMV, ncol=(ncol(X)+1))
colnames(EMV) <- c("gama",  replicate(ncol(X),"betas "))
}
return(EMV)
}

Exemplo:

Especifique conforme abaixo:

dados <- cbind(y,n,v,1,x1,x2)
         y: variáveis resposta
         n: número de indivíduos em cada cluster
         v: valor da função dos indivíduos
         1, x1, x2: lista de covariáveis dos clusters

fun.p = 1 - função de ligação logito
```

```
        2 - função de ligação complementar log-log
        3 - função de ligação log-log
        4 - função de ligação probito

fun.rho = 1 - estrutura de correlação exponencial
          2 - estrutura de correlação AR contínua
          3 - estrutura de correlação Gaussiana

chutes: lista de valores iniciais dos parâmetros

erro: erro admissível no critério de parada do algoritmo

EMV.MRBC(dados=cbind(y,n,v,1,x1,x2), fun.p=1, fun.rho=1, chutes=c(0,0,0,0))

Saída:
- EMV: vetor de estimadores de máxima verossimilhança
```

C.2 Intervalos de confiança

Função para obter os intervalos de confiança para o vetor de parâmetros $\boldsymbol{\theta}$, usando as quatro formas de construção dos intervalos de confiança descritos na Seção 3.2.

```
IC.MRBC <- function(dados, fun.p, fun.rho, nsim.mc, nsim.boot, nsim.perf,
                    EMV, conf=0.95, digt=3) {

y <- dados[,1]
n <- dados[,2]
v <- dados[,3]
X <- dados[,4:ncol(dados)]
m <- nrow(dados)

# Verossimilhança usual

J_bb <- function(EMV,xs,xr) {
preditor = X%*%EMV[-1]

if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }

if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }

if(fun.p==1){ dp1s = xs*exp(preditor)*(1+exp(preditor))^(-2) }
if(fun.p==2){ dp1s = xs*exp(preditor-exp(preditor)) }
if(fun.p==3){ dp1s = xs*exp(-preditor-exp(-preditor)) }
```

```
if(fun.p==4){ dp1s = xs*dnorm(preditor) }

if(fun.p==1){ dp1r = xr*exp(preditor)*(1+exp(preditor))^(-2) }
if(fun.p==2){ dp1r = xr*exp(preditor-exp(preditor)) }
if(fun.p==3){ dp1r = xr*exp(-preditor-exp(-preditor)) }
if(fun.p==4){ dp1r = xr*dnorm(preditor) }

if(fun.p==1){ dp2 = -xr*xs*exp(preditor)*(exp(preditor)-1)*
                    (1+exp(preditor))^(-3) }
if(fun.p==2){ dp2 = xr*xs*(exp(preditor-exp(preditor))
                    -exp(2*preditor-exp(preditor))) }
if(fun.p==3){ dp2 = -xr*xs*(exp(-preditor-exp(-preditor))
                    -exp(-2*preditor-exp(-preditor))) }
if(fun.p==4){ dp2 = -xr*xs*preditor*(2*pi)^(1/2)*exp(-(preditor^2)/2) }

A = choose(n,y) * p^y * (1-p)^(n-y)
B = p^(y/n) * (1-p)^((n-y)/n)
B[!((y==n)|(y==0))] = 0

res <- - sum((A*y^2*dp1s/p^2*dp1r*(1-rho)+A*y*dp2/p*(1-rho)-A*y*dp1r/p^2*
        (1-rho)*dp1s-2*A*y*dp1r/p*(n-y)*dp1s/(1-p)*(1-rho)+A*(n-y)^2*dp1s/
        (1-p)^2*dp1r*(1-rho)-A*(n-y)*dp2/(1-p)*(1-rho)-A*(n-y)*dp1r/(1-p)^2*
        (1-rho)*dp1s+B*y^2/n^2*dp1s/p^2*dp1r*rho+B*y/n*dp2/p*rho-B*y/n*dp1r/
        p^2*rho*dp1s-2*B*y/n^2*dp1r/p*(n-y)*dp1s/(1-p)*rho+B*(n-y)^2/n^2*dp1s/
        (1-p)^2*dp1r*rho-B*(n-y)/n*dp2/(1-p)*rho-B*(n-y)/n*dp1r/(1-p)^2*rho*dp1s)
        /(A*(1-rho)+B*rho)-(A*y*dp1r/p*(1-rho)-A*(n-y)*dp1r/(1-p)*(1-rho)+B*y/n*
        dp1r/p*rho-B*(n-y)/n*dp1r/(1-p)*rho)/(A*(1-rho)+B*rho)^2*(A*y*dp1s/p*(1-
        rho)-A*(n-y)*dp1s/(1-p)*(1-rho)+B*y/n*dp1s/p*rho-B*(n-y)/n*dp1s/(1-p)*rho))
return(res)
}

J_bcov <- function(EMV,xr) {
preditor = X%*%EMV[-1]

if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }

if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }

if(fun.p==1){ dp1r = xr*exp(preditor)*(1+exp(preditor))^(-2) }
if(fun.p==2){ dp1r = xr*exp(preditor-exp(preditor)) }
if(fun.p==3){ dp1r = xr*exp(-preditor-exp(-preditor)) }
if(fun.p==4){ dp1r = xr*dnorm(preditor) }

if(fun.rho==1) { drho1 = -v*exp(-EMV[1]*v) }
if(fun.rho==2) { drho1 = v*(EMV[1]^v)*EMV[1]^(-1) }
if(fun.rho==3) { drho1 = -2*EMV[1]*v^(2)*exp(-(EMV[1]*v)^2) }
```

```
A = choose(n,y) * p^y * (1-p)^(n-y)
B = p^(y/n) * (1-p)^((n-y)/n)
B[!((y==n)|(y==0))] = 0

res  <- - sum((-A*y*dp1r/p*drho1+A*(n-y)*dp1r/(1-p)*drho1+B*y/n*dp1r/p*
        drho1-B*(n-y)/n*dp1r/(1-p)*drho1)/(A*(1-rho)+B*rho)-(A*y*dp1r/p*
        (1-rho)-A*(n-y)*dp1r/(1-p)*(1-rho)+B*y/n*dp1r/p*rho-B*(n-y)/n*
        dp1r/(1-p)*rho)/(A*(1-rho)+B*rho)^2*(-A*drho1+B*drho1))
return(res)
}

J_cov <- function(EMV) {

preditor = X%*%EMV[-1]

if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }

if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }

if(fun.rho==1) { drho1 = -v*exp(-EMV[1]*v) }
if(fun.rho==2) { drho1 = v*(EMV[1]^v)*EMV[1]^(-1) }
if(fun.rho==3) { drho1 = -2*EMV[1]*v^(2)*exp(-(EMV[1]*v)^2) }

if(fun.rho==1) { drho2 = v^2*exp(-EMV[1]*v) }
if(fun.rho==2) { drho2 = (EMV[1]^(v-2))*v*(v-1) }
if(fun.rho==3) { drho2 = 2*v^2*exp(-(EMV[1]*v)^2)*(2*((EMV[1]*v)^2)-1) }

A = choose(n,y) * p^y * (1-p)^(n-y)
B = p^(y/n) * (1-p)^((n-y)/n)
B[!((y==n)|(y==0))] = 0

res <- - sum((-A*drho2+B*drho2)/(A*(1-rho)+B*rho)-
         (-A*drho1+B*drho1)^2/(A*(1-rho)+B*rho)^2)
return(res)
}

J <- matrix(ncol=ncol(EMV),nrow=ncol(EMV),0)
colnames(J) <- c("gama",  replicate((ncol(EMV)-1),"betas"))
rownames(J) <- c("gama",  replicate((ncol(EMV)-1),"betas"))

J[1,1] <- J_cov(EMV)

for (i in 1:ncol(EMV)) {
for (j in 1:ncol(EMV)) {
if( i>1 & (j>=i) ) {
```

```
J[i,j] <- J_bb(EMV,X[,i-1],X[,j-1]) ; J[j,i] <- J[i,j]  }
else { if(i==1 & j>1) {
J[i,j] <- J_bcov(EMV,X[,j-1]) ; J[j,i] <- J[i,j] } }
}}

varbeta <- diag(solve(J))

# Verossimilhança aumentada

J_bba <- function(EMV,xs,xr) {
preditor = X%*%EMV[-1]

if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }

if(fun.p==1){ dp1s = xs*exp(preditor)*(1+exp(preditor))^(-2) }
if(fun.p==2){ dp1s = xs*exp(preditor-exp(preditor)) }
if(fun.p==3){ dp1s = xs*exp(-preditor-exp(-preditor)) }
if(fun.p==4){ dp1s = xs*dnorm(preditor) }

if(fun.p==1){ dp1r = xr*exp(preditor)*(1+exp(preditor))^(-2) }
if(fun.p==2){ dp1r = xr*exp(preditor-exp(preditor)) }
if(fun.p==3){ dp1r = xr*exp(-preditor-exp(-preditor)) }
if(fun.p==4){ dp1r = xr*dnorm(preditor) }

if(fun.p==1){ dp2 = -xr*xs*exp(preditor)*(exp(preditor)-1)
                    *(1+exp(preditor))^(-3) }
if(fun.p==2){ dp2 = xr*xs*(exp(preditor-exp(preditor))
                    -exp(2*preditor-exp(preditor))) }
if(fun.p==3){ dp2 = -xr*xs*(exp(-preditor-exp(-preditor))
                    -exp(-2*preditor-exp(-preditor))) }
if(fun.p==4){ dp2 = -xr*xs*preditor*(2*pi)^(1/2)*exp(-(preditor^2)/2) }

res <- -sum((y/n)*(z+n-n*z)*(p^(-1)*dp2-p^(-2)*dp1s*dp1r)-(n-y)*
       ((z/n)+1-z)*((1-p)^(-1)*dp2 + (1-p)^(-2)*dp1s*dp1r))
return(res)
}

J_cova <- function(EMV) {

if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }

if(fun.rho==1) { drho1 = -v*exp(-EMV[1]*v) }
if(fun.rho==2) { drho1 = v*(EMV[1]^v)*EMV[1]^(-1) }
if(fun.rho==3) { drho1 = -2*EMV[1]*v^(2)*exp(-(EMV[1]*v)^2) }

if(fun.rho==1) { drho2 = v^2*exp(-EMV[1]*v) }
```

```
if(fun.rho==2) { drho2 = (EMV[1]^(v-2))*v*(v-1) }
if(fun.rho==3) { drho2 = 2*v^2*exp(-(EMV[1]*v)^2)*(2*(EMV[1]*v)^2-1) }

res1 <- z*((rho^(-1))*drho2 - (rho^(-2))*drho1^2)
res2 <- -(1-z)*(((1-rho)^(-1))*drho2 + ((1-rho)^(-2))*drho1^2)
res <- -sum(res1+res2)
return(res)
}

U_bba <- function(EMV,xr) {
preditor = X%*%EMV[-1]

if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }

if(fun.p==1){ dp1r = xr*exp(preditor)*(1+exp(preditor))^(-2) }
if(fun.p==2){ dp1r = xr*exp(preditor-exp(preditor)) }
if(fun.p==3){ dp1r = xr*exp(-preditor-exp(-preditor)) }
if(fun.p==4){ dp1r = xr*dnorm(preditor) }

res1 <- (y/n)*(z+n-n*z)*(p^(-1)*dp1r)
res2 <- -(n-y)*((z/n)+1-z)*((1-p)^(-1)*dp1r)
res <- sum(res1+res2)
return(res)
}

U_cova <- function(EMV) {

if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }

if(fun.rho==1) { drho1 = -v*exp(-EMV[1]*v) }
if(fun.rho==2) { drho1 = v*(EMV[1]^v)*EMV[1]^(-1) }
if(fun.rho==3) { drho1 = -2*EMV[1]*v^(2)*exp(-(EMV[1]*v)^2) }

res1 <- z*((rho^(-1))*drho1)
res2 <- -(1-z)*((1-rho)^(-1)*drho1)
res <- sum(res1+res2)
return(res)
}

J_betaa <- matrix(ncol=(ncol(EMV)-1),nrow=(ncol(EMV)-1),0)
colnames(J_betaa) <- c(replicate((ncol(EMV)-1),"betas"))
rownames(J_betaa) <- c(replicate((ncol(EMV)-1),"betas"))

preditor = X%*%EMV[-1]

if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
```

```
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }

if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }

v1 <- rho*p^(y/n)*(1-p)^((n-y)/n)
v2 <- (1-rho)*choose(n,y)*p^y*(1-p)^(n-y)
z <-  v1 /(v1+v2)
z[!((y==0)|(y==n))]=0

for (i in 1:(ncol(EMV)-1)) {
for (j in 1:(ncol(EMV)-1)) {
J_betaa[i,j] <- J_bba(EMV,X[,i],X[,j])
}}

Jaux <- matrix(ncol=ncol(EMV), nrow=ncol(EMV), 0)
Jaux[2:ncol(EMV),2:ncol(EMV)] <- J_betaa
Jaux[1,1] <- J_cova(EMV)

U_a <- matrix(nrow=ncol(EMV), ncol=ncol(EMV), 0)

MC.Ua <- list()

for(i in 1:nsim.mc){ MC.Ua[[i]] <- U_a }

preditor = X%*%EMV[-1]

if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }

if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }

v1 <- rho*p^(y/n)*(1-p)^((n-y)/n)
v2 <- (1-rho)*choose(n,y)*p^y*(1-p)^(n-y)
prob.z <- v1 / (v1+v2)

for (q in 1:nsim.mc){

z <- sapply(prob.z, rbinom, n=1, size=1)
z[!((y==0)|(y==n))]=0

U_a[1,1] <- U_cova(EMV)*U_cova(EMV)

for (k in 1:ncol(EMV)){
```

```
for (i in 1:ncol(EMV)){

if(i>1 & (k>=i)) {
U_a[i,k] <- U_bba(EMV,X[,k-1])*U_bba(EMV,X[,i-1]) ; U_a[k,i] <- U_a[i,k]}
else { if(i==1 & k>1) {
U_a[i,k] <- U_cova(EMV)*U_bba(EMV,X[,k-1]) ; U_a[k,i] <- U_a[i,k]} }
}}


MC.Ua[[q]] <- U_a
}

MC.Ua.m <- matrix(ncol=ncol(EMV), nrow=ncol(EMV),0)
aux <- numeric(nsim.mc)

for (i in 1:ncol(EMV)){
for (j in 1:ncol(EMV)){
   for (k in 1:nsim.mc){
   aux[k] <- MC.Ua[[k]][i,j]
   }
MC.Ua.m[i,j] <- mean(aux)
}}

Ja <- Jaux - MC.Ua.m

varbetaa <- diag(solve(Ja))

# Bootstrap não-paramétrico

BETAS = as.matrix(EMV)

EMVS <- matrix(ncol=ncol(EMV),nrow=nsim.boot,0)

for (j in 1:nsim.boot){

ordem <- sample(1:m, m, replace=T)
dados_novo = dados
for (i in 1:m){ dados_novo[i,] <- dados[ordem[i],] }

EMVS[j,] <- EMV.MRBC(dados=dados_novo, fun.p=fun.p, fun.rho=fun.rho, chutes=EMV)
}

# Perfilado

vero.usual <- function(beta){
preditor = X%*%(as.matrix(beta[-1]))
if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }
if(fun.rho==1) { rho = exp(-exp(beta[1])*v) }
```

```
if(fun.rho==2) { rho = (exp(beta[1])/(1+exp(beta[1])))^v }
if(fun.rho==3) { rho = exp(-(exp(beta[1])*v)^2) }
bin <- choose(n,y) * p^y * (1-p)^(n-y) * (1-rho)
ber <- p^(y/n) * (1-p)^((n-y)/n) * rho
ll <-  log(bin+ber)
ll[!(y==0 | y==n)] = log(bin)[!(y==0 | y==n)]
return(sum(ll))
}

vero.emvs <- matrix(ncol=(ncol(X)+1),nrow=nsim.perf,0)
teste <- matrix(ncol=(ncol(X)+1),nrow=nsim.perf,0)

EMV.aux <- EMV

if(fun.rho==1) { EMV.aux[1] <- log(EMV[1]) }
if(fun.rho==2) { EMV.aux[1] <- log(EMV[1]/(1-EMV[1])) }
if(fun.rho==3) { EMV.aux[1] <- log(EMV[1]) }

EMV.a <- EMV.aux

for (j in 1:(ncol(X)+1)){
for (i in 1:nsim.perf){
teste[,j] <- seq(EMV[j]-8, EMV[j]+8, length.out=nsim.perf)
EMV.aux <- EMV.a
EMV.aux[j] <- teste[i,j]
vero.emvs[i,j] <- vero.usual(EMV.aux)
}}

dif <- vero.emvs - (vero.usual(EMV.a) - (1/2)*qchisq(conf,1))

IC <- matrix(ncol=2, nrow=(ncol(X)+1), 0)
for (i in 1:(ncol(X)+1)) {
IC[i,] <- teste[c(which( c(sign(dif[,i]),0) * c(0,sign(dif[,i])) ==-1)),i]
}

if(fun.rho==1) { IC[1,] <- exp(IC[1,]) }
if(fun.rho==2) { IC[1,] <- exp(IC[1,])/(1+ exp(IC[1,])) }
if(fun.rho==3) { IC[1,] <- exp(IC[1,]) }

# resumos

confiança <- c((1-conf)/2, 1-(1-conf)/2)

resumo <- matrix(ncol=9, nrow=ncol(EMV))

resumo[,1] <- EMV
resumo[,2] <- EMV + qnorm((1-conf)/2)*sqrt(varbeta)
resumo[,3] <- EMV + qnorm(1-(1-conf)/2)*sqrt(varbeta)
resumo[,4] <- EMV + qnorm((1-conf)/2)*sqrt(varbetaa)
resumo[,5] <- EMV + qnorm(1-(1-conf)/2)*sqrt(varbetaa)
resumo[,6] <- apply(EMVS,2,quantile, probs=(1-conf)/2)
```

```
resumo[,7] <- apply(EMVS,2,quantile, probs=1-(1-conf)/2)
resumo[,8] <- IC[,1]
resumo[,9] <- IC[,2]

colnames(resumo) <- paste(c("EMV", "ICi", "ICs", "ICi_a", "ICs_a",
                    "ICi_b", "ICs_b", "ICi_p", "ICs_p"))
rownames(resumo) <- paste(c("gamma", replicate((ncol(EMV)-1),"betas")))

return(round(resumo,digt))
}

Exemplo:

Especifique conforme abaixo:

dados <- cbind(y,n,v,1,x1,x2)
         y: variáveis resposta
         n: número de indivíduos em cada cluster
         v: valor da função dos indivíduos
         1, x1, x2: lista de covariáveis dos clusters

fun.p = 1 - função de ligação logito
        2 - função de ligação complementar log-log
        3 - função de ligação log-log
        4 - função de ligação probito

fun.rho = 1 - estrutura de correlação exponencial
          2 - estrutura de correlação AR contínua
          3 - estrutura de correlação Gaussiana

EMV <- EMV.MRBC(dados=cbind(y,n,v,1,x1,x2), fun.p=1, fun.rho=1, chutes=c(0,0,0,0))

nsim.mc: número de simulações Monte Carlo

nsim.boot: número de simulações bootstrap

nsim.perf: número de valores do parâmetro para o intervalo perfilado

conf: nível de confiança para construção dos intervalos

digt: número de casas decimais na saída

IC.MRBC(dados = cbind(y,n,v,1,x1,x2), fun.p=1, fun.rho=1, nsim.mc=3000,
        nsim.boot=100, nsim.perf=5000, EMV, conf=0.95, digt=3)

Saída:
- EMV: vetor de estimadores de máxima verossimilhança
- (ICi,ICs): intervalo de confiança assintótico construído
             com a verossimilhança usual
- (ICi_a,ICs_a): intervalo de confiança assintótico construído
             com a verossimilhança aumentada
```

```
- (ICi_b,ICs_b): intervalo de confiança bootstrap
- (ICi_p, ICs_p): intervalo de confiança perfilado
```

C.3 Resíduos

Função para obter os resíduos padronizados, conforme descrito na Seção 3.4.1.

```
res.MRBC <- function(dados, fun.p, fun.rho, EMV) {

y <- dados[,1]
n <- dados[,2]
v <- dados[,3]
X <- dados[,4:ncol(dados)]
m <- nrow(dados)

hatpreditor = X%*%EMV[-1]
if(fun.p==1){ hatp = exp(hatpreditor)/(1+exp(hatpreditor)) }
if(fun.p==2){ hatp = 1 - exp(-exp(hatpreditor)) }
if(fun.p==3){ hatp = exp(-exp(-hatpreditor)) }
if(fun.p==4){ hatp = pnorm(hatpreditor) }
if(fun.rho==1) { hatrho = exp(-EMV[1]*v) }
if(fun.rho==2) { hatrho = EMV[1]^v }
if(fun.rho==3) { hatrho = exp(-(EMV[1]*v)^2) }
residuo.pad = (y - n*hatp)/sqrt(hatp*(1-hatp)*(n+hatrho*n*(n-1)))

index = 1:m
numero1 = length(which(abs(residuo.pad)>2))

log.vero <- function(EMV, tipo) {
if(tipo=="saturada") { p = y/n }
if(tipo=="MRBC") {
   preditor = X%*%EMV[-1]
   if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
   if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
   if(fun.p==3){ p = exp(-exp(-preditor)) }
   if(fun.p==4){ p = pnorm(preditor) }
   }
if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }
IA2 <- rep(0,m)
IA2[(y==n)|(y==0)] <- 1
bin <- choose(n,y) * p^y * (1-p)^(n-y) * (1-rho)
ber <- p^(y/n) * (1-p)^((n-y)/n) * rho * IA2
ll = log(bin+ber)
return(ll)
}

sat <- log.vero(EMV, tipo="saturada")
```

```
pred <- log.vero(EMV, tipo="MRBC")

residuo.dev = (sign(y - n*hatp)*sqrt( 2*sat - 2*pred ))/
              sqrt(hatp*(1-hatp)*(n+hatrho*n*(n-1)))

numero = length(which(abs(residuo.dev)>2))

par(mfrow=c(1,2),mar=c(1.5,1.5,1.5,1.5))

lim <- c(min(residuo.dev, residuo.pad)-0.5, max(residuo.dev, residuo.pad)+0.5)
plot(n*hatp, residuo.pad, xlab="valores preditos", ylab="", main="(a)",
     ylim=lim, cex.lab=1.2, lwd=2, cex.axis=1.2)
showLabels(n*hatp, residuo.pad, labels=index, id.method=list("y"),
     id.n = numero1, id.cex=1.2)
abline(h=c(1,-1), col="grey90")
abline(h=c(-2,2), col="grey70")

plot(n*hatp, residuo.dev, xlab="valores preditos", ylab="", main="(b)",
    ylim=lim, cex.lab=1.2, lwd=2, cex.axis=1.2)
showLabels(n*hatp, residuo.dev, labels=index, id.method=list("y"),
    id.n = numero, id.cex=1.2)
abline(h=c(1,-1), col="grey90")
abline(h=c(-2,2), col="grey70")

return(list(res.pad=residuo.pad,res.dev=residuo.dev))
}

Exemplo:

Especifique conforme abaixo:

dados <- cbind(y,n,v,1,x1,x2)
         y: variáveis resposta
         n: número de indivíduos em cada cluster
         v: valor da função dos indivíduos
         1, x1, x2: lista de covariáveis dos clusters

fun.p = 1 - função de ligação logito
        2 - função de ligação complementar log-log
        3 - função de ligação log-log
        4 - função de ligação probito

fun.rho = 1 - estrutura de correlação exponencial
          2 - estrutura de correlação AR contínua
          3 - estrutura de correlação Gaussiana

EMV <- EMV.MRBC(dados=cbind(y,n,v,1,x1,x2), fun.p=1, fun.rho=1, chutes=c(0,0,0,0))

res.MRBC(cbind(y,n,v,1,x1,x2), fun.p=1, fun.rho=1, EMV)

Saída:
```

```
- gráficos dos resíduos padronizados baseados nos valores preditos
  e resíduos padronizados via deviance
- res.pad: valores dos resíduos padronizados baseados nos valores preditos
- res.dev: valores dos resíduos padronizados via deviance
```

C.4 Influência global

Função para obter os valores das métricas distância de Cook generalizada e distância da verossimilhança para os modelos de regressão binomial correlacionada, conforme descrito na Seção 3.4.2.

```
inf.MRBC <- function(dados, fun.p, fun.rho, EMV) {

y <- dados[,1]
n <- dados[,2]
v <- dados[,3]
X <- dados[,4:ncol(dados)]
m <- nrow(dados)

U.f <- function(EMV, xr, pos) {
preditor = X%*%EMV[-1]

if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }

if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }

if(fun.p==1){ dp1r = xr*exp(preditor)*(1+exp(preditor))^(-2) }
if(fun.p==2){ dp1r = xr*exp(preditor-exp(preditor)) }
if(fun.p==3){ dp1r = xr*exp(-preditor-exp(-preditor)) }
if(fun.p==4){ dp1r = xr*dnorm(preditor) }

if(fun.rho==1) { drho1 = -v*exp(-EMV[1]*v) }
if(fun.rho==2) { drho1 = v*(EMV[1]^v)*EMV[1]^(-1) }
if(fun.rho==3) { drho1 = -2*EMV[1]*v^(2)*exp(-(EMV[1]*v)^2) }

IA2 <- rep(0,m)
IA2[(y==n)|(y==0)] <- 1

A = choose(n,y) * p^y * (1-p)^(n-y)
B = p^(y/n) * (1-p)^((n-y)/n)
B[!((y==n)|(y==0))] = 0

if(pos=="gama") { res <- drho1*(-A+B)/(A*(1-rho)+B*(rho)) }
```

```
if(pos=="betas"){ res <- dp1r*(A*(1-rho)*(y/p-(n-y)/(1-p))+B*(rho)*
                        ((y/n)/p-((n-y)/n)/(1-p)))/(A*(1-rho)+B*(rho))}
return(res)
}

U <- matrix(ncol=length(EMV), nrow=m,0)
colnames(U) <- c("gama", replicate(ncol(X),"betas"))

X.aux <- cbind(1,X)

for(j in 1:length(EMV)) {
aux <- U.f(EMV, X.aux[,j], pos=colnames(U)[j])
for (i in 1:m) { U[i,j] <- sum(aux[-i]) }
}

J_bb <- function(EMV,xs,xr) {
preditor = X%*%EMV[-1]

if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }

if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }

if(fun.p==1){ dp1s = xs*exp(preditor)*(1+exp(preditor))^(-2) }
if(fun.p==2){ dp1s = xs*exp(preditor-exp(preditor)) }
if(fun.p==3){ dp1s = xs*exp(-preditor-exp(-preditor)) }
if(fun.p==4){ dp1s = xs*dnorm(preditor) }

if(fun.p==1){ dp1r = xr*exp(preditor)*(1+exp(preditor))^(-2) }
if(fun.p==2){ dp1r = xr*exp(preditor-exp(preditor)) }
if(fun.p==3){ dp1r = xr*exp(-preditor-exp(-preditor)) }
if(fun.p==4){ dp1r = xr*dnorm(preditor) }

if(fun.p==1){ dp2 = -xr*xs*exp(preditor)*(exp(preditor)-1)
                    *(1+exp(preditor))^(-3) }
if(fun.p==2){ dp2 = xr*xs*(exp(preditor-exp(preditor))
                    -exp(2*preditor-exp(preditor))) }
if(fun.p==3){ dp2 = -xr*xs*(exp(-preditor-exp(-preditor))
                    -exp(-2*preditor-exp(-preditor))) }
if(fun.p==4){ dp2 = -xr*xs*preditor*(2*pi)^(1/2)*exp(-(preditor^2)/2) }

A = choose(n,y) * p^y * (1-p)^(n-y)
B = p^(y/n) * (1-p)^((n-y)/n)
B[!((y==n)|(y==0))] = 0

res <- - sum((A*y^2*dp1s/p^2*dp1r*(1-rho)+A*y*dp2/p*(1-rho)-A*y*dp1r/p^2*
        (1-rho)*dp1s-2*A*y*dp1r/p*(n-y)*dp1s/(1-p)*(1-rho)+A*(n-y)^2*dp1s/
```

```
        (1-p)^2*dp1r*(1-rho)-A*(n-y)*dp2/(1-p)*(1-rho)-A*(n-y)*dp1r/(1-p)^2*
        (1-rho)*dp1s+B*y^2/n^2*dp1s/p^2*dp1r*rho+B*y/n*dp2/p*rho-B*y/n*dp1r/
         p^2*rho*dp1s-2*B*y/n^2*dp1r/p*(n-y)*dp1s/(1-p)*rho+B*(n-y)^2/n^2*dp1s/
        (1-p)^2*dp1r*rho-B*(n-y)/n*dp2/(1-p)*rho-B*(n-y)/n*dp1r/(1-p)^2*
         rho*dp1s)/(A*(1-rho)+B*rho)-(A*y*dp1r/p*(1-rho)-A*(n-y)*dp1r/(1-p)*
        (1-rho)+B*y/n*dp1r/p*rho-B*(n-y)/n*dp1r/(1-p)*rho)/(A*(1-rho)+B*rho)^2*
        (A*y*dp1s/p*(1-rho)-A*(n-y)*dp1s/(1-p)*(1-rho)+B*y/n*dp1s/p*rho-B*(n-y)/n
        *dp1s/(1-p)*rho))
return(res)
}

J_bcov <- function(EMV,xr) {
preditor = X%*%EMV[-1]

if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }

if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }

if(fun.p==1){ dp1r = xr*exp(preditor)*(1+exp(preditor))^(-2) }
if(fun.p==2){ dp1r = xr*exp(preditor-exp(preditor)) }
if(fun.p==3){ dp1r = xr*exp(-preditor-exp(-preditor)) }
if(fun.p==4){ dp1r = xr*dnorm(preditor) }

if(fun.rho==1) { drho1 = -v*exp(-EMV[1]*v) }
if(fun.rho==2) { drho1 = v*(EMV[1]^v)*EMV[1]^(-1) }
if(fun.rho==3) { drho1 = -2*EMV[1]*v^(2)*exp(-(EMV[1]*v)^2) }

A = choose(n,y) * p^y * (1-p)^(n-y)
B = p^(y/n) * (1-p)^((n-y)/n)
B[!((y==n)|(y==0))] = 0

res  <- - sum((-A*y*dp1r/p*drho1+A*(n-y)*dp1r/(1-p)*drho1+B*y/n*dp1r/p*
          drho1-B*(n-y)/n*dp1r/(1-p)*drho1)/(A*(1-rho)+B*rho)-(A*y*dp1r/
          p*(1-rho)-A*(n-y)*dp1r/(1-p)*(1-rho)+B*y/n*dp1r/p*rho-B*(n-y)/
          n*dp1r/(1-p)*rho)/(A*(1-rho)+B*rho)^2*(-A*drho1+B*drho1))
return(res)
}

J_cov <- function(EMV) {

preditor = X%*%EMV[-1]

if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }
```

```
if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }

if(fun.rho==1) { drho1 = -v*exp(-EMV[1]*v) }
if(fun.rho==2) { drho1 = v*(EMV[1]^v)*EMV[1]^(-1) }
if(fun.rho==3) { drho1 = -2*EMV[1]*v^(2)*exp(-(EMV[1]*v)^2) }

if(fun.rho==1) { drho2 = v^2*exp(-EMV[1]*v) }
if(fun.rho==2) { drho2 = (EMV[1]^(v-2))*v*(v-1) }
if(fun.rho==3) { drho2 = 2*v^2*exp(-(EMV[1]*v)^2)*(2*((EMV[1]*v)^2)-1) }

A = choose(n,y) * p^y * (1-p)^(n-y)
B = p^(y/n) * (1-p)^((n-y)/n)
B[!((y==n)|(y==0))] = 0

res <- - sum((-A*drho2+B*drho2)/(A*(1-rho)+B*rho)-
        (-A*drho1+B*drho1)^2/(A*(1-rho)+B*rho)^2)
return(res)
}

J <- matrix(ncol=length(EMV),nrow=length(EMV),0)
colnames(J) <- c("gama", replicate((length(EMV)-1),"betas"))
rownames(J) <- c("gama", replicate((length(EMV)-1),"betas"))

J[1,1] <- J_cov(EMV)

for (i in 1:length(EMV)) {
for (j in 1:length(EMV)) {
if( i>1 & (j>=i) ) {
J[i,j] <- J_bb(EMV,X[,i-1],X[,j-1]) ; J[j,i] <- J[i,j]  }
else { if(i==1 & j>1) {
J[i,j] <- J_bcov(EMV,X[,j-1]) ; J[j,i] <- J[i,j] } }
}}

thetai <- matrix(ncol=length(EMV), nrow=m,0)
for (i in 1:m) { thetai[i,] <- EMV + solve(J) %*% U[i,] }

cd <- numeric()
for (i in 1:m){
cd[i]  <- t(thetai[i,] - EMV) %*% J %*% (thetai[i,] - EMV) }

numero_cd = length(which(abs(cd)>1))

log.vero <- function(EMV) {
preditor = X%*%EMV[-1]
if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }
```

```
if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }
IA2 <- rep(0,m)
IA2[(y==n)|(y==0)] <- 1
bin <- choose(n,y) * p^y * (1-p)^(n-y) * (1-rho)
ber <- p^(y/n) * (1-p)^((n-y)/n) * rho * IA2
ll = log(bin+ber)
return(sum(ll))
}

vero.thetai <- numeric()
for (i in 1:m){ vero.thetai[i] <- log.vero(thetai[i,]) }
dv = log.vero(EMV) - vero.thetai

numero_dv = length(which(abs(dv)>1))

par(mfrow=c(1,2), c(1.5,1.5,1.5,1.5))
index = 1:m

lim <- c(0, max(cd,dv)+3)
plot(cd, ylab="", xlab="observação", main="(a)", pch=16,
     ylim=lim, cex.lab=1.2, cex.axis=1.2)
showLabels(index, cd, labels=index, id.method=list("y"),
     id.n = numero_cd, id.cex=1.2)

plot(dv, ylab="", xlab="observação", pch=16, ylim=lim,
    main="(d)", cex.lab=1.2, cex.axis=1.2)
showLabels(index, dv, labels=index, id.method=list("y"),
    id.n = numero_dv, id.cex=1.2)

return(list(cook=cd,dist.vero=dv))
}


Exemplo:

Especifique conforme abaixo:

dados <- cbind(y,n,v,1,x1,x2)
         y: variáveis resposta
         n: número de indivíduos em cada cluster
         v: valor da função dos indivíduos
         1, x1, x2: lista de covariáveis dos clusters

fun.p = 1 - função de ligação logito
        2 - função de ligação complementar log-log
        3 - função de ligação log-log
        4 - função de ligação probito

fun.rho = 1 - estrutura de correlação exponencial
```

```
          2 - estrutura de correlação AR contínua
          3 - estrutura de correlação Gaussiana

EMV <- EMV.MRBC(dados=cbind(y,n,v,1,x1,x2), fun.p=1, fun.rho=1, chutes=c(0,0,0,0))

inf.MRBC(cbind(y,n,v,1,x1,x2), fun.p=1, fun.rho=1, EMV)

Saída:
- gráficos dos valores das distâncias de cook generalizada e
  das distâncias da verossimilhança
- cook: valores das distâncias de cook generalizada
- dist.vero: valores das distâncias da verossimilhança
```

C.5 Seleção de modelos

Função para obter os valores dos critérios de seleção AIC e BIC para os modelos de regressão binomial correlacionada, conforme descrito na Seção 3.5.

```
crit.MRBC <- function(dados, fun.p, fun.rho, EMV) {

y <- dados[,1]
n <- dados[,2]
v <- dados[,3]
X <- dados[,4:ncol(dados)]
m <- nrow(dados)

log.vero <- function(EMV) {
preditor = X%*%EMV[-1]
if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }
if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }
IA2 <- rep(0,m)
IA2[(y==n)|(y==0)] <- 1
bin <- choose(n,y) * p^y * (1-p)^(n-y) * (1-rho)
ber <- p^(y/n) * (1-p)^((n-y)/n) * rho * IA2
ll = log(bin+ber)
return(sum(ll))
}

aic = -2*log.vero(EMV) + 2*length(EMV)
bic = -2*log.vero(EMV) + 2*length(EMV)*log(m)

return(list(AIC=aic,BIC=bic))
}
```

```
Exemplo:

Especifique conforme abaixo:

dados <- cbind(y,n,v,1,x1,x2)
         y: variáveis resposta
         n: número de indivíduos em cada cluster
         v: valor da função dos indivíduos
         1, x1, x2: lista de covariáveis dos clusters

fun.p = 1 - função de ligação logito
        2 - função de ligação complementar log-log
        3 - função de ligação log-log
        4 - função de ligação probito

fun.rho = 1 - estrutura de correlação exponencial
          2 - estrutura de correlação AR contínua
          3 - estrutura de correlação Gaussiana

EMV <- EMV.MRBC(dados=cbind(y,n,v,1,x1,x2), fun.p=1, fun.rho=1, chutes=c(0,0,0,0))

crit.MRBC(cbind(y,n,v,1,x1,x2), fun.p=1, fun.rho=1, EMV)

Saída:
- valores dos critérios AIC e BIC.
```

Apêndice D

Programas para o MRBC: Abordagem Bayesiana

Neste capítulo, apresentamos os principais programas computacionais utilizados no MRBC via abordagem Bayesiana. As funções foram implementadas em R (R Development Core Team, 2011) para obter as amostras MCMC dos parâmetros a *posteriori*, o valor da CPO e dos valores preditos para a variável resposta via CPO, os valores dos resíduos Bayesianos, o valor da divergência de K-L e de sua calibração e o valor do critério de seleção de modelo DIC, discutidos no Capítulo 4.

D.1 Gibbs com passo de Metropolis

Função para gerar cadeias do vetor de parâmetros $\boldsymbol{\theta}$ via algoritmo de Gibbs com passo de Metropolis e dados aumentados, fazendo uso do passeio aleatório, conforme descrito no início do Capítulo 4. Para rodar esta função é necessário instalar o pacote mvtnorm.

```
MCMC.MRBC <- function(dados, fun.p, fun.rho, chutes, n.mcmc, hiper.var, sd.theta){

require(mvtnorm)
BETAS = as.matrix(chutes)
y <- dados[,1]
n <- dados[,2]
v <- dados[,3]
X <- dados[,4:ncol(dados)]
m <- nrow(dados)

cont.theta <- rep(0, length(BETAS))

vero <- function(beta){
preditor = X%*%(as.matrix(beta[-1]))
if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }
```

```
if(fun.rho==1) { rho = exp(-exp(beta[1])*v) }
if(fun.rho==2) { rho = (exp(beta[1])/(1+exp(beta[1])))^v }
if(fun.rho==3) { rho = exp(-(exp(beta[1])*v)^2) }
ll = sum ( log(p^((y/n)*(z+n-n*z))) + log((1-p)^((n-y)*((z/n)+1-z)))
return(ll)
}

Sigma <- diag(hiper.var)
post <- function(beta) {
ll = vero(beta)
priori = -(1/2)*t(beta)%*%solve(Sigma)%*%beta
post = priori + ll
return(post)
}

theta.sim  <- matrix(ncol=length(BETAS), nrow=n.mcmc, 0)
theta.sim[1,] <- chutes

IA2 <- rep(0,m)
IA2[(y==n)|(y==0)] <- 1

for (i in 1:(n.mcmc-1)){

new.theta <- rmvnorm(1,theta.sim[i,],diag(sd.theta))

preditor = X%*%theta.sim[i,-1]

if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }

if(fun.rho==1) { rho = exp(-exp(theta.sim[i,1])*v) }
if(fun.rho==2) { rho = (exp(theta.sim[i,1])/(1+exp(theta.sim[i,1])))^v }
if(fun.rho==3) { rho = exp(-(exp(theta.sim[i,1])*v)^2) }

bern <- p^(y/n)*(1-p)^((n-y)/n)*rho*IA2
bin  <- choose(n,y)*p^y*(1-p)^(n-y)*(1-rho)
prob.z <- bern/(bern+bin)
z <- sapply(prob.z, rbinom, n=1, size=1)

teste <- theta.sim[i,]
for(j in 1:length(BETAS)){
  teste.aux <- teste
  teste.aux[j] <- new.theta[j]
  t.theta <- exp( post(teste.aux) - post(teste) )
  if(runif(1) <= min(1, t.theta)) { teste[j] <- new.theta[j] }
  else { cont.theta[j] <- cont.theta[j]+1 }
  }

theta.sim[i+1,] <- teste
```

```
}

if(fun.rho==1) { theta.sim[,1] <- exp(theta.sim[,1]) }
if(fun.rho==2) { theta.sim[,1] <- exp(theta.sim[,1])/(1+ exp(theta.sim[,1])) }
if(fun.rho==3) { theta.sim[,1] <- exp(theta.sim[,1]) }

return(list(cadeia=theta.sim, taxa.aceitação=1-(cont.theta/n.mcmc)))
}

Exemplo:

Especifique conforme abaixo:

dados <- cbind(y,n,v,1,x1,x2)
         y: variáveis resposta
         n: número de indivíduos em cada cluster
         v: valor da função dos indivíduos
         1, x1, x2: lista de covariáveis dos clusters

fun.p = 1 - função de ligação logito
        2 - função de ligação complementar log-log
        3 - função de ligação log-log
        4 - função de ligação probito

fun.rho = 1 - estrutura de correlação exponencial
          2 - estrutura de correlação AR contínua
          3 - estrutura de correlação Gaussiana

chutes: lista de valores iniciais dos parâmetros

n.mcmc: número de simulações MCMC na cadeia

hiper.var: valores dos hiperparâmetros da priori para cada
           parâmetro

sd.theta: valores dos erros para cada parâmetro no gerador
          de candidatos do passeio aleatório

MCMC.MRBC(dados=cbind(y,n,v,1,x1,x2), fun.p=1, fun.rho=1, chutes=c(0,0,0,0),
        n.mcmc=40000, hiper.var=rep(10000,4), sd.theta=c(0.5,0.5,0.5,0.5))

Saída:
- cadeia mcmc para gama, beta0, beta1, ...
- taxa de aceitação dos candidatos
```

D.2 CPO e valores preditos via CPO

Função para obter os valores da CPO e os valores preditos para a variável resposta para cada cluster, conforme descrito na Seção 4.2.

```
CPO.MRBC <- function(dados, fun.p, fun.rho, theta.final){

y <- dados[,1]
n <- dados[,2]
v <- dados[,3]
X <- dados[,4:ncol(dados)]
m <- nrow(dados)
n.cad <- nrow(theta.final)

IA2 <- rep(0,m)
IA2[y==n|y==0]<-1

g <- function(beta,y,n,v,X,IA2) {
preditor = X%*%(as.matrix(beta[-1]))
if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }
if(fun.rho==1) { rho = exp(-exp(beta[1])*v) }
if(fun.rho==2) { rho = (exp(beta[1])/(1+exp(beta[1])))^v }
if(fun.rho==3) { rho = exp(-(exp(beta[1])*v)^2) }
lli = choose(n,y)*p^(y)*(1-p)^(n-y)*(1-rho) + p^(y/n)*(1-p)^((n-y)/n)*rho*IA2
return(lli)
}

inv.g <- matrix(ncol=m, nrow=n.cad,0)
for (j in 1:n1) { inv.g[j,] <-  1/g(theta.final[j,],y,n,v,X,IA2) }
cpoi <- 1/(apply(inv.g,2,mean))

pred.y.cpo <- matrix(ncol=max(n)+1, nrow=n.cad,0)
dens.max.n.cpo <- matrix(ncol=max(n)+1, nrow=m,0)
pred.y.final.cpo <- numeric()

for (i in 1:m) {
  teste.y = 0:n[i]
  for (j in 1:n.cad) {
  pred.y.cpo[j,] <- c(1/g(theta.final[j,],teste.y,n[i],v[i],X[i,],IA2[i]),
                         rep(1000,(max(n)+1-length(teste.y))))
  }

dens.max.n.cpo[i,] <- 1/(apply(pred.y.cpo,2,mean))
pred.y.final.cpo[i] = which.max(dens.max.n.cpo[i,])-1
}

return(list(CPOi=cpoi, pred.resp=pred.y.final.cpo))
```

```
}

Exemplo:

Especifique conforme abaixo:

dados <- cbind(y,n,v,1,x1,x2)
         y: variáveis resposta
         n: número de indivíduos em cada cluster
         v: valor da função dos indivíduos
         1, x1, x2: lista de covariáveis dos clusters

fun.p = 1 - função de ligação logito
        2 - função de ligação complementar log-log
        3 - função de ligação log-log
        4 - função de ligação probito

fun.rho = 1 - estrutura de correlação exponencial
          2 - estrutura de correlação AR contínua
          3 - estrutura de correlação Gaussiana

theta.final: matriz de valores provientes do MCMC,
             considerando o burn in e os saltos necessários

CPO.MRBC(dados=cbind(y,n,d,1,x1,x2), fun.p=1, fun.rho=1, theta.final)

Saída:
- CPOi: o valores da CPO para cada observação
- pred.res: o valores predito para cada observação via CPO
```

D.3 Resíduos Bayesianos

Função para obter os valores dos resíduos Bayesianos padronizados, conforme descrito na Seção 4.3.1. Para rodar esta função é necessário carregar a função CPO.MRBC, mostrada no Apêndice D.2.

```
resb.MRBC <- function(dados, fun.p, fun.rho, theta.final, c1, hiper.var){

y <- dados[,1]
n <- dados[,2]
v <- dados[,3]
X <- dados[,4:ncol(dados)]
m <- nrow(dados)
n.cad <- nrow(theta.final)
index = 1:m

# Amostra da distribuição dos resíduos a posteriori
```

```
dist.residuo.pad <- matrix(nrow=n.cad, ncol=m, 0)
dist.residuo <- matrix(nrow=n.cad, ncol=m, 0)

for (i in 1:n.cad) {
preditori = X%*%theta.final[i,-1]
if(fun.p==1){ hatpi = exp(preditori)/(1+exp(preditori)) }
if(fun.p==2){ hatpi = 1 - exp(-exp(preditori)) }
if(fun.p==3){ hatpi = exp(-exp(-preditori)) }
if(fun.p==4){ hatpi = pnorm(preditori) }
if(fun.rho==1) { hatrhoi = exp(-theta.final[i,1]*v) }
if(fun.rho==2) { hatrhoi = theta.final[i,1]^v }
if(fun.rho==3) { hatrhoi = exp(-(theta.final[i,1]*v)^2) }
dist.residuo.pad[i,] = (y - n*hatpi)/ sqrt(hatpi*(1-hatpi)*(n+hatrhoi*n*(n-1)))
dist.residuo[i,] = n*hatpi
}
media.pi = matrix(nrow=n.cad, ncol=m, rep(apply(dist.residuo, 2,mean),n.cad),byrow = T)

media.residuo = apply(dist.residuo.pad, 2, mean)

numero1 = length(which(abs(media.residuo)>2))

# Resíduos via CPO

res1 <- CPO.MRBC(dados, fun.p, fun.rho, theta.final)
residuo = y - res1$pred.resp

theta.med <- apply(theta.final,2,median)
hatpreditor = X%*%theta.med[-1]
if(fun.p==1){ hatp = exp(hatpreditor)/(1+exp(hatpreditor)) }
if(fun.p==2){ hatp = 1 - exp(-exp(hatpreditor)) }
if(fun.p==3){ hatp = exp(-exp(-hatpreditor)) }
if(fun.p==4){ hatp = pnorm(hatpreditor) }
if(fun.rho==1) { hatrho = exp(-theta.med[1]*v) }
if(fun.rho==2) { hatrho = theta.med[1]^v }
if(fun.rho==3) { hatrho = exp(-(theta.med[1]*v)^2) }

residuo.pad = residuo/sqrt(hatp*(1-hatp)*(n+hatrho*n*(n-1)))

numero = length(which(abs(residuo.pad)>2))

# Resíduos via deviance

IA2 <- rep(0,m)
IA2[y==n|y==0]<-1

g <- function(beta,y,n,v,X,IA2) {
preditor = X%*%beta[-1]
if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
```

```
if(fun.p==4){ p = pnorm(preditor) }
if(fun.rho==1) { rho = exp(-exp(beta[1])*v) }
if(fun.rho==2) { rho = (exp(beta[1])/(1+exp(beta[1])))^v }
if(fun.rho==3) { rho = exp(-(exp(beta[1])*v)^2) }
lli = choose(n,y)*p^(y)*(1-p)^(n-y)*(1-rho) + p^(y/n)*(1-p)^((n-y)/n)*rho*IA2
return(lli)
}

priori <- function(beta) {
Sigma <- diag(hiper.var)
pri = (2*pi)^(-(length(beta)/2))*(det(Sigma))^(-(1/2))*
      exp(-(1/2)*t(beta)%*%solve(Sigma)%*%beta)
return(as.numeric(pri))
}

esp <- matrix(ncol=m, nrow=n.cad,0)
g.priori <- numeric()
log.gm <- numeric()

for (j in 1:n.cad) {
esp[j,] <- log(g(theta.final[j,],y,n,v,X,IA2)/priori(theta.final[j,]))
}

esp.me <- apply(esp, 2, mean)
theta.mean <- apply(theta.final,2,mean)

for (i in 1:m) {
g.priori[i] <- log(g(theta.mean,y[i],n[i],v[i],X[i,],IA2[i])/priori(theta.mean))
log.gm[i] <- log(g(theta.mean, y[i],n[i],v[i],X[i,],IA2[i]))
}

pDi = -2*(esp.me-g.priori)

Di = -2*log.gm
raiz <- Di+pDi
DIC = sum(raiz) + sum(pDi)
res.di = sign(y-n*hatp)*sqrt(raiz)
res.pad.di = res.di / sqrt(hatp*(1-hatp)*(n+hatrho*n*(n-1)))

numero2 = length(which(abs(res.pad.di)>2))


par(mfrow=c(2,2), mar=c(1.5,1.5,1.5,1.5))

lim1=c(min(residuo.pad)-0.5, max(residuo.pad)+0.5)
plot(pred.y.final.cpo, residuo.pad, xlab="valor predito", ylab="", main="(a)",
     ylim=lim1, cex.lab=1.2, lwd=2, cex.axis=1.2)
showLabels(pred.y.final.cpo, residuo.pad, labels=index, id.method=list("y"),
     id.n = numero, id.cex=1.2)
abline(h=c(1,-1), col="grey90")
abline(h=c(-2,2), col="grey70")
```

```
ordem <- order(media.pi[1,])
lim2=c(min(media.residuo)-0.5, max(media.residuo)+0.5)
plot(media.pi[1,ordem], media.residuo[ordem], ylim=lim2, ylab="", xlab="valor esperado",
    main="(b)", cex.lab=1.2, lwd=2, cex.axis=1.2)
for (i in 1:m){
segments(media.pi[1,ordem[i]], HPDinterval(mcmc(dist.residuo.pad[,ordem[i]]))[1],
media.pi[1,ordem[i]],HPDinterval(mcmc(dist.residuo.pad[,ordem[i]]))[2], lty=2) }
showLabels(media.pi[1,ordem], media.residuo[ordem], labels=index, id.method=list("y"),
          id.n = numero1, id.cex=1.2)
abline(h=c(1,-1), col="grey90")
abline(h=c(-2,2), col="grey70")

lim3=c(min(res.pad.di)-0.5, max(res.pad.di)+0.5)
plot(pred.y.final.cpo, res.pad.di, xlab="valor predito", ylab="", main="(c)",
     ylim=lim3, cex.lab=1.2, lwd=2, cex.axis=1.2)
showLabels(pred.y.final, res.pad.di, labels=index, id.method=list("y"),
     id.n = numero2, id.cex=1.2)
abline(h=c(1,-1), col="grey90")
abline(h=c(-2,2), col="grey70")

numero3 = length(which(abs(pDi)>0.6))
lim4=c(min(res.di)-0.5, max(res.di)+0.5)
plot(res.di,pDi, xlim=lim4, xlab="", ylab="", ylim=c(0,1), main="(d)", cex.lab=1.2,
     lwd=2, cex.axis=1.2)
showLabels(res.di, pDi, labels=index, id.method=list("y"), id.n = numero3, id.cex=1.2)
x <- seq(-10,10,length.out=500)
lines(x,c1[1]-x^2, lty=2)
lines(x,c1[2]-x^2, lty=2)
lines(x,c1[3]-x^2, lty=2)

return(list(respad.post=media.residuo, respad.cpo=as.numeric(residuo.pad),
       respad.dev=as.numeric(res.pad.di), res.dev=as.numeric(res.di),
       leverage=pDi, DIC=DIC))
}

Exemplo:

Especifique conforme abaixo:

dados <- cbind(y,n,v,1,x1,x2)
         y: variáveis resposta
         n: número de indivíduos em cada cluster
         v: valor da função dos indivíduos
         1, x1, x2: lista de covariáveis dos clusters

fun.p = 1 - função de ligação logito
        2 - função de ligação complementar log-log
        3 - função de ligação log-log
        4 - função de ligação probito
```

```
fun.rho = 1 - estrutura de correlação exponencial
          2 - estrutura de correlação AR contínua
          3 - estrutura de correlação Gaussiana

theta.final: matriz de valores provientes do MCMC,
             considerando o burn in e os saltos necessários

c1: valores de contribuição no valor global do DIC

hiper.var: valores dos hiperparâmetros da priori para cada
           parâmetro

resb.MRBC(dados=cbind(y,n,d,1,x1,x2), fun.p=1, fun.rho=1, theta.final,
                  c1=c(1,2,4), hiper.var=rep(10000,4))

Saída:
- respad.post: amostra da distribuição a posteriori dos resíduos padronizados
- respad.cpo: valores dos resíduos padronizados baseado na CPO
- respad.dev: valores dos resíduos padronizados baseado na deviance
- res.dev: valores dos resíduos baseado na deviance
- leverage: alavancagem das obervações
- DIC: valor de critério DIC
```

D.4 Divergência de K-L

Função para obter os valores da divergência de K-L e a calibração da divergência de K-L, conforme descrito na Seção 4.3.2. Para rodar esta função é necessário carregar a função CPO.MRBC, mostrada no Apêndice D.2.

```
DKL.MRBC <- function(dados, fun.p, fun.rho, theta.final){

dados=cbind(y,n,v,1,x1,x2)
fun.p=1
fun.rho=1
y <- dados[,1]
n <- dados[,2]
v <- dados[,3]
X <- dados[,4:ncol(dados)]
m <- nrow(dados)
n.cad <- nrow(theta.final)

IA2 <- rep(0,m)
IA2[y==n|y==0]<-1

g <- function(beta,y,n,v,X,IA2) {
preditor = X%*%beta[-1]
if(fun.p==1){ p = exp(preditor)/(1+exp(preditor)) }
```

```
if(fun.p==2){ p = 1 - exp(-exp(preditor)) }
if(fun.p==3){ p = exp(-exp(-preditor)) }
if(fun.p==4){ p = pnorm(preditor) }
if(fun.rho==1) { rho = exp(-exp(beta[1])*v) }
if(fun.rho==2) { rho = (exp(beta[1])/(1+exp(beta[1])))^v }
if(fun.rho==3) { rho = exp(-(exp(beta[1])*v)^2) }
lli = choose(n,y)*p^(y)*(1-p)^(n-y)*(1-rho) + p^(y/n)*(1-p)^((n-y)/n)*rho*IA2
return(lli)
}

log.gt <- matrix(ncol=m, nrow=n.cad,0)

for (j in 1:n1) {
log.gt[j,] <- -2*log(g(theta.final[j,],y,n,v,X,IA2))
}

res1 <- CPO.MRBC(dados, fun.p, fun.rho, theta.final)

DKL = -log(res1$CPOi) + apply((log.gt/(-2)),2,mean)
cal = 0.5*(1+sqrt(1-exp(-2*DKL)))


par(mfrow=c(1,2), mar=c(1.5,1.5,1.5,1.5))
lim=c(0, max(DKL)+1)
plot(DKL, xlab="", ylab="divergência K-L", main="(a)", type="h",ylim=lim)
plot(cal, xlab="", ylab="Calibração", main="(b)", type="h",ylim=c(0.5,1))

return(list(DKL=DKL, CAL=cal))
}

Exemplo:

Especifique conforme abaixo:

dados <- cbind(y,n,v,1,x1,x2)
         y: variáveis resposta
         n: número de indivíduos em cada cluster
         v: valor da função dos indivíduos
         1, x1, x2: lista de covariáveis dos clusters

fun.p = 1 - função de ligação logito
        2 - função de ligação complementar log-log
        3 - função de ligação log-log
        4 - função de ligação probito

fun.rho = 1 - estrutura de correlação exponencial
          2 - estrutura de correlação AR contínua
          3 - estrutura de correlação Gaussiana

theta.final: matriz de valores provientes do MCMC,
             considerando o burn in e os saltos necessários
```

```
DKL.MRBC(dados=cbind(y,n,v,1,x1,x2), fun.p=1, fun.rho=1, theta.final)

Saída:
- gráficos da divergência de K-L e da calibração da divergência de K-L
- DKL: valor da divergência de K-L para cada cluster
- CAL: calibração da divergência de K-L para cada cluster
```

Apêndice E

Programas para o MRBCAEN

Neste capítulo, apresentamos os principais programas computacionais utilizados no MRBCAEN via abordagem clássica. As funções foram implementadas em R (R Development Core Team, 2011) para obter os estimadores em máxima verossimilhança, os intervalos de confiança discutidos no Capítulo 6. As demais rotinas, não apresentados aqui, são muito similares ao desenvolvidos no Apêndice C.

E.1 Estimadores de máxima verossimilhança

Função para obter os estimadores de máxima verossimilhança, $\hat{\boldsymbol{\theta}}$, do vetor de parâmetros $\boldsymbol{\theta}$, usando o algoritmo EM, conforme descrito na Seção 6.3.

```
EMV.MRBCEAN <- function(dados, fun.rho, chutes, b, erro = 1e-08){

BETAS = as.matrix(chutes)
y <- dados[,1]
n <- dados[,2]
v <- dados[,3]
w <- dados[,4]
X <- dados[,5:ncol(dados)]

vero <- function(beta){
hatx <- (b*beta[3]+w)/(1+b)
preditor = X%*%(as.matrix(beta[5:ncol(dados)])) + beta[2]*hatx
p = exp(preditor)/(1+exp(preditor))
if(fun.rho==1) { rho = exp(-exp(beta[1])*v) }
if(fun.rho==2) { rho = (exp(beta[1])/(1+exp(beta[1])))^v }
if(fun.rho==3) { rho = exp(-(exp(beta[1])*v)^2) }
ll = sum ( -(1/2)*log(exp(beta[4]))-(1/(2*exp(beta[4])))*((w^2)/b+beta[3]^2-
            ((b*beta[3]+w)^2)/(b*(1+b)))+z*log(rho)+(1-z)*log(1-rho)+
             log(p^((y/n)*(z+n-n*z)))+log((1-p)^((n-y)*(z/n+1-z))) )
return(ll)
}

dif <- 0.1
```

```
while(dif > erro) {
hatx <- (b*BETAS[3]+w)/(1+b)
preditor = X%*%(as.matrix(BETAS[5:ncol(dados)])) + BETAS[2]*hatx
p = exp(preditor)/(1+exp(preditor))
if(fun.rho==1) { rho = exp(-exp(BETAS[1])*v) }
if(fun.rho==2) { rho = (exp(BETAS[1])/(1+exp(BETAS[1])))^v }
if(fun.rho==3) { rho = exp(-(exp(BETAS[1])*v)^2) }
v1 <- rho*p^(y/n)*(1-p)^((n-y)/n)
v2 <- (1-rho)*choose(n,y)*p^y*(1-p)^(n-y)
z <-  v1 /(v1+v2)
z[!((y==0)|(y==n))]=0

res <- optim(BETAS, vero, method="BFGS", hessian=T, control=list(fnscale=-1))

if(fun.rho==1) { dif.gama <- abs(exp(BETAS[1]) - exp(res$par[1])) }
if(fun.rho==2) { dif.gama <- abs(exp(BETAS[1])/
                          (1+exp(BETAS[1])) - exp(res$par[1])/(1+exp(res$par[1]))) }
if(fun.rho==3) { dif.gama <- abs(exp(BETAS[1]) - exp(res$par[1])) }
dif.delta <- abs(BETAS[2] - res$par[2])
dif.mu <- abs(BETAS[3] - res$par[3])
dif.sigma2 <- abs(exp(BETAS[4]) - exp(res$par[4]))
dif.betas <- abs(BETAS[5:ncol(dados)] - res$par[5:ncol(dados)])

dif <- max(dif.gama, dif.delta, dif.mu, dif.sigma2, dif.betas)
BETAS <- res$par

if(fun.rho==1) { EMV <- c(exp(BETAS[1]), BETAS[2:3], exp(BETAS[4]), BETAS[5:ncol(dados)]) }
if(fun.rho==2) { EMV <- c(exp(BETAS[1])/(1+exp(BETAS[1])), BETAS[2:3],
                          exp(BETAS[4]), BETAS[5:ncol(dados)]) }
if(fun.rho==3) { EMV <- c(exp(BETAS[1]), BETAS[2:3], exp(BETAS[4]), BETAS[5:ncol(dados)]) }

EMV <- matrix(EMV, ncol=(ncol(X)+4))
colnames(EMV) <- c("gama", "delta", "mu", "sigma2", replicate(ncol(X),"betas"))
}
return(EMV)
}

Exemplo:

Especifique conforme abaixo:

dados <- cbind(y,n,v,w,1,x1,x2)
         y: variáveis resposta
         n: número de indivíduos em cada cluster
         v: valor da função dos indivíduos
         w: covariável medida com erro
         1, x1, x2: lista de covariáveis dos clusters

fun.rho = 1 - estrutura de correlação exponencial
          2 - estrutura de correlação AR contínua
```

```
          3 - estrutura de correlação Gaussiana

chutes: lista de valores iniciais dos parâmetros

b: proporção de erro observado em relação a variância da covariável não observada

erro: erro admissível no critério de parada do algoritmo

EMV.MRBCEAN(dados=cbind(y,n,v,w,1,x1), fun.rho=1, b=1, chutes=c(0,0,0,0,0,0))

Saída:
- EMV: vetor de estimadores de máxima verossimilhança
```

E.2 Intervalos de confiança

Função para obter os intervalos de confiança para o vetor de parâmetros $\boldsymbol{\theta}$, usando as três formas de construção dos intervalos de confiança descritos na Seção 6.4.

```
IC.MRBCEAN <- function(dados, fun.rho, nsim.mc, nsim.boot,
                          nsim.perf, EMV, conf=0.95, digt=3, b) {

y <- dados[,1]
n <- dados[,2]
v <- dados[,3]
w <- dados[,4]
X <- dados[,5:ncol(dados)]
m <- nrow(dados)

# Verossimilhança aumentada

J_bba <- function(EMV,xs,xr,pos1,pos2) {
hatx <- (b*EMV[3]+w)/(1+b)
preditor = X%*%(as.matrix(EMV[5:ncol(dados)]))+EMV[2]*hatx
p = exp(preditor)/(1+exp(preditor))
Alog <- exp(preditor)*(1+exp(preditor))^(-2)
Blog <- -exp(preditor)*(exp(preditor)-1)*(1+exp(preditor))^(-3)
Clog <- (exp(preditor)+1)/(exp(preditor)-1)

dp1s <- 0
if(pos1=='betas') { dp1s = xs*Alog }
if(pos1=='delta') { dp1s = ((b*EMV[3]+w)/(1+b))*Alog }
if(pos1=='mu')    { dp1s = (b*EMV[2]/(1+b))*Alog }

dp1r <- 0
if(pos2=='betas') { dp1r = xr*Alog }
if(pos2=='delta') { dp1r = ((b*EMV[3]+w)/(1+b))*Alog }
if(pos2=='mu')    { dp1r = (b*EMV[2]/(1+b))*Alog }
```

```
dp2 <- 0
if((pos1=='betas')&(pos2=='betas')) { dp2 = xr*xs*Blog }
if((pos1=='betas')&(pos2=='delta')) { dp2 = ((b*EMV[3]+w)/(1+b))*xs*Blog }
if((pos1=='betas')&(pos2=='mu'))    { dp2 = (b*EMV[2]/(1+b))*xs*Blog }
if((pos1=='delta')&(pos2=='delta')) { dp2 = ((b*EMV[3]+w)^2/(1+b)^2)*Blog }
if((pos1=='delta')&(pos2=='mu'))    { dp2 = (b/(1+b)^2)*(EMV[2]*(b*EMV[3]+w)
                                          -(1+b)*Clog)*Blog }
if((pos1=='mu')&(pos2=='mu'))       { dp2 = ((b^2*EMV[2]^2)/(1+b)^2)*Blog }

f2 <- 0
if((pos1=='mu')&(pos2=='mu'))         { f2 = -(1/(EMV[4]*(1+b))) }
if((pos1=='mu')&(pos2=='sigma2'))     { f2 = (EMV[3]-w)/(EMV[4]^2*(1+b)) }
if((pos1=='sigma2')&(pos2=='sigma2')) { f2 = (EMV[4]*(1+b)-2*w^2-2*EMV[3]^2+4*w*EMV[3])/
                                          (2*EMV[4]^3*(1+b)) }

res  <- -sum(f2+(p^(-1)*dp2-p^(-2)*dp1r*dp1s)*((y/n)*(z+n-n*z))-((1-p)^(-1)*dp2+
            (1-p)^(-2)*dp1r*dp1s)*((n-y)*((z/n)+1-z)))
return(res)
}

J_cova <- function(EMV) {

if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }

if(fun.rho==1) { drho1 = -v*exp(-EMV[1]*v) }
if(fun.rho==2) { drho1 = v*(EMV[1]^v)*EMV[1]^(-1) }
if(fun.rho==3) { drho1 = -2*EMV[1]*v^(2)*exp(-(EMV[1]*v)^2) }

if(fun.rho==1) { drho2 = v^2*exp(-EMV[1]*v) }
if(fun.rho==2) { drho2 = (EMV[1]^(v-2))*v*(v-1) }
if(fun.rho==3) { drho2 = 2*v^2*exp(-(EMV[1]*v)^2)*(2*(EMV[1]*v)^2-1) }

res <- -sum(z*((rho^(-1))*drho2 - (rho^(-2))*drho1^2)-(1-z)*
           (((1-rho)^(-1))*drho2+((1-rho)^(-2))*drho1^2))
return(res)
}

U_bba <- function(EMV,xs,pos1) {
hatx <- (b*EMV[3]+w)/(1+b)
preditor = X%*%(as.matrix(EMV[5:ncol(dados)])) + EMV[2]*hatx
p = exp(preditor)/(1+exp(preditor))
Alog <- exp(preditor)*(1+exp(preditor))^(-2)

dp1s <- 0
if(pos1=='betas') { dp1s = xs*Alog }
if(pos1=='delta') { dp1s = ((b*EMV[3]+w)/(1+b))*Alog }
if(pos1=='mu')    { dp1s = (b*EMV[2]/(1+b))*Alog }

f1 <- 0
```

```
if(pos1=='mu')     { f1 = (w-EMV[3])/(EMV[4]*(1+b)) }
if(pos1=='sigma2') { f1 = (-EMV[4]*(1+b)+w^2-2*w*EMV[3]+EMV[3]^2)/
                          (2*EMV[4]^2*(1+b)) }

res  <- sum(f1+dp1s*((y/n)*(z+n-n*z)*p^(-1)-(n-y)*((z/n)+1-z)*(1-p)^(-1)))
return(res)
}

U_cova <- function(EMV) {

if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }

if(fun.rho==1) { drho1 = -v*exp(-EMV[1]*v) }
if(fun.rho==2) { drho1 = v*(EMV[1]^v)*EMV[1]^(-1) }
if(fun.rho==3) { drho1 = -2*EMV[1]*v^(2)*exp(-(EMV[1]*v)^2) }

res <- sum(z*((rho^(-1))*drho1)-(1-z)*((1-rho)^(-1)*drho1))
return(res)
}

J_betaa <- matrix(ncol=(ncol(EMV)-1),nrow=(ncol(EMV)-1),0)
colnames(J_betaa) <- c("delta", "mu", "sigma2",replicate(ncol(X),"betas"))
rownames(J_betaa) <- c("delta", "mu", "sigma2",replicate(ncol(X),"betas"))

hatx <- (b*EMV[3]+w)/(1+b)
preditor = X%*%(as.matrix(EMV[5:ncol(dados)])) + EMV[2]*hatx
p = exp(preditor)/(1+exp(preditor))

if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }

v1 <- rho*p^(y/n)*(1-p)^((n-y)/n)
v2 <- (1-rho)*choose(n,y)*p^y*(1-p)^(n-y)
z <-  v1 /(v1+v2)
z[!((y==0)|(y==n))]=0

X.aux <- cbind(1,1,1,X)

for (i in 1:(ncol(EMV)-1)) {
for (j in i:(ncol(EMV)-1)) {
J_betaa[i,j] <- J_bba(EMV,X.aux[,i],X.aux[,j],
               pos1=colnames(J_betaa)[i],pos2=colnames(J_betaa)[j])
J_betaa[j,i] <- J_betaa[i,j]
}}

Jaux <- matrix(ncol=ncol(EMV), nrow=ncol(EMV), 0)
Jaux[2:ncol(EMV),2:ncol(EMV)] <- J_betaa
Jaux[1,1] <- J_cova(EMV)
```

```
U_a <- matrix(nrow=ncol(EMV), ncol=ncol(EMV), 0)

MC.Ua <- list()

for(i in 1:nsim.mc){ MC.Ua[[i]] <- U_a }

hatx <- (b*EMV[3]+w)/(1+b)
preditor = X%*%(as.matrix(EMV[5:ncol(dados)])) + EMV[2]*hatx
p = exp(preditor)/(1+exp(preditor))
Alog <- exp(preditor)*(1+exp(preditor))^(-2)

if(fun.rho==1) { rho = exp(-EMV[1]*v) }
if(fun.rho==2) { rho = EMV[1]^v }
if(fun.rho==3) { rho = exp(-(EMV[1]*v)^2) }

v1 <- rho*p^(y/n)*(1-p)^((n-y)/n)
v2 <- (1-rho)*choose(n,y)*p^y*(1-p)^(n-y)
prob.z <- v1 / (v1+v2)

for (q in 1:nsim.mc){

z <- sapply(prob.z, rbinom, n=1, size=1)
z[!((y==0)|(y==n))]=0

U_a[1,1] <- U_cova(EMV)*U_cova(EMV)

for (k in 1:ncol(EMV)){
for (i in 1:ncol(EMV)){

if(i>1 & (k>=i)) {
U_a[i,k] <- U_bba(EMV,X.aux[,i-1],pos1=colnames(J_betaa)[i-1])*
            U_bba(EMV,X.aux[,k-1],pos1=colnames(J_betaa)[k-1])
U_a[k,i] <- U_a[i,k] }

else { if(i==1 & k>1)
{ U_a[i,k] <- U_cova(EMV)*U_bba(EMV,X.aux[,k-1],pos1=colnames(J_betaa)[k-1]) ;
              U_a[k,i] <- U_a[i,k] }}
}}


MC.Ua[[q]] <- U_a
}

MC.Ua.m <- matrix(ncol=ncol(EMV), nrow=ncol(EMV),0)
aux <- numeric(nsim.mc)

for (i in 1:ncol(EMV)){
for (j in 1:ncol(EMV)){
   for (k in 1:nsim.mc){
   aux[k] <- MC.Ua[[k]][i,j]
```

```
   }
MC.Ua.m[i,j] <- mean(aux)
}}

Ja <- Jaux - MC.Ua.m

varbetaa <- diag(solve(Ja))

# Bootstrap não-paramétrico

BETAS = as.matrix(EMV)

EMVS <- matrix(ncol=ncol(EMV),nrow=nsim.boot,0)

for (j in 1:nsim.boot){

ordem <- sample(1:m, m, replace=T)
dados_novo = dados
for (i in 1:m){ dados_novo[i,] <- dados[ordem[i],] }

EMVS[j,] <- EMV.MRBCEAN(dados=dados_novo, fun.rho=fun.rho, chutes=EMV, b)
}

# Perfilado

vero.a.usual <- function(beta){
IA2 <- rep(0,m)
IA2[(y==n)|(y==0)] <- 1
hatx <- (b*beta[3]+w)/(1+b)
preditor = X%*%(as.matrix(beta[5:ncol(dados)])) + beta[2]*hatx
p = exp(preditor)/(1+exp(preditor))
if(fun.rho==1) { rho = exp(-exp(beta[1])*v) }
if(fun.rho==2) { rho = (exp(beta[1])/(1+exp(beta[1])))^v }
if(fun.rho==3) { rho = exp(-(exp(beta[1])*v)^2) }
bin <- choose(n,y) * p^y * (1-p)^(n-y) * (1-rho)
ber <- p^(y/n) * (1-p)^((n-y)/n) * rho * IA2
aux <- -(1/2)*log(2*pi*exp(beta[4])*(1+b))-(1/(2*exp(beta[4]))*((w^2)/b
       +beta[3]^2-(((b*beta[3]+w)^2)/(b*(1+b)))))
ll <-  aux+log(bin+ber)
return(sum(ll))
}

vero.emvs <- matrix(ncol=(ncol(X)+4),nrow=nsim.perf,0)
teste <- matrix(ncol=(ncol(X)+4),nrow=nsim.perf,0)

EMV.aux <- EMV

if(fun.rho==1) { EMV.aux[1] <- log(EMV[1]) ; EMV.aux[4] <- log(EMV[4]) }
if(fun.rho==2) { EMV.aux[1] <- log(EMV[1]/(1-EMV[1])) ;
                 EMV.aux[4] <- log(EMV[4]) }
if(fun.rho==3) { EMV.aux[1] <- log(EMV[1]) ; EMV.aux[4] <- log(EMV[4]) }
```

```
EMV.a <- EMV.aux

for (j in 1:(ncol(X)+4)){
for (i in 1:nsim.perf){
teste[,j] <- seq(EMV[j]-8, EMV[j]+8, length.out=nsim.perf)
EMV.aux <- EMV.a
EMV.aux[j] <- teste[i,j]
vero.emvs[i,j] <- vero.a.usual(EMV.aux)
}}

dif <- vero.emvs - (vero.a.usual(EMV.a) - (1/2)*qchisq(conf,1))

IC <- matrix(ncol=2, nrow=(ncol(X)+4), 0)
for (i in 1:(ncol(X)+4)) {
IC[i,] <- teste[c(which( c(sign(dif[,i]),0) * c(0,sign(dif[,i])) ==-1)),i]
}

if(fun.rho==1) { IC[1,] <- exp(IC[1,]) ; IC[4,] <- exp(IC[4,]) }
if(fun.rho==2) { IC[1,] <- exp(IC[1,])/(1+ exp(IC[1,])) ; IC[4,] <- exp(IC[4,]) }
if(fun.rho==3) { IC[1,] <- exp(IC[1,]) ; IC[4,] <- exp(IC[4,]) }

# resumos

confiança <- c((1-conf)/2, 1-(1-conf)/2)

resumo <- matrix(ncol=7, nrow=ncol(EMV))

resumo[,1] <- EMV
resumo[,2] <- EMV + qnorm((1-conf)/2)*sqrt(varbetaa)
resumo[,3] <- EMV + qnorm(1-(1-conf)/2)*sqrt(varbetaa)
resumo[,4] <- apply(EMVS,2,quantile, probs=(1-conf)/2)
resumo[,5] <- apply(EMVS,2,quantile, probs=1-(1-conf)/2)
resumo[,6] <- IC[,1]
resumo[,7] <- IC[,2]

colnames(resumo) <- paste(c("EMV", "ICi_a", "ICs_a", "ICi_b", "ICs_b",
                            "ICi_p", "ICs_p"))
rownames(resumo) <- paste(c("gamma", "delta", "mu", "sigma2",
                          replicate((ncol(EMV)-4),"betas")))

return(round(resumo,digt))
}

Exemplo:

Especifique conforme abaixo:

dados <- cbind(y,n,v,w,1,x1,x2)
         y: variáveis resposta
         n: número de indivíduos em cada cluster
```

```
         v: valor da função dos indivíduos
         w: covariável medida com erro
         1, x1, x2: lista de covariáveis dos clusters

fun.rho = 1 - estrutura de correlação exponencial
          2 - estrutura de correlação AR contínua
          3 - estrutura de correlação Gaussiana

nsim.mc: número de simulações Monte Carlo

nsim.boot: número de simulações bootstrap

nsim.perf: número de valores do parâmetro para o intervalo perfilado

EMV <- EMV.MRBCEAN(dados=cbind(y,n,v,w,1,x1), fun.rho=1, chutes=c(0,0,0,0,0,0))

conf: nível de confiança para construção dos intervalos

digt: número de casas decimáis na saída

b: proporção de erro observado em relação a variância da covariável não observada

IC.MRBCEAN(dados = cbind(y,n,v,w,1,x1), fun.rho=1, nsim.mc=3000,
          nsim.boot=100, nsim.perf=5000, EMV, conf=0.95, digt=3, b=1)

Saída:
- EMV: vetor de estimadores de máxima verossimilhança
- (ICi_a,ICs_a): intervalo de confiança assintótico construído
                 com a verossimilhança aumentada
- (ICi_b,ICs_b): intervalo de confiança bootstrap
- (ICi_p,ICs_p): intervalo de confiança perfilado
```