UNIVERSIDADE FEDERAL DE MINAS GERAIS

INSTITUTO DE CIÊNCIAS EXATAS

PROGRAMA DE PÓS-GRADUAÇÃO DO DEPARTAMENTO DE ESTATÍSTICA

EDUARDO FERNANDES E SILVA

Regressão logística binária com fator de cura para dados em conglomerados: Uma aplicação em traumatismo dentário

Belo Horizonte

2021

EDUARDO FERNANDES E SILVA

Regressão logística binária com fator de cura para dados em conglomerados: Uma aplicação em traumatismo dentário

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Estatística.

Orientador: Prof. Dr. Enrico Antônio Colosimo

Coorientador: Profª Dr. Juliana Vilela Bastos

Belo Horizonte

2021

Silva, Eduardo Fernandes e.

S586r Regressão logística binária com fator de cura para dados em conglomerados: uma aplicação em traumatismo dentário [manuscrito] / Eduardo Fernandes e Silva. - 2021.
53 f. il.

Orientador: Enrico Antônio Colosimo
Coorientadora: Juliana Vilela Bastos
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento Estatística.
Referências: f.42-43.
.
1. Estatística – Teses. 2. Análise de regressão – Teses. 3. Traumatismos dentários/realbilitação– Teses. 4. Cura – Estatística – Teses. I. Colosimo, Enrico Antônio. II. Bastos, Juliana Vilela. III. Universidade Federal de Minas Gerais; Instituto de Ciências Exatas, Departamento de Estatística. IV. Título.
CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6ª Região nº 1510

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

ATA DA DEFESA DE DISSERTAÇÃO DE MESTRADO DO ALUNO Eduardo Fernandes e Silva, MATRICULADO, SOB O Nº 2019661394, NO PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA, DO INSTITUTO DE CIÊNCIAS EXATAS, DA UNIVERSIDADE FEDERAL DE MINAS GERAIS, REALIZADA NO DIA 16 DE JUNHO DE 2021.

Aos 16 dias do mês de Junho de 2021, às 14h00, em reunião pública virtual 260 (conforme orientações para a atividade de defesa de dissertação durante a vigência da Portaria PRPG nº 1819) no Instituto de Ciências Exatas da UFMG, https://us02web.zoom.us/j/82695718134?pwd=a2M0bVVZeWFjVFp1djR2bGxuckhydz09, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pelo Colegiado do Programa de Pós-Graduação em Estatística, para julgar a defesa de dissertação do aluno Eduardo Fernandes e Silva, nº matrícula 2019661394, intitulada: "*Regressão logística binária com fator de cura para dados em conglomerados: Uma aplicação em traumatismo dentário"*, requisito final para obtenção do Grau de mestre em Estatística. Abrindo a sessão, o Senhor Presidente da Comissão, Prof. Enrico Antônio Colosimo, passou a palavra ao aluno para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do aluno. Após a defesa, os membros da banca examinadora reuniram-se reservadamente sem a presença do aluno e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação:

(X) Aprovada.

( ) Reprovada com resubmissão do texto em_____dias.

( ) Reprovada com resubmissão do texto e nova defesa em_____dias.

( ) Reprovada.

Prof. Enrico Antônio Colosimo –Orientador (EST/UFMG)

Profa. Juliana Vilela Bastos
Co-orientadora (Faculdade de Odontologia da UFMG)

Profa. Suely Ruiz Giolo
(DEST/UFPR)

Prof. Fábio Nogueira Demarqui
(DEST/UFMG)

Frederico Machado Almeida (Doutorando/UFMG)

O resultado final foi comunicado publicamente ao(à) aluno(a) pelo(a) Senhor(a) Presidente da Comissão. Nada mais havendo a tratar, o(a) Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 16 de Junho de 2021.

Observações: 1. No caso de aprovação da tese, a banca pode solicitar modificações a serem feitas na versão final do texto. Neste caso, o texto final deve ser aprovado pelo orientador da tese. O pedido de expedição do diploma do candidato fica condicionado à submissão e aprovação, pelo orientador, da versão final do texto.

2. No caso de reprovação da tese com resubmissão do texto, o candidato deve submeter o novo texto dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca que então decidirão pela aprovação ou reprovação da tese.

3. No caso de reprovação da tese com resubmissão do texto e nova defesa, o candidato deve submeter o novo texto com a antecedência à nova defesa que o orientador julgar adequada. A nova defesa, mediante todos os membros da anca, deve ser realizada dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca. Baseada no novo texto e na nova defesa, a banca decidirá pela aprovação ou reprovação da tese.

Resumo

Estudos transversais na área da saúde usualmente possuem desfecho binário e a regressão logística é a primeira opção para responder as perguntas do pesquisador. No entanto, algumas condições levam a não adequação do modelo, em sua forma padrão, para o tratamento do desfecho binário. A presença de cura, isso é, quando sabemos que uma parcela desconhecida da população não está mais em risco de desenvolver o evento de interesse, é um exemplo deste tipo de situação. O presente trabalho foi desenvolvido na linha de pesquisa "Metodologia e estatística na pesquisa em traumatismos dentários", parceria estabelecida, desde 2015, entre o Programa Traumatismos Dentários da Faculdade de Odontologia da UFMG (PTD FAO UFMG) e o Departamento de Estatística do ICEx-UFMG. O principal interesse dos pesquisadores foi estabelecer fatores de risco para a presença de Reabsorção Radicular Externa Inflamatória (RREI). A regressão logística foi a metodologia indicada para o estudo da associação entre fatores clínicos e radiográficos, medidos na primeira consulta do paciente no PTD FAO UFMG, e o desfecho de interesse, a RREI. Entretanto, a RREI só é esperada naqueles casos em que há necrose pulpar e infecção do canal radicular, ou seja, dentes cuja cicatrização pulpar for favorável não estão sob risco de desenvolver RREI. Como esta definição não é possível na consulta inicial, ou seja, no momento da coleta dos dados, caracteriza-se a presença de fração de cura latente, e que pode levar à inadequação do modelo logístico usual. Diop et al. (2011) afirmam que o problema de cura com resposta binária pode ser encarado como um modelo ZIB (Zero-inflated Binomial). Considerando-se que na casuística do projeto, um mesmo indivíduo pode apresentar mais de um dente traumatizado, o desafio do projeto foi ajustar um modelo logístico binário com fator de cura na presença de conglomerados. Para tanto, foi utilizado a metodologia apresentada por Hall e Zhang (2004), em que os autores flexibilizam o algoritmo **EM** (expectation-maximization) para acomodação de mais de uma medição por indivíduo em modelos zero inflacionados.

Neste trabalho, apresentamos o modelo de regressão logística com fator de cura latente, que tem a mesma forma do binomial inflacionados de zeros. Em seguida, estendemos o modelo para acomodar conglomerados, que representam um ou mais dentes por paciente, e, finalmente, apresentamos a aplicação desta metodologia para analisar a casuística da FO-UFMG.

Palavras chaves: Regressão Logística. ZIB. Fator de cura. Conglomerados.

Abstract

Cross-sectional studies in the health area usually have a binary outcome and logistic regression is the first option to answer the researcher's questions. However, some conditions lead to the model not being adequate, in its standard form, for the treatment of the binary outcome. The presence of a cure, that is, when we know that an unknown portion of the population is no longer at risk of developing the event of interest, is an example of this type of situation. This work was developed in the line of research "Methodology and statistics in research on dental trauma", a partnership established since 2015 between the Dental Trauma Program of the UFMG School of Dentistry (PTD FAO UFMG) and the ICEx-Statistics Department. UFMG. The researchers' main interest was to establish risk factors for the presence of Inflammatory External Root Resorption (IRR). Logistic regression was the methodology indicated for the study of the association between clinical and radiographic factors, measured in the patient's first consultation at the PTD FAO UFMG, and the outcome of interest, the RREI. However, RREI is only expected in those cases in which there is pulp necrosis and root canal infection, that is, teeth whose pulp healing is favorable are not at risk of developing RREI. As this definition is not possible in the initial consultation, that is, at the time of data collection, the presence of a latent cure fraction is characterized, which can lead to the inadequacy of the usual logistic model. Diop et al. (2011) claim that the binary response healing problem can be seen as a ZIB (Zero-inflated Binomial) model. Considering that in the project's casuistry, the same individual may have more than one traumatized tooth, the project challenge was to adjust a binary logistic model with a cure factor in the presence of conglomerates. For this purpose, the methodology presented by Hall e Zhang (2004), in which the authors make the EM algorithm more flexible to accommodate more than one measurement per individual in zero-inflated models.

In this work, we present the logistic regression model with latent cure factor, which has the same shape as the zero-inflated binomial. Then, we extend the model to accommodate clusters, which represent one or more teeth per patient, and, finally, we present the application of this methodology to analyze the case series at FO-UFMG.

Keywords: Logistic Regression. ZIB. Healing Factor. Cluster.

Sumário

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 INTRODUÇÃO

Um dos delineamentos mais frequentes na área da saúde consiste no estudo transversal, pois é uma ferramenta de grande utilidade para a descrição de características da população de interesse para a identificação de grupos de risco. Nesta modalidade de estudo investiga-se a "causa" e "efeito" de maneira simultânea e busca-se averiguar a associação existente entre a exposição e a doença. Fatores clínicos, demográficos, e terapêuticos entre outros, são candidatos a explicar as fontes de variabilidade dos desfechos estudados. Diante de respostas do tipo binária o modelo de regressão logística é o mais comum na literatura para se medir a associação entre os fatores medidos e o desfecho de interesse. Outros modelos podem ser aplicados nesta situação, mas a utilização da função de ligação *logit* traz a vantagem da interpretação por razão de chances, o que é de mais fácil entendimento para profissionais que não são da área estatística.

Algumas causas podem ser identificadas na literatura para a inadequação do modelo logístico, requerendo uma extensão do mesmo ou uso de um alternativo. Hall (2000) apresenta o modelo ZIB (Zero-Inflated Binomial) para o ajuste de dados binários nos casos em que há excessos de zeros para a distribuição Binomial e sua extensão incluindo efeitos aleatórios. Meyer e Mittag (2017) e Pires e Quinino (2019) incorporaram erros de classificações desconhecidos como parâmetros adicionais na função de verossimilhança. Hall e Zhang (2004) apresentam duas alternativas para lidar com modelos zero inflacionados envolvendo mais de uma medição para o mesmo indivíduo: flexibilizar o passo de maximização do algortimo **EM** e estender o GEE (Generalized Estimating Equations) para acomodar situações zero inflacionadas. Uma possível causa da inadequação do modelo logístico usual é a possibilidade de "cura". O uso deste modelo é adequado quando o pesquisador reconhece que uma fração desconhecida da população não está sujeita a certa característica ou evento em estudo, sendo importante estudar a parcela não suscetível e analisar o efeitos das covariáveis na resposta de interesse e na fração de cura.

O presente projeto, inserido na linha de pesquisa "Metodologia e estatística na pesquisa em traumatismos dentários", foi desenvolvido através da parceria estabelecida, desde 2015, entre o Programa Traumatismos Dentários da Faculdade de Odontologia da UFMG (PTD FAO UFMG) e o Departamento de Estatística do ICEx-UFMG.

O estudo atual foi motivado pelo interesse prático do PTD FAO UFMG em identificar fatores de risco para Reabsorções Radiculares Externas (RRE) após o reimplante de dentes permanentes avulsionados. A avulsão dentária é uma lesão traumática grave que consiste no deslocamento total do dente de seu alvéolo, causando a completa ruptura do feixe vásculo-nervoso apical, assim como das fibras do ligamento periodontal que ligam a raiz dentária ao osso alveolar. A recolocação imediata do dente no alvéolo, manobra conhecida como reimplante dentário, é o tratamento de escolha, e seu prognóstico no longo prazo apresenta grande variabilidade, pois depende de uma série de fatores relacionados ao manejo do dente avulsionado imediatamente após ao trauma e durante o tratamento emergencial. As RRE representam a principal sequela da cicatrização periodontal após o reimplante de dentes permanentes, podendo levar à perda do dente reimplantado. Estudos anteriores realizados pelo grupo de pesquisa do PTD FAO UFMG (Bastos et al. (2014) e Bastos et al. (2015)) mostraram que fatores demográficos e clínicos relacionados ao tratamento emergencial estão associados à ocorrência de formas graves de RRE. A metodologia dos estudos mencionados consistiu numa abordagem transversal na qual se realizou a coleta de dados clínicos e radiográficos para diagnóstico da presença, tipo e magnitude das RRE. Quando identificadas no exame radiográfico, as cavidades de reabsorção eram classificadas quanto ao tipo como reabsorções radiculares externas inflamatórias (RREI) e reabsorções radiculares externas por substituição (RRES). Para o presente estudo, o desfecho avaliado foi a presença e ausência de RREI, diagnosticada na consulta inicial de pacientes encaminhados para dar continuidade ao tratamento na Clínica de Traumatismos da FAO UFMG (CTD-FO-UFMG) após terem recebido o atendimento emergencial para o reimplante no Pronto Socorro Odontológico do Hospital Metropolitano Odilon Bherens. Considerando que a RREI é uma consequência da necrose pulpar, e consequente infecção do canal radicular, sabe-se que dentes cuja cicatrização pulpar for favorável, não desenvolverão RREI. Entretanto, no momento da coleta de informações na primeira consulta no CTD-FO-UFMG, não é possível definir o padrão de resposta pulpar uma vez que esta só pode ser diagnosticada no longo prazo. Esta característica clínica determina um fator de cura latente com resposta binária, que faz com que uma parcela da população em estudo não seja suscetível a desenvolver o desfecho, no caso a RREI, resultando em uma inadequação do modelo logístico convencional.

Considerando-se que o mesmo paciente pode apresentar mais de um dente reimplantado, as medidas oriundas de unidades amostrais agrupadas em conglomerados, ou seja, no mesmo indivíduo, são correlacionadas entre si e podem trazer vícios nas estimativas das quantidades de interesse. Sendo assim, os objetivos deste estudo são: apresentar o modelo logístico com fator de cura, estender esse modelo para tratar medidas em conglomerados e responder às perguntas clínicas ao identificar fatores associados com o desfecho de interesse. No Capítulo 2, apresentamos o modelo para uma amostra aleatória simples da população e a inferência para as quantidades desconhecidas. No Capítulo 3 abrangemos a extensão para o tratamento da amostra em conglomerados, isso é, com mais de uma medição por indivíduo, com a metodologia proposta por Hall e Zhang (2004). No Capítulo 4 serão discutidos os resultados da aplicação destes dois modelos numa situação real representada pela pesquisa conduzida no PTD FAO UFMG.

2 RESPOSTA BINÁRIA COM FATOR DE CURA: UMA AMOSTRA ALEATÓRIA SIMPLES

Considere uma resposta assumindo dois resultados possíveis, no nosso exemplo, ausência (0) ou presença (1) de RREI, sendo está última o evento de interesse. Um modelo de mistura de fração de cura assume que a população estudada é uma mistura de indivíduos suscetíveis, que experimentam o evento de interesse, e indivíduos não suscetíveis que nunca o experimentam, sendo considerados imunes ou curados. Segundo Maller e Zhou (1996), se em um estudo de sobrevivência existem indivíduos imunes presentes e os dados são modelados com base em modelos convencionais que ignoram a ocorrência destes indivíduos, os resultados podem ser enganosos. Resultados similares são esperados para o modelo de regressão logística com fator de cura.

Na seção 2.1 mostramos as definições básicas do modelo logístico com fator de cura, na seção 2.2 apresentamos a inferência para as quantidades desconhecidas de interesse utilizando a função de máxima verossimilhança observada e discutimos condições de regularidade e identificabilidade. Na seção 2.3 abrangemos uma forma alternativa de se fazer inferência utilizando a verossimilhança completa e o algoritmo **EM** *(Expectation–maximization)* e apresentamos um breve estudo de simulação na seção 2.4.

2.1 Modelo logístico com fator de cura

Considere Y uma varíavel binária. Defina W uma variável aleatória binária latente assumindo 1 caso ocorra certo evento de interesse com probabilidade p_1 e 0 caso contrário, com probabilidade $1 - p_1$, em que $0 < p_1 < 1$. Se $W = 1$, a variável aleatória binária Y é degenerada em 0, ou seja, assume este valor com probabilidade 1. Na situação motivadora deste projeto, W é a indicadora da ocorrência de cicatrização pulpar.

As probabilidades $P(Y=0)$ e $P(Y=1)$ são calculadas como segue:

$$\begin{aligned} P(Y=0) &= \sum_{w=0}^{1} P(Y=0, W=w) \\ &= P(Y=0, W=0) + P(Y=0, W=1) \\ &= P(Y=0 \mid W=0)\, P(W=0) + P(Y=0 \mid W=1)\, P(W=1) \\ &= P(Y=0 \mid W=0)\, P(W=0) + P(W=1) \\ &= P(Y=0 \mid W=0)\,(1-p_1) + p_1. \end{aligned} \tag{1}$$

e

$$\begin{aligned} P(Y=1) &= \sum_{w=0}^{1} P(Y=1, W=w) \\ &= P(Y=1, W=0) + P(Y=1, W=1) \\ &= P(Y=1 \mid W=0)\, P(W=0) + P(Y=1 \mid W=1)\, P(W=1) \\ &= P(Y=1 \mid W=0)\, P(W=0) \\ &= P(Y=1 \mid W=0)\,(1-p_1). \end{aligned} \tag{2}$$

Condicional a $W=0$, a variável aleatória de interesse segue uma distribuição Bernoulli com probabilidade π_1 de sucesso, isso é, $Y=1$, em que $0<\pi_1<1$.
Logo, Y pode ser definida como segue:

$$Y = \begin{cases} 0, & \text{com probabilidade } \quad p_1 \\ \text{Bernoulli}(\pi_1), & \text{com probabilidade } \quad (1-p_1). \end{cases}$$

A partir de (1) e (2), podemos definir Y da seguinte forma:

$$Y = \begin{cases} 0, & \text{com probabilidade } \quad p_1 \ + \ (1-p_1)\,(1-\pi_1) \\ 1, & \text{com probabilidade } \quad (1-p_1)\,\pi_1. \end{cases} \tag{3}$$

A partir da definição em (3) é possível determinar que $E[Y] = (1-p_1)\pi_1$ e $Var[Y] = (1-p_1)(\pi_1)(1-(1-p_1)\pi_1)$.

Na seção 2.2 vamos considerar a inferência no modelo logístico com fator de cura para amostra aleatória simples da população, incluindo covariáveis. Assim como em modelos lineares generalizados, são utilizadas funções de ligações com o objetivo de conectar a

média da distribuição com o preditor linear. De acordo com Yamaguchi et al. (1992), uma das vantagens em se utilizar um modelo de mistura padrão está na facilidade de se incluir variáveis regressoras através de uma função ligação.

2.2 Inferência no Modelo Logístico com fator de cura

Vamos construir a função de verossimilhança em termos de uma amostra aleatória de tamanho n, $\mathbf{Y} = (y_1, ..., y_n)'$.

Sejam $\mathbf{X}$ e $\mathbf{Z}$ matrizes contendo as covariáveis com dimensões $n \times (p+1)$ e $n \times (q+1)$ respectivamente, em que p e q são as quantidades de covariáveis e $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ vetores de dimensões $(p+1) \times 1$ e $(q+1) \times 1$, respectivamente. As matrizes $\mathbf{X}$ e $\mathbf{Z}$ não podem ser iguais, mais detalhes serão discutidos na sub-seção 2.2.1 sobre as condições de identificabilidade.

Considerando a função logística, temos que:

$$\pi_1(x) = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1 + \exp(\mathbf{X}\boldsymbol{\beta})} \tag{4}$$

e

$$p_1(z) = \frac{\exp(\mathbf{Z}\boldsymbol{\gamma})}{1 + \exp(\mathbf{Z}\boldsymbol{\gamma})} \tag{5}$$

Substituindo (4) e (5) em (3), é possível construir a função log verossimilhança como sendo:

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\gamma}; Y) = \sum_{i=1}^{n} \Bigg(& I_{(0)}(y_i) \log \left[\exp(\mathbf{Z}_i\boldsymbol{\gamma}) + \frac{1}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})} \right] \\ & - \log\left(1 \ + \exp(\mathbf{Z}_i\boldsymbol{\gamma})\right) + \left(1 - I_{(0)}(y_i)\right) \left(\mathbf{X}\boldsymbol{\beta} - \log\left(1 + \exp(\mathbf{X}_i\boldsymbol{\beta})\right)\right) \Bigg) \end{aligned} \tag{6}$$

em que $I_{(0)}(y_i)$ é uma função indicadora, que assume valor 1 se $y_i = 0$ e 0, caso contrário.

Hall (2000) define a função log verossimilhança do modelo ZIB (zero-inflated binomial) como sendo:

$$l(\boldsymbol{\beta},\boldsymbol{\gamma};Y)=\sum_{i=1}^{n}\Bigg(I_{(0)}(y_i)\log\left[\exp(\mathbf{Z}_i\boldsymbol{\gamma})+(1+\exp(\mathbf{X}_i\boldsymbol{\beta}))^{-n_i}\right]-\log\left(1\ +\exp(\mathbf{Z}_i\boldsymbol{\gamma})\right)$$
$$+\left(1-I_{(0)}(y_i)\right)\left(y_i\mathbf{X}_i\boldsymbol{\beta}-n_i\log\left(1+\exp(\mathbf{X}_i\boldsymbol{\beta})\right)+\binom{n_i}{y_i}\right)\Bigg) \tag{7}$$

A expressão em (6) é um caso particular da função log-verossimilhança em (7) considerando que todo n_i é igual a 1 e y_i é binário 0 ou 1. Portanto, pode-se dizer que a fração de cura está causando a inflação de zeros, conforme proposto por Diop et al. (2011).

A obtenção das estimativas de máxima verossimilhança e da matriz hessiana podem ser realizadas a partir do algoritmo Escore de Fisher, descrito no Apêndice A, ou por métodos numéricos interativos, como o L-BFGS (Limited-memory BFGS), proposto pela primeira vez em Nocedal (1980). O L-BFGS apresenta vantagens em relação a métodos quase-Newton tradicionais, pois não é necessário resolver um sistema de equações lineares em cada passo, armazenando informações das últimas iterações, pois provavelmente são mais relevantes na construção de aproximações, sendo mais factível em altas dimensões. Sanseverino (2014) destaca que uma das desvantagens do método L-BFGS é a convergência lenta. Mais detalhes sobre os passos do algoritmo L-BFGS podem ser encontrados em Sanseverino (2014).

Seja $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\gamma})'$, $\mathbf{K}\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right)$ o inverso da matriz de informação de Fisher esperada e o interesse em testar $H_0:\boldsymbol{\theta_0}=\boldsymbol{\theta_0}$ versus a alternativa $H_1:\boldsymbol{\theta_0}\neq\boldsymbol{\theta_0}$.

Sob certas condições de regularidade, que serão discutidas na sub-seção 2.2.1, os estimadores de máxima verossimilhança são assintóticamente consistentes e eficientes com a seguinte distribuição:

$$\widehat{\boldsymbol{\theta}}\overset{d}{\longmapsto}N_{p+q+2}\left(\boldsymbol{\theta},\mathbf{K}\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right)\right)$$

O teste de Wald é baseado nesta distribuição assintótica. Seja $\mathbf{K}\left(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}}\right)$ o inverso da matriz de informação de Fisher esperada estimada. (ENGLE, 1984) define que, sob H_0, a estatística:

$$W_d=\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)'\mathbf{K}\left(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}}\right)\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)$$

tem distribuição Qui-Quadrado com $(p+q+2)$ graus de liberdade quando n $\to \infty$.

De forma alternativa, as hipótese apresentadas podem ser concluídas a partir do teste baseado na comparação da função verossimilhança do modelo restrito, sob H_0, com o modelo irrestrito, sob H_1. Sob a hipótese nula, a estatística:

$$TRV = 2\left[l\left(\widehat{\boldsymbol{\theta}}\right) - l\left(\widehat{\boldsymbol{\theta}_0}\right)\right]$$

tem distribuição Qui-Quadrado quando $n \to \infty$, com graus de liberdade igual ao números de restrições em H_0.

2.2.1 Condições de regularidade e identificabilidade

Um problema comum da presença do fator de cura é a impossibilidade de se diferenciar indivíduos suscetíveis e aqueles curados.Ou seja, se $Y_i = 0$, o pesquisador não sabe se o *i-ésimo* indivíduo está na parcela da população em risco, mesmo sem desenvolver o evento, ou naquela curada.

Diop et al. (2011) apresentam quatro condições de regularidade que serão necessárias para garantir a identificabilidade e resultados assintóticos. Uma condição que se difere das tradicionais dos modelos de regressão logística, que é: existe uma covariável contínua V que está em $\mathbf{X}$, mas não em $\mathbf{Z}$. Se β_V e γ_V denotam os coeficientes de V nos dois preditores lineares, então $\beta_V \neq 0$ e $\gamma_V = 0$. De maneira igual, podemos assumir que V está em $\mathbf{Z}$ e não em $\mathbf{X}$.

Essa condição, de acordo Diop et al. (2011), é necessária para a indentificação de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$, e impõe restrições, já que os dois preditores lineares não podem ser iguais.

Pode-se relacionar com o problema de identificabilidade de misturas de modelos de regressão logística, que foi investigado por Follmann e Lambert (1991). Os autores mostraram que misturas finitas de regressões logísticas são identificáveis, desde que o número de combinação de valores observados das covariáveis seja suficientemente grande. Por exemplo: três covariáveis binárias produzem oito combinações possíveis. A condição pode ser vista como suficiente para atingir a exigência, segundo Diop et al. (2011). Condição semelhante aparece em Kelley e Anderson (2008). Neste artigo, os autores dizem que a

mistura de duas Bernoulli é identificável se o número de combinações for pelo menos sete. Se o foco for a comparação de dois grupos, pode-se incluir uma covariável continua em um e apenas um preditor linear, mesmo que não significativa, que o modelo em questão se torna identificável.

Se $\mathbf{X}_i$ for igual a $\mathbf{Z}_i$ e, se trocarmos $\boldsymbol{\beta}$ por $\boldsymbol{\gamma}$ em 6, observamos o mesmo valor de log-verossimilhança, o que segundo Diop et al. (2011), é uma causa de modelo não identificável. A condição elimina essa chance segundo os autores, pois impede que $\mathbf{X}\boldsymbol{\beta}$ e $\mathbf{Z}\boldsymbol{\gamma}$ sejam iguais e, portanto, os parâmetros não são permutáveis.

2.3 Verossimilhança completa e algoritmo **EM**

Uma forma útil de maximização é a utilização do algoritmo **EM**. O artigo de Dempster et al. (1977) apresenta os fundamentos do algoritmo e a eficácia em problemas estatísticos. A estratégia do algoritmo **EM** é realizar a maximização baseada na ideia de dados faltantes ou não observados, considerando a distribuição condicional aos dados realmente observados. A popularidade do algoritmo origina-se de que pode ser simples de implementar, trazendo simplificações para problemas complexos, obtendo com confiabilidade o máximo global através de passos estáveis e crescentes.

Seja $w_i = 1$ quando Y_i assume o valor 0 e $w_i = 0$ caso Y_i segue uma distribuição Bernoulli com probabilidade de sucesso π_1.

Considerando $W = (w_1, w_2, ..., w_n)^T$ como os dados faltantes, a função log-verossimilhança para os dados completos (Y, W) é:

$$\begin{aligned} l_c(\boldsymbol{\gamma}, \boldsymbol{\beta}; y, w) &= \log \prod_{i=1}^{n} P(Y_i = y_i, W_i = w_i) \\ &= \sum_{i=1}^{n} \left[w_i \mathbf{Z}_i \boldsymbol{\gamma} \; - \; \log\left(1 + \exp\left(\mathbf{Z}_i \boldsymbol{\gamma}\right)\right)\right] \\ &+ \sum_{i=1}^{n} (1 - w_i) \left[\left(1 - I_{(0)}(y_i)\right) \mathbf{X}_i \boldsymbol{\beta} - \log\left(1 + \exp\left(\mathbf{X}_i \boldsymbol{\beta}\right)\right)\right] \\ &= l_c(\boldsymbol{\gamma}; w) + l_c(\boldsymbol{\beta}; y, w) \end{aligned} \tag{8}$$

Nota-se que a log-verossimilhança em 8 se fatorou em duas partes, uma que depende apenas de $\boldsymbol{\gamma}$ e outra de $\boldsymbol{\beta}$. A função de log-verossimilhança em 6 não fatora o conjunto de parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$. Esta fatoração, obtida com a função de verossimilhança para dados completos, ajuda no tratamento de mais de uma medição para o mesmo indivíduo, que é o objetivo principal deste trabalho.

No *j-ésimo* passo do algoritmo **EM** é calculado:

$$Q\left(\boldsymbol{\beta},\boldsymbol{\gamma} \mid \boldsymbol{\beta}^{(j)},\boldsymbol{\gamma}^{(j)}\right) = E\left[l_c\left(\boldsymbol{\beta},\boldsymbol{\gamma}; y, w \mid y, \boldsymbol{\beta}^{(j)},\boldsymbol{\gamma}^{(j)}\right)\right] \tag{9}$$

Como $l_c\left(\boldsymbol{\beta},\boldsymbol{\gamma}; y, w\right)$ é linear em relação a w, (HALL; ZHANG, 2004) definem que a expressão em 9 se simplifica em:

$$\begin{aligned} Q\left(\boldsymbol{\beta},\boldsymbol{\gamma} \mid \boldsymbol{\beta}^{(j)},\boldsymbol{\gamma}^{(j)}\right) &= E\left[l_c\left(\boldsymbol{\beta},\boldsymbol{\gamma}; y, w \mid y, \boldsymbol{\beta}^{(j)},\boldsymbol{\gamma}^{(j)}\right)\right] \\ &= l_c\left(\boldsymbol{\gamma}; y, w^{(j)}\right) + l_c\left(\boldsymbol{\beta}; y, w^{(j)}\right) \end{aligned} \tag{10}$$

em que $w^{(j)} = E\left[w \mid y, \boldsymbol{\beta}^{(j)},\boldsymbol{\gamma}^{(j)}\right]$

O algoritmo **EM** é inicializado com um chute inicial $\boldsymbol{\theta}^{(0)}$ e então alterna entre os passos de esperança e maximização, sendo descrito como:

1. Iniciar $\boldsymbol{\theta}^{(0)}$ e fazer $j = 0$
2. **Passo E** = Estimar w_i da média condicional $\hat{w_i}^{(j)} = E\left[w_i \mid y_i, \boldsymbol{\beta}^{(j)},\boldsymbol{\gamma}^{(j)}\right]$.
   Tem-se que:

$$\begin{aligned} E\left[w_i \mid y, \boldsymbol{\beta}^{(j)},\boldsymbol{\gamma}^{(j)}\right] &= E\left[w_i \mid y = 0, \boldsymbol{\beta}^{(j)},\boldsymbol{\gamma}^{(j)}\right] I_{(0)}(y_i) \\ &\quad + E\left[w_i \mid y = 1, \boldsymbol{\beta}^{(j)},\boldsymbol{\gamma}^{(j)}\right]\left(1 - I_{(0)}(y_i)\right) \end{aligned} \tag{11}$$

   As esperanças em 11 são calculadas via Teorema de Bayes com as estimativas dos parâmetros naquele passo, da seguinte forma:

$$
\begin{aligned}
E\left[w_i \mid y=1, \boldsymbol{\beta}^{(j)}, \boldsymbol{\gamma}^{(j)}\right] &= 0 \times P\left[w_i=0 \mid y=1, \boldsymbol{\beta}^{(j)}, \boldsymbol{\gamma}^{(j)}\right] \\
&\quad + 1 \times P\left[w_i=1 \mid y=1, \boldsymbol{\beta}^{(j)}, \boldsymbol{\gamma}^{(j)}\right] \\
&= P\left[w_i=1 \mid y=1, \boldsymbol{\beta}^{(j)}, \boldsymbol{\gamma}^{(j)}\right] \\
&= \frac{P\left[y=1 \mid w_i=1, \boldsymbol{\beta}^{(j)}, \boldsymbol{\gamma}^{(j)}\right] P\left[w_i=1\right]}{P\left[y=1 \mid \boldsymbol{\beta}^{(j)}, \boldsymbol{\gamma}^{(j)}\right]} \\
&= 0
\end{aligned}
$$

e

$$
\begin{aligned}
E\left[w_i \mid y=0, \boldsymbol{\beta}^{(j)}, \boldsymbol{\gamma}^{(j)}\right] &= 0 \times P\left[w_i=0 \mid y=0, \boldsymbol{\beta}^{(j)}, \boldsymbol{\gamma}^{(j)}\right] \\
&\quad + 1 \times P\left[w_i=1 \mid y=0, \boldsymbol{\beta}^{(j)}, \boldsymbol{\gamma}^{(j)}\right] \\
&= P\left[w_i=1 \mid y=0, \boldsymbol{\beta}^{(j)}, \boldsymbol{\gamma}^{(j)}\right] \\
&= \frac{P\left[y=0 \mid w_i=1, , \boldsymbol{\beta}^{(j)}, \boldsymbol{\gamma}^{(j)}\right] P\left[w_i=1\right]}{P\left[y=0 \mid \boldsymbol{\beta}^{(j)}, \boldsymbol{\gamma}^{(j)}\right]} \\
&= \frac{p_1}{p_1 + (1-p_1)(1-\pi_1)}
\end{aligned}
$$

Logo, o *j-ésimo* elemento de $\hat{w_i}^{(j)}$ é dado por:

$$
\hat{w_i}^{(j)} = I_0(y_i)\left[1 + \exp\left(-\mathbf{Z}_i\boldsymbol{\gamma}^{(j)}\right)\left(\frac{1}{1+\exp\left(\mathbf{X}_i\boldsymbol{\beta}^{(j)}\right)}\right)\right]^{-1} \tag{12}
$$

3. **Passo M para $\boldsymbol{\gamma}$** : A estimativa $\boldsymbol{\gamma}^{(j+1)}$ é encontrada maximizando $l_c\left(\boldsymbol{\gamma}; w^{(j)}\right)$ em 8. (HALL, 2000) define que é uma regressão binomial não ponderada com o vetor de respostas $W^{(j)}$.
4. **Passo M para $\boldsymbol{\beta}$** : A estimativa $\boldsymbol{\beta}^{(j+1)}$ é encontrada maximizando $l_c\left(\boldsymbol{\beta}; w^{(j)}\right)$ em 8. Hall (2000) define a expressão como uma regressão logística com pesos $(1 - w_i^{(j)}), i : 1, ..., n$ e o vetor de resposta Y_i
5. Repetir os passos **E** e **M** até convergência de acordo com algum critério pré estabelecido.

Um critério de convêrgencia que pode ser utilizado para parar o algoritmo interativo **EM** é:

$$\max_r \left(\frac{|\boldsymbol{\theta}_r^{j+1} - \boldsymbol{\theta}_r^j|}{|\boldsymbol{\theta}_r^j| + \delta_1} \right) < \delta_2 \tag{13}$$

em que δ_1 e δ_2 são pré especifícados e $max_r A_r$ representa o maior valor de A_r com sub escrito r, em que $r = 1, \ldots, p$. Existem vários critérios conhecidos e já estabelecidos de convergências possíveis. Porém, não é objetivo desta dissertação analisar diferentes métodos de parada e de valores pré especifícados, pois, se o algoritmo está bem implementando, o critério é coerente e os valores necessários para pré especificar são pequenos, os resultados finais devem ser bastante parecidos a partir de certo momento.

A forma usual de obter a variância dos estimadores obtidos via algoritmo **EM** é usar a fórmula de Louis (1982). Seja $C = (Y, W)$ os dados completos e $I_c(\widehat{\theta})$, $I_Y(\widehat{\theta})$ a matriz de informação para dados completos e observados. (LOUIS, 1982) mostrou que:

$$I_Y(\widehat{\boldsymbol{\theta}}) = I_c(\widehat{\boldsymbol{\theta}}) - I_{W|Y}(\widehat{\boldsymbol{\theta}}) \tag{14}$$

A expressão em 14 é interpretada da seguinte forma: a matriz de informação dos dados observados é a diferença entre a informação dos dados completos e dos dados ausentes. A obtenção da matriz de informação é um procedimento complexo pois requer encontrar a distribuição da variável latente condicionada aos dados observados. Sendo assim, Sy e Taylor (2001) propuseram uma alternativa, descrita a seguir. Denote por $l_{obs}(.)$ e l_c as funções de verossimilhanças baseadas nos dados observados e completos, respectivamente. Então,

$$\begin{aligned} \frac{\partial l_{obs}}{\partial \gamma_s} &= \frac{\partial l_c\left(\boldsymbol{\beta}, \boldsymbol{\gamma}, y, w\right)}{\partial \gamma_s} | w_i = w_i^* \\ &= \sum_{i=1}^{n} \mathbf{Z}_{is} \left(w_i^* - p_{1i}\right) \end{aligned}$$

onde w_i^* representa o valor de w_i avaliado nos valores de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$. Para encontrar a matriz de informação, tem-se que:

$$-\frac{\partial^2 l_{obs}}{\partial\gamma_s\partial\gamma_{\,b}} = -\sum_{i=1}^{n}\mathbf{Z}_{is}\left(\frac{\partial w_i^*}{\partial\gamma_b} - \frac{\partial p_{1i}}{\partial\gamma_b}\right) \tag{15}$$

Na expressão em 15:

$$\frac{\partial w_i^*}{\partial\gamma_b} = I_0(y_i)\mathbf{Z}_{ib}\left[\frac{1+\exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)+\exp\left(-\mathbf{Z}_i\boldsymbol{\gamma}\right)}{1+\exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)}\right]^{-2}\frac{\exp\left(-\mathbf{Z}_i\boldsymbol{\gamma}\right)}{1+exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)}$$

e

$$\frac{\partial p_{1i}}{\partial\gamma_b} = \mathbf{Z}_{ib}\left[p_{1i}\left(1-p_{1i}\right)\right]$$

Similarmente,

$$\begin{aligned}\frac{\partial l_{obs}}{\partial\beta_r} &= \frac{\partial l_c\left(\boldsymbol{\beta},\boldsymbol{\gamma},y,w\right)}{\partial\beta_r}|w_i = w_i^* \\ &= \sum_{i=1}^{n}\mathbf{X}_{ir}\left(1-w_i^*\right)\left[y_i-\pi_{1i}\right]\end{aligned}$$

e

$$\begin{aligned}-\frac{\partial^2 l_{obs}}{\partial\beta_r\partial\beta l} = &-\sum_{i=1}^{n}\mathbf{X}_{ir}\left(\frac{\partial\left(1-w_i^*\right)}{\partial\beta_l}\right)\left[y_i-\pi_{1i}\right] \\ &-\sum_{i=1}^{n}\mathbf{X}_{ir}\left(1-w_i^*\right)\left[\frac{\partial\left(y_i-\pi_{1i}\right)}{\partial\beta_l}\right]\end{aligned} \tag{16}$$

Na expressão em 16, as derivadas em função de β_l são:

$$\begin{aligned}\frac{\partial\left(1-w_i^*\right)}{\partial\beta_l} &= -\frac{\partial w_i^*}{\partial\beta_{\,l}} \\ &= -I_0(y_i)\mathbf{X}_{il}\left[1+\frac{\exp\left(-\mathbf{Z}_i\boldsymbol{\gamma}\right)}{1+\exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)}\right]^{-2}\exp\left(-\mathbf{Z}_i\boldsymbol{\gamma}\right)\pi_{1i}\left(1-\pi_{1i}\right)\end{aligned} \tag{17}$$

e

$$\frac{\partial\left(y_i-\pi_{1i}\right)}{\partial\beta_l} - \mathbf{X}_{il}\pi_{1i}\left(1-\pi_{1i}\right)$$

e por fim

$$-\frac{\partial^2 l_{obs}}{\partial\gamma_s\partial\beta a} = -\sum_{i=1}^{n} \mathbf{Z}_{is}\left(\frac{\partial w_i^*}{\partial\beta_a}\right)$$

sendo que a derivada de w_i^* em função de β é dada pela expressão em 17.

Para testar os parâmetros $\boldsymbol{\theta}$, pode-se utilizar o teste de Wald. A próxima seção apresenta estudo de simulação para testar o comportamento do modelo com diferentes tamanhos amostrais. A implementação deste algoritmo não é complexa. O passo **E** tem sua forma fechada dada em 12 e os dois passos **M** são ajustes de modelos lineares generalizados, encontrados em vários software estatísticos.

2.4 Estudo de Simulação

Nessa Seção, simulações de Monte Carlo são realizadas. Os resultados apresentados foram obtidos utilizando o algortimo **EM**, descrito na seção 2.3, fixando δ_1 e δ_2 em 0.001 em 13. Existe um interesse prático maior nesse metodologia, pois apresenta uma fatoração da função de verossimilhança completa, sendo útil no próximo capítulo.Foram simulados três cenários, variando a fração de cura em aproximadamente 25%, 50% e 75%, sendo controlados pelos valores reais do vetor $\boldsymbol{\gamma}$. Foi gerada uma covariável X_1 de uma distribuição Normal Padrão e X_2 de uma Bernoulli com probabilidade de sucesso 0,5, relacionados com o preditor linear da resposta observada. Para o preditor linear da cura, foi gerado Z_1 de uma distribuição Normal Padrão. O vetor de valores reais para $\boldsymbol{\beta}$ é $(\beta_0 = 1, \beta_1 = 2, \beta_2 = 1)^{'}$ em todas os cenários. Para obter uma fração de cura de aproximadamente 25%, foram escolhidos os valores reais para o vetor $\boldsymbol{\gamma}$ de $(\gamma_0 = -2,0, \gamma_1 = -2,5)^{'}$, para 50% os valores reais são $(\gamma_0 = 0,10, \gamma_1 = 1,0)^{'}$ e para 75% assumimos $(\gamma_0 = 2,0, \gamma_1 = 2,5)^{'}$. Os passos para gerar o banco são:

1. Gerar aleatoriamente n valores das covariáveis das distribuições pré-especificadas e construir as matrizes $\mathbf{X}$ e $\mathbf{Z}$.
2. Obter o preditor linear $\mathbf{Z}\boldsymbol{\gamma}$ e gerar aleatoriamente o fator de cura W_i, para $i = 1, ..., n$, da distribuição Bernoulli com probabilidade de sucesso sendo o inverso da ligação *logit* associada a este preditor linear.

3. Se o valor W_i for 1, a variável resposta Y_i será 0 com probabilidade 1. Caso contrário, é gerado aleatoriamente o valor para Y_i de uma distribuição Bernoulli com probabilidade de sucesso sendo o inverso da *logit* associada ao preditor linear $\mathbf{X}\boldsymbol{\beta}$.

Os passos são repetidos B = 200 vezes. São guardadas as estimativas pontuais e variabilidade dos parâmetros de interesse, além da fração de cura, em cada repetição. Em cada simulação é calculado o intervalo baseado na distribuição assintótica dos estimadores de máxima verossimilhança com confiança de $(1-\alpha)100\%$, com $0 < \alpha < 1$ e o vício relativo absoluto. Também foi calculado a quantidade de vezes que o verdadeiro valor do parâmetro pertence ao intervalo assintótico. Seja $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})'$. O vício relativo absoluto do *j-ésimo* elemento de $\boldsymbol{\theta}$, expresso em porcentagem, é:

$$\frac{\widehat{\boldsymbol{\theta}_j} - \boldsymbol{\theta}_j}{\boldsymbol{\theta}_j} \times 100\%$$

2.4.1 Resultados

Os resultados são apresentados em tabelas. A coluna VR(%) indica a média dos 200 valores de vícios relativos, medido em %, e PC(%) é a probabilidade de cobertura, isso é, indica a proporção de vezes em que os intervalos de confiança contiveram o verdadeiro valor do parâmetro.

A Tabela 1 apresenta os resultados para a fração de cura de aproximadamente 25%. Com o aumento do tamanho amostral, as médias das estimativas ficam mais próximas dos valores reais, as médias dos erros-padrão e vícios relativos diminuem.

Problemas de convergência foram encontrados com n = 100. Neste caso, foi excluído essa amostra e gerada outra, com uma taxa de ocorrência de aproximadamente 3%. Esses problemas não foram observados para n = 500 e 1000.

Tabela 1 – Simulações Monte Carlo pelo algortimo EM- Cenário com fração de cura 25% de cura

n	Par	Média	Média Erro-padrão	Desvio-padrão das estimativas	VR(%)	PC(%)
1000	β_0	1,012	0,198	0,203	1,20%	95,4%
	β_1	2,026	0,213	0,200	1,30%	96,3%
	β_2	1,029	0,247	0,256	2,90%	93,0%
	γ_0	-2,037	0,281	0,290	1,85%	94,4%
	γ_1	-2,550	0,297	0,313	2,00%	94,4%
500	β_0	1,041	0,288	0,298	4,10%	95,3%
	β_1	2,077	0,312	0,335	3,85%	94,5%
	β_2	1,055	0,358	0,373	5,50%	95,1%
	γ_0	-2,079	0,409	0,437	3,95%	95,1%
	γ_1	-2,690	0,434	0,484	7,60%	93,7%
100	β_0	1,364	0,868	1,118	36,40%	96,4%
	β_1	2,595	0,979	1,236	29,75%	96,7%
	β_2	1,224	1,010	1,185	22,40%	96,7%
	γ_0	-2,369	1,137	1,185	18,45%	91,5%
	γ_1	-2,972	1,202	1,309	18,88%	95,1%

A seguir as Tabela 2 e Tabela 3 apresentam os resultados encontrados nos cenários com fração de cura de 50% e 75%. Assim como na simulação anterior, com o aumento do tamanho amostral, a média das estimativas se aproxima do valor real e o erro-padrão diminui. Em comparação com a Tabela 1, nota-se que, para o preditor linear da resposta observada, o erro padrão é cerca de 2 vezes maior com a fração de cura de 50%. Na Tabela 2 nota-se também médias dos vícios relativos maiores para o γ_0 em comparação com os demais. Na Tabela 3, em comparação com a Tabela 2, para o preditor linear da resposta observada, as médias dos erros-padrão e vícios relativos são levemente superiores para $n = 1000$ e $n = 500$ com a fração de cura aproximadamente de 75%. Nota-se menores coberturas dos valores reais dos intervalos de confiança para β_0 e β_1 para os tamanhos amostrais de $n = 100$. Em contrapartida, a cobertura para γ_1 é bastante elevada. Problemas de convergência foram encontrados com n = 100, com taxa de ocorrência 21,0% no cenário 2 e aproximadamente 24% para o terceiro.

Tabela 2 – Simulações Monte Carlo pelo algortimo EM- Cenário com fração de cura 50% de cura

n	Par	Média	Média Erro-padrão	Desvio-padrão das estimativas	VR(%)	PC(%)
	β_0	1,02	0,412	0,396	2,00%	96,8%
	β_1	2,089	0,407	0,388	4,45%	91,1%
1000	β_2	1,042	0,399	0,406	4,20%	92,7%
	γ_0	0,094	0,145	0,140	-0,06%	94,4%
	γ_1	1,014	0,124	0,115	1,40%	96,1%
	β_0	1,134	0,651	0,682	13,40%	94,5%
	β_1	2,185	0,641	0,678	9,25%	95,4%
500	β_2	1,072	0,606	0,652	7,20%	93,6%
	γ_0	0,073	0,213	0,243	-27,00%	94,5%
	γ_1	1,029	0,183	0,195	2,90%	98,1%
	β_0	1,317	1,665	1,383	31,70%	96,6%
	β_1	2,731	1,772	1,700	36,55%	94,0%
100	β_2	1,165	1,673	1,796	16,50%	96,1%
	γ_0	-0,060	0,558	0,400	-160,00%	98,7%
	γ_1	1,211	0,502	0,495	21,00%	97,4%

Tabela 3 – Simulações Monte Carlo pelo algortimo EM- Cenário com fração de cura 75% de cura

n	Par	Média	Média Erro-padrão	Desvio-padrão das estimativas	VR(%)	PC(%)
	β_0	1,116	0,510	0,518	11,60%	95,8%
	β_1	2,208	0,532	0,543	10,40%	97,2%
1000	β_2	1,086	0,544	0,539	8,60%	96,7%
	γ_0	2,015	0,177	0,183	0,75%	94,5%
	γ_1	2,519	0,256	0,251	0,76%	94,9%
	β_0	1,100	0,728	0,736	10,00%	93,1%
	β_1	2,262	0,761	0,784	13,10%	95,0%
500	β_2	1,131	0,781	0,777	13,10%	96,0%
	γ_0	2,026	0,258	0,248	1,30%	96,5%
	γ_1	2,619	0,384	0,389	4,76%	96,5%
	β_0	0,753	1,500	1,698	-24,70%	86,0%
	β_1	2,337	1,628	1,785	16,85%	84,4%
100	β_2	1,087	2,083	1,971	8,70%	94,1%
	γ_0	1,988	0,770	0,850	-0,60%	96,8%
	γ_1	3,515	1,451	1,475	40,60%	99,6%

Nos três cenários observou-se o comportamento padrão e esperado, ou seja: com o aumento do tamanho amostral, a média das estimativas pontuais estão mais próximas dos valores reais e os erros-padrão diminuem. As simulações para os mesmos cenários foram feitas utilizando a função de verossimilhança observada, dada em 6. Os resultados estão no Apêndice B e são consistentes com os observados nas simulações apresentadas nessa Seção.

Problemas numéricos foram observados nas simulações com a verossimilhança completa e observada em n = 100, mas com uma frequência menor aplicando o **EM**,

podendo ser uma consequência da fatoração da função de verossimilhança completa. Utilizando a verossimilhança observada, foi necessário gerar outra amostra com frequência de 6,7% para fração de cura de 25%, aproximadamente 25% com fração de cura de 50% e 31,6% com cura de 75%.

Não existiu nenhum critério para escolha dos valores reais do vetor $\boldsymbol{\beta}$. Para o vetor $\boldsymbol{\gamma}$, o único critério de escolha dos valores é para controlar a fração de cura. Para o tamanho amostral, o único critério foi escolher n = 100, pois esse é aproximadamente o tamanho do banco de dados utilizado na aplicação real e depois observar o comportamento do modelo com maiores tamanho amostrais. Não existe a necessidade de simular com um n superior a 1000, as diferenças encontradas serão pequenas em todos os resultados.

3 RESPOSTA BINÁRIA COM FATOR DE CURA: UMA AMOSTRA EM CONGLOMERADOS

A metodologia apresentada no capítulo anterior não é apropriada quando temos uma amostra em conglomerados, ou seja, mais de uma medição por indivíduo. Na prática, isso pode ocorrer em muitas situações. Ignorar a presença de dados agrupados na análise pode acarretar vícios nas conclusões, pois os erros-padrão das estimativas tendem a ficar mal estimados. Além disso, podemos estar descartando informações, pois é necessário sortear apenas uma medição do indivíduo para a metodologia do Capítulo 2 ser válida.

Para tratar de maneira adequada mais de uma medição no mesmo indivíduo, Hall (2000) apresentou modelos mistos inflacionados de zero como uma possível solução. Porém, como em modelos mistos, a interpretação direta não é em nível populacional, o que não é de interesse prático para o pesquisador. Hall e Zhang (2004) apresentaram duas soluções: a primeira consiste em alterar o passo **M** do algoritmo **EM** para acomodar mais de uma medição no mesmo indivíduo, incluindo equações bastante parecidas com o *Generalized Estimating Equation* (GEE), denominado de algoritmo **ES** (*The Expectation-Solution*). A segunda opção propõe alterar as equações do GEE para acomodar inflação de zero. Segundo os autores, a partir de estudo de simulações realizados, a primeira solução é mais eficiente e, portanto, foi a adotada neste projeto.

A Seção 3.1 apresenta o algoritmo **ES** e um estudo de simulação é realizado na Seção 3.2.

3.1 Algortimo **ES**

Seja K a quantidade de conglomerados e $\boldsymbol{y_i}$ um vetor de dimensão $n_i \times 1$ contendo a resposta de interesse para o *i-ésimo* indivíduo ou cluster, $i = 1, ..., K$. Assume-se que Y_{ij} é variável aleatória resposta associada a observação y_{ij}. Assuma também que Y_{ij} segue uma distribuição degenerada no 0 com probabilidade p_{ij} se ocorreu a cura e alguma distribuição de contagem com probabilidade $1 - p_{ij}$, em que $0 < p_{ij} < 1$, com $j = 1, ..., n_i$.

Assim como em Modelos Lineares Generalizados, é necessário que a distribuição de contagem pertença à família exponencial. Hall e Zhang (2004) apresentam que uma

distribuição pertence à família de dispersão exponencial se podemos escrever a densidade ou função de probabilidade da seguinte forma:

$$f_2\left(y_{ij},\theta_{ij},\phi\right)=h_2\left(y_{ij},\phi\right)\exp\left[\left(\theta_{ij}y_{ij}-k\left(\theta_{ij}\right)\right)u_{ij}/\phi\right] \tag{18}$$

em que θ_{ij} é um parâmetro de locação canônico e u_{ij} é constante conhecida. Os autores definem k como sendo uma função geradora acumulada, portanto a distribuição tem média e variância condicionais $\zeta_{ij}=k^{'}\left(\theta_{ij}\right)$ e $\nu\left(\zeta_{ij}\right)\phi/u_{ij}$ respectivamente, em que $\nu\left(\zeta_{ij}\right)=k^{''}\left(\theta_{ij}\right)$ é a função de variância. ϕ é um parâmetro de dispersão que é utilizado, por exemplo, para o tratamento de superdispersão. Nesta dissertação, foi considerado $\phi=1$ e que Y_{ij} segue uma distribuição Bernoulli com probabilidade de sucesso π_{1ij}.

As quantidades de interesse em 18 são:

$$\theta_{ij}=\log\frac{\pi_{1ij}}{1-\pi_{1ij}};k\left(\theta_{ij}\right)=\log\left(1+\exp\left(\theta_{ij}\right)\right);\phi=u_{ij}=h_2\left(y_{ij},\phi\right)=1. \tag{19}$$

Com isso,

$$E\left[Y_{ij}\right]=\zeta_{ij}=k^{'}\left(\theta_{ij}\right)=\pi_{1ij}$$
$$\nu\left(\zeta_{ij}\right)=k^{''}\left(\theta_{ij}\right)=\pi_{1ij}\left(1-\pi_{1ij}\right)$$

Utilizando a função de ligação canônica, temos que $\theta_i\left(\zeta_i\right)=\eta_i=\mathbf{X}_i\boldsymbol{\beta}$ ou $\zeta_{ij}=\theta^{-1}\left(\eta_{ij}\right)$. Em relação às probabilidades de mistura p_{1i} ,o mais usual é relacionar com covariáveis a partir de uma função de ligação *logit*, isso é, $logit(p_{1i})=\mathbf{Z}_i\boldsymbol{\gamma}$.

Seja $\mathbf{Y}=(Y_{11},...,Y_{Kn_k})$ e $\mathbf{W}=(w_{11},...,w_{Kn_k})$, em que $w_{ij}=0$, se a variável aleatória de interesse segue a distribuição Bernoulli, e $w_{ij}=1$, caso contrário, *i = 1, ..., K, j = 1, ..., n_i*. Hall e Zhang (2004) definem a função log verossimilhança para os dados completos (**Y**,**W**) como sendo:

$$\begin{aligned}l_c\left(\boldsymbol{\gamma},\boldsymbol{\beta};y,w\right)=&\sum_{i,j}\left[w_{ij}\mathbf{Z}_{ij}\boldsymbol{\gamma}^{'}\ -\ \log\left(1+\exp\left(\mathbf{Z}_{ij}\boldsymbol{\gamma}^{'}\right)\right)\right]\\&+\sum_{i,j}\left(1-w_{ij}\right)\left[\left(1-I_{(0)}(y_{ij})\right)\mathbf{X}_{ij}\boldsymbol{\beta}^{'}-\log\left(1+\exp\left(\mathbf{X}_{ij}\boldsymbol{\beta}^{'}\right)\right)\right]\end{aligned} \tag{20}$$

De forma similar ao raciocínio apresentado na Seção 2.3 para o cálculo do passo **E** no **EM**, Hall e Zhang (2004) apresentam a forma deste passo no algoritmo **ES**, que consiste na estimação de w_{ij} no *h-ésimo* passo, dado por:

$$\hat{w_{ij}}^{(h)} = I_0(y_{ij}) \left[1 + \exp\left(-\mathbf{Z}_i \boldsymbol{\gamma}'^{(h)}\right) \left(\frac{1}{1 + \exp\left(\mathbf{X}_i \boldsymbol{\beta}'^{(h)}\right)}\right)\right]^{-1} \tag{21}$$

em que $I_0(y_{ij})$ é a indicadora se $y_{ij} = 0$.

Hall e Zhang (2004) definem a maximização em relação a $\boldsymbol{\gamma}$ como sendo:

$$\sum_{i=1}^{K} \left[\frac{\partial \boldsymbol{p_{1i}}(\boldsymbol{\gamma})^T}{\partial \boldsymbol{\gamma}}\right] \left[\boldsymbol{A_i}^{1/2}\left[\boldsymbol{p_{1i}}(\boldsymbol{\gamma})\right] \mathbf{I} \boldsymbol{A_i}^{1/2}\left[\boldsymbol{p_{1i}}(\boldsymbol{\gamma})\right]\right]^{-1} \left[\boldsymbol{w_i}^{(h)} - \boldsymbol{p_{1i}}(\boldsymbol{\gamma})\right] = 0 \tag{22}$$

em que $\boldsymbol{A_i}(\boldsymbol{p_{1i}}) = \text{diag}\left[p_{1i1}(1 - p_{i1}), ..., p_{1in_i}(1 - p_{1in_i})\right]$ e em relação a $\boldsymbol{\beta}$:

$$\sum_{i=1}^{K} \left[\frac{\partial \boldsymbol{\zeta_i}(\boldsymbol{\beta})^T}{\partial \boldsymbol{\beta}}\right] \left[\boldsymbol{D_i}^{1/2}\left[\boldsymbol{\zeta_i}(\boldsymbol{\beta})\right] \mathbf{I} \boldsymbol{D_i}^{1/2}\left[\boldsymbol{\zeta_i}(\boldsymbol{\beta})\right]\right]^{-1} \boldsymbol{W_i}^{(h)}\left[\boldsymbol{y_i} - \boldsymbol{\zeta_i}(\boldsymbol{\beta})\right] = 0 \tag{23}$$

em que $\boldsymbol{W}_i^{(h)} = \text{diag}\left[\left(1 - w_{i1}^{(h)}\right), ..., \left(1 - w_{in_i}^{(h)}\right)\right]$ e $\boldsymbol{D_i}(\zeta_i) = \text{diag}\left[\pi_{1i1}(1 - \pi_{i1}), ..., \pi_{1in_i}(1 - \pi_{1in_i})\right]$

As expressões em 22 e 23 possuem a mesma forma de um GEE com matriz de correlação de trabalho igual a identidade. Uma extensão natural é flexibilizar para acomodar outras estruturas de trabalho, como AR(1), não estruturada e simetria composta. Hall e Zhang (2004) substituem as equações em 22 e 23 por:

$$\sum_{i=1}^{K} \left[\frac{\partial \boldsymbol{p_{1i}}(\boldsymbol{\gamma})^T}{\partial \boldsymbol{\gamma}}\right] \left[\boldsymbol{A_i}^{1/2}\left[\boldsymbol{p_{1i}}(\boldsymbol{\gamma})\right] \mathbf{R}(\boldsymbol{\delta}) \boldsymbol{A_i}^{1/2}\left[\boldsymbol{p_{1i}}(\boldsymbol{\gamma})\right]\right]^{-1} \left[\boldsymbol{w_i}^{(h)} - \boldsymbol{p_{1i}}(\boldsymbol{\gamma})\right] = 0 \tag{24}$$

e

$$\sum_{i=1}^{K} \left[\frac{\partial \boldsymbol{\zeta_i}(\boldsymbol{\beta})^T}{\partial \beta}\right] \left[\boldsymbol{D_i}^{1/2}\left[\boldsymbol{\zeta_i}(\boldsymbol{\beta})\right] \mathbf{P}(\boldsymbol{\rho}) \boldsymbol{D_i}^{1/2}\left[\boldsymbol{\zeta_i}(\boldsymbol{\beta})\right]\right]^{-1} \boldsymbol{W_i}^{(j)}\left[\boldsymbol{y_i} - \boldsymbol{\zeta_i}(\boldsymbol{\beta})\right] = 0 \tag{25}$$

em que $\mathbf{R}(\boldsymbol{\delta})$ e $\mathbf{P}(\boldsymbol{\rho})$ são matrizes de correlação de trabalho, que devem ser especificadas. Nas expressões, $\boldsymbol{\delta}$ e $\boldsymbol{\rho}$ são parâmetros de correlação e devem ser estimados.

O artigo de Hall e Zhang (2004) apresenta uma estratégia para obter maior eficiência no processo de estimação, substituindo as equações em 24 e 25 por equações de estimativas combinadas. Inicialmente, iremos ajustar as equações apresentadas aqui, observando na simulação se os resultados encontrados já são satisfatórios.

Portanto, no passo **E** é calculada a expressão em 21 e agora, no passo **S**, foram utilizadas as equações em 24 e 25. O critério de parada do algoritmo é o mesmo usado no **EM**, dado em 13.

A dissertação de Xu (2013) é uma aplicação do algoritmo **ES** no contexto de um modelo ZINB (*Zero-Inflated Negative Binomial*). O autor detalha passos de estimação, apresenta as equações para obter a variância "sanduíche" e disponibiliza os códigos em R, que foram adaptados para a situação do presente trabalho. Os parâmetros de correlação também devem ser estimados em cada iteração. O autor definiu os estimadores de momento para δ, quando temos a simetria composta:

$$\hat{\delta} = \sum_{i=1}^{K} \sum_{s<t} \frac{(w_{is} - p_{1is})(w_{it} - p_{1it})}{\sqrt{p_{1is}p_{1it}(1 - p_{1is})(1 - p_{1it})}} \tag{26}$$

e para ρ, no caso em que a resposta é contagem e o ajuste é realizado com a Binominal Negativa. No presente trabalho, temos:

$$\hat{\rho} = \sum_{i=1}^{K} \sum_{s<t} \frac{(1 - w_{is})(1 - w_{it})(y_{is} - \pi_{1is})(y_{it} - \pi_{1it})}{\sqrt{\pi_{1is}\pi_{1it}(1 - \pi_{1is})(1 - \pi_{1it})}} \tag{27}$$

Os passos de estimação apresentados em Xu (2013) são:

1. Seja h = 0. Além de um chute inicial para β e γ, os autores exigem um chute inicial para os parâmetros de correlação δ e ρ.
2. Passo **E**: Computar a esperança da variável latente condicionada aos dados observados, pela expressão dada em 21.
3. Passo **S**: Encontrar a solução em função de β e γ das expressões em 24 e 25.
4. Com os valores β^{h+1}, γ^{h+1} e $\hat{w_{ij}}^{(h)}$, atualizar o valor de δ, dado em 26.
5. Com os valores β^{h+1}, γ^{h+1} e $\hat{w_{ij}}^{(h)}$, atualizar o valor de ρ, dado em 27.
6. Repetir os passos 2-5 até a convergência.

Sejam $V_{\gamma i}$ e $V_{\beta i}$ iguais a $\boldsymbol{A_i}^{1/2}\left[\boldsymbol{p_{1i}}\left(\boldsymbol{\gamma}\right)\right]\mathbf{R}\left(\boldsymbol{\delta}\right)\boldsymbol{A_i}^{1/2}\left[\boldsymbol{p_{1i}}\left(\boldsymbol{\gamma}\right)\right]$ e $\boldsymbol{D_i}^{1/2}\left[\boldsymbol{\zeta_i}\left(\boldsymbol{\beta}\right)\right]\mathbf{P}\left(\boldsymbol{\rho}\right)\boldsymbol{D_i}^{1/2}\left[\boldsymbol{\zeta_i}\left(\boldsymbol{\beta}\right)\right]$, respectivamente. Os autores definem a variância "sanduíche" como sendo $\boldsymbol{B^{-1}M_1B^{-1}}$, em que:

$$\boldsymbol{B}=\begin{pmatrix}\sum_{i=1}^{K}\frac{\partial p_{1i}\left(\boldsymbol{\gamma}\right)^{T}}{\partial\boldsymbol{\gamma}}V_{\gamma i}^{-1}\frac{\partial p_{1i}\left(\boldsymbol{\gamma}\right)}{\partial\boldsymbol{\gamma^{T}}} & 0\\ 0 & \sum_{i=1}^{K}\left[\frac{\partial\zeta_i\left(\boldsymbol{\beta}\right)^{T}}{\partial\beta}\right]\mathrm{diag}\left(1-\hat{\boldsymbol{w}}_{\boldsymbol{i}}\right)^{1/2}V_{\boldsymbol{\beta i}}^{-1}\mathrm{diag}\left(1-\hat{\boldsymbol{w}}_{\boldsymbol{i}}\right)\frac{\partial\zeta_i\left(\boldsymbol{\beta}\right)}{\partial\beta^{T}}\end{pmatrix}$$

e

$$\boldsymbol{M_1}=\boldsymbol{M}\boldsymbol{M}^{T}$$

em que

$$\boldsymbol{M}=\begin{pmatrix}\sum_{i=1}^{K}\frac{\partial p_{1i}\left(\boldsymbol{\gamma}\right)^{T}}{\partial\boldsymbol{\gamma}}V_{\boldsymbol{\gamma i}}^{-1}\left(\hat{\boldsymbol{w}_{\boldsymbol{i}}}-\hat{\boldsymbol{p_{1i}}}\right)\\ \sum_{i=1}^{K}\frac{\partial\boldsymbol{\zeta_i}\left(\boldsymbol{\beta}\right)^{T}}{\partial\beta}V_{\boldsymbol{\beta i}}^{-1}\mathrm{diag}\left(1-\hat{\boldsymbol{w}}_{\boldsymbol{i}}\right)\left[\boldsymbol{y_i}-\boldsymbol{\zeta_i}\left(\boldsymbol{\beta}\right)\right]\end{pmatrix}$$

3.2 Estudo de simulação

Nessa seção vamos avaliar a metodologia apresentada na Seção 3.1 através de um pequeno estudo de simulação. Os dados simulados foram gerados com a função *rbin* do pacote *SimCorMultRes* de Touloumis (2016). O processo de geração dos dados utiliza o método NORTA, proposto por Cario e Nelson (1997), que generalizou a abordagem apresentada em Li e Hammond (1975) para gerar valores de distribuições marginais contínuas, discretas ou mistas, a partir de uma matriz de correlação especificada. Temos o seguinte modelo marginal, com $i = 1, ..., K$ e $j = 1, ..., n_i$, em que K é a quantidade de conglomerados:

$$P\left(Y_{ij}=1\mid x_{ij}\right)=F\left(\beta_{j0}+\beta_{j}^{'}x_{ij}\right)$$

em que β_{j0} é o intercepto relacionado no instante j, β_j é o vetor de parâmetros relacionado com as covariáveis medidas e F é alguma função de distribuição acumulada. O limiar a seguir gera respostas binárias agrupadas

$$Y_{ij} = 1 \Longleftrightarrow U_{ij}^{B} \leq \beta_{j0} + 2\beta_{j}^{'} x_{ij}$$

em que $U_{ij} = \beta_{j}^{'} x_{ij} + e_{ij}^{B}$. Temos aqui que e_{ij} são variáveis aleatórias, tais que:

- $e_{ij} \sim F$, para todo i e j
- $e_{i_1 j_1}$ é independente de $e_{i_2 j_2}$ para todo $i_1 \neq i_2$.

O objetivo é então gerar valores de F a partir de uma matriz de correlação especificada, usando o método NORTA. Mais detalhes podem ser encontrados em Li e Hammond (1975), Cario e Nelson (1997) e na documentação do pacote *SimCorMultRes* Touloumis (2016).

A simulação foi conduzida supondo como verdadeira a estrutura de correlação simetria composta. Esta escolha é justificada pela aplicação prática, pois usamos essa estrutura de trabalho na modelagem. Além disso, no GEE existe a garantia de que os estimadores são consistentes, mesmo com a má especificação da matriz de trabalho.

Assim como na simulação para o caso de uma amostra independente, foram geradas 3 covariáveis: X1 de uma Normal Padrão e X3 uma Bernoulli com probabilidade de sucesso 0,5. O vetor com valores reais é $\boldsymbol{\beta} = (\beta_0 = 1, \beta_1 = 2, \beta_2 = 1.2)$. Essas duas covariáveis estão relacionadas ao preditor linear da resposta observada, que é $1 + 2X_1 + 1.2X_2$. A terceira covariável Z_1 foi gerada de uma Normal Padrão e está relacionada ao preditor linear da cura. Assim como no Capítulo 2, os valores reais de $\boldsymbol{\gamma}$ foram variados para obter diferentes frações de cura.

A simulação inicialmente foi conduzida considerando dois tamanhos de conglomerados, 250 e 100, e cada com 4 medições. A matriz de correlação para gerar a cura e a resposta observada são idênticas, com dimensão 4x4, com o valor de correlação igual a 0,50. Segue-se a lógica da presença de cura para formar o vetor final de resposta para o *í-ésimo* indivíduo.

Para obter uma fração de cura média de 25%, o vetor real do vetor $\boldsymbol{\gamma}$ é $= (\gamma_0 = -2.0, \gamma_1 = -2.5)$, e com 75% é $(\gamma_0 = 2.0, \gamma_1 = 2.5)$. Devido ao tempo computacional gasto, não foi realizada a simulação considerando a fração de cura de 50%, pois, como observado no Capítulo 2, existem poucas diferenças entre essa fração de cura com a de

75%, em principal diferença é a tendência o erro-padrão ser levemente superior com 75% de cura.

A Tabela 4 apresenta os resultados encontrados para o Monte Carlo com 200 repetições com fração de cura de 25%. As médias das estimativas se aproximam do valor real e os desvios-padrão diminuem com o aumento da quantidade de grupos. Devido ao tempo computacional, não foi simulado valores de K superiores de 250, pois com este tamanho, as médias das estimativas pontuais já estão bastante próximas dos valores reais. As simulações com 75% de cura estão no Apêndice B. A principal diferença é que são necessárias mais interações para o algortimo convergir com maior cura. Considerando a cura de 75%, aproximadamente 39% das simulações foram necessárias mais de 50 iterações para convergência.

Tabela 4 – Simulações Monte Carlo pelo algortimo ES- Cenário com fração de cura 25%

K	Par	Média	Média Erro-padrão	Desvio-padrão das estimativas	VR(%)
250	β_0	1,049	0,168	0,321	4,90%
	β_1	2,076	0,181	0,289	3,38%
	β_2	1,248	0,251	0,349	24,80%
	γ_0	-2,047	0,115	0,392	2,35%
	γ_1	-2,502	0,154	0,403	0,08%
100	β_0	1,371	0,304	0,792	37,10%
	β_1	2,378	0,332	0,863	18,90%
	β_2	1,352	0,446	1,230	35,2%
	γ_0	-1,939	0,182	0,464	-3,5%
	γ_1	-2,546	0,246	0,502	1,96%

A média dos erros-padrão está diferente do desvio-padrão das estimativas. Esse fato também foi encontrado em Xu (2013). Isso se deve ao fato de que a variância "sanduíche" apresentada trata a variável latente como conhecida, ignorando a incerteza sobre a sua estimativa. Uma forma de contornar essa dificuldade é utilizar o *bootstrap* a nível de conglomerado para estimar o intervalo de confiança. Ou seja, é sorteado com reposição a identificação do grupo, carregando todas as 4 medições para formar o banco de dados para o *bootstrap*. Foram consideradas 200 repetições *bootstrap* para cada banco de dados gerado, sendo calculado o desvio das estimativas obtidas e o intervalo de confiança com os quantis de ordem 2,5% e 97,5%. A Tabela 5 apresenta os resultados encontrados para K = 250, mostrando a média dos desvios-padrão e estimativas pontuais obtidas no bootstrap, e PC(%), que é a probabilidade de cobertura dos intervalos de confiança.

Tabela 5 – Resultados do *bootstrap*. Fração de cura 25% e K = 250 grupos.

Par	Média do desvio bootstrap	Desvio-padrão das estimativas	Média da estimativa pontual	PC(%)
β_0	0,318	0,321	1,03	91,3%
β_1	0,336	0,289	2,12	94,8%
β_2	0,390	0,349	1,25	93,1%
γ_0	0,384	0,392	-2,14	91,3%
γ_1	0,361	0,403	-2,61	91,3%

A média do desvio das estimativas *bootstrap* estão próximas do desvio padrão das estimativas e a tendência é o intervalo de confiança se tornar mais simétrico em torno do valor real com o aumento do número de repetições. A vantagem deste procedimento é o fato da implementação computacional não ser complexa, pois é uma reamostragem com reposição e não exige contas adicionais.

Não foi realizada mais nenhuma simulação variando a quantidade de medições em cada conglomerado ou a correlação entre os mesmos, devido ao tempo computacional. Pois, além do tempo já existente no processo de estimação, houve um acréscimo de tempo devido as replicas *bootstrap*. Essa é a principal desvantagem. Portanto, uma perspectiva de investigação futura seria propor uma flexibilização da fórmula de Loius para o algoritmo **ES**, facilitando assim as conclusões em uma situação prática.

Foram também realizadas simulações com a fração de cura média de aproximadamente 75%, com resultados no Apêndice B.

4 APLICAÇÃO REAL

O modelo logístico com fator de cura foi aplicado em um banco de dados proveniente da Programa Traumatismos Dentários da Faculdade de Odontologia da UFMG, composto por 90 pacientes portadores de 104 dentes reimplantados após avulsão traumática, distribuídos de acordo com a Tabela 6.

Tabela 6 – Quantidade de pacientes x Dentes

Quantidade de pacientes	Número de dentes
78	1 dente
10	2 dentes
2	3 dentes

Dados clínicos e radiográficos foram coletados na primeira consulta feita na CTD FAO UFMG. O objetivo foi identificar quais dentre os fatores de risco/covariáveis descritos a seguir, estão relacionados com a RREI inicial:

- Meio de armazenamento do dente após a avulsão e antes do reimplante, categorizado em armazenamento a seco ou em meios úmidos (saliva, água, soro e leite).
- Período extra-oral, isto é, o tempo em minutos que o dente permanece fora da boca após a avulsão.
- Grau de rizogênese: estágio de formação da raiz do dente avulsionado categorizado de acordo com o diâmetro do forame apical.
- O tempo em dias entre a data do reimplante e ínicio do tratamento endodôntico na CTD FAO UFMG, denominado tempo de infecção.
- A prescrição ou não de antibióticoterapia sistêmica após o reimplante.
- Sexo do paciente.

A seguir, a Tabela 7 apresenta medidas descritivas das covariáveis contínuas. Período extra-oral e tempo de infecção, devido à grande dispersão, foram consideradas na escala logarítmica no modelo.

Tabela 7 – Descritivas das covariáveis contínuas

Covariável	Mediana	Média	Desvio padrão
Período Extra-Oral	120,0	219,0	310,7
Tempo de Infecção	55,0	85,1	89,71

A Tabela 8 ilustra a caracterização de meio de armazenamento do dente, grau de rizogênese e tipo de reabsorção, que são realizados a nível dos 104 dentes. Aproximadamente, 35% dos dentes avulsionados foram armazenados em ambientes secos, 60% possuem grau de rizogênese 2/3/4 e 71,2% da amostra apresentou RREI.

Tabela 8 – Descritivas das covariáveis categóricas a nível do dente

Covariável	Categorias	Frequência
Meio	Líquido	66 (63,4%)
	Seco	36 (34,6%)
	Sem informação	2 (2,0%)
Grau de Rizo	2/3/4	62 (59,6%)
	5	42 (40,4%)
Tipo de reabsorção	Ausente	30 (28,8%)
	RREI	74 (71,2%)

A Tabela 9 apresenta as covariáveis que devem ser descritivas a nível dos 90 indivíduos, que são sexo e prescrição de antibiótico. Aproximadamente 69% da amostra é formada por pessoas do sexo masculino e 27% não tinham a informação sobre a prescrição de antibióticos. Não foi feita nenhum tratamento dos dados ausentes nessa dissertação.

Tabela 9 – Descritivas das covariáveis categóricas a nível do indivíduo

Covariável	Categorias	Frequência
Prescrição de antibiótico	Sim	13 (14,4%)
	Não	53 (58,9%)
	Sem informação	24 (26,7%)
Sexo	Masculino	62 (68,9%)
	Feminino	28 (31,1%)

A Seção 4.1 apresenta o ajuste do modelo de regressão logística com fator de cura para uma amostra independente, sendo realizado o sorteio de um dente para aqueles pacientes com mais de uma medição, e a Seção 4.2 apresenta o ajuste para a amostra completa, isso é, com conglomerados.

4.1 Amostra independente

O ajuste é realizado pelo algoritmo **EM**, utilizado o teste de Wald para testar a significância dos coeficientes. O ajuste pelo **EM** foi escolhido devido à menor proporção de problemas numéricos encontrados na simulação, fatoração da função de verossimilhança e devido ao fato da metodologia do tratamento de dados com conglomerados ser uma flexibilização do algoritmo.

Não foi possível ajustar os modelos simples, isto é, com apenas uma covariável, por causa de problemas de convergência nos fatores de risco: antibiótico e período extra oral.

Para encontrar um método de seleção das covariáveis, foi utilizado o resultado de Bastos et al. (2014), que apresenta associação entre tempo de infecção com tipo de reabsorção. As etapas de seleção do modelo final foram:

1. Fixou-se o logaritmo do tempo de infecção no preditor linear da resposta observada. Então ajustou-se no preditor da cura todas as outras covariáveis, separadamente, passando para a segunda etapa todas as covariáveis com p-valores inferiores a 0.25.
2. Ajustou-se todas as covariáveis que passaram na etapa 1 para o preditor da cura conjuntamente. No caso de covariável não significativa em um nível fixado, retirou-se aquele com maior p-valor e ajustou-se o modelo novamente. Este processo foi repetido até que todas as covariáveis ficassem significativas.
3. Todas as covariáveis retiradas em (2) retornaram, separadamente, para confirmar se não eram estatisticamente significativas.
4. E, por fim, ajustou-se no modelo final obtido em (3) todas as outras covariáveis no preditor linear da resposta observada, separadamente, passando para próxima etapa todas as covariáveis com p-valores inferiores a 0.25.
5. Repetiu-se os passos 2 e 3 para o preditor da resposta observada.

Na primeira etapa, foi necessário incluir a covariável período extra-oral nos passos com meio de armazenamento, sexo e grau de rizogênese, seguindo o que foi discutido na Seção 2.2.1: quando o objetivo for a comparação entre grupos, pode-se incluir uma covariável continua em apenas um preditor linear, mesmo que não significativa, a fim de que o modelo se torne identificável.

O detalhamento de todas as etapas estão no Apêndice C. O modelo final é apresentado na Tabela 10, considerando um nível de significância de 5%. A categoria base para antibiótico foi a prescrição. EP corresponde ao erro-padrão obtido, OR a razão de chances e IC o intervalo com 95% de confiança. Para o preditor linear da resposta, o tempo de infecção foi considerado na escala logarítmica e a interpretação deve ser feita na escala original, tornando o efeito multiplicativo.

Tabela 10 – Modelo I- EM com Tempo de infecção, antibiótico e 5% de significância

Preditor	Parâmetro	Estimativa	EP	P-Valor	OR	IC 95% (OR)
	β_0	-10,613	5,363	-	-	-
RREI	β_1 (Log Tempo Infecção)	3,447	1,637	0,035	1,87	[1,04; 3,36]
	γ_0	-0,207	0,671	-	-	-
Cura	γ_1 (Antibiótico-Não)	-2,352	0,933	0,01	0,095	[0,01; 0.59]

Uma possível interpretação para o tempo de infecção seria: com o aumento de 20% do tempo de infecção, a chance da ocorrência de RREI é 1,87 vezes a de não ocorrência, ou seja, um aumento de 87%. Para o preditor da cura, a interpretação é: pacientes para os quais a antibióticoterapia sistêmica não foi prescrita tiveram aproximadamente 90.5% menos chance de ter cicratização pulpar. O mesmo modelo foi ajustado com a verossimilhança observada. Os resultados são muito parecidos e são apresentados no Apêndice C, fato que era esperado de acordo com o observado nas simulações.

Na etapa 2 foi ajustado um modelo com antibiótico e período extra-oral (PerEO), no preditor linear da cura. Esse pode ser visto como um modelo concorrente, se consideramos o nível de significância de 10% na etapa 2, já que oferece uma interpretação útil para o pesquisador. O modelo está apresentado na Tabela 11. A coluna IC é o intervalo com 90% de confiança.

Tabela 11 – Modelo II-EM com Tempo de infecção, antibiótico, PerEO e 10% de significância

Preditor	Parâmetro	Estimativa	EP	P-Valor	OR	IC 90% (OR)
	β_0	-14,623	8,622	-	-	-
RREI	β_1 (Log Tempo Infecção)	4,691	2,680	0,080	2,352	[1,05; 5,23]
	γ_0	2,773	1,872	-	-	-
Cura	γ_1 (Antibiótico-Não)	-2,641	0,956	0,005	0,071	[0,01; 0,34]
	γ_2 (Log PerEO)	-0,591	0,354	0,095	0,897	[0,80; 0,99]

Como período extra-oral deve ser interpretado em sua escala original, pode-se concluir que, com o aumento de 20% do período extra-oral, a chance de cura diminui em

aproximadamente 11%. Em relação ao antibiótico, pacientes que não tiveram a prescrição tem aproximadamente 93% menos chance de cura (cicatrização pulpar). Para o preditor linear da resposta observada, com um aumento de 20% do tempo de infecção, a chance de RREI é 2,35 vezes maior. O ajuste com a verossimilhança observada está no Apêndice C. Os resultados são muito parecidos.

Do ponto de vista clínico o modelo II é mais interessante. A presença de período extra-oral é coerente neste modelo com o que já foi encontrado em estudos prévios, além de reforçar a recomendação de que o reimplante deve ser realizado o mais rápido possível, idealmente dentro de 15 minutos.

4.2 Amostra com conglomerados

No ajuste do modelo com conglomerados, adotou-se uma estratégia parecida com a Seção 4.1. A conclusão deriva do intervalo de confiança *bootstrap*, baseados nos quantis, com 200 repetições, considerando-se significativo quando o valor 0 não pertence ao intervalo. Foi utilizado o intervalo baseados nos quantis por ser mais robusto do que métodos que utilizam a suposição de normalidade, como o p-valor.

Os passos de ajuste do modelo utilizados foram:

1. Fixou-se o logaritmo do tempo de infecção no preditor linear da resposta observada. Então ajustou-se no preditor da cura todas as outras covariáveis, separadamente. Passaram para a etapa seguinte as covariáveis significativas.
2. Ajustou-se todas as covariáveis que passaram na etapa (1) para o preditor da cura conjuntamente. Retirou-se todas as covariáveis não significativas até se obter um modelo com todos os marcadores significativos.
3. Todas as covariáveis retiradas em (2) retornam, separadamente, para confirmar se não são estatisticamente significativas.
4. Ajustou-se um modelo incluindo-se todas as covariáveis significativas em (3).
5. Ajustou-se no modelo em (4) as covariáveis separadamente no preditor linear da resposta observada e repetiu-se os passos anteriores.

O detalhamento das etapas estão no Apêndice C. Os intervalos de confiança *bootstrap* de 95% e 90% foram utilizados, ficando a cargo do pesquisador a escolha do modelo mais interessante para discussão de resultados dentro da sua área.

A estrutura de correlação simetria composta foi utilizada por ser, entre as implementadas no software R, a mais razoável nessa situação. O Auto-Regressivo de ordem 1 não pode ser utilizado pois não tem componente temporal na casuística. A não estruturada não é viável pela quantidade de parâmetros de correlação a serem estimados e a independência ignora a presença de medidas repetidas, o que pode acarretar vícios no erro padrão. Ainda que, na presente amostra poucos indivíduos apresentassem mais de um dente traumatizado, esta é uma característica frequente nos bancos de dados do PTD FAO UFMG. Sendo assim, o objetivo foi ajustar o modelo estatístico mais apropriado, visando sua aplicação em estudos futuros.

A Tabela 12 apresenta o modelo final,com intervalos com 95% de confiança *bootstrap* para a razão de chances, com logaritmo do tempo de infecção no preditor linear da resposta observada e antibiótico para o preditor linear da cura. Os intervalos de confiança são os quantis de ordem 2.5% e 97.5%. DP indica o desvio padrão das estimativas obtidas pelo *bootstrap* e OR é a razão de chances.

Tabela 12 – Modelo I-ES com Tempo de infecção, antibiótico e 5% de significância

Preditor	Parâmetro	Estimativa	DP Boot	OR	IC Boot OR (95%)
	β_0	-12,183	3,051	-	-
RREI	β_1 (Log Tempo Infecção)	3,877	0,883	2,027	[1,75; 3,15]
	γ_0	-0,686	0,718	-	-
Cura	γ_1 (Antibiótico-Não)	-1,960	0,759	0,140	[0,04; 0,93]

Uma possível interpretação para o tempo de infecção é: Com um aumento de 20% do tempo de infecção, a chance de ocorrência de RREI foi 2,027 vezes maior. Para a cura, o paciente que não teve a prescrição de antibiótico teve 86% menos chance de ter cicratização pulpar. É o mesmo modelo encontrado anteriormente, o que era esperado, pois tem-se poucos agrupamentos. A estimativa para o parâmetro de correlação na estrutura da resposta observada é -0,012 (Erro-padrão = 4,726) e na estrutura da cura, é -0,186 (Erro-padrão = 0,232).

Considerando o nível de significância de 10%, nos passos descritos de seleção do modelo final, encontra-se um ajuste com coeficientes significativos com logaritmo do tempo

de infecção no preditor linear da resposta observada e prescrição de antibiótico com logaritmo do período extra-oral no preditor da cura. Os intervalos de confiança *bootstrap* são com os quantis de ordem 5% e 95%. Os resultados são apresentados na Tabela 13:

Tabela 13 – Modelo II-ES com Tempo de infecção, antibiótico, ´PerEO e 10% de significância

Preditor	Parâmetro	Estimativa	DP Boot	OR	IC Boot OR (90%)
RREI	β_0	-13,984	7,021	-	-
	β_1 (Log Tempo Infecção)	4,321	2,023	2,199	[1,17; 3,56]
Cura	γ_0	1,872	1,879	-	-
	γ_1 (Antibiótico-Não)	-2,187	0,872	0,112	[0,02; 0,53]
	γ_2 (Log PerEO)	-0,514	0,314	0,910	[0,80; 0,99]

Uma possível interpretação para o período extra-oral é que, com o aumento de 20% do tempo do dente fora da boca após a avulsão, a chance de cura diminuiu em aproximadamente 9%. Pacientes que não tiveram a prescrição de antibiótico após o reimplante tem aproximadamente 89% menos chance de cura. Já para o preditor linear da resposta observada, com o aumento de 20% no tempo de infecção, a chance de se observar RREI foi 2.20 vezes maior. A estimativa do parâmetro de correlação para o preditor linear da resposta observada é -0,228 (Erro-padrão de 9,817) e para a cura é -0,176 (Erro-padrão 0,212).

Os modelos II encontrados nas Seções 4.1 e 4.2 são parecidos. Esse fato era esperado devido a pouca quantidade de pacientes com mais de uma medição. Estudos adicionais são importantes para avaliar as situações em que há ganhos substanciais em considerar a metodologia com conglomerados. Como discutido na Seção anterior,o modelo II corrobora com resultados obtidos nos estudos anteriores da literatura, inclusive aqueles realizados pelo grupo de pesquisa do PTD FAO UFMG

Assim como na simulação, a utilização do bootstrap tem vantagem e desvantagens. A vantagem é que foi possível identificar fatores associados à ocorrência de reabsorção e da cura, que era o objetivo prático do pesquisador, com um implementação computacional que não é complexa. As desvantagens e limitações são o tempo computacional pela quantidade de repetições necessárias e possíveis erros numéricos que podem acontecer. Esses problemas dependem do sorteio obtido no bootstrap. Em alguns casos no sorteio, tem-se poucos eventos na variável resposta de interesse e problemas de convergência ocorreram para todas as covariáveis, com mais frequência nas covariáveis categóricas, como meio de

armazenamento e grau de rizogênese. Nesses casos, a amostra bootstrap é descartada. Esta é uma possibilidade real na área de saúde em que muitas vezes trabalha-se com amostras relativamente pequenas.

Não foi possível calcular a estimativa pontual e intervalar para a proporção de cura por questões de não identificabilidade, já que quando ajustamos apenas o intercepto no preditor linear da fração de cura, o erro padrão estimado é excessivamente grande. Os parâmetros de correlação estimados são baixos, e era esperado, dado que tem-se poucos indivíduos com mais de uma medição.

5 Considerações finais

Modelos ZI (Zero-Inflated) são adequados quando existem mais zeros do que o esperado para um distribuição de contagem. Porém, em muitos casos, não se sabe qual o fenômeno que causa a inflação de zeros. Ao determinar a função de verossimilhança da regressão logística binária com fator de cura, notamos que é um caso particular dessa função para modelo ZIB. Portanto, podemos inferir que a presença de cura é um dos efeitos que causam a inflação de zeros. Este fato é reforçado por Diop et al. (2011): a presença de cura em resposta binária pode ser encarada como um ajuste zero inflacionado binomial.

O desafio dessa dissertação foi ajustar de maneira adequada situações em que temos mais de uma medição por indivíduo em modelos de cura para resposta binária. Usamos a metodologia presente em Hall e Zhang (2004). Os autores alteram o passo de maximização do algoritmo **EM** em modelos ZI, para acomodação de medidas amostrais em conglomerados, ao incluir as equações do GEE. A grande vantagem desse processo de estimação é a fatoração da função de verossimilhança para dados completos. Também é apresentada pelos autores uma estratégia para obter maior eficiência no processo de estimação, substituindo as equações de estimação em 24 e 25, apresentadas na Seção 3, por equações de estimativas combinadas. Porém, não foi necessário a implementação nessa dissertação, pois os resultados encontrados nas simulações já são satisfatórios, isso é, a média das estimativas pontuais estão próximas dos valores reais, com pequenos vícios relativos, e o desvio das estimativas diminuem com o aumento da quantidade de conglomerados.

Xu (2013) define a variância "sanduíche"para os parâmetros de interesse. As equações apresentadas não retiram a influência da inclusão da variável latente no processo de estimação, o que subestima o erro padrão. Esse fato também foi encontrado nas simulações feitas por Xu (2013). Quando temos uma amostra independente da população de interesse, a solução é a utilização da fórmula de Louis (1982). Não foi encontrada na literatura uma metodologia, para as formas apresentadas, com o objetivo de retirar a influência da variável latente. A solução utilizada para contornar essa dificuldade foi o *bootstrap* a nível do conglomerado. Os resultados encontrados foram satisfatórios, o desvio-padrão das estimativas encontradas pelo *bootstrap* são parecidos com os desvios padrão das simulações Monte Carlo, que é considerado como sendo o "erro padrão real".

O efeito da prescrição de antibiótico como preditor de cura é um achado inédito, uma vez que esta variável não foi identificada como significativa em trabalhos prévios sobre este tema. Considerando-se que a base de dados é praticamente a mesma, uma possível explicação seria as diferentes metodologias estatísticas empregadas, podendo-se conjecturar que o presente modelo seria mais apropriado para as característica dos dados. Não foi encontrado na literatura trabalhos que levam em consideração o fato da cicratização pulpar ser um fator de cura latente na primeira consulta do paciente para o desfecho de interesse.

As limitações dos modelos estão relacionadas com menores tamanhos amostrais. Foram observados problemas númericos nas simulações no Capítulo 2 e Capítulo 3, com $n = 100$ e $K = 100$ grupos, respectivamente. Nenhum problema númerico foi identificado para maiores tamanhos amostrais e quantidade de grupos. Além disso, dificultades númericas também foram encontrados no banco de dados que ilustra a metodologia, que contém 90 pacientes e 104 dentes. Em muitos situações reais, principalmente na área da saúde, um estudo com esses tamanhos amostrais já são considerados grandes. Como solução para encontrar uma estratégia de seleção das covariáveis, pode ser fundamental manter um canal de comunicação com o pesquisador utilizando o resultado já existente na literatura. Nesta dissertação foi utilizado o estudo de Bastos et al. (2015), que aponta evidências de associação do tempo de infecção com o desfecho de interesse.

Desafios para futuros trabalhos são encontrar formas para retirar o efeito da variável latente na estimação da variância "sanduíche" apresentada em Xu (2013), o tratamento de medidas ausentes/dados perdidos nas covariáveis, já que faz parte da casuística da aplicação real, análise de resíduos para verificar a qualidade de ajuste e medir o percentual de conglomerados em que haveria ganhos substanciais em usar a abordagem da metodologia do Capítulo 3.

Referências[1]

BASTOS, J. et al. Age and timing of pulp extirpation as major factors associated with inflammatory root resorption in replanted permanent teeth. *Journal of endodontics*, Elsevier, v. 40, n. 3, p. 366–371, 2014. Citado 2 vezes nas páginas 8 e 35.

BASTOS, J. et al. A study of the interleukin-1 gene cluster polymorphisms and inflammatory external root resorption in replanted permanent teeth. *International endodontic journal*, Wiley Online Library, v. 48, n. 9, p. 878–887, 2015. Citado 2 vezes nas páginas 8 e 42.

CARIO, M. C.; NELSON, B. L. *Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix*. [S.l.], 1997. Citado 2 vezes nas páginas 29 e 30.

DEMPSTER, A. P. et al. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 39, n. 1, p. 1–22, 1977. Citado na página 15.

DIOP, A. et al. Maximum likelihood estimation in the logistic regression model with a cure fraction. *Electronic journal of statistics*, The Institute of Mathematical Statistics and the Bernoulli Society, v. 5, p. 460–483, 2011. Citado 6 vezes nas páginas 4, 5, 13, 14, 15 e 41.

ENGLE, R. F. Wald, likelihood ratio, and lagrange multiplier tests in econometrics. *Handbook of econometrics*, Elsevier, v. 2, p. 775–826, 1984. Citado na página 13.

FOLLMANN, D. A.; LAMBERT, D. Identifiability of finite mixtures of logistic regression models. *Journal of Statistical Planning and Inference*, Elsevier, v. 27, n. 3, p. 375–381, 1991. Citado na página 14.

HALL, D. B. Zero-inflated poisson and binomial regression with random effects: a case study. *Biometrics*, Wiley Online Library, v. 56, n. 4, p. 1030–1039, 2000. Citado 4 vezes nas páginas 7, 12, 17 e 25.

HALL, D. B.; ZHANG, Z. Marginal models for zero inflated clustered data. *Statistical Modelling*, Sage Publications Sage CA: Thousand Oaks, CA, v. 4, n. 3, p. 161–180, 2004. Citado 10 vezes nas páginas 4, 5, 7, 9, 16, 25, 26, 27, 28 e 41.

KELLEY, M. E.; ANDERSON, S. J. Zero inflation in ordinal data: incorporating susceptibility to response through the use of a mixture model. *Statistics in medicine*, Wiley Online Library, v. 27, n. 18, p. 3674–3688, 2008. Citado na página 14.

LI, S. T.; HAMMOND, J. L. Generation of pseudorandom numbers with specified univariate distributions and correlation coefficients. *IEEE Transactions on Systems, Man, and Cybernetics*, IEEE, n. 5, p. 557–561, 1975. Citado 2 vezes nas páginas 29 e 30.

LOUIS, T. A. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 44, n. 2, p. 226–233, 1982. Citado 2 vezes nas páginas 18 e 41.

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

MALLER, R. A.; ZHOU, X. *Survival analysis with long-term survivors*. [S.l.]: John Wiley & Sons, 1996. Citado na página 10.

MEYER, B. D.; MITTAG, N. Misclassification in binary choice models. *Journal of Econometrics*, Elsevier, v. 200, n. 2, p. 295–311, 2017. Citado na página 7.

NOCEDAL, J. Updating quasi-newton matrices with limited storage. *Mathematics of computation*, v. 35, n. 151, p. 773–782, 1980. Citado na página 13.

PIRES, M. C.; QUININO, R. d. C. Repeated responses in misclassification binary regression: A bayesian approach. *Statistical Modelling*, SAGE Publications Sage India: New Delhi, India, v. 19, n. 4, p. 412–443, 2019. Citado na página 7.

SANSEVERINO, E. B. *Aceleração Do Método De Otimização L-BFGS Usando Tecnologia CUDA*. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2014. Citado na página 13.

SY, J.; TAYLOR, J. Standard errors for the cox proportional hazards cure model. *Mathematical and computer modelling*, Elsevier, v. 33, n. 12-13, p. 1237–1251, 2001. Citado na página 18.

TOULOUMIS, A. Simulating correlated binary and multinomial responses under marginal model specification: The simcormultres package. *The R Journal*, v. 8, n. 2, p. 79–91, 2016. Disponível em: <https://journal.r-project.org/archive/2016/RJ-2016-034/index.html>. Citado 2 vezes nas páginas 29 e 30.

XU, S. Generalized estimating equation based zero-inflated models with application to examining the relationship between dental caries and fluoride exposures. 2013. Citado 4 vezes nas páginas 28, 31, 41 e 42.

YAMAGUCHI, M. et al. Preliminary criteria for classification of adult still's disease. *The Journal of rheumatology*, v. 19, n. 3, p. 424, 1992. Citado na página 12.

Apêndice A – Inferência para o modelo binário com fator de cura.

Esse Apêndice apresenta as contas para inferência das quantidades desconhecidas $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ na Seção 2.2.

A função escore é definida como:

$$\mathbf{U}\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right)=\begin{pmatrix}\mathbf{U}\left(\boldsymbol{\beta}\right)\\ \mathbf{U}\left(\boldsymbol{\gamma}\right)\end{pmatrix}=\begin{pmatrix}\dfrac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma};Y)}{\partial\boldsymbol{\beta}}\\ \dfrac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma};Y)}{\partial\boldsymbol{\gamma}}\end{pmatrix}$$

para *j = 1, ... , p+1* e *r = 1, ... , q+1*, temos que:

$$\mathbf{U}_j\left(\boldsymbol{\beta}\right)=\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma};Y)}{\partial\beta_j} \tag{28}$$

e

$$\mathbf{U}_r\left(\boldsymbol{\gamma}\right)=\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma};Y)}{\partial\gamma_r} \tag{29}$$

Para *j = 1, ... , p+1*, temos que:

$$\begin{aligned}\mathbf{U}_j\left(\boldsymbol{\beta}\right)&=\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma};Y)}{\partial\beta_j}\\ &=-\sum_{i=1}^{n}\Bigg(I_{(0)}(y_i)\left[\frac{exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)}{1+exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)}\frac{\mathbf{X}_{ij}}{\left(1+exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)\right)exp\left(\mathbf{Z}_i\boldsymbol{\gamma}\right)+1}\right]\\ &+\left(1-I_{(0)}(y_i)\right)\left[\mathbf{X}_{ij}\left(1-\frac{exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)}{1+exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)}\right)\right]\Bigg)\end{aligned} \tag{30}$$

em que $I_{(0)}(y_i)$ é a indicadora se y_i é igual a 0.

Reescrevendo (A.3) em função de π_1 e p_1, temos:

$$\mathbf{U}_j\left(\boldsymbol{\beta}\right)=\sum_{i=1}^{n}\left(s_i\mathbf{X}_{ij}\right)$$

em que:

$$s_i=-I_{(0)}(y_i)\frac{\pi_{1i}\left(1-\pi_{1i}\right)\left(1-p_{1i}\right)}{p_{1i}+\left(1-p_{1i}\right)\left(1-\pi_{1i}\right)}+\left(1-I_{(0)}(y_i)\right)\left(1-\pi_{1i}\right)$$

para *r = 1, ... , q+1*, temos que:

$$\begin{aligned}\mathbf{U}_j(\boldsymbol{\gamma}) &= \frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma};Y)}{\partial \gamma_r} \\ &= \sum_{i=1}^{n}\left[I_{(0)}(y_i)\left(\frac{\mathbf{Z}_{ir} exp(\mathbf{Z}_i\boldsymbol{\gamma})(1+exp(\mathbf{X}_i\boldsymbol{\beta}))}{(1+exp(\mathbf{X}_i\boldsymbol{\beta}))exp(\mathbf{Z}_i\boldsymbol{\gamma})+1}\right)\right. \\ &\left. - \frac{exp(\mathbf{Z}_i\boldsymbol{\gamma})}{1+exp(\mathbf{Z}_i\boldsymbol{\gamma})}\mathbf{Z}_{ir}\right] \end{aligned} \tag{31}$$

Reescrevendo (A.4) em função de π_1 e p_1, temos:

$$\mathbf{U}_j(\gamma) = \sum_{i=1}^{n}(t_i\mathbf{Z}_{ir})$$

em que:

$$t_i = I_{(0)}(y_i)\frac{p_{1i}}{p_{1i}+(1-\pi_{1i})(1-p_{1i})} - p_{1i}$$

Seja $\mathbf{M}_{\boldsymbol{\beta}} = (s_1,...,s_n)^T$ e $\mathbf{M}_{\boldsymbol{\gamma}} = (t_1,...,t_n)^T$ e defina-se $\mathbf{M}$ um vetor de dimensão $2n \times 1$ e $\mathbf{G}$ uma matriz de dimensão $2n \times (p+q+2)$ como:

$$\mathbf{M} = \begin{pmatrix}\mathbf{M}_{\boldsymbol{\beta}} \\ \mathbf{M}_{\boldsymbol{\gamma}}\end{pmatrix} \text{ e } \mathbf{G} = \begin{pmatrix} X & 0 \\ 0 & Z\end{pmatrix}$$

A função escore pode ser escrita como:

$$\mathbf{U}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \begin{pmatrix} X'\mathbf{M}_{\boldsymbol{\beta}} \\ Z'\mathbf{M}_{\boldsymbol{\gamma}}\end{pmatrix} = \mathbf{G}'\mathbf{M}$$

A matriz de informação de Fisher $\mathbf{I}(\boldsymbol{\beta},\boldsymbol{\gamma})$ pode ser dividida em:

$$\mathbf{I}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \begin{pmatrix}\mathbf{I}_{\boldsymbol{\beta\beta}} & \mathbf{I}_{\boldsymbol{\beta\gamma}} \\ \mathbf{I}_{\boldsymbol{\gamma\beta}} & \mathbf{I}_{\boldsymbol{\gamma\gamma}}\end{pmatrix} \tag{32}$$

e seus elementos são dados por:

$$\mathbf{I}_{\boldsymbol{\beta\beta}} = -E\left[\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\gamma};Y)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right]$$

$$\mathbf{I}_{\boldsymbol{\gamma\gamma}} = -E\left[\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\gamma};Y)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}\right]$$

$$\mathbf{I}_{\boldsymbol{\beta\gamma}} = -E\left[\frac{\partial^2 l\left(\boldsymbol{\beta},\boldsymbol{\gamma};Y\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}'}\right]$$
$$\mathbf{I}_{\boldsymbol{\gamma\beta}} = \mathbf{I}'_{\boldsymbol{\beta\gamma}}$$

Para r e s de *1, ... , q+1*, j e k de *1, ... , p+1* e $j \neq k$, $r \neq s$, temos que:

$$\frac{\partial^2 l\left(\boldsymbol{\beta},\boldsymbol{\gamma};Y\right)}{\partial\gamma_s\partial\gamma_r} = \sum_{i=1}^{n}\left(\mathbf{Z}_{ir}\mathbf{Z}_{is}\exp\left(\mathbf{Z}_i\boldsymbol{\gamma}\right)\left[I_{(0)}(y_i)\frac{\left(1+\exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)\right)}{\left[\left(1+\exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)\right)\exp\left(\mathbf{Z}_i\gamma\right)+1\right]^2}\right]\right.$$
$$\left. - \frac{1}{\left[1+exp\left(\mathbf{Z}_i\boldsymbol{\gamma}\right)\right]^2}\right) \tag{33}$$

$$\frac{\partial^2 l\left(\boldsymbol{\beta},\boldsymbol{\gamma};Y\right)}{\partial\beta_k\partial\beta_j} = -\sum_{i=1}^{n}\left(\frac{\mathbf{X}_{ik}\mathbf{X}_{ij}\exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)}{\left(1+\exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)\right)^2}\left[\left(I_0(y_i)\frac{exp\left(\mathbf{Z}_i\boldsymbol{\gamma}\right)-\exp\left(\mathbf{Z}_i\boldsymbol{\gamma}\right)\left[\exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)\right]^2+1}{\left[\left(1+\exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)\right)\exp\left(\mathbf{Z}_i\boldsymbol{\gamma}\right)+1\right]^2}\right)\right.\right.$$
$$\left.\left. - \left(1-I_0(y_i)\right)\right]\right) \tag{34}$$

e

$$\frac{\partial^2 l\left(\boldsymbol{\beta},\boldsymbol{\gamma};Y\right)}{\partial\beta_j\partial\gamma_s} = \sum_{i=1}^{n}\left[\mathbf{X}_{ij}\mathbf{Z}_{is}I_0(y_i)\left(\frac{\exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)\exp\left(\mathbf{Z}_i\boldsymbol{\gamma}\right)}{\left[\left(1+\exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)\right)\exp\left(\mathbf{Z}_i\boldsymbol{\gamma}\right)+1\right]^2}\right)\right] \tag{35}$$

Reescrevendo (A.6), (A.7) e (A.8) em função de π_1 e p_1, temos que:

$$\frac{\partial^2 l\left(\boldsymbol{\beta},\boldsymbol{\gamma};Y\right)}{\partial\gamma_s\partial\gamma_r} = \sum_{i=1}^{n}\left(\mathbf{Z}_{ir}\mathbf{Z}_{is}\frac{p_{1i}}{1-p_{1i}}\left[I_{(0)}(y_i)(1-\pi_{1i})\left(\frac{1-p_{1i}}{p_{1i}+(1-p_{1i})(1-\pi_{1i})}\right)^2\right]-(1-p_{1i})^2\right)$$

$$\frac{\partial^2 l\left(\boldsymbol{\beta},\boldsymbol{\gamma};Y\right)}{\partial\beta_k\partial\beta_j} = -\sum_{i=1}^{n}\left[\mathbf{X}_{ik}\mathbf{X}_{ij}\pi_{1i}(1-\pi_{1i})\left[I_0(y_i)\left(\left((1-\pi_{1i})^2-p_{1i}\pi_{1i}^2\right)\left(\frac{1-p_{1i}}{\left[p_{1i}+(1-p_{1i})(1-\pi_{1i})\right]^2}\right)\right)\right.\right.$$
$$\left.\left. - \left(1-I_0(y_i)\right)\right]\right]$$

e

$$\frac{\partial^2 l\left(\boldsymbol{\beta},\boldsymbol{\gamma};Y\right)}{\partial\beta_j\partial\gamma_s} = \sum_{i=1}^{n}\mathbf{X}_{ij}\mathbf{Z}_{is}\left[I_0(y_i)\left(\frac{p_{1i}\pi_{1i}(1-\pi_{1i})(1-p_{1i})^2}{\left[p_{1i}+(1-p_{1i})(1-\pi_{1i})\right]^2}\right)\right]$$

Como $E\left[I_0(y_i)\right] = P(Y=0) = p_i + (1-p_1)\left(1-\pi_1\right)$ e $E\left[1-I_0(y_i)\right] = P(Y=1) = (1-p_1)\pi_1$ temos que :

$$I_{\beta\beta_{jk}} = \sum_{i=1}^{n}\left[u_i \mathbf{X}_{ij}\mathbf{X}_{ik}\right]$$

$$I_{\gamma\gamma_{rs}} = \sum_{i=1}^{n}\left[l_i \mathbf{Z}_{ir}\mathbf{Z}_{is}\right]$$

$$I_{\beta\gamma_{js}} = \sum_{i=1}^{n}\left[d_i \mathbf{X}_{ij}\mathbf{Z}_{is}\right]$$

em que

$$u_i = \frac{\left[(1-\pi_{1i})^2 - p_{1i}\pi_{i1}^2\right](1-p_{1i})(1-\pi_{1i})\pi_{1i}}{p_{1i}+(1-p_1)(1-\pi_1)} - (1-p_{1i})\pi_{1i}$$

$$l_i = \frac{p_{1i}}{1-p_{1i}}\left[\frac{(1-\pi_{1i})(1-p_{1i})^2}{p_{1i}+(1-p_1 i)(1-\pi_{1i})} - (1-p_{1i})^2\right]$$

e

$$d_i = -\frac{p_{1i}\pi_{1i}(1-\pi_{1i})(1-p_1 i)^2}{p_{1i}+(1-p_{1i})(1-\pi_{1i})}$$

Seja $\mathbf{C}_{\boldsymbol{\beta\beta}} = \operatorname{diag}(u_1,...,u_n)$, $\mathbf{C}_{\boldsymbol{\gamma\gamma}} = \operatorname{diag}(l_1,...,l_n)$ e $\mathbf{C}_{\beta\gamma} = \operatorname{diag}(d_1,...,d_n)$ e $\mathbf{C}$ uma matriz de dimensão $2n \times 2n$ como sendo:

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{\boldsymbol{\beta\beta}} & \mathbf{C}_{\boldsymbol{\beta\gamma}} \\ \mathbf{C}_{\boldsymbol{\gamma\beta}} & \mathbf{C}_{\boldsymbol{\gamma\gamma}} \end{pmatrix}$$

A matriz de informação de Fisher pode ser escrita como:

$$\mathbf{I}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \mathbf{G}^{'}\mathbf{C}\mathbf{G}$$

Logo, as estimativas dos parâmetros podem ser obtidas pelo método de escore de Fisher. Considerando um chute inicial $\left(\boldsymbol{\beta}^0,\boldsymbol{\gamma}^0\right)$ e $m=0$. O passo $m+1$ do algortimo interativo é:

$$\begin{pmatrix} \boldsymbol{\beta}^{m+1} \\ \gamma^{m+1} \end{pmatrix} = \begin{pmatrix} \beta^m \\ \gamma^m \end{pmatrix} + \left[\mathbf{I}(\beta^m,\gamma^m)\right]^{-1}\mathbf{U}(\beta^m,\gamma^m)$$

A matriz de variância e covariância é obtida por $\mathbf{K}\left(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}}\right) = \left[\mathbf{I}(\boldsymbol{\beta^m},\boldsymbol{\gamma^m})\right]^{-1}$ no último passo.

Apêndice B – Estudo de simulação

Apresentamos nesse Apêndice nas Tabelas B.1, B.2 e B.3, os resultados das simulações no Capítulo 2, obtidos utilizando a verossimilhança observada.

Tabela 1 – Monte Carlo via Verossimilhança observada- 25% de cura

n	Par	Média	Média Erro padrão	Desvio padrão das estimativas	VR(%)	P(%)
1000	β_0	1,006	0,134	0,140	0,06%	95,9%
	β_1	2,010	0,269	0,272	0,05%	96,0%
	β_2	1,015	0,127	0,141	1,50%	95,0%
	γ_0	-2,090	0,344	0,437	4,50%	90,0%
	γ_1	-2,590	0,477	0,615	3,60%	89,0%
500	β_0	1,071	0,193	0,188	7,10%	96,0%
	β_1	2,120	0,470	0,527	6,60%	96,0%
	β_2	1,038	0,185	0,188	3,80%	97,0%
	γ_0	-2,140	0,527	0,683	7,00%	92,0%
	γ_1	-2,700	0,735	0,903	8,00%	89,0%
100	β_0	1,200	0,519	0,60	20,00%	94,0%
	β_1	2,250	1,120	0,886	12,50%	97,0%
	β_2	1,220	0,505	0,631	22,00%	97,0%
	γ_0	-2,190	1,260	1,220	9,50%	91,0%
	γ_1	-3,160	1,810	1,710	-226,40%	89,0%

Tabela 2 – Monte Carlo via Verossimilhança observada- 50% de cura

n	Par	Média	Média Erro padrão	Desvio padrão das estimativas	VR(%)	P(%)
1000	β_0	1,020	0,678	0,643	2,00%	94,4%
	β_1	2,030	1,280	0,980	1,50%	91,0%
	β_2	1,100	0,486	0,488	10,00%	92,1%
	γ_0	0,018	0,262	0,273	-82,00%	94,9%
	γ_1	1,040	0,135	0,148	4,00%	95,5%
500	β_0	0,910	1,150	1,020	-9,0%	78,5%
	β_1	1,870	1,310	1,390	-6,50%	80,0%
	β_2	1,150	0,697	0,863	15,00%	82,3%
	γ_0	-0,218	0,483	0,574	-318,00%	94,6%
	γ_1	1,100	0,236	0,263	10,00%	95,4%
100	β_0	1,040	1,090	1,150	4,00%	85,0%
	β_1	2,040	1,630	1,340	2,00%	85,0%
	β_2	1,220	0,754	0,873	22,00%	87,4%
	γ_0	-0,093	0,439	0,426	193,00%	96,9%
	γ_1	1,050	0,215	0,206	5,00%	96,1%

Tabela 3 – Monte Carlo via Verossimilhança observada- 75% de cura

n	Par	Média	Média Erro padrão	Desvio padrão das estimativas	VR(%)	P(%)
	β_0	1,030	1,310	1,710	3,00%	86,6%
	β_1	1,910	1,930	1,740	-4,50%	85,1%
1000	β_2	1,220	0,976	1,310	22,00%	83,0%
	γ_0	1,840	0,323	0,313	-8,00%	97,0%
	γ_1	2,670	0,403	0,456	6,80%	97,0%
	β_0	1,060	1,060	2,530	6,00%	86,6%
	β_1	1,660	2,040	2,330	-17,00%	84,8%
500	β_2	1,540	1,390	2,100	54,00%	87,5%
	γ_0	1,74	0,455	0,353	-13,00%	97,3%
	γ_1	3,010	0,682	0,858	20,40%	97,3%
	β_0	0,606	4,230	4,790	-39,40%	76,6%
	β_1	1,020	5,400	5,040	-49,00%	97,4%
100	β_2	3,130	6,300	6,720	213,00%	99,7%
	γ_0	1,190	1,830	2,230	-40,50%	94,8%
	γ_1	5,500	4,040	3,260	120,00%	96,1%

A Tabela B.4 apresenta os resultados obtidos na simulação no Capítulo 3, utilizando o algoritmo **ES** e com a fração de cura média de aproximadamente 75%.

Tabela 4 – Monte Carlo via ES- Fração de cura 75%

K	Par	Média	Média Erro padrão	Desvio padrão das estimativas	VR(%)
	β_0	1,038	0,287	0,705	3,80%
	β_1	2,240	0,345	0,730	12,00%
250	β_2	1,330	0,453	0,928	33,00%
	γ_0	2,08	0,192	0,296	4,00%
	γ_1	2,64	0,224	0,435	5,60%
	β_0	1,280	0,243	0,972	28,00%
	β_1	2,470	0,288	1,04	23,50%
100	β_2	1,93	0,405	2,160	93,00%
	γ_0	2,060	0,149	0,199	3,00%
	γ_1	2,60	0,175	0,324	4,00%

Apêndice C – Aplicação real

A Tabela a seguir apresenta o ajuste do modelo I do Capítulo 4 via verossimilhança observada.

Tabela 1 – Modelo I- Ajuste via verossimilhança observada

Preditor	Parâmetro	Estimativa	EP	P-Valor	OR	IC 95% (OR)
	β_0	-10,619	5,366	-	-	-
RREI	β_1 (Log Tempo Infecção)	3,449	1,639	0,035	1,875	[1,20; 1,85]
	γ_0	-0,206	0,671	-	-	-
Cura	γ_1 (Antibiótico-Não)	-2,352	0,932	0,011	0,095	[0,01; 0,59]

A Tabela a seguir apresenta o ajuste do modelo II do Capítulo 4 via verossimilhança observada.

Tabela 2 – Modelo II-Ajuste via verossimilhança observada

Preditor	Parâmetro	Estimativa	EP	P-Valor	OR	IC 90% (OR)
	β_0	-14,661	8,667	-	-	-
RREI	β_1 (Log Tempo Infecção)	4,704	2,696	0,081	2,358	[1,05; 5,28]
	γ_0	2,774	1,871	-	-	-
Cura	γ_1 (Antibiótico-Não)	-2,640	0,955	0,005	0,071	[0,01; 0,34]
	γ_2 (Log PerEO)	-0,591	0,354	0,095	0,897	[0,80; 0,99]

As Tabelas C.3, C.4 e C.5 apresentam o detalhamento das etapas para encontrar o modelo final do Capítulo 4, seção 4.1

Tabela 3 – Parte 1 do ajuste de modelo- EM

Covariável	P-Valor
Log-Tempo Infec-Preditor da resposta obs	0,006
PereO-Preditor da cura	0,29
Meio-Preditor da cura	0,72
Log-Tempo Infec-Preditor da resposta obs	0,035
Antibiótico-Preditor da cura	0,011
Log-Tempo Infec-Preditor da resposta obs	0,014
PereO-Preditor da cura	0,19
Sexo-Preditor da cura	0,65
Log-Tempo Infec-Preditor da resposta obs	0,006
PereO-Preditor da cura	0,18
Log-Tempo Infec-Preditor da resposta obs	$< 0{,}001$
Grau de rizogênese-Preditor da cura	0,76
PereO-Preditor da cura	0,17

Tabela 4 – Parte 2 do ajuste de modelo- EM

Covariável	P-Valor
Log-Tempo Infec-Preditor da resposta obs	0,08
Antibiótico-Preditor da cura	0,005
PereO-Preditor da cura	0,095
Log-Tempo Infec-Preditor da resposta obs	0,03
Antibiótico-Preditor da cura	0,01

Tabela 5 – Parte 3 do ajuste de modelo- EM

Covariável	P-Valor
Log-Tempo Infec-Preditor da resposta obs	0,04
Sexo-Preditor da resposta obs	0,96
Antibiótico-Preditor da cura	0,01
Log-Tempo Infec-Preditor da resposta obs	0,03
Meio-Preditor da resposta obs	0,60
Antibiótico-Preditor da cura	0,01
Log-Tempo Infec-Preditor da resposta obs	0,09
PereO-Preditor da resposta obs	0,14
Antibiótico-Preditor da cura	0,007
Log-Tempo Infec-Preditor da resposta obs	0,04
Grau de Rizo-Preditor da resposta obs	0,93
Antibiótico-Preditor da cura	0,01

As Tabelas C.6, C.7 e C.8 a seguir apresentam os passos realizados para ajuste do modelo no Capítulo 4, seção 4.2, com intervalos de 95 % de confiança *bootstrap*.

Tabela 6 – Parte 1 do ajuste de modelo- ES

Covariável	IC(95%)-Bootstrap
Log-Tempo Infec-Preditor da resposta obs	[0,64; 3,03]
PereO-Preditor da cura	[-4,85; 3,59]
Meio-Preditor da cura	[-1,72; 0,44]
Log-Tempo Infec-Preditor da resposta obs	[3,07; 6,30]
Antibiótico-Preditor da cura	[-3,17; -0,067]
Log-Tempo Infec-Preditor da resposta obs	[0,63; 2,89]
PereO-Preditor da cura	[-1,94; 0,35]
Sexo-Preditor da cura	[-7,68; 5,00]
Log-Tempo Infec-Preditor da resposta obs	[1,75; 2,97]
PereO-Preditor da cura	[-1,71; 0,30]
Log-Tempo Infec-Preditor da resposta obs	[0,62; 3,02]
Grau de rizogênese-Preditor da cura	[-3,07; 3,38]
PereO-Preditor da cura	[-1,70; 0,29]

Tabela 7 – Parte 2 do ajuste de modelo- ES

Covariável	IC(95%)-Bootstrap
Log-Tempo Infec-Preditor da resposta obs	[0,44; 9,96]
Antibiótico- Preditor da cura	[-4,05; -0,16]
Meio-Preditor da cura	[-2,72; 1,41]
Log-Tempo Infec-Preditor da resposta obs	[0,75; 6,75]
Antibiótico- Preditor da cura	[-4,35; -0,25]
Sexo-Preditor da cura	[-3,27; 0,92]
Log-Tempo Infec-Preditor da resposta obs	[-0,30; 10,13]
Antibiótico- Preditor da cura	[-3,96; -0,45]
PereO-Preditor da cura	[-1,25; 0,149]
Log-Tempo Infec-Preditor da resposta obs	[0,61; 7,13]
Antibiótico- Preditor da cura	[-3,74; -0,18]
Grau de rizogênese-Preditor da cura	[-1,84; 1,85]

Tabela 8 – Parte 3 do ajuste de modelo- ES

Covariável	IC(95%)-Bootstrap
Log-Tempo Infec-Preditor da resposta obs	[0,52; 10,58]
Meio- Preditor da resposta obs	[-2,76; 2,63]
Antibiótico- Preditor da cura	[-3,61; -0,19]
Log-Tempo Infec-Preditor da resposta obs	[0,50; 7,23]
Sexo- Preditor da resposta obs	[-3,25; 3,20]
Antibiótico- Preditor da cura	[3,73; -0,19]
Log-Tempo Infec-Preditor da resposta obs	[-1,94; 27,80]
PereO-Preditor da resposta obs	[-1,23; 8,85]
Antibiótico- Preditor da cura	[-3,61; -0,35]
Log-Tempo Infec-Preditor da resposta obs	[0,55; 7,42]
Grau de rizogênese-Preditor da resposta obs	[-2,25; 3,25]
Antibiótico- Preditor da cura	[-3,68; -0,24]

As Tabelas C.9, C.10 e C.11 a seguir apresentam os passos realizados para ajuste do modelo no Capítulo 4, seção 4.2, com intervalos de 90 % de confiança Bootstrap.

Tabela 9 – Parte 1 do ajuste de modelo- ES

Covariável	IC(90%)-Bootstrap
Log-Tempo Infec-Preditor da resposta obs	[0,84; 4,83]
PereO-Preditor da cura	[-1,54; 1,26]
Meio-Preditor da cura	[-4,16; 2,90]
Log-Tempo Infec-Preditor da resposta obs	[1,18; 6,57]
Antibiótico-Preditor da cura	[-3,44; -0,48]
Log-Tempo Infec-Preditor da resposta obs	[0,82; 4,71]
PereO-Preditor da cura	[-1,75; 0,16]
Sexo-Preditor da cura	[-6,64; 3,97]
Log-Tempo Infec-Preditor da resposta obs	[1,10; 5,57]
PereO-Preditor da cura	[-1,48; 0,70]
Log-Tempo Infec-Preditor da resposta obs	[0,82; 3,82]
Grau de rizogênese-Preditor da cura	[-2,54; 4,85]
PereO-Preditor da cura	[-1,53; 0,13]

Tabela 10 – Parte 2 do ajuste de modelo- ES

Covariável	IC(90%)-Bootstrap
Log-Tempo Infec-Preditor da resposta obs	[0,84; 8,19]
Antibiótico- Preditor da cura	[-3,78; -0,49]
Meio-Preditor da cura	[-2,32; 1,19]
Log-Tempo Infec-Preditor da resposta obs	[1,24; 6,26]
Antibiótico- Preditor da cura	[-4,02; -0,58]
Sexo-Preditor da cura	[-2,93; 0,57]
Log-Tempo Infec-Preditor da resposta obs	[0,86; 6,97]
Antibiótico- Preditor da cura	[-3,62; -0,62]
PereO-Preditor da cura	[-1,14; -0,05]
Log-Tempo Infec-Preditor da resposta obs	[0,96; 8,86]
Antibiótico- Preditor da cura	[-3,33; -0,52]
Grau de rizogênese-Preditor da cura	[-1,29; 2,62]

Tabela 11 – Parte 3 do ajuste de modelo- ES

Covariável	IC(90%)-Bootstrap
Log-Tempo Infec-Preditor da resposta obs	[1,04; 7,71]
Meio- Preditor da resposta obs	[-4,75; 1,49]
Antibiótico- Preditor da cura	[-3,45; -0,50]
Log-Tempo Infec-Preditor da resposta obs	[1,02; 7,48]
Sexo- Preditor da resposta obs	[-2,75; 2,95]
Antibiótico- Preditor da cura	[-3,33; -0,49]
Log-Tempo Infec-Preditor da resposta obs	[0,74; 19,41]
PereO-Preditor da resposta obs	[-0,18; 5,62]
Antibiótico- Preditor da cura	[-3,40; -0,10]
Log-Tempo Infec-Preditor da resposta obs	[-0,52; 14,01]
Grau de rizogênese-Preditor da resposta obs	[-5,41; 2,44]
Antibiótico- Preditor da cura	[-3,31; -0,42]