UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Modelos de Riscos Competitivos no Estudo de Evasão Discente

Fabiana Arca Cruz Tortorelli

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGES)

ICMC USP SÃO CARLOS

Fabiana Arca Cruz Tortorelli

Modelos de Riscos Competitivos no Estudo de Evasão Discente

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

Área de Concentração: Estatística

Orientadora: Prof. Dra. Juliana Cobre

USP – São Carlos
Março de 2020

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

T712m

Tortorelli, Fabiana Arca Cruz
Modelos de Riscos Competitivos no Estudo de Evasão Discente / Fabiana Arca Cruz Tortorelli; orientadora Juliana Cobre. -- São Carlos, 2020.
56 p.

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

1. Evasão Universitária. 2. Riscos Competitivos. 3. Modelo log-log Complementar de Riscos Competitivos. 4. Inferência Bayesiana. I. Cobre, Juliana , orient. II. Título.

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

Fabiana Arca Cruz Tortorelli

Study of University Dropout using Competing Risks Models

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

Concentration Area: Statistics

Advisor: Prof. Dra. Juliana Cobre

USP – São Carlos
March 2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

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

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado da candidata Fabiana Arca Cruz Tortorelli, realizada em 11/02/2020:

Profa. Dra. Juliana Cobre
ICMC/USP

Prof. Dr. Eduardo Yoshio Nakano
UnB

Profa. Dra. Silvia Emiko Shimakura
UFPR

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

Profa. Dra. Juliana Cobre

A meus pais.

AGRADECIMENTOS

Agradeço aos meus pais e meus irmãos, que sempre me apoiam e me ajudam em minhas escolhas.

Agradeço ao Paulo e à Erika que me ofereceram uma oportunidade de conseguir um estudo melhor.

Agradeço aos meus amigos que sempre me apoiam e me motivam a buscar meus sonhos.

Gostaria de agradecer meu namorado que também sempre me apoia e me incentiva a não parar de buscar conhecimentos.

Dedico um agradecimento especial aos meus professores de graduação Sueli Mieko Tanaka Aki, Paulo Leandro Dattori da Silva, Fernando Manfio e Leandro Fiorini Aurichi. Eles foram fundamentais para a minha formação.

Agradeço também ao professor Mário de Castro por dar um suporte na parte de programação e ajudar com inúmeras dúvidas.

À minha professora orientadora Juliana Cobre pela paciência e dedicação e principalmente por me auxiliar nas decisões que foram surgindo no período em que trabalhamos juntas.

Agradeço à CAPES pelo apoio financeiro.

RESUMO

TORTORELLI, F. A. C. **Modelos de Riscos Competitivos no Estudo de Evasão Discente**. 2020. 56 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

Um dos problemas enfrentado por universidades públicas e privadas é a evasão discente. Desta forma, a motivação deste trabalho foi investigar as características que levam um aluno matriculado em um curso regular de graduação aos desfechos evadir e graduar. Para o estudo utilizamos modelos de riscos competitivos para dados discretos, e propomos utilizar a transformação log-log complementar na função de risco para estudar o risco da causa específica (evadir, graduar) ao longo de um período de tempo calculado em semestres. As estimativas dos parâmetros e seleção do modelo foram obtidas através da inferência bayesiana. Verificamos que o modelo proposto neste trabalho consegue ser ajustado aos dados teóricos. Na aplicação em dados reais do curso de Matemática Aplicada do Instituto de Ciências Matemáticas e de Computação (ICMC) concluímos que o grau de instrução dos pais e forma de ingresso podem contribuir para evasão discente.

Palavras-chave: Evasão Universitária, Riscos Competitivos, Modelo log-log Complementar de Riscos Competitivos, Inferência Bayesiana.

ABSTRACT

TORTORELLI, F. A. C. **Study of University Dropout using Competing Risks Models** . 2020. 56 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

Dropout is a problem faced by public and private universities. Therefore, the motivation of this work is to investigate the characteristics of students enrolled in a regular undergraduate course and verify which one is related with dropout and/or graduate. For the study we consider competitive risk models, and we propose to use the transformation of complementary log-log for the risk function into discrete data to study specific cause risk at a period of time (calculated in semesters). For the parameters estimations we used Bayesian Inference. We verified that the model proposed in this work can be adjusted to the theoretical data. In real data from the Applied Mathematics course to the Institute of Mathematical and Computer Sciences (ICMC), we concluded that parental education and admission type influence in dropout outcome.

Keywords: University Dropout, Competing Risks, Competing Risks Complementary log-log Model, Bayesian Inference.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

A evasão discente no Ensino Superior é um problema enfrentado por universidades públicas e privadas. Segundo o Censo do Ensino Superior divulgado pelo Ministério da Educação em 2015, em 2009 o Brasil teve uma perda de aproximadamente nove bilhões de reais com a evasão, pois as instituições, mesmo com poucos alunos, têm que manter a sua infraestrutura (bibliotecas, materiais de ensino, equipamentos, salas de aula, pagamento de professores e de técnicos administrativos, etc). No ensino privado, a evasão pode provocar perdas financeiras para o aluno. Também é possível que cursos com alta taxa de evasão tenham a imagem denegrida. Portanto, além de desperdícios sociais e econômicos, a evasão discente em universidades públicas leva a desperdícios acadêmicos, pois compromete a qualidade do ensino e, por consequência, da pesquisa e da extensão de serviços à comunidade.

Diversas propostas têm sido realizadas para analisar dados de evasão. O trabalho baseado em teoria social e econômica de Tinto (1975) discute como o processo de interação entre os alunos e o sistema acadêmico e social pode contribuir para a permanência ou evasão do estudante no Ensino Superior, mostrando assim que o problema não é recente. Lehmann (2007) analisou através de dados qualitativos e entrevistas o que leva os estudantes definidos como first-generation (aqueles que são os primeiros da família à ingressarem na univerdidade) a abandonarem o curso que ingressaram. Pietro e Cutillo (2008) mostram que, em uma Universidade Italiana, a reestruturação no programa de graduação contribuiu para a redução da evasão.

Willett e Singer (1991) apresentam o quanto se pode aprender sobre os indivíduos ao analisar *quando* um evento ocorre e não apenas *se* o evento irá ocorrer, e modelos de análise de sobrevivência nos permitem saber qual o período de maior risco para um indivíduo. A proposta de Júnior, Silveira e Ostermann (2012) considerou dados longitudinais para ilustrar o uso da Análise de Sobrevivência no fluxo escolar de graduação do curso de Física da Universidade federal de Rio Grande do Sul utilizando tabelas de contingências e técnicas não paramétricas e semi-paramétricas. Paura e Arhipova (2014) também consideram modelos semi paramétricos

para entender os fatores que levam os estudantes a evadir de um curso após o primeiro ano de graduação. Juajibioy (2016) estudou os fatores que levam os estudantes à evasão através do conjunto de dados da Fundación Universidad Autónoma de Colombia usando o método de Análise de Sobrevivência com tempo discreto e modelos semi paramétricos de risco proporcional de Cox (COX, 1972).

Ao ingressar no Ensino Superior, o estudante pode seguir diversos desfechos como abandonar o curso, ser expulso da universidade, mudar de curso (transferir) ou conseguir o diploma e a ocorrência de um desses desfechos impede a ocorrência dos outros em relação ao curso que o aluno se matriculou, um caso de eventos competitivos. Além disso, em geral, a situação dos alunos é atualizada a cada semestre, ou seja, se o indivíduo decide abandonar o curso ou transferir para outro, na mesma ou em outra instituição, ou quando o aluno gradua, a sua situação acadêmica será formalizada ao final do semestre, então os tempos observados são discretos. Desta forma, modelos de análise de sobrevivência com riscos competitivos para dados discretos também foram propostos para estudar o problema da evasão (ORTIS; DEHON, 2011; MEGGIOLARO; GIRALDO; CLERICI, 2017).

Vallejos e Steel (2017) propuseram um modelo de riscos competitivos para dados discretos baseado no modelo de chances proporcionais de Cox (COX, 1972) para identificar fatores que influenciam no risco de evasão e de graduação dos estudantes dos programas de graduação da Pontifícia Universidade Católica do Chile. Neste trabalho propomos um modelo de riscos competitivos para dados discretos baseado na transformação log-log complementar. Um ponto destacado no trabalho deles é a possível violação da suposição da proporcionalidade das chances. Portanto, também investigaremos quais são as consequências de considerar o modelo por nós proposto quando os dados seguem o modelo proposto por Vallejos e Steel (2017) e *vice-versa*. Consideramos também a estimativa através da inferência bayesiana. O objetivo final é obter um modelo que identifique fatores relacionados ao processo da evasão e de que forma isso ocorre.

O trabalho está organizado da seguinte forma: No Capítulo 2 apresentamos algumas definições sobre Análise de Sobrevivência para dados discretos incluindo riscos competitivos e a descrição do modelo de Vallejos e Steel (2017). No Capílulo 3 descrevemos o modelo proposto. No Capítulo 4 ilustramos um estudo de simulação. O Capítulo 5 contém a aplicação da proposta aos dados do Bacharelado em Matemática Aplicada e Computação Científica do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, *campus* São Carlos. Por fim, no Capítulo 6 temos as propostas de trabalhos futuros.

CAPÍTULO

2

CONCEITOS

Análise de Sobrevivência é uma área da Estatística que estuda o tempo, chamado tempo de vida, até observar um evento de interesse. No nosso trabalho estamos interessados em estudar o tempo de permanência do estudante na universidade e este pode apresentar diversos desfechos ao longo de sua trajetória no meio acadêmico como abandonar o curso, ser expulso, transferir de curso ou de universidade, caracterizados como evasão, e graduar. Pode ser que para um indivíduo não observamos nenhum dos desfechos durante o período estudado, que na prática são os matriculados, caracterizando, no contexto de análise de sobrevivência, uma observação censurada. Uma vez que a situação do aluno é, em geral, atualizada a cada semestre, fica mais próximo da realidade se considerarmos em nossa análise tempos discretos.

Importantes livros-texto sobre Análise de Sobrevivência não abordam a análise de dados discretos. Sendo assim, a seguir apresentaremos alguns conceitos fundamentais para o desenvolvimento deste trabalho. Caso o leitor esteja familiarizado com o assunto, é possível avançar diretamente para a Seção 2.2. Como referências citamos, por exemplo, Jenkins (2005), Lawless (2011) e Tutz e Schmid (2016).

2.1 Análise de sobrevivência para dados discretos

Suponha que T assuma valores discretos, $T = 1,2,3,\ldots,q$, e sua função de probabilidade seja $p(t) = P(T = t)$. A *função de risco*, dado o vetor das variáveis preditoras $\mathbf{x}$, é definida por

$$\lambda(t|\mathbf{x}) = P(T = t|T \geq t,\mathbf{x}), \quad t = 1,2,\ldots,q.$$

Esta função representa a probabilidade condicional de falhar no tempo t, ou seja, observar o evento de interesse, dado que o indivíduo está em risco em t.

A *função de sobrevivência* é formulada por

$$S(t|\mathbf{x}) = P(T > t|\mathbf{x}) = \sum_{t_s > t} p(t_s). \tag{2.1}$$

Assim, para o indivíduo sobreviver ao tempo t, ele tem que sobreviver ao tempo $t-1$ e ao tempo $t-2,\ldots,$ e ao tempo $t=1$, ou seja, a todos os tempos anteriores a t.
Note que $P(T \geq t) = S(t-1)$ e $p(t) = S(t-1) - S(t)$, assim

$$\lambda(t|\mathbf{x}) = P(T=t|T \geq t,\mathbf{x}) = \frac{P(T=t,T \geq t|\mathbf{x})}{P(T \geq t|\mathbf{x})} = \frac{P(T=t|\mathbf{x})}{P(T \geq t_j|\mathbf{x})} = \frac{P(T=t|\mathbf{x})}{S(t-1)|\mathbf{x})} = 1 - \frac{S(t|\mathbf{x})}{S(t-1|\mathbf{x})}. \tag{2.2}$$

Note que quando $t=1$, temos que para $j<t$, $S(j|\boldsymbol{x}) = 1$.

Usando a última igualdade da expressão (2.2) e indução temos que

$$S(t|\mathbf{x}) = \prod_{t_s \leq t} \{1 - \lambda(t_s|\mathbf{x})\}, \tag{2.3}$$

e $P(T=t|\mathbf{x})$ pode ser escrita como

$$P(T=t|\mathbf{x}) = \lambda(t|\mathbf{x})S(t-1|\mathbf{x}) = \frac{\lambda(t|\mathbf{x})}{1-\lambda(t|\mathbf{x})}S(t|\mathbf{x}). \tag{2.4}$$

Logo, para dados discretos, a função de risco é uma probabilidade ($0 \leq \lambda(t|\mathbf{x}) \leq 1$).

A *função de verossimilhança*, considerando censura à direita e não informativa, é expressa por

$$L(\theta) \propto \prod_{i=1}^{n} P(T_i = t_i|\mathbf{x})^{c_i} P(T_i > t_i|\mathbf{x})^{1-c_i}, \tag{2.5}$$

em que $c_i = 1$ indica observação do evento de interesse e $c_i = 0$ indica observação da censura. Substituindo (2.3) e (2.4) em (2.5) temos

$$L(\theta) \propto \prod_{i=1}^{n} \left\{ \left[\frac{\lambda(t_i|\mathbf{x})}{1-\lambda(t_i|\mathbf{x})} \right]^{c_i} \prod_{s=1}^{t_i} [1 - \lambda(t_s|\mathbf{x})] \right\}, \tag{2.6}$$

em que t_i é o tempo observado do indivíduo i.

2.2 Riscos Competitivos

Riscos competitivos no contexto de Análise de Sobrevivência são caracterizados pela possível ocorrência de mais de um evento e a observação de um dos eventos impede a dos outros. Carvalho *et al.* (2011) abordam três métodos para estudos que envolvem riscos competitivos: sobrevivência até a ocorrência do primeiro evento, não importando qual deles tenha ocorrido, usado em situações que se desejam verificar vários efeitos de um mesmo fator de risco. O segundo é o risco da causa específica, usado para estimar o efeito de uma variável sobre um evento específico. O último é a modelagem da função de subdistribuição do risco.

Seguindo as definições de Tutz e Schmid (2016), sejam $R \in \{1,2,\ldots,m\}$ os eventos de interesse que competem entre si. Para o tempo discreto $T \in \{1,2,\ldots,q+1\}$, a *função de risco da causa específica* é definida por

$$\lambda_r(t|\mathbf{x}) = P(T=t,R=r|T \geq t,\mathbf{x}), \quad r=1,2,\ldots,m, \tag{2.7}$$

em que $\mathbf{x}$ é o vetor de covariáveis.

A *função de risco geral* é dada por

$$\lambda(t|\mathbf{x}) = \sum_{r=1}^{m} \lambda_r(t|\mathbf{x}) = P(T = t|T \geq t, \mathbf{x}). \tag{2.8}$$

Para cada tempo t o indivíduo está em risco de sofrer um dos m eventos com função de risco $\lambda_1(t|\mathbf{x}), \lambda_2(t|\mathbf{x}), ..., \lambda_m(t|\mathbf{x})$, e também em risco de não sofrer nenhum deles, $1 - \lambda(t|\mathbf{x})$.

A probabilidade de ocorrer o evento r no tempo t é

$$P(T = t, R = r|\mathbf{x}) = \lambda_r(t|\mathbf{x})S(t-1|\mathbf{x}), \tag{2.9}$$

em que $S(\cdot)$ é a função dada por (2.3).

A contribuição da verossimilhança para a observação do i-ésimo indivíduo, considerando censura não informativa é dada por

$$\begin{aligned} L_i(\theta) &\propto P(T_i = t_i, R_i = r_i|\mathbf{x}_i)^{c_i} P(T_i > t_i)^{1-c_i} \\ &\propto \lambda_{r_i}(t_i|\mathbf{x}i)^{c_i}(1 - \lambda(t_i|\mathbf{x}_i))^{1-c_i} \prod_{t=1}^{t_i-1}(1 - \lambda(t|\mathbf{x}_i)). \end{aligned} \tag{2.10}$$

Assim, para cada indivíduo i defina, para $t < t_i$,

$$\mathbf{y}_{it}^T = (y_{it0}, y_{it1}, \ldots, y_{itm}) = (1, 0, \ldots, 0),$$

em que y_{it0} = 1 indica a ocorrência de nenhum dos eventos em todos os tempos antes de t_i.

Para $t = t_i$ e $c_i = 1$, defina

$$\mathbf{y}_{it_i}^T = (y_{it_i0}, y_{it_i1}, \ldots, y_{it_im}) = (0, \ldots, 1, \ldots, 0),$$

com $y_{it_ir_i} = 1$ e para as demais 0, ou seja, para o indivíduo i observou-se o evento r, denotado por r_i, em $t = t_i$ e para $c_i = 0$ considere

$$\mathbf{y}_{it_i}^T = (y_{it_i0}, y_{it_i1}, \ldots, y_{it_im}) = (1, 0, \ldots, 0).$$

Desta forma, dados as variáveis indicadoras, (2.10) pode ser escrita como

$$L_i(\theta) \propto \prod_{t=1}^{t_i}\{\prod_{r=1}^{m} \lambda_r(t|\mathbf{x}_i)^{y_{itr}}\}\{(1 - \lambda(t|\mathbf{x}_i))\}^{y_{it0}} \tag{2.11}$$

$$\propto \prod_{t=1}^{t_i}\{\prod_{r=1}^{m} \lambda_r(t|\mathbf{x}_i)^{y_{itr}}\}\{(1 - \sum_{r=1}^{m} \lambda_r(t|\mathbf{x}_i))\}^{y_{it0}}. \tag{2.12}$$

Portanto, a função log-verossimilhança considerando todas as observações é expressa por

$$l(\theta) \propto \sum_{i=1}^{n}\sum_{t=1}^{t_i}\left(\sum_{r=1}^{m} y_{itr}\log(\lambda_r(t|\mathbf{x}_i)) + y_{it0}\log(1 - \sum_{r=1}^{m}\lambda_r(t|\mathbf{x}_i))\right). \tag{2.13}$$

Em análises da evasão no Ensino Superior é interessante saber o perfil do indivíduo que abandona o curso (um evento específico) para que medidas de prevenção possam ser tomadas. Este perfil pode ser caracterizado pela identificação dos fatores de risco e período de maior risco. Desta forma, nosso foco é modelar o risco da causa específica.

2.3 Modelo de Vallejos e Steel (2017)

O modelo de riscos competitivos de Vallejos e Steel (2017) se baseia no modelo de chances proporcionais de Cox (COX, 1972) para dados discretos e é dado por

$$\log\left\{\frac{\lambda_r(t|\delta,\mathbf{B};x_i)}{\lambda_0(t|\delta,\mathbf{B};x_i)}\right\} = \delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}, \quad r = 1,\ldots,m, \quad i = 1,\ldots,n, \tag{2.14}$$

em que $r = 1,2,\ldots,m$ são os m eventos que competem, δ_{rt} é definido como o logaritmo do risco basal de observar o evento r com respeito ao risco de observar nenhum evento no tempo t, $\mathbf{B} = \{\boldsymbol{\beta}_{(1)},\boldsymbol{\beta}_{(2)},\ldots,\boldsymbol{\beta}_{(m)}\}$, com $\boldsymbol{\beta}_{(r)}$ sendo o vetor dos coeficientes de regressão relacionado ao evento r, $\delta = \{\delta_{11},\ldots,\delta_{m1},\delta_{12},\ldots,\delta_{m2},\ldots\}$ e

$$\lambda_0(t|\delta,\mathbf{B};x_i) = 1 - \sum_{r=1}^{m}\lambda_r(t|\delta,\mathbf{B};x_i) \tag{2.15}$$

é o risco de nenhum evento ser observado no tempo t. Desta forma, a expressão (2.7) para este modelo é dada por

$$\lambda_r(t|\delta,\mathbf{B};x_i) = \frac{\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)})}{1+\sum_{s=1}^{m}\exp(\delta_{st} + x_i'\boldsymbol{\beta}_{(s)})} \tag{2.16}$$

A função log-verossimilhança considerando todas as observações fica caracterizada por

$$l(\theta) \propto \sum_{i=1}^{n}\sum_{t=1}^{t_i}\left(\sum_{r=1}^{m} y_{itr}\log(\lambda_r(t|\mathbf{x}_i)) + y_{it0}\log(1-\sum_{r=1}^{m}\lambda_r(t|\mathbf{x}_i))\right) \tag{2.17}$$

$$\propto \sum_{i=1}^{n}\sum_{t=1}^{t_i}\Big(\sum_{r=1}^{m} y_{itr}\log\Big(\frac{\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)})}{1+\sum_{s=1}^{m}\exp(\delta_{st} + x_i'\boldsymbol{\beta}_{(s)})}\Big) + y_{it0}\log\Big(1-\sum_{r=1}^{m}\frac{\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)})}{1+\sum_{s=1}^{m}\exp(\delta_{st} + x_i'\boldsymbol{\beta}_{(s)})}\Big)\Big), \tag{2.18}$$

em que n é o número de indivíduos, t_i e r é o tempo e evento observado do indivíduo i, respectivamente, com $i = 1,\ldots,n$ e $r = 1,\ldots,m$.

Na próxima seção apresentaremos a nossa proposta, o método de estimação e seleção de variáveis abordado no estudo de simulação e aplicação nos dados reais e interpretação do modelo de Vallejos e Steel (2017) e da proposta.

CAPÍTULO

3

MODELO

3.1 O modelo

O modelo que apresentamos baseia-se na versão discreta do modelo de riscos proporcionais para tempos contínuos chamado log-log complementar, o qual pode ser aplicado para tempos intrinsicamente discretos (JENKINS, 2005). Desta forma, no cenário de riscos proporcionais temos

$$S(t|\mathbf{x}) = S_b(t|\mathbf{x})^{\exp(\mathbf{x}'\boldsymbol{\beta})}, \tag{3.1}$$

em que $S_b(t|\mathbf{x})$ é a função de sobrevivência basal, e considerando (2.2) e indução temos que (3.1) pode ser expressa por

$$1-\lambda(t|\mathbf{x}) = (1-\lambda_b(t))^{\exp(\mathbf{x}'\boldsymbol{\beta})},$$

e então

$$\lambda(t|\mathbf{x}) = 1-(1-\lambda_b(t))^{\exp(\mathbf{x}'\boldsymbol{\beta})}. \tag{3.2}$$

Aplicando a transformação log-log complementar em (3.2) temos

$$\log(-\log(1-\lambda(t|\mathbf{x}))) = \alpha_t + \mathbf{x}'\boldsymbol{\beta}, \tag{3.3}$$

em que $\alpha_t = \log(-\log(1-\lambda_b(t)))$ é o log-log complementar do risco basal.

Portanto, estendemos o modelo (3.3) para acomodar riscos competitivos e assim, o modelo que abordamos nesse estudo é dado por

$$\log\left\{-\log\left\{1-\frac{\lambda_r(t|\delta,\mathbf{B};x_i)}{\lambda_r(t|\delta,\mathbf{B};x_i)+\lambda_0(t|\delta,\mathbf{B};x_i)}\right\}\right\} = \delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}, \tag{3.4}$$

em que $r = 1,2,...,m$ são os m eventos que competem, δ_{rt} é definido como o log-log complementar do risco basal de observar o evento r com respeito ao complementar do risco de observar todos os eventos menos o evento r no tempo t, $\mathbf{B} = \{\boldsymbol{\beta}_{(1)}, \boldsymbol{\beta}_{(2)}, \ldots, \boldsymbol{\beta}_{(m)}\}$ com $\boldsymbol{\beta}_{(r)}$ sendo o

vetor dos coeficientes de regressão relacionado ao evento r e $\delta = \{\delta_{11}, \ldots, \delta_{m1}, \delta_{12}, \ldots, \delta_{m2}, \ldots\}$. Temos que

$$\lambda_0(t|\delta, \mathbf{B}; x_i) = 1 - \sum_{r=1}^{m} \lambda_r(t|\delta, \mathbf{B}; x_i) \tag{3.5}$$

é o risco de observar nenhum evento no tempo t.

Assim, de (3.4) temos

$$\begin{aligned}
1 - \frac{\lambda_r(t|\delta,\mathbf{B};x_i)}{\lambda_r(t|\delta,\mathbf{B};x_i)+\lambda_0(t|\delta,\mathbf{B};x_i)} &= \exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)})) \\
\Rightarrow \frac{\lambda_r(t|\delta,\mathbf{B};x_i)}{\lambda_r(t|\delta,\mathbf{B};x_i)+\lambda_0(t|\delta,\mathbf{B};x_i)} &= 1 - \exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)})) \\
\Rightarrow \lambda_r(t|\delta,\mathbf{B};x_i) &= \lambda_0(t|\delta,\mathbf{B};x_i)\frac{1-\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}{\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))} \qquad (3.6) \\
\Rightarrow \sum_{r=1}^{m} \lambda_r(t|\delta,\mathbf{B};x_i) &= \lambda_0(t|\delta,\mathbf{B};x_i)\left(\sum_{r=1}^{m}\frac{1-\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}{\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}\right) \\
\Rightarrow \lambda_0(t|\delta,\mathbf{B};x_i) + \sum_{r=1}^{m} \lambda_r(t|\delta,\mathbf{B};x_i) &= \lambda_0(t|\delta,\mathbf{B};x_i) + \lambda_0(t|\delta,\mathbf{B};x_i)\left(\sum_{r=1}^{m}\frac{1-\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}{\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}\right) \\
\Rightarrow 1 &= \lambda_0(t|\delta,\mathbf{B};x_i)\left(1 + \sum_{r=1}^{m}\frac{1-\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}{\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}\right) \\
\Rightarrow \lambda_0(t|\delta,\mathbf{B};x_i) &= \frac{1}{1+\sum_{r=1}^{m}\frac{1-\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}{\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}}. \qquad (3.7)
\end{aligned}$$

Desta forma, usando (3.6) temos

$$\lambda_r(t|\delta,\mathbf{B};x_i) = \frac{1}{1+\sum_{r=1}^{m}\frac{1-\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}{\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}} \times \frac{1-\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}{\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}. \tag{3.8}$$

As funções $\lambda_r(t|\delta,\mathbf{B};x_i)$, $r = 1, \ldots, m$, representam os *riscos da causa específica*, ou seja, a probabilidade de observar o evento r em t dado que nenhum evento foi observado antes de t.

Substituindo (3.7) e (3.8) na função log-verossimilhança em (2.13) temos que

$$\begin{aligned}
l(\theta) = \sum_{i=1}^{n}\sum_{t=1}^{t_i}\Bigg\{ \sum_{r=1}^{m} y_{itr} \log &\left(\frac{1}{1+\sum_{r=1}^{m}\frac{1-\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}{\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}} \times \frac{1-\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}{\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}\right) \\
&+ y_{it0} \log\left(\frac{1}{1+\sum_{r=1}^{m}\frac{1-\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}{\exp(-\exp(\delta_{rt} + x_i'\boldsymbol{\beta}_{(r)}))}}\right)\Bigg\}, \qquad (3.9)
\end{aligned}$$

em que y_{itr} e y_{it0} são definidos na Seção 2.2.

O modelo (3.4) pode envolver um grande número de parâmetros, pois existem $m \times T$ diferentes δ_{rt}. Por exemplo, se estivermos estudando um período de 16 semestres e 3 eventos

competitivos, teríamos $3 \times 16 = 48$ parâmetros para estimar além dos parâmetros dos coeficientes de regressão $\boldsymbol{\beta}_{(r)}$ para $r = 1,2,3$. Além disso, Vallejos e Steel (2017) apontam que fazer inferência por máxima verossimilhança no modelo que propuseram pode ser problemático, pois em um cenário de estudos da evasão com riscos competitivos é difícil observar indivíduos que conseguem diploma no primeiro ou segundo semestre, de modo que a estimativa via máxima verossimilhança para os parâmetros δ_{rt} nestes primeiros períodos seria $-\infty$. Esta situação é caracterizada por problemas de separação que ocorrem quando a resposta pode ser separada quase ou completamente por uma covariável (SOUZA, 2010). De forma similar, é possível observar essa situação no modelo proposto, ou seja, as estimativas δ_{rt} nestes primeiros períodos seriam $-\infty$ e então consideramos o método bayesiano para estimar os parâmetros envolvidos, o qual será explicado na Seção 3.3.

3.2 Interpretação dos modelos

Para o modelo de Vallejos e Steel (2017) em (2.13) temos que

$$\frac{\frac{\lambda_r(t|x_1)}{\lambda_0(t|x_1)}}{\frac{\lambda_r(t|x_2)}{\lambda_0(t|x_2)}} = \exp(\boldsymbol{\beta}_{(r)}(x_1 - x_2)). \tag{3.10}$$

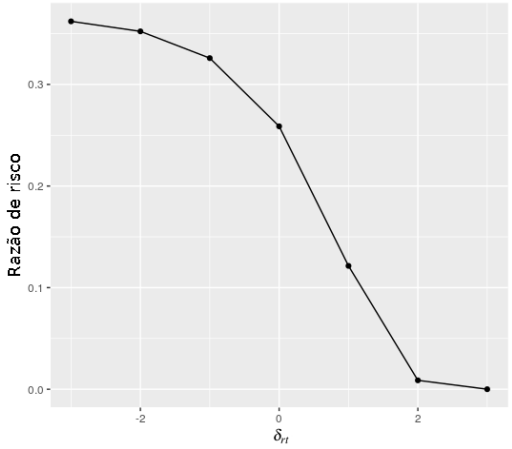
Logo, a razão dos riscos do evento r sob o risco de nenhum evento não depende do tempo t. Para o modelos por nós proposto temos o seguinte resultado:

$$\frac{\frac{\lambda_r(t|x_1)}{\lambda_0(t|x_1)}}{\frac{\lambda_r(t|x_2)}{\lambda_0(t|x_2)}} = \frac{\exp(-\exp(\delta_{rt} + x_2\boldsymbol{\beta}_{(r)}))}{\exp(-\exp(\delta_{rt} + x_1\boldsymbol{\beta}_{(r)}))} \times \frac{1-\exp(-\exp(\delta_{rt} + x_1\boldsymbol{\beta}_{(r)}))}{1-\exp(-\exp(\delta_{rt} + x_2\boldsymbol{\beta}_{(r)}))} \tag{3.11}$$

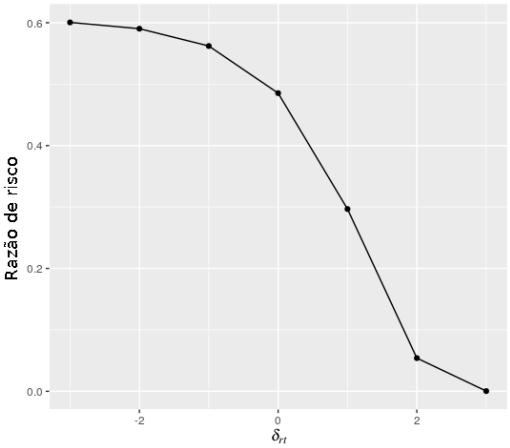
Portanto, a razão depende do valor de δ_{rt}.

A Figura 1 mostra a razão dos riscos da expressão (3.11) considerando valor de x_1 igual a 1, x_2 igual a zero e variando os valores de δ_{rt} (entre -3 e 3) e $\beta_{(r)}$. O resultado mostra que a razão não é proporcional ao longo do tempo.

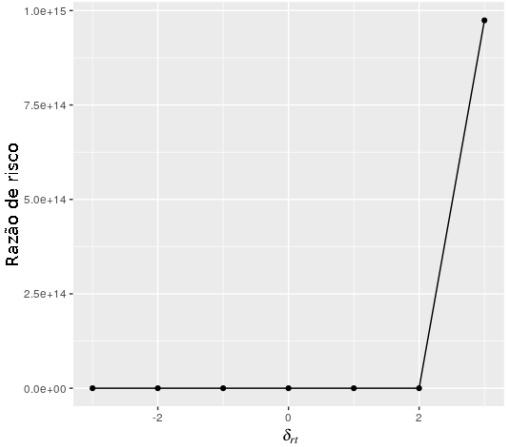
Figura 1 – Razão dos riscos da expressão (3.11) considerando x_1 igual a 1, x_2 igual a zero e variando os valores de δ_{rt} e $\beta_{(r)}$.



(a) $\beta_{(r)}$ igual a -1.



(b) $\beta_{(r)}$ igual a -0.5.



(c) $\beta_{(r)}$ igual a 1.

3.3 Estimação dos parâmetros

Ao lidar com modelos de regressão binária existe uma situação que pode dificultar o processo de estimação dos parâmetros: a separabilidade. A definição de separabilidade dada por Albert e Anderson (1984) estabelece a ideia de que é possível identificar a resposta binária através das covariáveis. Neste caso, a maximização via verossimilhança pode ser afetada, resultando em estimativas não únicas ou finitas (LI; CLYDE, 2018; GHOSH *et al.*, 2018). Para contornar o problema da separabilidade, Ghosh *et al.* (2018) mostraram uma condição necessária e suficiente para existir a média da distribuição a *posteriori* dos coeficientes de regressão quando estes assumem Cauchy independentes como distribuições a *priori*. Este resultado é compreensível quando não assumimos uma função de ligação específica, porém a presença da separabilidade não satisfaz a condição necessária.

No nosso cenário, graduar nos primeiros períodos de matrícula é um evento raro, implicando um risco da causa específica igual a zero. O risco em (3.7) é próximo de zero quando δ_{rt} assume valores grandes negativamente, similarmente o que ocorre com a função de ligação logística em Vallejos e Steel (2017). Assim, a correção do estimador de máxima verossimilhança proposto por King e Zeng (2001) não é apropriado para nosso conjunto de dados. Ainda, eventos raros são a causa da separabilidade, logo não preenche os requisitos necessários para o teorema de Ghosh *et al.* (2018). Uma solução proposta por Vallejos e Steel (2017) é assumir como a *priori* para cada evento a distribuição Cauchy multivariada. Tal proposta é baseada nos resultados do trabalho de Polson, Scott e Windle (2013), o qual não é válido para nossa proposta.

Assim, para a estimação dos parâmetros do modelo em (3.3), sugerimos o uso do JAGS (Just Another Gibbs Sampler), um sistema de código aberto, o qual fornece a distribuição a *posteriori* pelo método de Monte Carlo via Cadeia de Markov sem ser necessário declarar um amostrador (PLUMMER *et al.*, 2003) e sugerimos distribuições a *priori* independentes para os parâmetros $\boldsymbol{\delta}$ e $\boldsymbol{\beta}_{(r)}$, ou seja

$$p(\boldsymbol{\delta}, \boldsymbol{B}) = \prod_{r=1}^{m} \left(\prod_{t=1}^{T} p(\delta_{rt}) \prod_{s=1}^{k} p(\beta_{rs}) \right). \tag{3.12}$$

E para completar nossa proposta, escolhemos para todos $p(\cdot)$ em (3.12) a distribuição normal com média 0 e variância 100, refletindo um fraco conhecimento prévio sobre os parâmetros.

Utilizamos o pacote *R2jags* (SU; YAJIMA, 2015) dentro do ambiente *R* (TEAM *et al.*, 2013) e para cada iteração foram amostrados os parâmetros δ_{rt} e $\beta_{(r)}$ e então calculado o valor dos riscos estimados.

3.4 Critério para seleção do modelo

Tendo em vista o objetivo de identificar o perfil do indivíduo que abandona um curso do Ensino Superior, é importante descobrir características relacionadas ao evento. Desta forma, para

seleção dos modelos consideramos o *Deviance Information Criterion* (DIC) (SPIEGELHALTER *et al.*, 2002) no qual tem-se:

$$DIC = 2\bar{D} - D(\hat{\boldsymbol{\theta}}),$$

em que $\bar{D} = E(D(\boldsymbol{x}, \boldsymbol{\theta})|\boldsymbol{x})$ e $D(\hat{\boldsymbol{\theta}}) = D(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$, sendo $D(\boldsymbol{x}, \boldsymbol{\theta}) = -2\log(f(\boldsymbol{x}|\boldsymbol{\theta}))$, $f(x|\boldsymbol{\theta})$ função de verossimilhança de um conjunto de observações $\boldsymbol{x}$ dado parâmetros desconhecidos $\boldsymbol{\theta}$.

Como a distribuição a posteriori é obtida pelo método de Monte Carlo via cadeia de Markov, utilizamos o valor do DIC através das amostras geradas pelo MCMC. Com isso, consideramos o melhor modelo aquele com menor valor de DIC.

CAPÍTULO

4

ESTUDO DE SIMULAÇÃO

Neste capítulo realizamos os estudos de simulação, nos quais investigamos as consequências de considerar um modelo ou o outro na geração dos dados e estimação dos parâmetros. Desta forma, os estudos compreendem na geração do conjunto de dados através do modelo proposto neste trabalho e os parâmetros estimados através do modelo proposto por Vallejos e Steel (2017) e *vice-versa*.

Em cada simulação, consideramos quatro períodos e geramos uma covariável binária usando distribuição binomial com parâmetro igual a 0,5 e a partir dela e dos parâmetros $\beta_{(r)}$ e δ_{rt} fixos (Tabela 1) para número de evento igual a 2, sou seja, $r = 1, 2$, que correspondem aos verdadeiros parâmetros, calculamos o risco da causa específica, o qual chamaremos de risco gerado, de cada evento considerado. Com esses riscos gerados obtemos o vetor de respostas e, a partir disso, ignoramos os valores conhecidos dos parâmetros e partimos para o processo de estimação conforme as etapas da Seção 3.3. Para este estudo fizemos 500 simulações considerando tamanho de amostra igual a 400.

Tabela 1 – Valores dos parâmetros verdadeiros para cada evento ($r = 1, 2$).

	$r = 1$	$r = 2$
$\beta_{(r)}$	1	0,5
δ_{r1}	-0,8	0,2
δ_{r2}	0,3	-0,3
δ_{r3}	-1	0,1
δ_{r4}	3	-5

Para a inferência bayesiana geramos uma cadeia de 100000 iterações com salto de 20 e descartamos as primeiras 1000, construindo uma amostra de tamanho 4000. Todos os cálculos computacionais foram realizados através do *R* (R Core Team, 2019), utilizando o pacote *R2jags* (SU; YAJIMA, 2015) e adaptações dos códigos do trabalho de Vallejos e Steel (2017) disponíveis em https://github.com/catavallejos/Univer sitySurvival.

4.1 Resultados - primeiro estudo

As Tabelas 2 e 3 mostram os valores dos riscos gerados e estimados de cada evento (1 e 2) em cada período, considerando para a geração dos dados e estimação dos parâmetros o modelo log-log complementar proposto. As Tabelas 4 e 5 mostram os valores dos riscos gerados e estimados de cada evento (1 e 2) em cada período, considerando para a geração dos dados o modelo log-log complementar proposto e o modelo de Vallejos e Steel (2017) para estimação dos parâmetros.

Tabela 2 – Riscos gerados e estimados para o evento 1 assumindo o modelo log-log complementar proposto para estimação dos parâmetros, quando o verdadeiro modelo é o log-log complementar.

Período	Riscos gerados	Riscos estimados
1	0,1927	0,1916
2	0,3692	0,3670
3	0,0511	0,0426
4	0,9999	0,9826

Tabela 3 – Riscos gerados e estimados para o evento 2 assumindo o modelo log-log complementar proposto para estimação dos parâmetros, quando o verdadeiro modelo é o log-log complementar.

Período	Riscos gerados	Riscos Estimados
1	0,6304	0,6278
2	0,4218	0,4183
3	0,7741	0,7746
4	0,0000	0,0131

Tabela 4 – Riscos gerados e estimados para o evento 1 assumindo modelo de Vallejos e Steel (2017) para estimação dos parâmetros, quando o verdadeiro modelo é o log-log complementar.

Período	Riscos gerados	Riscos estimados
1	0,1927	0,1928
2	0,3692	0,3529
3	0,0511	0,0507
4	0,9999	0,9868

Tabela 5 – Riscos gerados e estimados para o evento 2 assumindo modelo de Vallejos e Steel (2017) para estimação dos parâmetros, quando o verdadeiro modelo é o log-log complementar.

Período	Riscos gerados	Riscos estimados
1	0,6304	0,6296
2	0,4218	0,4419
3	0,7741	0,7769
4	0,0000	0,0101

A Tabela 6 mostra os valores dos parâmetros de regressão estimados ($\hat{\beta}_{(r)}$). Considere a notação **nn** a situação em que os dados são gerados e os parâmetros estimados pelo modelo proposto no trabalho (log-log complementar) e **nc** a situação em que os dados são gerados pelo modelo proposto mas os parâmetros estimados pelo modelo da referência principal.

Tabela 6 – Valores dos coeficientes de regressão estimados, $\hat{\beta}_{(r)}$, através do modelo proposto por nós (**nn**) e modelo de Vallejos e Steel (2017) (**nc**) para os eventos 1 e 2, quando o verdadeiro modelo é o log-log complementar.

		$\hat{\beta}_{(r)}$	
r	$\beta_{(r)}$	**nn**	**nc**
1	1,0000	1,0023	1,5078
2	0,5000	0,5102	1,0281

Para as 500 réplicas calculamos o valor do DIC e verificamos que em 57% das vezes o caso em que os parâmetros são estimados pelo modelo proposto teve o valor do DIC menor, indicando uma leve preferência no sentido de o modelo mais adequado ser o log-log complementar. Além disso, pela Tabela 6 notamos uma melhor aproximação aos verdadeiros parâmetros a situação **nn**.

4.2 Resultados - segundo estudo

No segundo estudo, assumimos o modelo de Vallejos e Steel (2017) para a geração dos dados. As Tabelas 7, 8, 9 e 10 mostram os valores dos riscos gerados e estimados de cada evento (1 e 2) em cada período, considerando log-log complementar proposto e o modelo de Vallejos e Steel (2017) para estimação dos parâmetros.

Tabela 7 – Riscos gerados e estimados para o evento 1 assumindo modelo de Vallejos e Steel (2017) para estimação dos parâmetros, quando o mesmo é o verdadeiro modelo.

Perído	**Riscos gerados**	**Riscos estimados**
1	0,2284	0,2281
2	0,3200	0,3207
3	0,0940	0,0883
4	0,9299	0,9373

Tabela 8 – Riscos gerados e estimados para o evento 2 assumindo modelo de Vallejos e Steel (2017) para estimação dos parâmetros, quando o mesmo é o verdadeiro modelo.

Período	**Riscos gerados**	**Riscos Estimados**
1	0,4664	0,4671
2	0,3602	0,3595
3	0,5712	0,5789
4	0,0006	0,0006

Tabela 9 – Riscos gerados e estimados para o evento 1 assumindo o modelo log-log complementar proposto para estimação dos parâmetros, quando modelo de Vallejos e Steel (2017) é o modelo verdadeiro.

Período	**Riscos gerados**	**Riscos estimados**
1	0,2284	0,2274
2	0,3200	0,3183
3	0,0940	0,0830
4	0,9299	0,9325

Tabela 10 – Riscos gerados e estimados para o evento 2 assumindo o modelo log-log complementar proposto para estimação dos parâmetros, quando modelo de Vallejos e Steel (2017) é o modelo verdadeiro.

Período	**Riscos gerados**	**Riscos estimados**
1	0,4664	0,4659
2	0,3602	0,3565
3	0,5712	0,5746
4	0,0006	0,0003

A Tabela 11 mostra os valores dos parâmetros de regressão estimados ($\hat{\beta}_{(r)}$). Considere a notação **cc** a situação em que os dados são gerados e os parâmetros estimados pelo modelo de Vallejos e Steel (2017) e **cn** a situação em que os dados são gerados pelo modelo da referência principal e estimados pelo proposto no trabalho (log-log complementar).

Tabela 11 – Valores dos coeficientes de regressão estimados ($\hat{\beta}_{(r)}$) através do modelo de Vallejos e Steel (2017) (**cc**) e modelo por nós proposto (**cn**) para os eventos 1 e 2.

		$\hat{\beta}_{(r)}$	
r	$\beta_{(r)}$	**cc**	**cn**
1	1,0000	1,0286	0,7558
2	0,5000	0,5195	0,3273

Para as 500 réplicas calculamos o valor do DIC e verificamos que em 67% das vezes o caso em que os parâmetros são estimados pelo nosso modelo teve o valor do DIC menor, indicando que, ao considerar o modelo de Vallejos e Steel (2017) para a geração do conjunto de dados, ainda há uma preferência no sentido de modelo mais adequado ser o proposto neste trabalho, embora os valores das estimativas dos coeficientes de regressão da situação *cc* se ajustarem melhor aos verdadeiros parâmetros.

4.3 Conclusões

Nesse Capítulo investigamos através de um estudo de simulação a qualidade das estimativas da inferência bayesiana descrita na Seção 3.3. As estimativas se aproximam dos verdadeiros valores dos parâmetros, tanto para nosso caso como para o modelo de Vallejos e Steel (2017). Estes resultados, por sua vez, mostram-nos que é viável o uso da inferência por nós proposta por retornar valores próximos aos reais.

Através do valor do DIC, avaliamos para as 500 réplicas a quatidade de vezes que o modelo correto foi selecionado e notamos que para o primeiro estudo o modelo verdadeiro é selecionado 57% da vezes e para o segundo estudo, o verdadeiro modelo é selecionado 33%. Ainda, notamos que os modelos se aproximam em algumas situações, levando a obtenção de estimativas aproximadas.

Considerando os riscos das expressões (2.16) e (3.8) e os valores da Tabela 1, temos que para alguns períodos os riscos se aproximam. Tal resultado explica porque os modelos se confundem, não sendo identificados quando deveriam. Vale ressaltar que, nesse caso, não há prejuízo, pois os riscos competitivos são bem estimados mesmo não sendo adotado o modelo correto.

CAPÍTULO

5

ANÁLISE DE DADOS

5.1 Descrição dos dados

O conjunto de dados reais é proveniente do curso de Bacharelado em Matemática Aplicada e Computação Científica (BMACC) do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC/USP), *campus* de São Carlos. Os dados contêm características acadêmicas como data e forma de ingresso (primeira chamada, segunda chamada,...), data de encerramento, motivo do encerramento (três semestres sem matrícula, jubilamento, transferência externa, transferência interna, graduado, outros), média com reprovação, média sem reprovação e características socioeconômicas como raça, escolaridade, número de pessoas sustentadas pela renda, instrução educacional dos pais. Essas informações correspondem a 400 alunos matriculados desde 2000 até o ano de 2016, porém como existem dados faltantes conseguimos fazer a aplicação com 343 alunos.

Devido à existência de dados faltantes, as covariáveis utilizadas estão descritas na Tabela 12.

Tabela 12 – Descrição das variáveis utilizadas.

Notação	Variável	Descrição
x_1	Ingresso	Forma de ingresso categorizada por 0 para os alunos que ingressaram na primeira chamada, e 1 para os alunos que ingressaram a partir da segunda chamada
x_2	Renda	Número de pessoas sustentadas pela renda categorizada por 0 se o número é até quatro pessoas, e 1 se o número é mais de quatro pessoas
x_3	Ensino Médio	Escolaridade categorizada em 0 para alunos que estudaram toda ou maior parte do ensino médio em escola pública, e 1 para alunos que estudaram toda ou maior parte do ensino médio em escola particular
x_4	Raça	Raça categorizada por 0 se os alunos se consideram brancos, e 1 se o alunos se consideram de outras etnias diferentes da branca
x_5	Instrução mãe	Grau de instrução da mãe categorizada por 0 para as mães que não estudaram ou estudaram até o ensino médio (completo ou incompleto), e 1 para as mães que estudaram ensino superior completo ou incompleto
x_6	Instrução pai	Grau de instrução do pai categorizada por 0 para os pais que não estudaram ou estudaram até o ensino médio (completo ou incompleto), e 1 para os pais que estudaram ensino superior completo ou incompleto

Ao final do ano de 2016, dentre os 343 alunos, 33 (9,6%) ainda estavam matriculados, 88 (25,7%) graduados e 222 (64,7%) evadidos. As Tabelas 13 e 14 mostram as características dos alunos graduados e evadidos respectivamente. Percebemos que tanto para os graduados quanto para os evadidos a porcentagem aparecem próximas para as variáveis Ensino Médio, Instrução da mãe e do pai, raça e renda, ficando por volta de 50% para as três primeiras variáveis e 65%, 75% versus 35%, 25% para renda e raça. Já para a variável ingresso, percebemos que 70% dos alunos graduados ingressaram na primeira chamada, enquanto 30% nas demais chamadas . As Figuras 2 e 3 mostram a relação entre o tempo de permanência medido em semestres e a porcentagem de alunos evadidos e formados, respectivamente. Verificamos que muitos estudantes permanecem pouco tempo no curso (5% um semestre, 10% dois semestres e 13% três semestres). Em relação aos graduados, 14% permaneceram oito semestres, ou seja, graduam no tempo ideal do curso (quatro anos).

Tabela 13 – Porcentagem dos alunos graduados dentro das categorias.

Categorias	Ensino médio	Ingresso	Instrução mãe	Instrução pai	Raça	Renda
0	56,81%	70,45%	56,82%	54,54%	76,13%	63,64%
1	43,19%	29,55%	43,18%	45,46%	23,87%	36,36%

Tabela 14 – Porcentagem dos alunos evadidos dentro das categorias.

Categorias	Ensino médio	Ingresso	Instrução mãe	Instrução pai	Raça	Renda
0	59,45%	45,04%	49,55%	53,15%	74,77%	67,11%
1	40,55%	54,96%	50,45%	46,85%	25,23%	32,89%

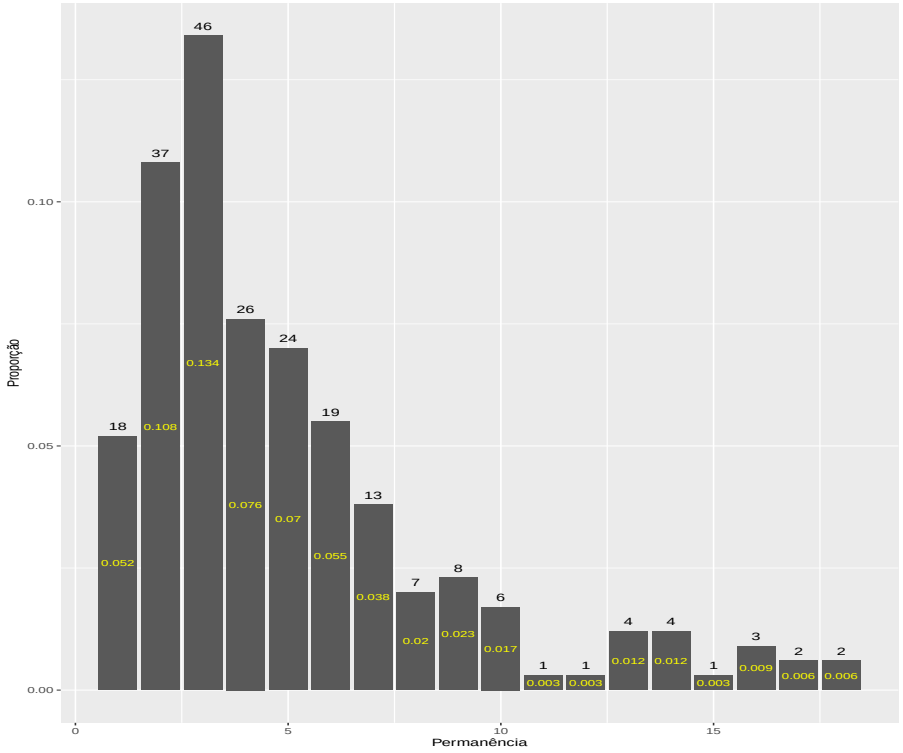


Figura 2 – Proporção de evadidos do total de alunos relacionado ao tempo de permanência.

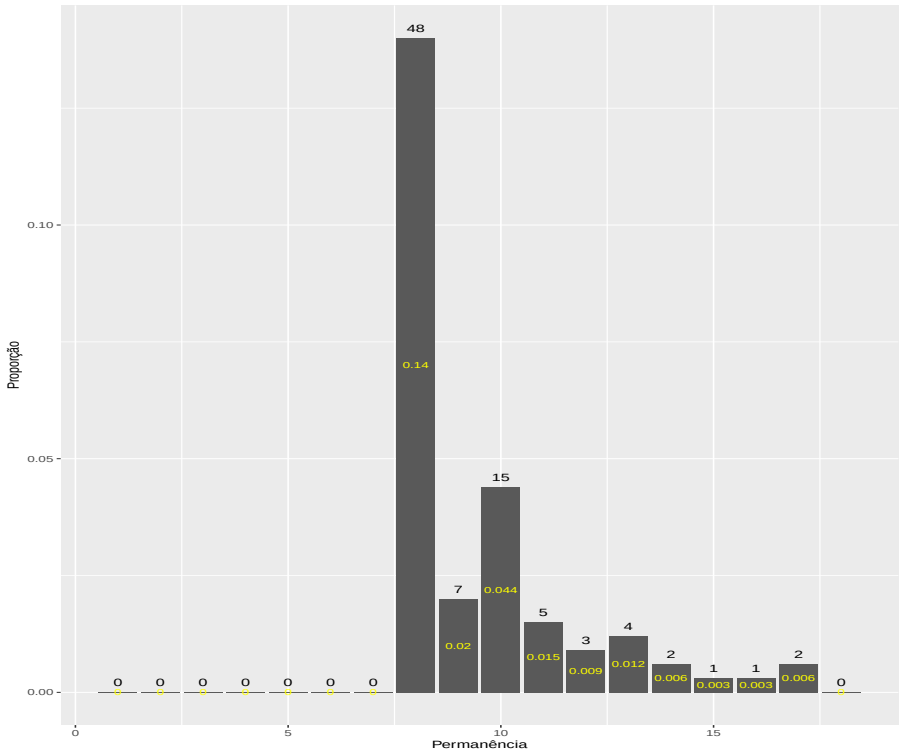


Figura 3 – Proporção de graduados do total de alunos relacionado ao tempo de permanência.

Os eventos competitivos considerados foram evadido (por motivo de três semestres sem matrícula, jubilamento, transferência externa, transferência interna) e graduado. Para os alunos que ainda estavam matriculados nesse período, o evento matriculado foi considerado censura. Para cada estudante calculamos o tempo de permanência em semestre dele no curso. A variável tempo, T, é medida em semestres e, portanto, discreta. O número de semestres considerados para o tempo foi 18. Para a estimação dos parâmetros, foram considerados 50000 iterações com passo de 10 e 10 e descartamos as primeiras 10000, totalizando uma amostra de tamanho 4000. Os gráficos das cadeias encontram-se no Apêndice B. Para a análise de convergência, consideramos o gráfico do traço.

5.2 Resultados

A Tabela 15 mostra o valor do DIC dos modelos utilizados na seleção. Neste método, selecionamos o modelo que compreende as variáveis ingresso (x_1) e grau de instrução da mãe (x_5), resultado esperado conforme descritivas das Tabelas 13 e 14.

Tabela 15 – Seleção do modelo através do cálculo do DIC.

Modelo	**Covariáveis**	**DIC**
1	Nenhuma	41561794
2	Todas	41561796
3	x_1	**41561782**
4	x_2	41561799
5	x_3	41561797
6	x_4	41561800
7	x_5	41561792
8	x_6	41561797
9	$x_1 + x_2$	41561786
10	$x_1 + x_3$	41561786
11	$x_1 + x_4$	41561786
12	$x_1 + x_5$	**41561782**
13	$x_1 + x_6$	41561784
14	$x_1 + x_5 + x_2$	41561786
15	$x_1 + x_5 + x_3$	41561785
16	$x_1 + x_5 + x_4$	41561783
17	$x_1 + x_5 + x_6$	41561783
18	$x_1 + x_5 + x_2 + x_3$	41561790
19	$x_1 + x_5 + x_2 + x_4$	41561787
20	$x_1 + x_5 + x_2 + x_6$	41561788
21	$x_1 + x_5 + x_3 + x_4$	41561789
22	$x_1 + x_5 + x_3 + x_6$	41561790
23	$x_1 + x_5 + x_4 + x_6$	41561791

A Tabela 16 mostra os valores dos betas estimados da variável ingresso e instrução mãe

para os eventos graduar e evadir e o intervalo de credibilidade de cada parâmetro.

Tabela 16 – Valores dos coeficientes de regressão estimados de cada variável do modelo selecionado para os eventos evadir e graduar.

Variável	Evadir	Intervalo de Credibilidade	Graduar	Intervalo de Credibilidade
Ingresso	0,482	(0,218; 0,763)	-0,465	(-0,948; -0,025)
Instrução mãe	0,320	(0,035; 0,587)	0,155	(-0,286; 0,574)

A Figura 4 mostra o grágico do log-log complementar dos risco basal estimado para os eventos evadir e graduar para o conjunto de dados reais do curso de Bacharelado em Matemática Aplicada e Computação Científica (BMACC) do Instituto de Ciências Matemáticas e de Computaçãoda Universidade de São Paulo (ICMC/USP).

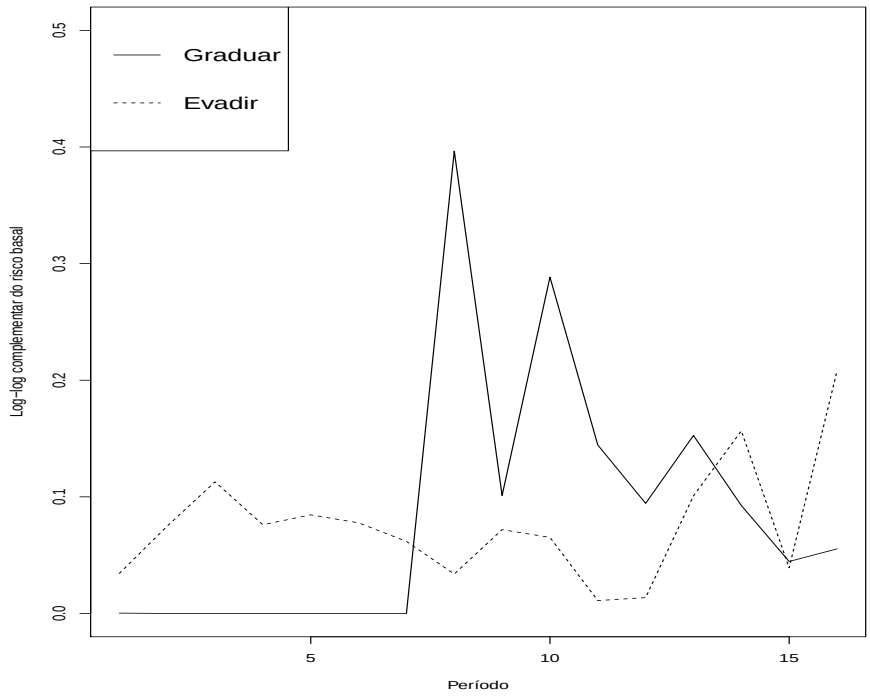


Figura 4 – Log-log complementar do risco basal estimado para os eventos evadir e graduar.

5.3 Conclusões

Com a aplicação ao conjunto de dados reais, nossa proposta considera as covariáveis Ingresso (x_1) e Instrução mãe (x_5) como as mais relevantes considerando o menor valor do DIC.

Através da expressão do risco da causa específica dada por (3.8), verificamos que:

1. O risco de graduar é menor para estudantes que ingressaram a partir da segunda chamada ($x_1 = 1$) e grau de instrução da mãe superior completo ou incompleto ($x_5 = 1$) se

comparado com estudantes que ingressaram de primeira chamada ($x_1 = 0$) e grau de instrução da mãe superior completo ou incompleto ($x_5 = 1$). Ou seja, considerando os alunos cuja mãe possui grau de instrução superior completo ou incompleto, o risco de graduar é maior para os que ingressaram de primeira chamada. A Figura 5 mostra os riscos estimados. Notamos que, nos dados da Tabela 13, 70% dos alunos graduados ingressaram na primeira chamada.

2. O risco de graduar é menor para estudantes que ingressaram a partir da segunda chamada ($x_1 = 1$) e grau de instrução da mãe ensino médio completo ou incompleto ($x_5 = 0$) se comparado com estudantes que ingressaram de primeira chamada ($x_1 = 0$) e grau de instrução da mãe é até ensino médio completo ou incompleto ($x_5 = 0$). Ou seja, considerando os alunos cuja mãe possui grau de instrução ensino médio completo ou incompleto, o risco de graduar é maior para os que ingressaram de primeira chamada. A Figura 5 mostra os riscos estimados.

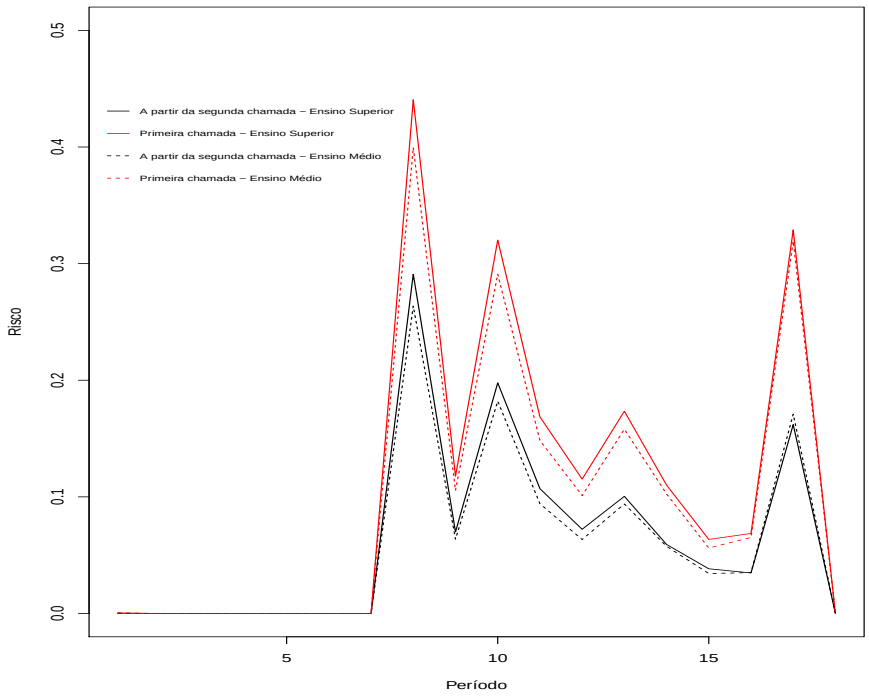


Figura 5 – Risco de graduar para os estudantes de primeira chamada e a partir da segunda chamada considerando grau de instrução da mãe Ensino Superior completo ou incompleto (linha contínua) e Ensino Médio completo ou incompleto (linha tracejada).

3. O risco de evadir é maior para para estudantes que ingressaram a partir da segunda chamada ($x_1 = 1$) e grau de instrução da mãe superior completo ou incompleto ($x_5 = 1$) se comparado com estudantes que ingressaram de primeira chamada ($x_1 = 0$) e grau de instrução da mãe superior completo ou incompleto ($x_5 = 1$). Ou seja, considerando os alunos cuja mãe possui grau de instrução superior completo ou incompleto, o risco de evadir é maior para os alunos que ingressaram a partir da segunda chamada. A Figura 6 mostra os riscos estimados.

4. O risco de evadir é maior para estudantes que ingressaram a partir da segunda chamada

($x_1 = 1$) e grau de instrução da mãe ensino médio completo ou incompleto ($x_5 = 0$) se comparado com estudantes que ingressaram de primeira chamada ($x_1 = 0$) e grau de instrução da mãe é ensino médio completo ou incompleto ($x_5 = 0$). Ou seja, considerando os alunos cuja mãe possui grau de instrução ensino médio completo ou incompleto, o risco de evadir é maior para os alunos que ingressaram a partir da segunda chamada. A Figura 6 mostra os riscos estimados.

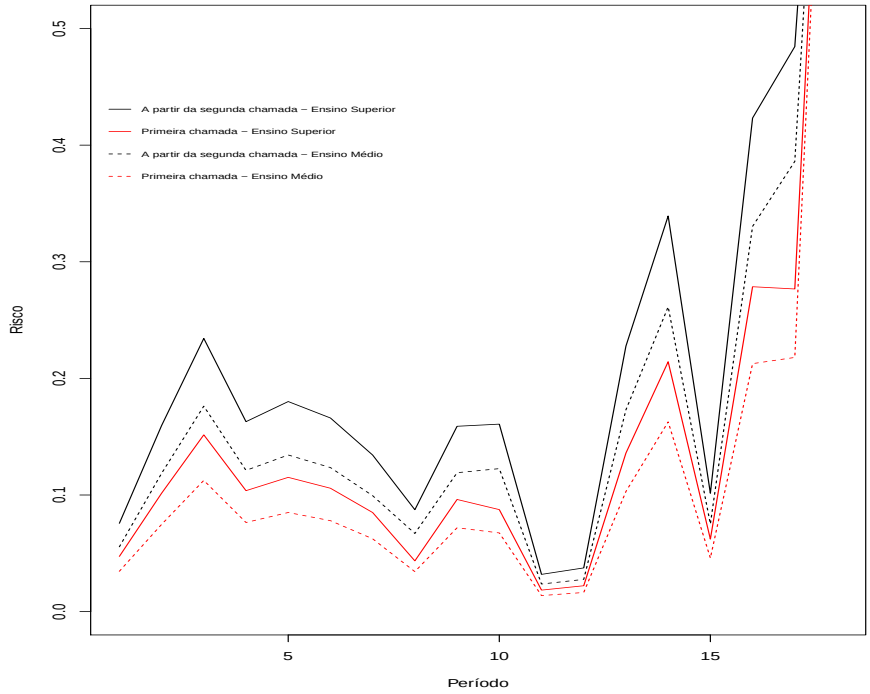


Figura 6 – Risco de evadir para os estudantes de primeira chamada e a partir da segunda chamada considerando grau de instrução da mãe Ensino Superior completo ou incompleto (linha contínua) e Ensino Médio completo ou incompleto (linha tracejada).

CAPÍTULO

6

PROPOSTAS DE TRABALHOS FUTUROS

Nosso trabalho consistiu no estudo da referência principal (VALLEJOS; STEEL, 2017), na proposta de um modelo seguindo suposições diferentes, em um resultado analítico que verifica a interpretação dos modelos (proposto e estudado), em um estudo de simulação e em uma análise de dados reais.

A proposta de um trabalho futuro compreende em:

- Realizar estudo de simulação considerando diferentes percentuais de censuras e outros valores para os betas.
- Analisar outras funções de ligação, por exemplo a probit e a Cauchy;
- Analisar os modelos em outro contexto. Por exemplo, contrato de previdência privada, em que a retirada do valor total ao fim do contrato é um evento que compete com a conversão mensal;
- Fazer uma comparação entre um modelo contínuo de causas competitivas e as abordagens discretas visando mensurar as consequências de adotar um modelo contínuo quando os dados são discretos;
- Análise de resíduos.

REFERÊNCIAS

ALBERT, A.; ANDERSON, J. A. On the existence of maximum likelihood estimates in logistic regression models. **Biometrika**, Oxford University Press, v. 71, n. 1, p. 1–10, 1984. Citado na página 29.

CARVALHO, M. S.; ANDREOZZI, V. L.; CODEÇO, C. T.; CAMPOS, D. P.; BARBOSA, M. T. S.; SHIMAKURA, S. E. **Análise de Sobrevivência: Teoria e Aplicações em Saúde**. Rio de Janeiro: SciELO-Editora FIOCRUZ, 2011. Citado na página 22.

COX, D. Regression models and life tables (with discussion). **Journal of the Royal Statistical Society**, v. 34, p. 187–220, 1972. Citado nas páginas 20 e 24.

GHOSH, J.; LI, Y.; MITRA, R. *et al.* On the use of cauchy prior distributions for bayesian logistic regression. **Bayesian Analysis**, International Society for Bayesian Analysis, v. 13, n. 2, p. 359–383, 2018. Citado na página 29.

JENKINS, S. P. **Survival Analysis**. Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK, 2005. Citado nas páginas 21 e 25.

JUAJIBIOY, J. C. Study of university dropout reason based on survival model. **Open Journal of Statistics**, Scientific Research Publishing, v. 6, n. 5, p. 908–916, 2016. Citado na página 20.

JÚNIOR, P. L.; SILVEIRA, F. L. d.; OSTERMANN, F. Análise de sobrevivência aplicada ao estudo do fluxo escolar nos cursos de graduação em Física: um exemplo de uma universidade brasileira. **Revista Brasileira de Ensino de Física**, v. 34, 1403, 10 p., 2012. Citado na página 19.

KING, G.; ZENG, L. Logistic regression in rare events data. **Political analysis**, Cambridge University Press, v. 9, n. 2, p. 137–163, 2001. Citado na página 29.

LAWLESS, J. F. **Statistical Models and Methods for Lifetime Data**. New York City: John Wiley & Sons, 2011. Citado na página 21.

LEHMANN, W. "i just didn't feel like i fit in": The role of habitus in university dropout decisions. **Canadian Journal of Higher Education**, v. 37, n. 2, 2007. Citado na página 19.

LI, Y.; CLYDE, M. A. Mixtures of g-priors in generalized linear models. **Journal of the American Statistical Association**, Taylor & Francis, v. 113, n. 524, p. 1828–1845, 2018. Citado na página 29.

MEGGIOLARO, S.; GIRALDO, A.; CLERICI, R. A multilevel competing risks model for analysis of university students' careers in Italy. **Studies in Higher Education**, v. 42, p. 1259–1274, 2017. Citado na página 20.

ORTIS, E. A.; DEHON, C. **The roads to success: Analyzing dropout and degree completion at university**. Working Papers ECARES 2011-025, ULB-Universite Libre de Bruxelles, 2011. Citado na página 20.

PAURA, L.; ARHIPOVA, I. Cause analysis of students' dropout rate in higher education study program. **Procedia-Social and Behavioral Sciences**, v. 109, p. 1282–1286, 2014. Citado na página 19.

PIETRO, G. D.; CUTILLO, A. Degree flexibility and university drop-out: The italian experience. **Economics of Education Review**, Elsevier, v. 27, n. 5, p. 546–555, 2008. Citado na página 19.

PLUMMER, M. *et al.* Jags: A program for analysis of bayesian graphical models using gibbs sampling. In: VIENNA, AUSTRIA. **Proceedings of the 3rd international workshop on distributed statistical computing**. [S.l.], 2003. v. 124, n. 125, p. 10. Citado na página 29.

POLSON, N. G.; SCOTT, J. G.; WINDLE, J. Bayesian inference for logistic models using pólya–gamma latent variables. **Journal of the American statistical Association**, Taylor & Francis, v. 108, n. 504, p. 1339–1349, 2013. Citado na página 29.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2019. Disponível em: <http://www.R-project.org/>. Citado na página 31.

SOUZA, A. O. **Testes estatísticos em regressão logística sob a condição de separabilidade**. Tese (Doutorado) — Universidade Federal de Viçosa, 2010. Citado na página 27.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. V. D. Bayesian measures of model complexity and fit. **Journal of the royal statistical society: Series b (statistical methodology)**, Wiley Online Library, v. 64, n. 4, p. 583–639, 2002. Citado na página 30.

SU, Y.-S.; YAJIMA, M. R2jags: Using r to run 'jags'. **R package version 0.5-7**, v. 34, 2015. Citado nas páginas 29 e 31.

TEAM, R. C. *et al.* R: A language and environment for statistical computing. Vienna, Austria, 2013. Citado na página 29.

TINTO, V. Dropout from higher education: A theoretical synthesis of recent research. **Review of Educational Research**, v. 45, p. 89–125, 1975. Citado na página 19.

TUTZ, G.; SCHMID, M. **Modeling Discrete Time-To-Event Data**. New York: Springer, 2016. Citado nas páginas 21 e 22.

VALLEJOS, C. A.; STEEL, M. F. Bayesian survival modelling of university outcomes. **Journal of the Royal Statistical Society: Series A (Statistics in Society)**, v. 180, p. 613–631, 2017. Citado nas páginas 15, 17, 20, 24, 27, 29, 31, 32, 33, 34, 35 e 45.

WILLETT, J. B.; SINGER, J. D. From whether to when: New methods for studying student dropout and teacher attrition. **Review of Educational Research**, v. 61, p. 407–450, 1991. Citado na página 19.

APÊNDICE

A

GRÁFICOS DAS SIMULAÇÕES

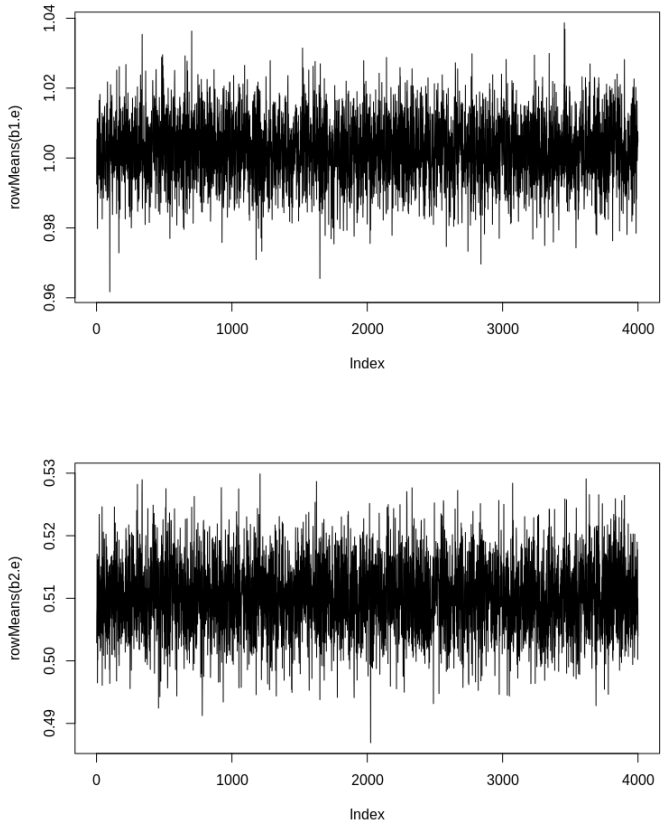


Figura 7 – Convergência dos coeficientes de regressão para os eventos 1 e 2.

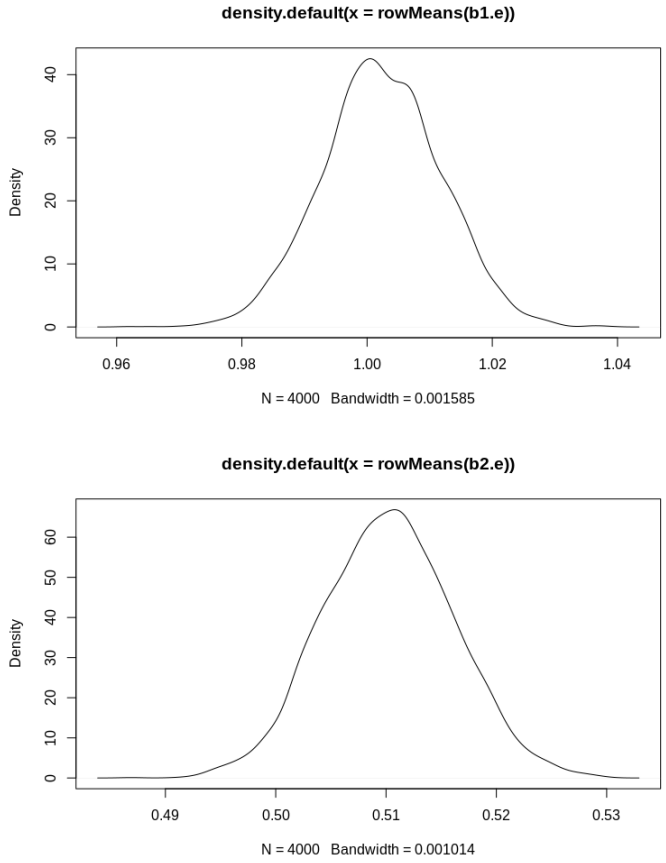


Figura 8 – Densidade dos coeficientes de regressão para os eventos 1 e 2.

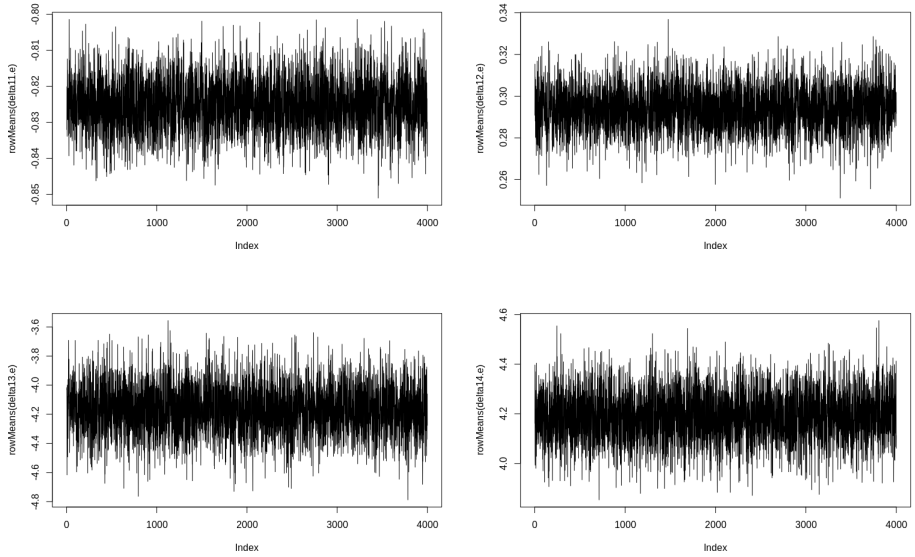


Figura 9 – Convergência do log-log complementar do risco basal de observar o evento 1 com respeito ao complementar do risco de observar evento 2 para todos os períodos considerando os dados gerados e estimados pelo modelo proposto.

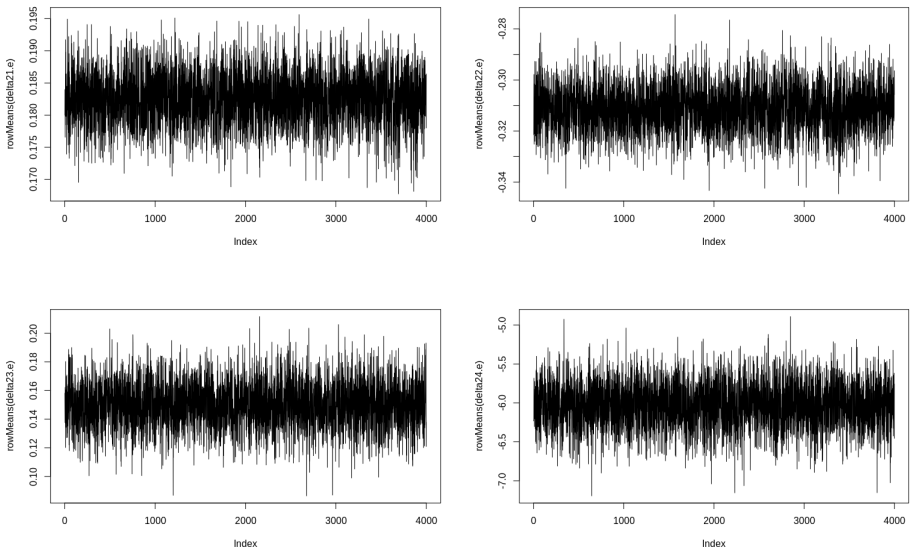


Figura 10 – Convergência do log-log complementar do risco basal de observar o evento 2 com respeito ao complementar do risco de observar evento 1 para todos os períodos considerando os dados gerados e estimados pelo modelo proposto.

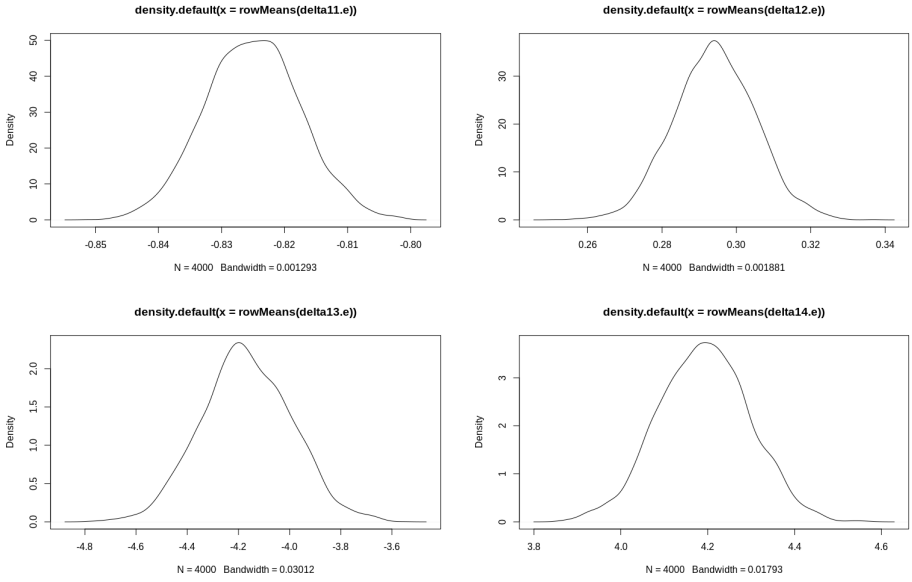


Figura 11 – Densidade do log-log complementar do risco basal de observar o evento 1 com respeito ao complementar do risco de observar evento 2 para todos os períodos considerando os dados gerados e estimados pelo modelo proposto.

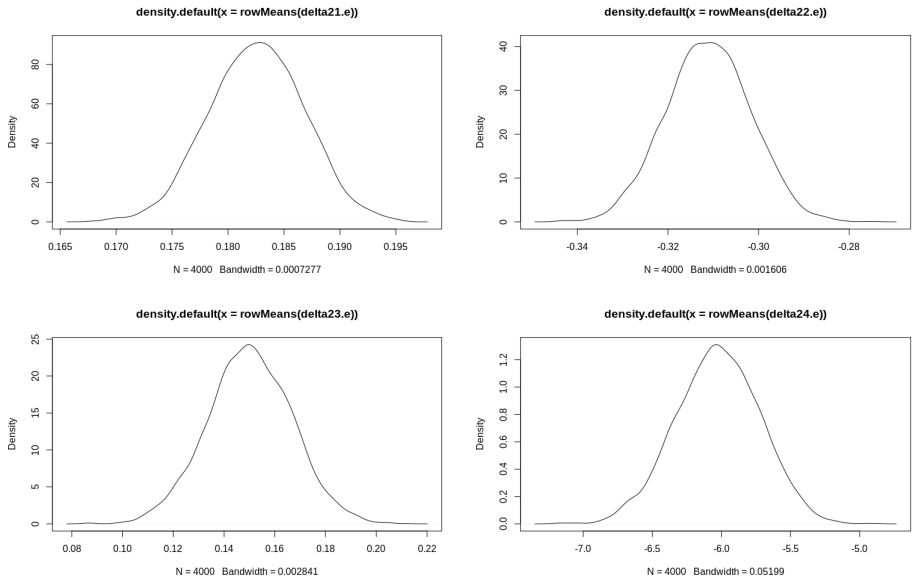


Figura 12 – Densidade do log-log complementar do risco basal de observar o evento 2 com respeito ao complementar do risco de observar evento 1 para todos os períodos considerando os dados gerados e estimados pelo modelo proposto.

APÊNDICE

B

GRÁFICOS DA APLICAÇÃO AOS DADOS REAIS

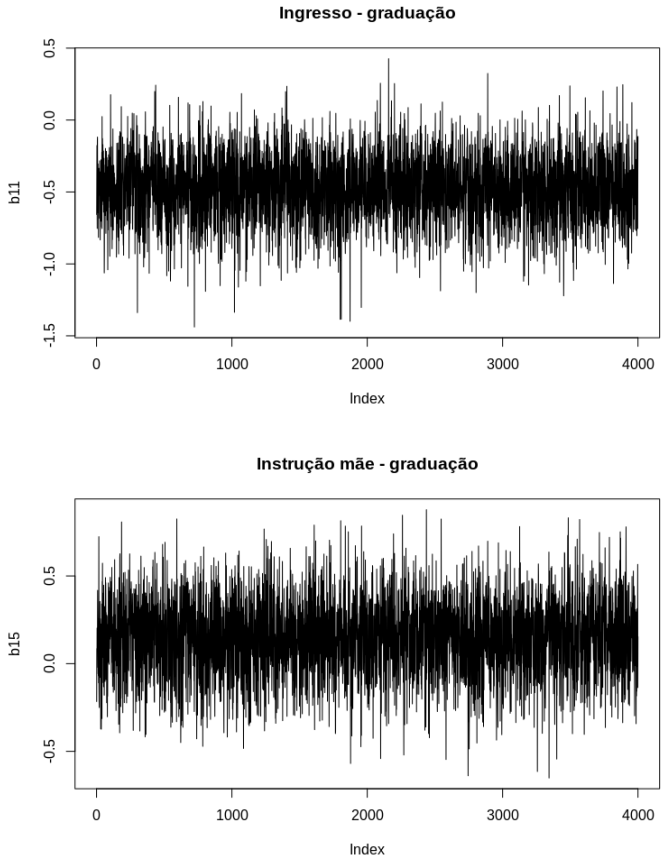


Figura 13 – Convergência dos coeficientes de regressão das variáveis Ingresso e Instrução mãe para o evento graduação.

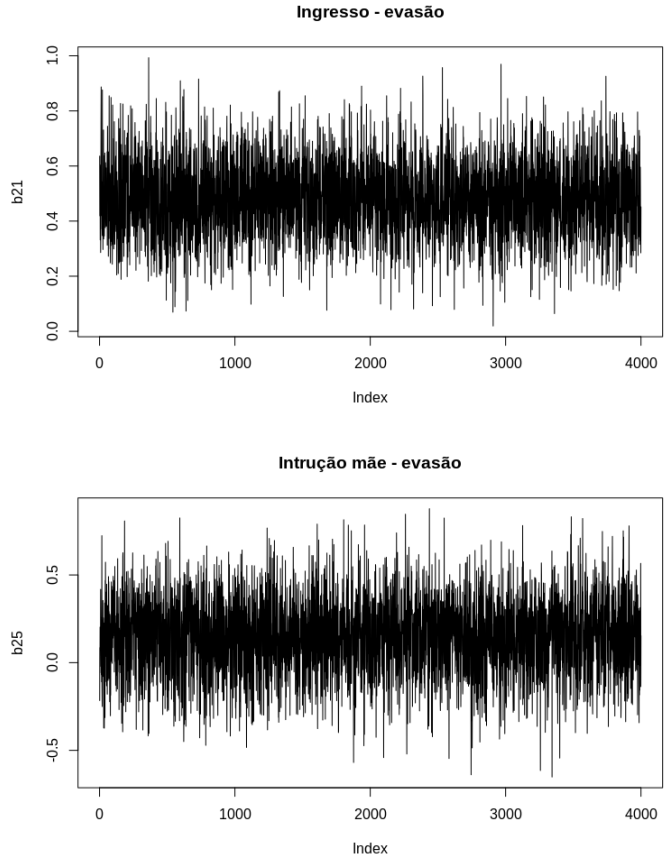


Figura 14 – Convergência dos coeficientes de regressão das variáveis Ingresso e Instrução mãe para o evento evasão.

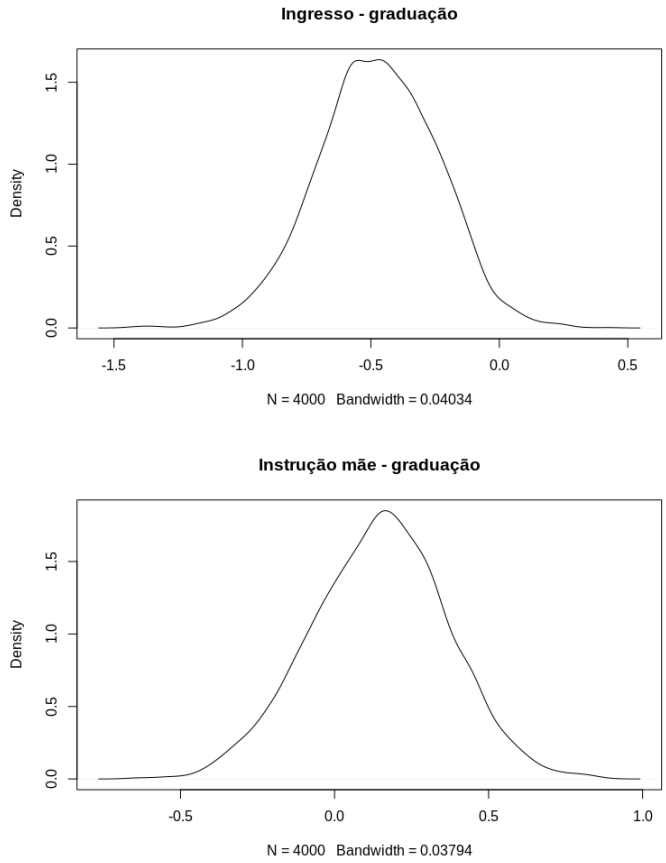


Figura 15 – Densidade das estimativas dos coeficientes de regressão das variáveis Ingresso e Instrução mãe para o evento graduação.

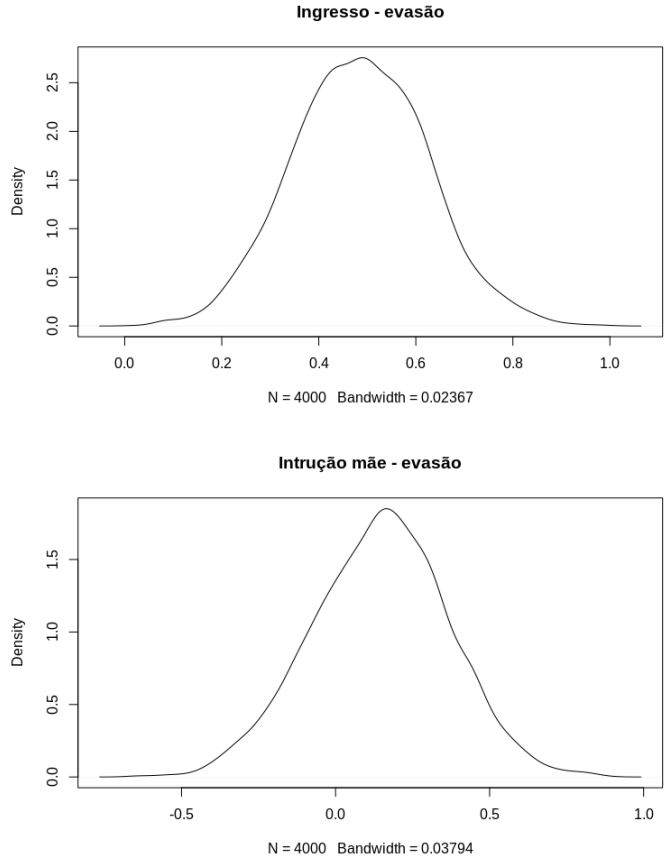


Figura 16 – Densidade das estimativas dos coeficientes de regressão das variáveis Ingresso e Instrução mãe para o evento evasão.