Thalita Paula Monteiro

Modelos de regressão logística para inadimplência entre doadores filantrópicos: variáveis determinantes & predição

Niterói – RJ, Brasil
2015

Universidade Federal Fluminense

Thalita Paula Monteiro

Modelos de regressão logística para inadimplência entre doadores filantrópicos: variáveis determinantes & predição

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador: Prof. Adrian Heringer Pizzinga

Niterói - RJ, Brasil

2015

Universidade Federal Fluminense

Thalita Paula Monteiro

Modelos de regressão logística para inadimplência entre doadores filantrópicos: variáveis determinantes & predição

Monografia de Projeto Final de Graduação sob o título *"Modelos de regressão logística para inadimplência entre doadores filantrópicos: variáveis determinantes & predição",* defendida por Thalita Paula Monteiro e aprovada em 07 de julho de 2015, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

_____

Prof. Adrian Heringer Pizzinga
Orientador
Departamento de Estatística – UFF

_____

Prof. Marco Aurélio dos Santos Sanfins
Departamento de Estatística – UFF

_____

Prof. Valentin Sisko
Departamento de Estatística – UFF

Niterói, 07 julho de 2015

Monteiro, Thalita Paula
Modelos de regressão logística para inadimplência entre doadores filantrópicos: variáveis determinantes & predição / Thalita Paula Monteiro; Adrian Heringer Pizzinga, orientador. Niterói, 2015.
37 f. : il.

Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2015.

1. Doadores. 2. Inadimplência. 3. Instituição filantrópica. 4. Modelo de regressão logística. 5. Predição. I. Pizzinga, Adrian Heringer, orientador. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.

CDD -

Resumo

Este Trabalho de Conclusão de Curso oferece uma solução para o seguinte problema: entender e otimizar (reduzir) as taxas de inadimplência de pessoas físicas quanto às suas comprometidas doações para instituições filantrópicas. O esforço é concentrado na busca de um adequado modelo de regressão logística para desvendar as variáveis que influenciem as probabilidades de ocorrerem inadimplência e adimplência por parte de uma dada pessoa física e, principalmente, para predizer tais casos a partir do conhecimento dessas variáveis. O estudo de caso envolve um conjunto de dados de pessoas físicas doadoras da Junta de Missões Nacionais da Convenção Batista Brasileira oriundos das cinco regiões do Brasil. Os principais resultados obtidos: (i) as variáveis idade, estado civil, sexo, região, classificação (do tipo de doador), tipo de cobrança e vencimento se mostraram como os principais determinantes da inadimplência; e (ii) a taxa de acerto de predição, em validação fora da amostra, mostrou que o modelo final é capaz de excelente antecipação dos casos de inadimplência, os quais são os mais importantes de serem preditos.

Palavras-chaves: doadores; inadimplência; instituição filantrópica; modelo de regressão logística, predição.

Dedicatória

Dedico este TCC, a todo trabalho missionário feito no Brasil, em especial a Junta de Missões Nacionais, que nestes 108 anos de história tem mostrado tamanha seriedade, amor ao próximo e obediência aos dois maiores mandamentos nos deixado por Jesus, como descrito em Mateus 22: 37 a 39 :

> *"E Jesus disse-lhe: Amarás o Senhor teu Deus de todo o teu coração, e de toda a tua alma, e de todo o teu pensamento. Este é o primeiro e grande mandamento. E o segundo, semelhante a este, é:*
> *Amarás o teu próximo como a ti mesmo."*

Agradecimentos

Agradeço primeiramente a Deus pela alegria e privilégio de estar concretizando mais um de seus sonhos em minha vida, pensando perfeitamente em cada detalhe desde a faculdade, curso, professores, colegas; levando-me além do que poderia imaginar, assim como afirma o apóstolo Paulo em 1 Coríntios 2:9:

> *"Olho nenhum viu, ouvido nenhum ouviu, mente nenhuma imaginou o que Deus preparou para aqueles que o amam"*

Aos meus pais, Luis e Antônia, por todo apoio, sustento e oração durante todo este tempo de faculdade. Sempre incentivando, apoiando, chorando junto em meio às dificuldades, vibrando em meio às conquistas e muitas das vezes anulando os seus próprios sonhos para que o meus pudessem se tornar realidade. Muito obrigada! Toda essa força e garra que vêem em mim não existiriam se vocês não estivessem na base e fossem a minha referência.

Ao meu irmão Carlos Augusto por todo incentivo e companheirismo durante todo curso, segurando as cordas quando elas pareciam se romper. Obrigada pelo exemplo de homem, caráter e principalmente fidelidade a Deus.

Ao meu sobrinho Isaque, que nasceu exatamente neste período e veio para despertar em mim um sentimento desconhecido. Você veio para somar e incentivar ainda mais a busca por uma boa formação, um futuro de sucesso podendo assim lhe propiciar o melhor desta vida.

Aos amigos, que não são poucos, por isso não citarei nomes, obrigada pelo apoio em todo tempo. Aos de Niterói que se tornaram minha segunda família, muito obrigada pelo companheirismo, e por fazerem de alguma forma que a saudade e a vida longe de casa se tornasse mais fácil. Aos amigos que estão longe, obrigada porque em todo tempo estiveram me apoiando, acreditando e torcendo por mim.

Aos colegas de curso, por cada experiência que vivemos juntos, cada aprendizado, estudo, conhecimentos, desafios, que foram muitos, mas graças a Deus vencemos todos, prova disso que estamos aqui para contar cada um deles.

Aos queridos professores que marcaram este tempo, sempre com muito profissionalismo e empenho para que os cursos oferecidos fossem os melhores e mais completos possíveis. Em especial nesta reta final de TCC, sempre se mostrando solidários e prontos para contribuir e ajudar no que fosse preciso.

A Junta de Missões Nacionais e a gerência pela confiança e credibilidade de estar realizando este estudo em busca de respostas para um sério desafio da empresa. Obrigada por todas as conversas esclarecedoras e toda a disponibilidade de dados e informações. Em especial, agradeço aos amigos da GEDEM pelas orações e apoio.

Ao orientador Adrian por ter aceitado o desafio de me orientar, pelo empenho e dedicação em todas as etapas deste projeto, sempre com ótimas dicas e sugestões e principalmente por sempre prezar pela excelência. Esse foi um dos maiores motivos por ter escolhido você.

A todos que direta ou indiretamente contribuíram para que esse sonho se tornasse realidade. Muito obrigada!

Sumário

1. Introdução

1.1. Motivação

A Junta de Missões Nacionais (JMN) é uma agência missionária filantrópica que pertence às Igrejas Batistas da Convenção Batista Brasileira (CBB). De acordo com Oliveira (2007), a JMN iniciou suas atividades no dia 25 de junho de 1907 durante a primeira assembléia da CBB na cidade de Salvador na Bahia, *"tomada pelo amor ao Brasil e aos brasileiros que estão sem Deus e sem esperança"*.Sua missão é Conquistar a Pátria para Cristo, de acordo com as palavras de L.M.Bratcher, o apóstolo do sertão: *"Aconteça o que acontecer, os planos para evangelização da pátria precisam ser cumpridos... sei que vamos ganhar esta terra para Cristo"*.

Como declarado no site oficial (www.missoesnacionais.org.br), o maior desafio da JMN é fazer com que o evangelho seja proclamado em cada lugar desse país: "Não importa aonde for. Seja no campo, na floresta, nas tribos, nas regiões ribeirinhas, na cidade, no interior, na periferia das grandes cidades, nos condomínios fechados, nos bairros, nas vilas, nas grandes metrópoles etc."

Existe uma grande preocupação de caráter social por parte da JMN, como publicado no seu site oficial: *"O que fazer em favor dos moradores de rua, daqueles que são escravos das drogas, das tribos urbanas e tantos outros grupos que precisam de Cristo? Precisamos continuar enviando missionários que sejam apaixonados por Cristo e seus por semelhantes, para que, o evangelho seja proclamado em todos os lugares e a todas as pessoas sem distinção de raça, cor ou religião"*.

Para tanto, a JMN trabalha sustentando missionários em vários projetos. Dentre estes, destaca-se a Cristolândia, um programa de recuperação de dependentes químicos. Somente neste projeto, existem 34 unidades distribuídas em sete estados do Brasil, acolhendo 1.202 contemplados – dentre esses últimos, destaca-se o testemunho, publicado por Santos (2015), de uma pessoa (eis usuário em recuperação) atendida no projeto, com 30 anos, e com histórico de consumo de drogas por 12 anos, 4 deles morando nas ruas da Cracolândia em São Paulo:

"Hoje casada e feliz só tenho a agradecer porque até aqui o Senhor me sustentou e nunca desistiu de mim e desenvolveu a alegria de poder sonhar novamente".

Como afirma Brandão (2015):

"É preciso avançar e, portanto, não podemos deixar de contribuir para que mais vidas sejam alcançadas e transformadas".

1.2 Relevância

Ao ser analisada a distribuição das fontes de receita da JMN, fica evidenciada uma concentração de receitas em duas fontes, quais sejam, o Programa de Adoção Missionária (PAM) e a Campanha Anual. Dentre as duas, destaca-se, tanto pelo volume da receita quanto pelo crescimento histórico, a fonte de receita advinda do PAM. A distribuição de receitas no período de 2014 pode ser verificada na Figura 1:

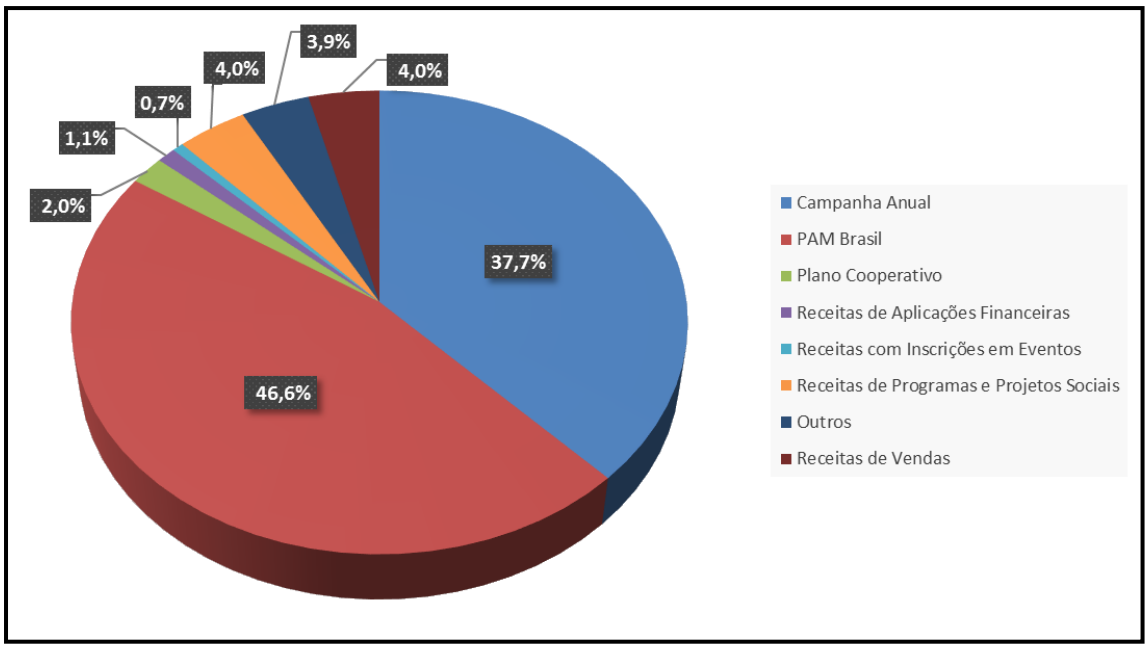

Figura 1 – Distribuição de receitas globais da JMN
(Fonte: Sistema Financeiro e Contábil da JMN 2015).

Se a análise considerar o crescimento histórico, detalhado na Figura 2, fica ainda mais evidente que o PAM é uma estratégia de captação de receitas que se tornou fator de importância central para a sobrevivência da JMN.

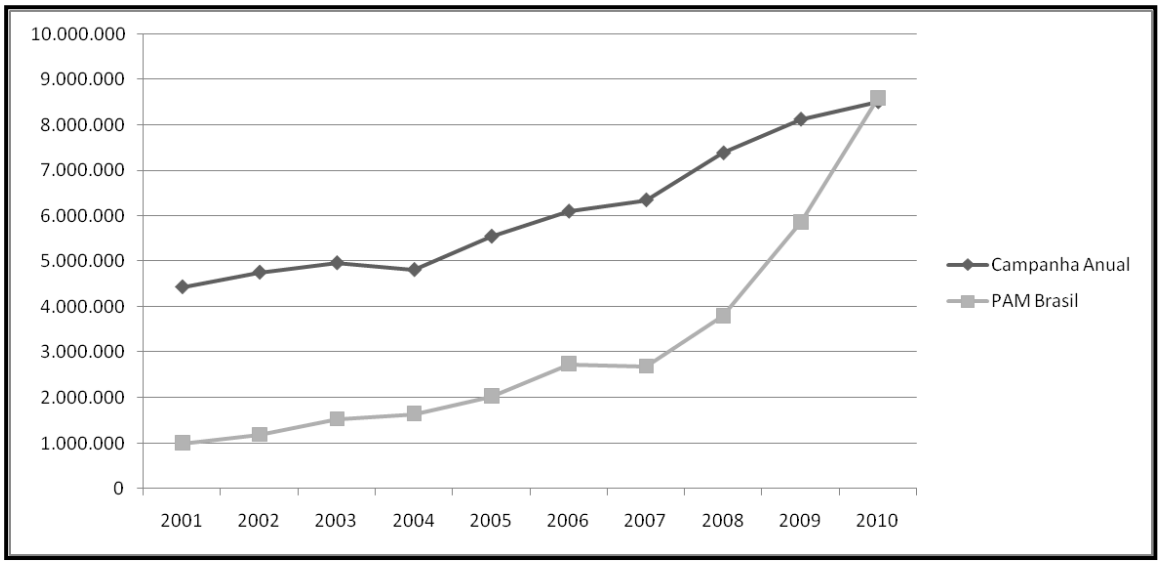

Figura 2 – Crescimento histórico do PAM Brasil

(Fonte: Sistema Financeiro e Contábil da JMN 2015).

O PAM permite mais avanços em novos projetos, ministérios sociais, evangelização em todo país e ainda um envolvimento pessoal e mais direto dos doadores, adotando um missionário específico e seu projeto por meio de ofertas designadas, oração e apoio. Este trabalho em parceria funciona como uma empresa com vários sócios ou parceiros.

Pela Figura 3, e considerando a importância do PAM, os resultados não mostraram-se satisfatórios no corrente período, pois as correspondentes doações não alcançaram o valor esperado.

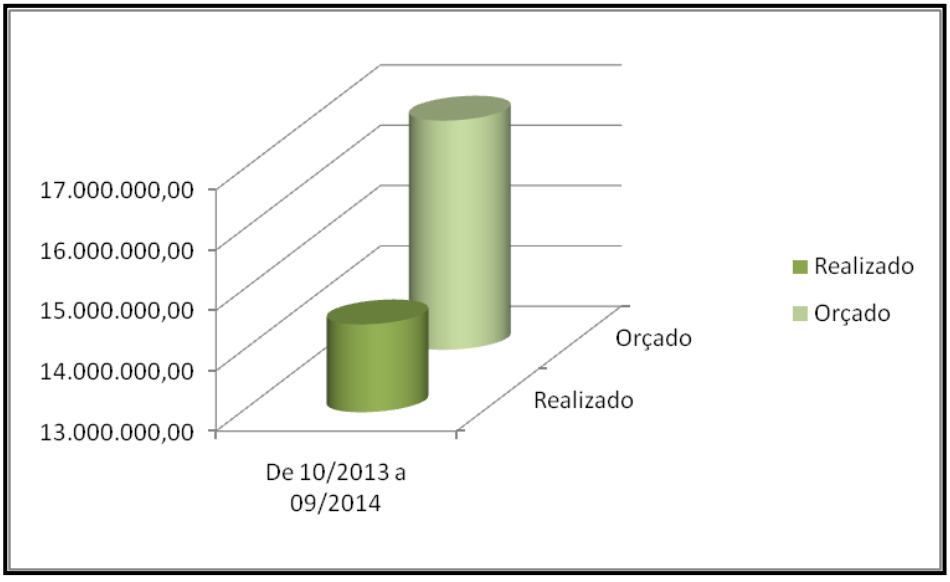

Figura 3 – Dado orçamentário do PAM Brasil

(Fonte: Sistema Financeiro e Contábil da JMN 2015).

Este resultado se deu por conta do alto índice de inadimplência dos doadores em seus respectivos projetos, o que trouxe, por conseguinte, uma preocupação quanto à estabilidade financeira da JMN.

1.3 Analogia com o risco de crédito

Blatt (1999) enfatiza que a palavra crédito se originada palavra latina *credere* - crer, confiar, acreditar – e do substantivo *creditum* –ou seja, confiança. O crédito geralmente se refere à expectativa de se receber um valor em um determinado período de tempo. De acordo com Caouette et al. (1999), página 1, o risco de crédito é definido como sendo a probabilidade de que essa expectativa não se cumpra.

Especificamente, pode-se dizer que o risco de crédito é a probabilidade de o credor não receber, quando as obrigações assumidas pelo tomador não forem liquidadas da forma acordada por ambos. Bessis (1998), página 81, por seu turno, define risco de crédito como sendo as perdas geradas por um evento de *default* do tomador ou pela deterioração da sua qualidade de crédito; esta última não necessariamente resulta em perda imediata para o credor, mas, sim, em um aumento da probabilidade de que um evento de default venha a acontecer.

Como exemplo, Bessis (1998) cita que um evento de default de um tomador pode ser entendido como o atraso no pagamento de uma obrigação, o descumprimento de uma cláusula contratual restritiva (*covenant*), o início de um procedimento legal como a concordata e a falência, a inadimplência de natureza econômica etc. Cada empresa delimita qual é o evento de default mais pertinente.

Tendo em mente os objetivos deste Trabalho de Conclusão de Curso (TCC), pode-se considerar, como efeito de default, a inadimplência de natureza econômica e o atraso no pagamento de uma obrigação por parte do doador (em particular para este estudo, se refere ao tomador) no período de 90 dias.

O risco de crédito pode ser avaliado pelos seus componentes que são o risco de default, o risco de exposição e o risco de recuperação (cf. Bessis, 1998). O risco de default é a probabilidade de o devedor não honrar seus compromissos assumidos; o risco de exposição é a incerteza do valor do crédito no momento do default; e o risco de recuperação é a incerteza do valor que pode ser recuperado pelo credor no caso de um default do doador.

Os modelos de risco de crédito são ferramentas usadas para mensurar o risco de tomadores e transações. Entre estes, citam-se os modelos de previsão de insolvência que têm como objetivo calcular/estimar a probabilidade de uma empresa ou indivíduo cair em um evento de default em um determinado período de tempo. Esses modelos se baseiam em uma amostra histórica de empresas tomadoras de crédito consideradas insolventes, empresas e indivíduos que caíram em eventos de default, ou solventes, quando não caíram. Com a amostra, são identificadas as variáveis que melhor discriminam as empresas em insolventes ou solventes no período da análise. Esse conjunto de variáveis é usado para classificar as empresas com suas futuras operações de créditos como prováveis solventes ou insolventes. Estes modelos de previsão de insolvência se baseiam em técnicas estatísticas de Análise Multivariada, tais quais a análise discriminante e classificatória (cf. Johnson & Wichern, 2007), e de Análise de Regressão, de cujos possíveis exemplos assinalam-se os modelos de regressão logística (cf. Hosmer & Lemeshow, 1989; Dobson, 1990; e Neter et al., 1996). Para este TCC, foi usada uma amostra histórica de doadores inadimplentes e adimplentes, a fim de mensurar o risco de inadimplência e estimar um modelo de regressão logística para predição.

1.4 Organização do texto

O restante deste TCC assim se organiza. O capítulo 2 descreve os objetivos do TCC. O capítulo 3 é reservado aos pormenores essenciais da teoria dos modelos de regressão logística. No capítulo 4, é desenvolvido, por completo, todo o exercício empírico com os dados sobre pessoas físicas doadoras da JMN – ao longo deste, serão abordados: o banco de dados, estatísticas descritivas, o ciclo percorrido de modelagem, interpretações dos resultados, e análise do poder preditivo com o modelo definitivo. O capítulo 5 destina-se à conclusão e às sugestões de estudos futuros dentro do tema.

2. Objetivos

Neste TCC, são estimados modelos de regressão logística para que sejam atendidos dois objetivos: (i) identificar as variáveis explicativas estatisticamente significantes que determinam as probabilidades de ocorrerem inadimplência e adimplência pelas pessoas físicas doadoras do PAM; (ii) como tarefa principal, predizer/antever casos de inadimplência e adimplência, a partir do conhecimento das variáveis explicativas previamente identificadas,

correspondentes a novas pessoas físicas doadoras.Visa-se, com estes dois objetivos devidamente atingidos, à ampliação das arrecadações e do desenvolvimento de ações de mobilização de pessoas, com o objetivo de aumentar a receita da JMN – e, por conseguinte, alcançar mais eficiência na implementação dos projetos.

O foco, pelo menos nas contribuições empíricas deste TCC, recairá sobre dados com informações de *pessoas físicas*, pois elas representam quase 7 em cada 10 contribuições ao PAM (cf. Amaral, 2013). Além do mais, cerca de 7 em cada 10 pessoas que assinam um PAM o deixam após três meses (cf. Souza, 2013).

3. Modelos de regressão logística

3.1 Modelos Lineares Generalizados para respostas binárias

Uma frequente aplicação da teoria dos Modelos Lineares Generalizados (MLGs) – cf. Cordeiro (1986), McCullagh & Nelder (1989), Dobson (1990) e Hardin & Hilbe (2007) – é sua utilização na modelagem de variáveis dependentes do tipo binário 0-1 (este é o caso particular mais simples de situações em que a variável resposta representa atributos categóricos nominais ou, no máximo, ordinais). Existem algumas possibilidades, já bem estabelecidas, de se formular um MLG para uma análise de regressão, na qual a variável resposta só assume dois valores possíveis, indicando a ocorrência ou não de um determinado evento de interesse – este último é comumente chamado de "sucesso" (cf. DeGroot, 1986). Ao longo desta seção, dar-se-á foco no *modelo de regressão logística*.

3.2 Características dos componentes básicos dos MLGs para respostas binárias

De acordo com Cordeiro (1986), McCullagh & Nelder (1989), Dobson (1990) e Hardin & Hilbe (2007), um MLG é definido por um trinômio de componentes básicos, sendo um *aleatório*, outro *sistemático* e, por fim, a *função de ligação*. As principais características destes componentes para os MLGs destinados à modelagem de respostas binárias são apresentadas abaixo:

- **Componente aleatório:** Devido à natureza dos dados, admite-se que:

$$Y_i \sim Ber(p_i),$$

na qual p_i é a probabilidade de que ocorra sucesso na observação da i-ésima unidade amostral (isto é: $Y_i = 1$). Pode ser provado que a distribuição de Bernoulli possui expressão analítica pertencente à família exponencial na forma canônica (cf. DeGroot, 1986; e Dobson, 1990). Logo, o componente aleatório está de acordo com os pressupostos básicos de um MLG.

- **Componente sistemática:** Define-se uma transformação linear $T: \mathbb{R}^p \longrightarrow \mathbb{R}^n$ do vetor de parâmetros $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1})'$. A correspondente matriz de transformação é dada por uma matriz X, cuja primeira coluna é um vetor de 1´s (para contabilizar um efeito de intercepto β_0) e as suas demais $p-1$ colunas são formadas por valores de variáveis explicativas. O vetor-imagem em $\mathbb{R}^n$ gerado é composto dos *preditores lineares* definidos a expressão já muito conhecida $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1}$.

- **Função de ligação:** Nota-se que a média que Y_i é dada pela probabilidade de sucesso p_i, a qual deve pertencer ao intervalo $(0,1)$, e o preditor linear η_i é um número real, uma maneira conveniente de se interligar os dois é escolher uma função de ligação $g(.)$, que tenha como domínio o intervalo citado e como imagem a reta real. Matematicamente,

$g: (0,1) \longrightarrow \mathrm{R}$

ou, equivalentemente,

$$g^{-1}: \mathbb{R} \longrightarrow (0,1),$$

posto que tal função deve ser monótona e diferenciável, para não gerar conflitos com o processo de estimação do vetor de coeficientes β. Escolhida a função de ligação, tem-se:

$$g(p_i) = \eta_i \Leftrightarrow p_i = g^{-1}(\eta_i) \quad , \quad i = 1, 2, \ldots, n.$$

3.3 Função de ligação logito

Uma opção frequentemente adotada para a função de ligação para MLGs é a dada pela função de ligação *logit*:

$$\log_e\left(\frac{p_i}{1-p_i}\right)=\eta_i \ , \ i=1,2,\ldots,n$$

Invertendo-se esta relação, obtém-se:

$$p_i=\frac{\exp(\eta_i)}{1+\exp(\eta_i)}=\frac{1}{1+\exp(-\eta_i)} \ , \ i=1,2,\ldots,n\,.$$

Muitas vezes, a expressão "logito da probabilidade" é usada, quando esta transformação é utilizada como ligação em modelagem logit[1].

A taxa de variação de p_i em relação a k-ésima variável independente é dada por:

$$\frac{dp_i}{dx_i}=\beta_k p_i(1-p_i)\,.$$

O modelo logit apresenta diversas propriedades que o tornam bastante atrativo. Por exemplo, o processo de estimação não envolve maiores complicações computacionais. Outra vantagem deste modelo é a sua maior amplitude de interpretação, já que pode ser de interesse não só analisar a probabilidade p_i, mas também a *chance* definida por $\frac{p_i}{1-p_i}$, a qual, no modelo logit, é facilmente obtida tomando-se anti-logaritmos nos dois lados da expressão que envolve o logito.

3.4 A interpretação do parâmetro β_k no modelo logit

Apresenta-se, agora, um desenvolvimento matemático relativamente simples, mas que possibilitará a compreensão dos significados dos parâmetros de regressão no modelo logit. Especificamente, cada um destes coeficientes fornece informação de quanto são alterados o logito da probabilidade e a chance, caso a variável explicativa correspondente seja aumentada em uma unidade. Para facilitar a notação, admite-se que o modelo possui uma variável independente (regressão logística *simples*), sendo que as conclusões podem ser estendidas diretamente para o caso com mais variáveis (regressão logística *múltipla*). Também será omitido o sub-índice i, correspondente à identificação da unidade amostral. Logo, utilizar-se-á a escrita $\eta=\beta_0+\beta_1 x$.

Inicialmente, considere-se a expressão do modelo logit que evidencia a chance:

[1]Frequentemente, o ajuste do modelo logit é chamado de *regressão logística*.

$$\frac{p}{1-p} = \exp(\eta) = \exp(\beta_0 + \beta_1 x) = \exp(\beta_0)\exp(\beta_1 x).$$

Convenciona-se, a partir de agora, que o modelo acima, quando aplicado em x, é chamado de Chance 1 e, quando aplicado em $x+1$, é chamado de Chance 2. Ou seja:

$$\text{Chance}1 = \exp(\beta_0)\exp(\beta_1 x),$$
$$\text{Chance}2 = \exp(\beta_0)\exp(\beta_1(x+1)) = \exp(\beta_0)\exp(\beta_1 x + \beta_1) = \exp(\beta_0)\exp(\beta_1 x)\exp(\beta_1).$$

Das duas expressões acima, obtém-se a relação desejada entre as duas chances:

$$\text{Chance}2 = \text{Chance}1\left[\exp(\beta_1)\right].$$

Como pode ser visto acima, a exponencial de β_1 representa uma constante de proporcionalidade entre as Chances 1 e 2 – e, como esta constante também é recuperada por uma divisão entre as Chances 1 e 2, é muito frequente a atribuição a esta da expressão *razão de chance*[2]. Assim, podem ser tecidas interpretações a respeito da relação entre as chances – ou probabilidades, ou ainda, os valores observados das respostas – e os diferentes níveis de x:

1º) $\beta_1 > 0 \Leftrightarrow \exp(\beta_1) > 1$: Isto implica a conclusão de que, quando se aumenta x, a chance também é aumentada. A relação de p com x é dita ser *positiva*.

2º) $\beta_1 < 0 \Leftrightarrow \exp(\beta_1) < 1$: Daí, percebe-se que a chance é diminuída, caso se aumente x. A relação de p com x é dita ser *negativa*.

3º) $\beta_1 = 0 \Leftrightarrow \exp(\beta_1) = 1$: Quando isto ocorre, pode ser dito que, aumentando-se ou diminuindo-se x, a chance – ou, equivalentemente, a probabilidade – não se altera. A relação de p com x é dita ser *inexistente*.

Foi visto, portanto, que o parâmetro $\exp(\beta_1)$ desempenha um papel fundamental na interpretação de um modelo ajustado via regressão logística. Seu estimador de máxima verossimilhança é dado por $\exp(B_1)$, no qual B_1 é o estimador de máxima verossimilhança de β_1.

[2]Em inglês: *odds ratio*.

Cordeiro (1986), McCullagh & Nelder (1989) e Dobson (1990) demonstram que, se o conjunto de dados for relativamente grande, B_1 terá distribuição aproximadamente Normal, implicando que $\exp(B_1)$ terá uma distribuição aproximadamente Log-Normal. Esta última distribuição assintótica é assimétrica. Logo, inferências geralmente são feitas, inicialmente, para β_1 e, após isto, obtêm-se as conclusões para $\exp(\beta_1)$ – vide Hosmer (1989) para obtenção de mais detalhes.

3.5 Inferência e seleção de modelos

O critério de seleção do modelo estatístico visa encontrar o modelo que envolva o mínimo de parâmetros possíveis a serem estimados e que explique de maneira eficaz a incerteza da variável resposta. Para isso, são estabelecidas estratégias para selecionar o melhor modelo. Para chegar neste modelo parcimonioso, foram adotadas, neste TCC, as seguintes estratégias: testes de significância individuais, teste da razão de verossimilhança e critérios de informação de Akaike.

3.5.1 Testes de significância individuais

Geralmente, um teste de significância individual (Dobson, 1990) é utilizado para se ter uma primeira percepção de quais são as variáveis explicativas relevantes em um modelo de regressão. As hipóteses do teste, para o caso de se testar a significância da *j*-ésima variável explicativa, são:

$$\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases}$$

A estatística para este teste é a divisão entre o estimador de máxima verossimilhança do coeficiente correspondente e seu erro padrão; ou seja,

$$t = \frac{\hat{\beta}_j}{s.e.\left(\hat{\beta}_j\right)}.$$

A hipótese nula é rejeitada se o módulo do valor observado da estatística é superior ao quantil de probabilidade 1- $\alpha/2$ de uma distribuição normal padrão, sob um nível de significância α.

3.5.2 Teste da razão de verossimilhança

Este teste, também conhecido como teste de significância conjunta, é adequado quando o objetivo é comparar dois modelos (Kleinbaum & Klein, 2010). Estes devem ser *encaixados*, isto é, todas as q variáveis explicativas do *modelo reduzido* (MR) estão também entre as p variáveis explicativas do *modelo completo* (MC), sendo que $q < p = q + k$; veja que k é a diferença entre número de parâmetros nos dois modelos. Supondo sem perda de generalidade que os coeficientes cuja significância conjunta está sendo testada são os k últimos, as hipóteses do teste são:

$$\begin{cases} H_0: \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0 & \rightarrow modelo\ reduzido \\ H_1: algum\ \beta_k \neq 0, \quad k = q, q+1, \ldots, p-1 & \rightarrow modelo\ completo \end{cases}$$

A estatística de teste usada é

$$\text{LR} = -2\ \log(MR) - (-\ 2\log(MC)) = -2\ \log\left(\frac{MR}{MC}\right),$$

a qual tem distribuição nula assintótica Qui Quadrado com p-q graus de liberdade. Dessa forma, a hipótese nula é rejeitada se o valor observado da estatística LR for superior ao quantil de probabilidade 1- α da referida distribuição nula, sob um nível de significância α.

3.5.3 Critério de informação de Akaike

O critério de informação de Akaike (AIC) (Hardin & Hilbe, 2007) se baseia na função de log-verossimilhança, com introdução de um fator de correção como modo de penalização da complexidade do modelo. A estatística deste critério é:

$$AIC = -2\ [\log(L) - k]\ ,$$

na qual: $\begin{cases} L = toma\ o\ valor\ da\ verossimilhança\ para\ o\ modelo\ estimado \\ k = número\ de\ parâmetros\ do\ modelo\ estimado \end{cases}$

Na comparação entre dois modelos, aquele que apresentar menor valor de AIC será aquele considerado mais adequado aos dados.

3.6 Estimação de probabilidades e predição de futuros valores da variável reposta

É bastante comum, na prática, ajustar modelos para respostas de distribuição Bernoulli com o objetivo tanto de se estimar a média de cada resposta condicionada a níveis das variáveis explicativas, quanto também de se realizarem predições do modo mais eficaz possível.

Quanto ao processo de predição com um MLG para dados binários, elemento que está no âmbito deste TCC, há dois enfoques sobre o grau de dificuldade encontrado. Em outras palavras, a predição de novas observações de uma variável aleatória de Bernoulli modelada por MLG é extremamente fácil por um lado, e um tanto difícil por outro: a predição pode ser tida como fácil, se for levado em consideração que não demandados construções de intervalos de predição, sendo somente necessário o ato de observar a probabilidade estimada pontualmente como já explanado e reportando o valor predito: $\hat{Y}_h = 1$, se for predito que ocorrerá sucesso, e $\hat{Y}_h = 0$, se for predito o contrário. Mas existe a necessidade de se encontrar um *limiar* dentro do intervalo $(0,1)$, a partir do qual obtêm-se os valores preditos da variável resposta, através de comparações, respeitando o seguinte critério:

- Se a probabilidade estimada ultrapassar o limiar, prediz-se $\hat{Y}_h = 1$;
- Se a probabilidade estimada não ultrapassar o limiar, prediz-se $\hat{Y}_h = 0$.

A obtenção deste limiar é totalmente heurística, o que exige a investigação e comparação de procedimentos alternativos. Abaixo, são explicados brevemente dois destes procedimentos:

1. Utilizar o limiar 0,5 para o processo de predição: Este é o método mais fácil de se aplicar, além de ser muito intuitivo. Possui a desvantagem de, possivelmente, não ser o ponto que minimiza a porcentagem de predições erradas.
2. Buscar o limiar ótimo para minimização da taxa de predições incorretas: A partir do conjunto de dados utilizado no ajuste, é possível encontrar um limiar que torna o processo de predição o mais consistente possível, no seguinte sentido: encontrado tal limiar ótimo, obtêm-se os valores preditos de acordo com a regra de predição descrita anteriormente e comparam-se estes como os valores observados. O número de coincidências entre os valores observados e seus respectivos preditos (predições corretas) deve ser o maior possível. Uma maneira de se obter o

limiar ótimo pode ser a partir de um candidato inicial (exemplos: 0,5, média amostral da variável dependente etc.) e, em seguida, praticar-se uma análise de sensibilidade para valores próximos, sempre visando maximizar (minimizar) as predições corretas (incorretas). Outra forma de implementar essa busca é mais intensa, contudo muito mais exaustiva/abrangente: criar uma grade de valores no intervalo $(0,1)$ e, para cada valor da grade, fazer do mesmo um candidato a limiar e praticar a predição para cada valor da variável dependente dentro do banco de dados. Aqueles valores da grade que oferecerem taxas percentuais de acertos mais altas podem passar a um "segundo turno" de escolha, o qual levará em consideração as duas taxas de acertos, a dos sucessos (os 1's) e a dos não-sucessos (os 0's).

4. Estudo de Caso

Ao longo desta subseção, serão apresentadas as análises e os resultados obtidos de uma regressão logística feita com a base de dados sobre pessoas físicas doadoras da JMN. Pretende-se entender quais variáveis do banco de dados têm efeitos estatisticamente significantes para descrever a chance (relembre a definição de chance, enunciada na subseção 3.4) de um doador ser inadimplente, interpretar os efeitos estatisticamente significantes considerando a magnitude e sinal, e predizer a variável binária que caracteriza um doador inadimplente (sim ou não), a partir do melhor modelo logístico estimado e selecionado.

4.1 Descrição do banco de dados

O banco de dados, fornecido pelo Sistema Financeiro e Contábil da JMN no decorrente ano de 2015, possui diversas informações a respeito de 77.498 doadores, dentre eles pessoas físicas e jurídicas.

Ao analisar o banco de dados fornecido, foram detectados alguns problemas, tais como registros duplicados do mesmo doador e valores faltantes. Após a eliminação das linhas correspondentes aos problemas detectados e a seleção apenas de registros referentes a doadores classificados como pessoas físicas, o banco passou a ter 32.416 linhas.

Considerando que neste TCC se utilizará o método da validação *fora da amostra* do modelo, no qual é especialmente importante quando a finalidade é a previsão de resultados, foram selecionados de forma aleatória 10% dos dados para ser usado na validação. Portanto,

os 90% restantes, que corresponde a 29.174 registros dos doadores, estão sendo considerados no banco de dados base para os resultados a seguir.

As análises deste TCC consideram, além da variável dependente “inadimplente”, as seguintes variáveis explicativas: idade, estado civil, sexo, região, classificação, cobrança e vencimento. Todas as variáveis estão detalhadamente descritas na Tabela 1. As demais variáveis originalmente disponíveis no banco de dados foram descartadas pela forte incidência de campos nulos e/ou por se tratarem de variáveis não categorizáveis.

Tabela 1 – Variáveis utilizadas na análise.

Nome	Significado	Descrição
Inadimplente	Doador com mais de 3 meses sem efetuar pagamento do título no valor comprometido	(0) Não
		(1) Sim
Idade	Idade do doador	1 a 102 anos
Estado Civil	Estado Civil do doador	(C) Casado ou amasiado
		(D) Divorciado
		(S) Solteiro
		(V) Viúvo
Sexo	Sexo do doador	(0) Masculino
		(1) Feminino
Região	Região do Brasil que reside o doador	(1) Norte
		(2) Nordeste
		(3) Centro-Oeste
		(4) Sudeste
		(5) Sul
Classificação	Código interno da empresa para classificação do doador de acordo com o valor da sua doação	(1) Bronze 1
		(2) Bronze 2
		(3) Bronze 3
		(4) Prata 1
		(5) Prata 2, Ouro 1 e Ouro 3
Cobrança	Forma de pagamento do título de cobrança escolhida pelo doador	(1) Boleto
		(2) Crédito
		(3) Débito
Vencimento	Dia do vencimento do título de cobrança escolhido pelo doador	De 1 a 30

4.2 Estatísticas descritivas: análise de frequência dos inadimplentes

Ao longo desta subseção, estarão sendo apresentadas estatísticas descritivas das variáveis explicativas, usadas no modelo estimado, através dos gráficos de barras, agrupados pela variável dependente Inadimplente.

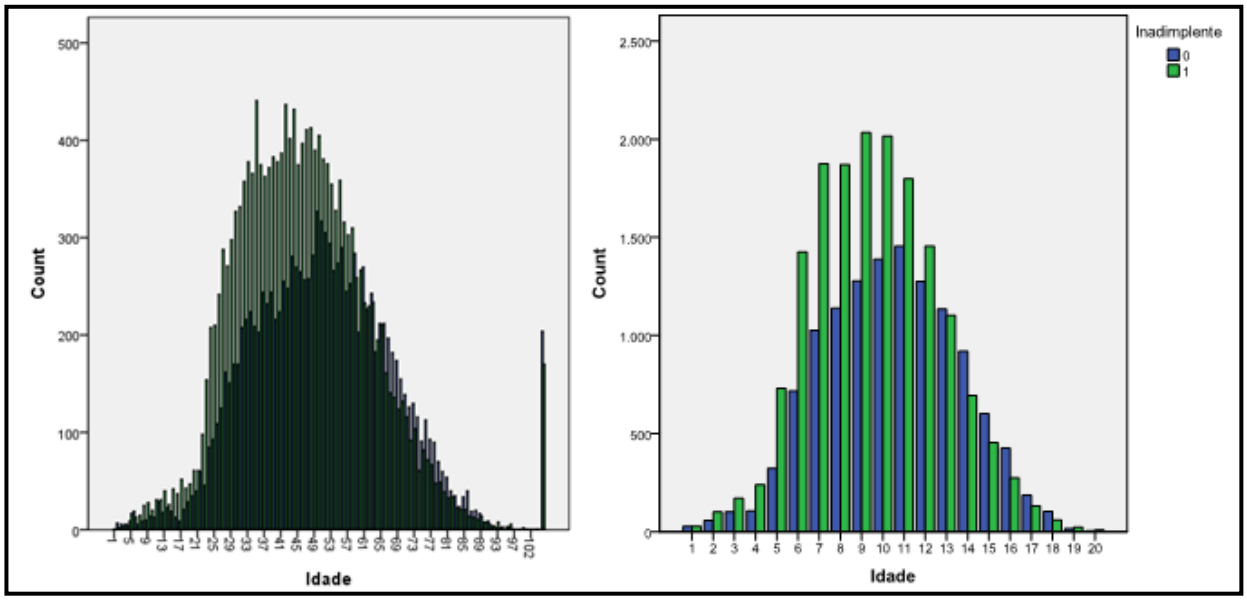

Figura 4 – Gráficos de barras para a variável “Idade”
(com diferentes categorizações).

Com base na Figura 4, é possível perceber que a grande parte dos doadores se concentram entre as idades de 25 a 70 anos. Os doadores com mais de 65 anos tendem a ser mais adimplentes, enquanto que, dentre os mais novos, observa-se maior freqüência de inadimplência. O gráfico da direita na Figura 4 corresponde à variável categorizada de 5 em 5 anos, e do da esquerda corresponde a todos os valores, sem tratamento da variável, incluindo outlier.

Com base na Figura 5, a maioria dos doadores é do sexo feminino e, para ambos os sexos, a frequência de doadores inadimplentes é maior, quando comparada aos adimplentes. Além disso, a maioria dos doadores são casados e a frequência de inadimplência é maior nos doadores casados, divorciados e solteiros, o que se inverte quando o doador é viúvo.

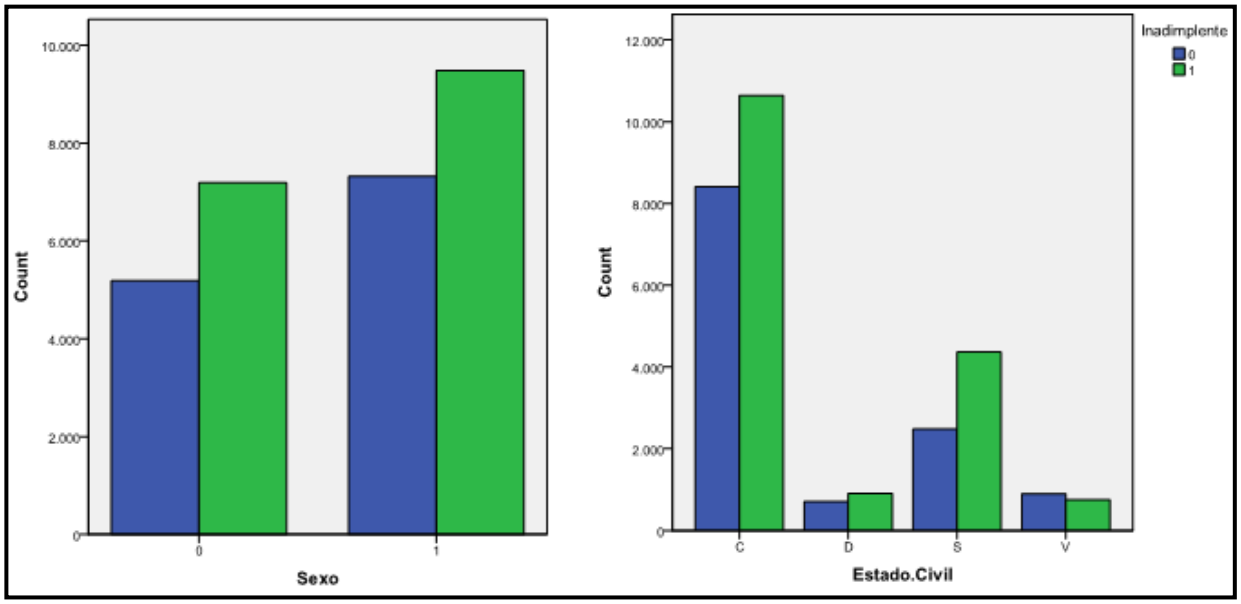

Figura 5 – Gráficos de barras para as variáveis “Sexo” e “Estado Civil”.

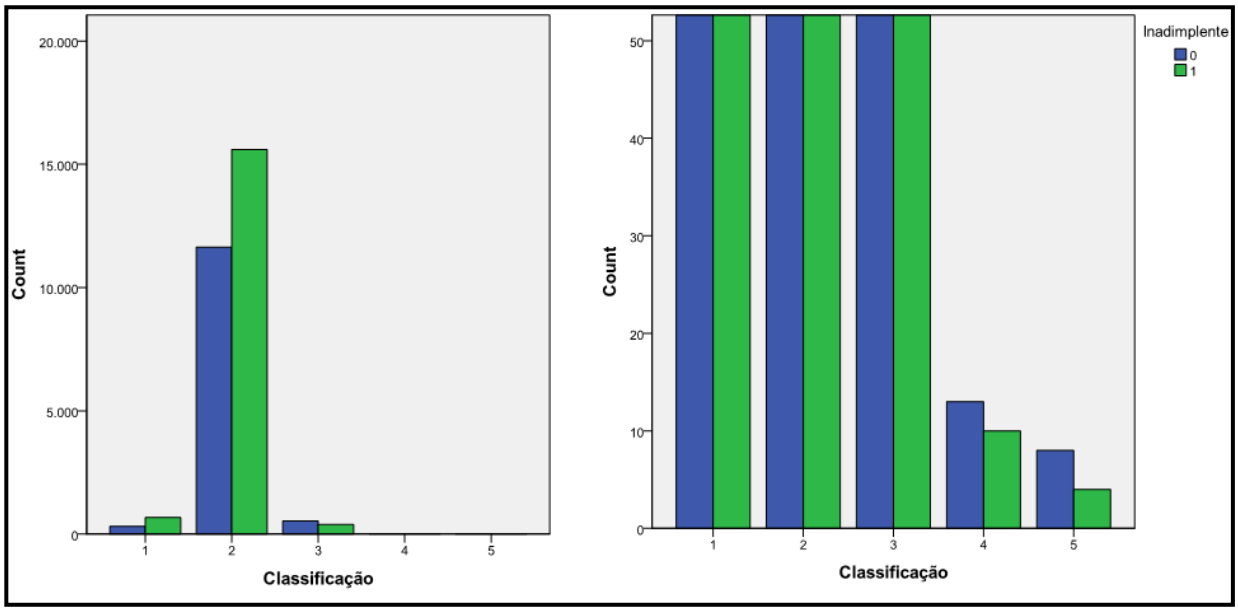

Figura 6 – Gráficos de barras para a variável “Classificação”
(em diferentes escalas).

Na Figura 6, é possível perceber que a grande maioria dos doadores se concentram na classificação Bronze 2 e que a frequência de inadimplência é maior para este grupo. Esta realidade se inverte quando o doador é classificado como Prata 1, Prata 2 e Ouro 1, quando, desta vez, a frequência dos doadores adimplentes é maior.

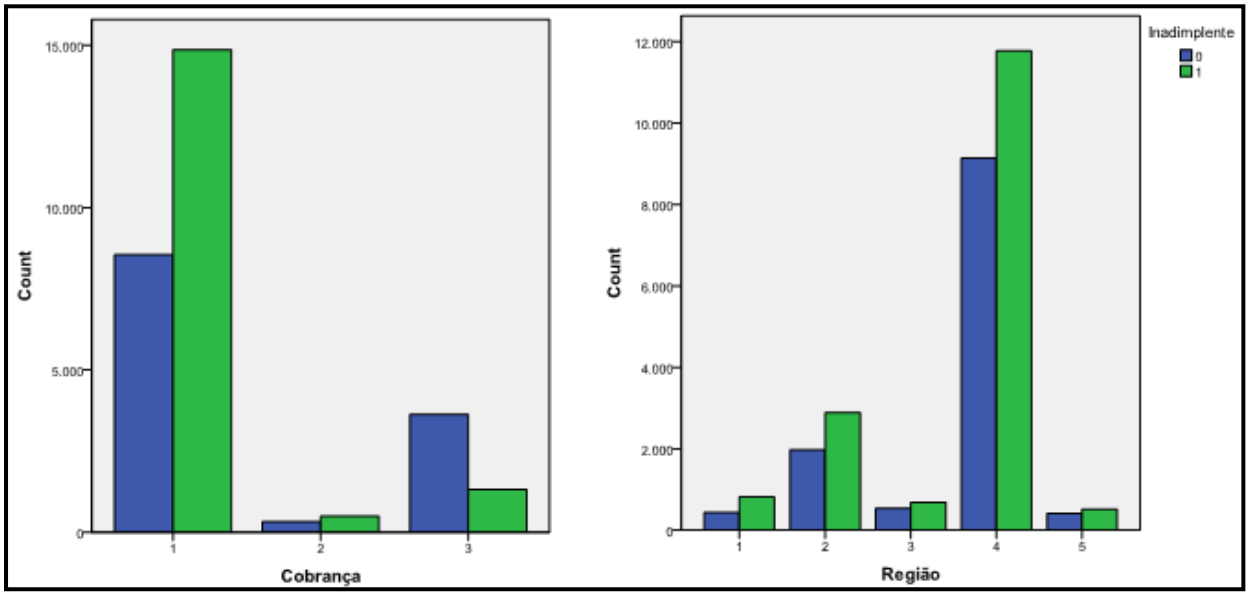

Figura 7 – Gráficos de barras para as variáveis “Tipo de Cobrança” e “Região”.

Com base na Figura 7, é possível observar que a grande parte dos doadores se concentra no tipo de cobrança por boleto e que é grande a frequência de inadimplentes neste grupo, o que se opõe quando observado os doadores com tipo de cobrança em débito, no qual a grande maioria é adimplente. Além disso, a maior parte dos doadores está na Região Sudeste do Brasil e, em seguida, na Região Nordeste. Para as 5 regiões a frequência de doadores inadimplentes se mostrou maior, quando comparada aos adimplentes; porém, para as duas supracitadas regiões com maior quantidade de doadores, a diferença entre as quantidades de adimplentes e inadimplentes aumenta.

Na Figura 8, é possível observar que os doadores estão bem distribuídos entre os vencimentos no início e final do mês. Porém, a frequência de inadimplência muda, pois no dia 5 a frequência de inadimplência é menor e no dia 30 foi consideravelmente maior, quando comparado a frequência dos doadores adimplentes.

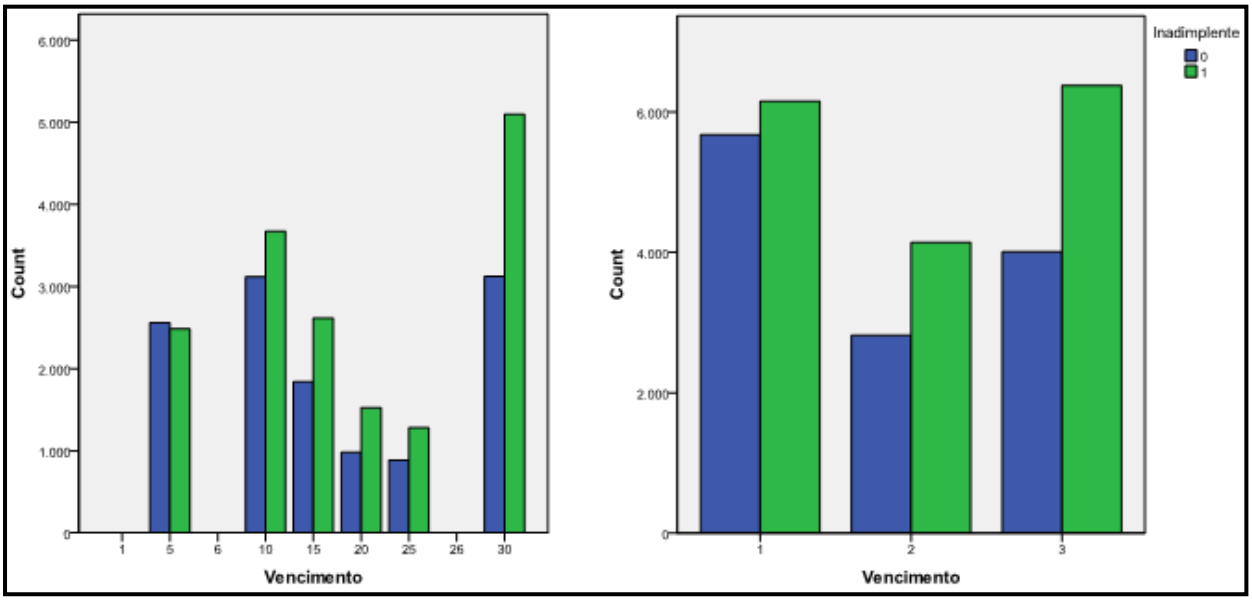

Figura 8 – Gráficos de barras para a variável "Vencimento"
(com diferentes categorizações).

4.3 Modelagem

Ao longo desta subseção, serão apresentados o ciclo percorrido da modelagem, as interpretações dos resultados e a análise do poder preditivo com validações dentro e fora da amostra.

Para cada variável explicativa categórica, dentre todas aquelas já pormenorizadas na Tabela 1, foram criadas variáveis do tipo *dummy*. Na Tabela, 2 estão indicadas as suas categorias de referência – ou seja: as categorias para as quais não foram criadas variáveis dummy, e com relação às quais os efeitos das demais categorias (dentro de um variável categórica específica) devem ser comparados (vide subseção 3.4).

Tabela 2: Categorias de referência para as variáveis do tipo dummy.

Variável Explicativa Categórica	**Categoria de Referência**
Estado Civil	Viúvo
Sexo	Masculino
Região	Norte
Classificação	Bronze 1
Cobrança	Boleto

4.3.1. Modelo com todas as variáveis

Na Tabela 3, estão os resultados consolidados da estimação do modelo de regressão logístico com todas as variáveis explicativas (Modelo 1).

Tabela 3: Modelo com todas as variáveis.

Variáveis	Estimativa do Coeficiente	Erro Padrão	Estatística t[1]	P-Valor	Razão de Chance
Idade	-,019	,001	521,573	,000	,981
Estado.Civil(Casado)	,116	,057	4,220	,040	1,123
Estado.Civil (Divorciado)	,258	,076	11,572	,001	1,294
Estado.Civil (Solteiro)	,148	,064	5,319	,021	1,159
Sexo	-,101	,026	14,875	,000	,904
Região (Norte)	-,302	,071	18,275	,000	,739
Região (Centro-Oeste)	-,356	,088	16,284	,000	,700
Região (Sudeste)	-,337	,065	26,644	,000	,714
Região (Sul)	-,375	,095	15,656	,000	,687
Classificação (Bronze 2)	-,020	,075	,072	,788	,980
Classificação (Bronze 3)	-,502	,103	23,862	,000	,605
Classificação (Prata 1)	-,469	,452	1,078	,299	,625
Classificação (Prata 2 ou mais)	-1,289	,624	4,264	,039	,276
Cobrança (Crédito)	-,142	,074	3,648	,056	,868
Cobrança (Débito)	-1,638	,036	2080,789	,000	,194
Vencimento	,015	,001	132,210	,000	1,015
Constante	1,538	,124	152,889	,000	4,656

[1] Estatística de Teste t estimada para o teste de significância individual.

É possível observar que os teste de significância individual, considerando um nível de significância de 5%, sugerem que as variáveis Classificação Bronze 2, Prata 1 e Cobrança (Crédito) não são significativas para este modelo.

4.3.2. Modelo com termos polinomiais

Na Tabela 4, estão os resultados consolidados da estimação do modelo de regressão logístico com todas as variáveis explicativas mais as variáveis com efeitos polinomiais (Modelo 2).

Tabela 4: Modelo com todas as variáveis e efeitos polinomiais.

Variáveis	Estimativa do Coeficiente	Erro Padrão	Estatística t[1]	P-Valor	Razão de Chance
Idade	,077	,009	75,294	,000	1,080
Idade²	-,002	,000	144,860	,000	,998
Idade³	,00001	,000	164,037	,000	1,000
Estado.Civil (Casado)	-,042	,058	,520	,471	,959
Estado.Civil (Divorciado)	,135	,077	3,076	,079	1,145
Estado.Civil (Solteiro)	,019	,065	,082	,775	1,019
Sexo	-,110	,026	17,497	,000	,895
Região (Norte)	-,294	,071	17,176	,000	,746
Região (Centro-Oeste)	-,350	,089	15,565	,000	,705
Região (Sudeste)	-,320	,065	23,969	,000	,726
Região (Sul)	-,374	,095	15,458	,000	,688
Classificação (Bronze 2)	-,241	,079	9,197	,002	,786
Classificação (Bronze 3)	-,726	,106	46,531	,000	,484
Classificação (Prata 1)	-,712	,452	2,487	,115	,491
Classificação (Prata 2 ou mais)	-1,545	,626	6,094	,014	,213
Cobrança (Crédito)	-,142	,075	3,623	,057	,867
Cobrança (Débito)	-1,655	,036	2098,699	,000	,191
Vencimento	,083	,026	9,719	,002	1,086
Vencimento²	-,004	,002	5,509	,019	,996
Vencimento³	,00007	,000	4,748	,029	1,000
Constante	,144	,204	,501	,479	1,155

[1]Estatística de Teste t estimada para o teste de significância individual.

É possível observar que os testes de significância individual, considerando um nível de significância de 5%, inicialmente indicam que as variáveis Estado Civil Casado,

Divorciado e Solteiro, Classificação Prata 1, Cobrança(Crédito) e a Constante não são significativas para este segundo modelo.

Para acessar a significância dos termos polinomiais adicionais, foi feito o teste da razão de verossimilhança, com o objetivo de comprar os Modelos 1 e 2 estimados.

Tabela 5: Comparação entre modelos.

	Estatística LR	**P-valor**	**AIC**
Modelo 1	-	-	36535,93
Modelo 2	180,25	$2{,}2^{-16}$	36363,68

Na Tabela 5, é possível observar a estatística do teste da razão de verossimilhança e o p-valor deste teste, o que leva a concluir que o modelo completo com todas as variáveis explicativas e os efeitos polinomiais (Modelo 2) é o mais adequado. O critério de informação de Akaike confirma essa conclusão, pois o valor da estatística para o Modelo 2 é menor.

Na Tabela 6, é feita a comparação do modelo selecionado até aqui, que é o Modelo 2, com outros modelos sem as variáveis que não foram significativas de acordo com os testes de significância individuais. O Modelo 3 se refere ao modelo com todas variáveis explicativas do Modelo 2, retirando somente a variável Classificação Prata 1. O Modelo 4 se refere ao modelo com todas as variáveis explicativas do Modelo 2, retirando somente as variáveis de Estado Civil Casado, Divorciado e Solteiro. O modelo 5 se refere ao modelo com todas as variáveis explicativas do Modelo 2, retirando somente a variável Cobrança Crédito. Já o Modelo 6 se refere ao modelo com todas as variáveis explicativas do Modelo 2, retirando as 5 variáveis (Classificação Prata 1, Estado Civil Casado, Divorciado e Viúvo, e a variável Cobrança (Crédito)) que não foram significativas no teste de significância individual.

Tabela 6: Comparação entre modelos.

	Estatística LR	**P-valor**	**AIC**
Modelo 2[1]	-	-	36363,68
Modelo 3	2,503	0,1136	36364,18
Modelo 4	11,513	0,0092	36369,19
Modelo 5	3.5904	0.05811	36365.27
Modelo 6	17.9	0.003075	36371.58

[1] Modelo completo para os testes da razão de verossimilhança para os demais modelos (vide subseção 3.5.2).

Ao observar a Tabela 6, é possível concluir, com base nos testes da razão de verossimilhança e nos critérios de informação de Akaike, que o modelo mais adequado é o modelo completo (Modelo 2). Apesar de o teste da razão de verossimilhança, que compara o Modelo 3 e Modelo 5 com o Modelo 2 , indicar, sob um nível de significância de 5%, que o modelo reduzido (Modelo 3) é tão adequado quanto o modelo completo (Modelo2), o critério de informação de Akaike aponta o Modelo 2 como sendo o modelo mais adequado. Portanto, o modelo selecionado para este TCC é o Modelo 2 com todas as variáveis explicativas e os efeitos polinomiais.

Os resultados para o modelo selecionado, observados na Tabela 4, propiciam as seguintes interpretações de razão de chance (cf. seção 3.4) para as variáveis categóricas:

- Quando comparado a doadores viúvos, é estimado para os doadores divorciados um acréscimo de 14% na chance de ser inadimplente, e um acréscimo de 2% para doadores solteiros. Já para os doadores casados, é estimado um decréscimo de 4% na chance de ser inadimplente.
- Quando comparado aos homens, é estimado para as mulheres um decréscimo de 10% na chance se ser inadimplente.
- Quando comparado aos doadores residentes da Região Norte, é estimado para os doadores das demais regiões um decréscimo na chance de ser inadimplente, destacando-se a Região Sul com o decréscimo de 31%.
- Quando comparado aos doadores de classificação Bronze 1, é estimado para os doadores das demais classificações um decréscimo na chance de ser inadimplente, destacando – se os doadores de Classificação Prata 2, Ouro 1 ou Ouro 3, com o decréscimo de 78%.

- Quando comparado aos doadores com o tipo de cobrança em forma de boleto, é estimado para os doadores com o demais tipo de cobrança um decréscimo na chance de ser inadimplente, destacando-se o tipo de cobrança em débito com um decréscimo de 80%.

Para as variáveis numéricas “Idade” e “Vencimento”, foram feitos gráficos das razões de chance conforme pode ser observado na Figura 9.

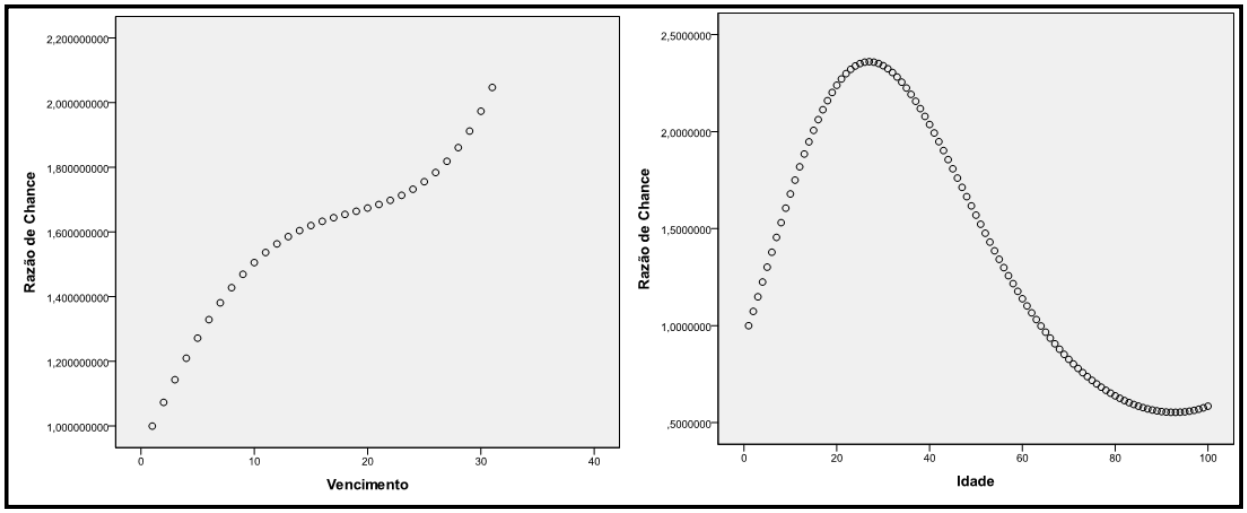

Figura 9 – Gráficos de dispersão para as Razões de Chance das variáveis “Vencimento” e “Idade”.

Eis algumas interpretações que complementam a informação da Figura 9:

- Quando comparado aos doadores com o Vencimento para o dia 5, é estimado para os doadores com Vencimento para o dia 30 um acréscimo de 55% na chance de ser inadimplente.
- Quando comparado aos doadores com Idade de 65 anos, é estimado para os doadores com Idade de 27 anos um acréscimo de 144,22% na chance de ser inadimplente.

Na Tabela 7, com informações sobre predição (vide subseção 3.6), é possível observar a uma taxa de 66,4% de acertos globais dos casos dentro da amostra e uma excelente taxa de 83,1% de acerto quando os doadores são inadimplentes.

Tabela7: Validação dentro da amostra.[1]

Observados		Estimados		
		Inadimplente		**Porcentagem de Acerto**
		0	1	
Inadimplente	0	5506	6995	44,0
	1	2812	13860	83,1
Porcentagem de Acerto Global				**66,4**

[1]Estimação usando o limiar 0,510 (limiar ótimo para este modelo – vide subseção 3.6).

Para validar esses resultados, foi feita uma validação fora na amostra com dados previamente retirados do banco de dados para tal fim (vide explicações na subseção 4.1). Conforme observa-se na Tabela 8, todas as taxa praticamente mantiveram-se nos mesmos patamares.

Tabela 8: Validação fora da amostra.[1]

Observados		Estimados		
		Inadimplente		**Porcentagem de Acerto**
		0	1	
Inadimplente	0	632	759	45,4
	1	328	1523	82,2
Porcentagem de Acerto Global				**66,5**

[1]Estimação usando o limiar 0,510 (o mesmo limiar ótimo previamente encontrado).

4.3.3. Modelo com efeitos interativos

Com o objetivo de melhorar o poder preditivo dos inadimplentes, foram estimados alguns modelos usando as interações entre as seguintes variáveis: idade e estado civil; idade e sexo; idade e vencimento; e estado civil e sexo. Os testes da razão de verossimilhança indicaram que o modelo com efeitos interativos é mais adequado. Porém, ao observarem-se os resultados de predição, não se notou uma melhora significativa; além do mais, modelos de regressão que abrangem variáveis com efeitos interativos em suas especificações são mais difíceis de serem interpretados. Portanto, o modelo escolhido continua sendo o Modelo 2, com todas as variáveis explicativas e com termos polinomiais, sem os efeitos interativos.

5. Conclusão

O estudo de caso visou entender e otimizar (reduzir) as taxas de inadimplência de pessoas físicas quanto às suas comprometidas doações para instituições filantrópicas. Este envolveu um conjunto de dados, da Junta de Missões Nacionais da Convenção Batista Brasileira, referentes a pessoas físicas doadoras e a forma de parceria estabelecida no instante que o doador adota um projeto da instituição e começa a contribuir. As variáveis analisadas que se mostraram como as principais determinantes da inadimplência foram: idade, estado civil, sexo, região, classificação (do tipo de doador), cobrança e vencimento. Além dessas, foram introduzidos alguns efeitos polinomiais com as variáveis "idade" e "vencimento". Os testes de significância individuais para cada uma dessas variáveis, considerando um nível de significância de 5%, inicialmente indicarem que algumas delas não seriam significativas para o modelo; porém, os testes de comparabilidade de modelos (teste da razão de verossimilhança) e os resultados do critério de informação de Akaike indicaram que o melhor modelo seria com todas as variáveis, mais os efeitos polinomiais. Para o modelo selecionado, foi feita a validação dentro e fora da amostra, levando a taxas de acerto de predição capaz de excelente antecipação dos casos de inadimplência, os quais são os mais importantes de serem preditos neste TCC.

Diante dos resultados obtidos, espera-se que sejam desenvolvidas ações específicas para o perfil de doadores inadimplentes que podem ser caracterizados por doadores jovens, do sexo masculino, solteiros, com doações de pequenos valores, pagando em forma de boleto, na Região Norte, com a data de vencimento para o final do mês; já os adimplentes podem ser caracterizados por doadores com mais de 65 anos, do sexo feminino, sendo viúva ou casada, com doações de maiores valores, pagando em forma de débito, residente na Região Sul e com o vencimento para o inicio do mês. Essas ações devem ser desenvolvidas com o objetivo de conscientizar os doadores inadimplentes da importância de tê-los como parceiros nos projetos da empresa e o que se poderia fazer com a soma dessas doações não recebidas. Com o perfil dos doadores adimplentes, pode ser feita uma pesquisa exaustiva sobre onde está este público e desenvolver ações específicas para ter mais doadores neste perfil. As predições futuras serão fundamentais para o planejamento orçamentário da empresa, pois através delas pode-se estimar o montante de doações com as quais a empresa poderá contar.

Estudos futuros com o objetivo de melhorar os resultados podem ser feitos, tais quais a estimação de modelos lineares generalizados baseados nas distribuições Poisson e/ou

Binomial Negativa) para contagem de meses de inadimplência de um doador. Desta forma, estaria sendo usada uma variável resposta mais completa, pois a variável resposta usada neste TCC é uma transformação não-biunívoca da contagem de meses de inadimplência, e este tipo de categorização implica inevitável perda de informação. Com um novo modelo baseado na contagem de meses de inadimplência, é possível fazer predição de inadimplência também com possível inclusão de efeitos interativos e buscar limiares ótimos. A predição, desta vez, pode melhorar, pois seria consequência de uma modelagem com mais informações na variável resposta.

Outra tarefa futura, tanto com um modelo de regressão logística quanto com um modelo para contagens de meses, é aprofundar a busca por efeitos interativos, testando efeitos com três variáveis, com o intuito de melhorar a predição dos adimplentes. Outra extensão certamente de interesse é, com o melhor modelo, usar algum procedimento de otimização para minimizar a probabilidade de inadimplência. Com isso, seriam encontrados os patamares ótimos das variáveis explicativas e, assim, o perfil ótimo para os doadores de instituições filantrópicas no Brasil.

Referências

AMARAL, O. Á. *Entrevista pessoal e por correspondência por David Allen Bledsoe entre 25 de setembro e 03 de outubro.* Rio de Janeiro, RJ, 2013

BESSIS, J. *Risk management in banking.* Chichester: John Wiley & Sons, 1998.

BLATT, A. *Avaliação de Riscos e Decisões de Crédito: um enfoque prático*. Nobel, 1999.

BLEDSOE, D. A. *Cooperação e Conexão Missionária dos Batistas Brasileiros,* 1ª edição. Convicção, 2014.

BRANDÃO, F. "Palavra do Diretor*". A Pátria Para Cristo, nº 268,* 2015

CAOUETTE, J. B. et al. *Gestão do risco de crédito: o próximo grande desafio financeiro*. Qualitymark, 1999.

CORDEIRO, G. *Modelos Lineares Generalizados*. Campinas: VII Simpósio Nacional de Probabilidade e Estatística, 1986.
DEGROOT, M. H. *Probability and Statistics*. 2ªedicão. Addison Wesley, 1986.

DOBSON, A. J. *An Introduction to Generalized Linear Models*. 2ª edição. Chapman & Hall/CRC, 1990.

HARDIN, J. W. e HILBE, J. M. *Generalized Linear Models and Extensions*. 2ª edição. Stata Press, 2007.

HOSMER, D. W. e LEMESHOW, S. *Applied Logistic Regression*. John Wiley & Sons, 1989.

JOHNSON & WICHERN, *Applied Multivariate Statistical Analysis,* 6ª edição, Prentice Hall, 2007.

KLEINBAUM, D. G. e KLEIN, M. *Logistic Regression: A Self-Learning Text*. 3ª edição. Springer, 2010.

MESQUITA, M. *Missionários & Recursos.* Betânia, 2011.

McCULLAGH, P. e NELDER, J. *Generalized Linear Models*. 2ª edição. Chapman and Hall, 1989.

NETER, J., KUTNER, M., NACHTSHEIM, C. e WASSERMAN, W. *Applied Linear Statistical Models*. 4ªedição. Times Mirror Higher Education Group, 1996.

OLIVEIRA, Z. M. *100 anos da Junta de Missões Nacionais da CBB*, 1ª edição. Convenção Batista Brasileira, 2007.

SANTOS, M. H. L. "Duas Vidas Transformadas, Um Lar Edificado, Jesus Glorificado!*". A Pátria Para Cristo,* nº 268, 2015

SOUZA, S. O. Entrevista em áudio por David Allen Bledsoe em 17 de setembro. Rio de Janeiro, RJ, 2013.