UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA

MODELAGEM DO VALOR VITALÍCIO DO CLIENTE VIA ABORDAGEM DE ANÁLISE DE SOBREVIVÊNCIA

MAXIMILIANO LUIS CHIAMULERA

Porto Alegre
2017

MAXIMILIANO LUIS CHIAMULERA

MODELAGEM DO VALOR VITALÍCIO DO CLIENTE VIA ABORDAGEM DE ANÁLISE DE SOBREVIVÊNCIA

Trabalho de Conclusão de Curso apresentado para obtenção do grau de Bacharel em Estatística

Orientador Metodológico
Doutor Danilo Marcondes Filho

Porto Alegre
2017

CIP - CATALOGAÇÃO NA PUBLICAÇÃO

Chiamulera, Maximiliano Luis

Modelagem do Valor Vitalício do Cliente via abordagem de Análise de Sobrevivência / Maximiliano Luis Chiamulera. -- 2017.

46 f.

Orientador: Danilo Marcondes Filho.

Trabalho de conclusão de curso (Graduação) -- Universidade Federal do Rio Grande do Sul, Instituto de Matemática, Curso de Estatística, Porto Alegre, BR - RS, 2017.

1. Valor vitalício cliente. 2. Regressão Cox. 3. Análise sobrevivência. 4. Customer lifetime value. I. Marcondes Filho, Danilo, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Instituto de Matemática e Estatística
Departamento de Estatística

Modelagem do Valor Vitalício do Cliente via abordagem de Análise de Sobrevivência

Maximiliano Luis Chiamulera

Banca examinadora:

Doutor Danilo Marcondes Filho (orientador)
Universidade Federal do Rio Grande do Sul

Doutor Rodrigo Citton Padilha dos Reis
Universidade Federal do Rio Grande do Sul

AGRADECIMENTOS

Depois de uma gigante jornada, enfim meus tempos no curso de Estatística se findam e nesse momento impar eu não poderia deixar de agradecer primeiramente a minha esposa Daiana, por ter me incentivado tanto a retomar o curso depois de ter evadido em 2010 e por toda a sua dedicação e compreensão nesses últimos meses.

Não posso deixar de agradecer ao meu irmão por ter sido o que ele é desde o dia que saiu meu listão no longínquo 19/01/2002.

A minha dinda Fernanda por ter sido crucial no começo da minha jornada acadêmica e por ter me acolhido como um filho.

Ao meu orientar professor Danilo, pelo seu esforço em me fazer compreender aspectos das análises e pelas diversas e extensas reuniões realizadas remotamente.

Ao Pedro e a Pmweb, por acreditar nos números e na estatística e ter compreendido os insanos horários da UFRGS ainda mais nas últimas semanas de TCC.

A todos os chefes que tive durante a minha vida profissional, Sabine, Carine, Graziela, Jader, Erbene, por terem contribuído de forma imensa no meu aprendizado e crescimento profissional, por terem sempre me instigado a construir análises mais elaboradas, que fugissem do esperado.

Aos queridos professores que tive o prazer de ter tido aula neste meu retorno a faculdade, Cleber, Pumi, Vigo, Gabriela, Vanessa, Patricia, Stela, Lisi, Liane, Hudson entre outros, por terem feito aulas de estatística mais atuais e atrativas, sempre se esforçando para que eu conseguisse entender algum axioma ou demonstração, à vocês meu muito obrigado.

DEDICATÓRIA

Dedico este trabalho sobre sobrevivência a minha querida mãe Núbia, que sempre se esforçou ao máximo para dar uma educação exemplar para mim e meu irmão, que não media esforços para que sempre tivéssemos tudo que fosse necessário para estudar e que sempre colocou nossos estudos acima de tudo. Infelizmente depois de tantos esforços, o listão foi uma alegria grande demais e minha mãe nunca pode me ver na UFRGS. Mãe obrigado por ter feito tudo por nós.

“Les statistiques sont une forme d'accomplissement de désir, tout comme les rêves.”

(Jean Baudrillard)

RESUMO

Umas das dificuldades na alocação dos recursos de marketing é definir o quanto será investido na aquisição e manutenção de cada cliente. Os investimentos representativos devem depender da capacidade da empresa de prever lucros futuros e os custos de uma classificação errada dos clientes pode ter resultados críticos. Fazendo uso do Valor Vitalício do Cliente (CLV, sigla em inglês de Customer Life Time), que pode ser definido como a diferença entre as aquisições que o cliente fará e o custo para gerar estas aquisições, pode-se diminuir as incertezas na hora de realizar os investimentos de Marketing. Para estimar o CLV é comum que além das estimativas das receitas e custos futuros, se estime o tempo de relacionamento do cliente. Esse contexto torna a análise de sobrevivência uma ferramenta importante para realizar tais estimativas. A partir de dados cadastrais e transacionais disponíveis, oriundos de uma empresa de venda de calçados e roupas, este trabalho se propõe a modelar o tempo de relacionamento dos clientes da empresa a partir da regressão de Cox ajustada aos dados. A partir do modelo ajustado obtemos uma estimativa do CLV dos clientes da empresa. Após modelar a sobrevivência e calcular o CLV dividiu-se os clientes em 80-20 com base no CLV para que então fosse estimada a assertividade do modelo, o valor comparativo foi o valor real gasto pelos clientes em um recorte temporal futuro, essa metodologia mostrou-se satisfatória, conseguindo prever corretamente a classificação 20-80 de 85% dos clientes.

Palavras-chave: Valor vitalício cliente. Regressão Cox. Análise sobrevivência.

ABSTRACT

One of the difficulties in allocating marketing resources is to define how much will be invested in the acquisition and maintenance of each customer. The representative investments must depend on the company's ability to predict future profits and the costs of misclassification of customers may have critical results. Making use of the Customer's Lifetime Value (CLV), which can be defined as the difference between the purchases that the customer will make and the cost to generate these acquisitions, it could reduce the uncertainties when making Marketing investments. In order to estimate the CLV, it is common that besides the estimation of future revenues and costs we also estimate the client's relationship time. This context makes survival analysis an important tool to make such estimates. Based on available cadastral and transactional data from a shoe and clothing sales company, this paper proposes to model the relationship time of the company's clients based on the Cox regression adjusted to the data. From the adjusted model we obtain an estimate of the CLV of the company's clients. After modeling the survival and calculating the CLV, the clients were divided into 80-20 based on the CLV so that the assertiveness of the model was then estimated, the comparative value was the actual value spent by the clients in a future time cut, this methodology showed satisfactory, being able to correctly predict the 80-20 rating of 85% of customers.

Keywords: Customer lifetime value. Cox regression. Survival analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

A dinâmica de mercado atualmente tem obrigado as empresas a trabalharem com margens de lucros cada vez mais estreitas. Neste contexto fica cada vez mais importante o conceito de Valor Vitalício do Cliente ou CLV, do termo em inglês *Customer Lifetime Value*, que pode ser definido, sem perder generalidade, como a diferença entre o valor de todas as aquisições realizadas e o custo para gerar todas estas aquisições (DWYER, 1997).

Umas das complexas tarefas na distribuição dos recursos de marketing é definir o quanto será investido na aquisição e manutenção de cada cliente, como o comportamento de consumo dos diferentes grupos de clientes é distinto, assim como o tempo de relacionamento que eles terão, em média, com a empresa (KOTLER e ARMSTRONG, 1996). Investimentos representativos devem depender da capacidade da empresa de prever lucros futuros, e os custos de uma classificação errada dos clientes tem diferentes níveis de relevância para diferentes tipos de investimentos em marketing (MALTHOUSE e BLATTBERG, 2005). A importância de entender o perfil de consumo dos diferentes grupos de clientes, aliado ao crescente volume de dados, faz inclusive que as empresas, muitas vezes, terceirizem a execução de seu CRM e DBM (do termo em inglês *data base management*, ou gestão de banco de dados).

Neste sentido, estimar o *Customer Lifetime Value* (CLV) se torna útil para que as empresas possam decidir quem são e quais as características de seus clientes e quais deles terão um relacionamento duradouro ou ainda decidir a alocação dos recursos de Marketing nos clientes certos (JAIN e SINGH, 2002). Podemos encontrar na literatura uma série de proposições para estimar o CLV a partir de uma ampla variedade de abordagens, incluindo a utilização da Análise de Sobrevivência e mais especificamente a Regressão de Cox.

O presente trabalho apresenta uma abordagem para modelagem do CLV de uma empresa que vende roupas e sapatos em lojas físicas ou através de seu site de vendas utilizando a Análise de Sobrevivência. Mais especificamente, num passo inicial, a partir de um modelo de Cox ajustado ao banco de dados, serão identificadas as variáveis associadas ao tempo de vida dos clientes. A seguir, a curva de sobrevida dos clientes do banco será obtida e o CLV será então estimado. Utilizando um recorte

temporal no banco de dados, a acurácia do CLV obtidos será avaliada para diferentes intervalos de tempo para frente.

2. OBJETIVOS

Objetivo Geral

- Verificar as variáveis transacionais e de cadastro que estão associadas ao tempo de permanência do cliente na empresa, por meio da regressão de Cox

Objetivo específico

- Calcular o Valor Vitalício dos Clientes (CLV) da empresa utilizando a função de sobrevivência estimada via modelo de Cox

3. REVISÃO DA LITERATURA

3.1. Customer Relationship Management - CRM

O gerenciamento do relacionamento do cliente ou *customer relationship management* é o processo de atrair e reter clientes rentáveis, construindo relações de longo prazo por meio da entrega de valor e satisfação aos clientes (KOTLER e KELLER, 2010). Clientes rentáveis são aqueles que ao longo do seu relacionamento com a empresa tem um saldo positivo entre o valor investido pela empresa no cliente e o valor que o cliente investiu na empresa.

Os modelos preditivos de cancelamento podem ser divididos em estáticos e dinâmicos (POEL e LARIVIÈRE, 2004), sendo que os modelos dinâmicos devem ser preferidos pois acompanham o comportamento do cliente ao longo do tempo ao passo que os estáticos apenas aferem os valores do cliente em um tempo específico, além de produzirem estimativas mais precisas.

Diversos modelos dinâmicos são construídos, especialmente nas áreas de serviços financeiros e telecomunicações, a fim de estimar o risco de um cliente parar de comprar o produto ou ainda modelos para identificar o comportamento relativamente anterior ao ato do cancelamento para que medidas remediadoras possam ser disparadas em tempo de evitar esse cancelamento. Dentre as técnicas estatísticas comumente utilizadas para a construção destes modelos estão a regressão logística, a análise de discriminante e a análise de sobrevivência (POEL e LARIVIÈRE, 2004). Neste cenário a análise de sobrevivência tem certa vantagem pois além de predizer o cancelamento ou não do cliente, resultado também obtido com a construção das duas outras análises, ele permite estimar o tempo até que tal cancelamento venha ocorrer. Essa informação será de suma importância para calcular o CLV, que será abordado na seção seguinte.

3.2. Valor Vitalício do Cliente – CLV

O Valor Vitalício do Cliente pode ser definido, sem perder generalidade, como a diferença entre o valor de todas as aquisições realizadas e o custo para gerar todas estas aquisições (DWYER, 1997).

Jain e Singh em seu trabalho (JAIN e SINGH, 2002) nos apresentam algumas opções para o cálculo de CLV, como os modelos estruturais básicos (BERGER e NASR,

1998), os modelos baseados na migração de clientes (DWYER, 1997), modelos baseados na alocação ótima dos recursos (BLATTERBER e DEIGHTON, 1996) e os modelos baseados no relacionamento com o cliente, tendo as probabilidades de transição estimadas por cadeias de Markov (PFEIFER e CARRAWAY, 2000).

Para do cálculo do CLV é preciso modelar a margem de contribuição financeira gerada pelos clientes, estimar a probabilidade de permanecer na carteira da empresa até um determinado tempo e determinar a taxa de desconto apropriada pela empresa para investimentos em marketing (FERREIRA, 2007). Baseado nos conceitos propostos por (BLATTBERG, *et al.*, 2001), Ferreira propõe para estimar o CLV de um cliente *i* até o tempo *T*, a fórmula:

$$CLV_i(T) = \sum_{t=0}^{T} S_i(t) M_i(t)(1+R)^{-t} \qquad (1)$$

onde temos:

$S_i(t)$: função de sobrevivência associada,

$M_i(t)$: modelo linear hierárquico utilizado para a previsão da margem e

R: taxa de desconto aplicada.

Na fórmula 1 $S_i(t)$ estima a probabilidade de um clientes *i* não deixar de comprar até um tempo t, ou a grosso modo ter um relacionamento sobrevivente até o tempo *t*, ao passo que $M_i(t)$ estima o valor financeiro que o cliente *i* trará para empresa até o tempo *t*, e já *R* é a taxa de desconto aplicada e deve levar em conta o custo de capital da empresa, as tendências gerais da economia e a taxa de inflação esperada e além disso deve relevar uma componente subjetiva de risco (PEPPERS e ROGERS, 2005).

Contudo os modelos de CLV também podem ser considerados sem a parcela dos custos decorrentes ao relacionamento, tornando a interpretação do CLV como o máximo valor do cliente (BERGER e NASR, 1998).

Neste trabalho o termo $S_i(t)$ será calculado a partir da Análise de Sobrevivência através de uma Regressão de Cox ajustada de um banco de dados com variáveis transacionais e cadastrais. O termo $M_i(t)$ será simplificado, sendo substituído pelo valor médio gasto por dia pelo cliente e o termo *R* será desconsiderado.

3.3. Modelos de Análise de Sobrevivência

A modelagem da sobrevivência apresenta uma série de métodos que buscam descrever o tempo até a ocorrência do evento de interesse, o que neste trabalho representa o tempo até o cliente deixar de realizar compras na empresa (ocorrência do evento de "morte"). Esta modelagem pode ser feita considerando a influência de uma série de variáveis preditoras através da utilização de modelos de regressão paramétricos (necessitamos assumir uma distribuição subjacente para o tempo de vida do indivíduo) ou semi-paramétricos. Adicionalmente, esta classe de métodos também permite o uso de informações dos clientes que ainda não tenham deixado de comprar e de clientes que sofreram o evento antes do término da janela temporal do estudo. Este cenário nos remete ao conceito de dados censurados. Os dados censurados podem ser à esquerda do período de estudo, quando não se conhece a data de início do relacionamento com a empresa, fato que não se aplica ao presente estudo; a direita do período de estudo, quando o cliente deixa de comprar após o período de estudo, e assim é contabilizado na análise apenas o seu comportamento até o término da janela de estudo, não considerando o tempo após essa janela até o evento; ou ainda a censura intervalar quando o cliente começa o relacionamento durante o período de estudo e deixa de comprar ainda dentro do período (CARVALHO, *et al.*, 2011). Adicionalmente, os métodos de análise de sobrevivência permitem a inclusão de indivíduos que já estavam sobre o risco de sofrer o evento antes do início da janela temporal do estudo. Neste caso, utilizamos a ideia do truncamento, mais especificamente trabalhamos com o truncamento à esquerda. Assim, consideramos apenas as informações disponíveis após o início do estudo e descartamos os dados disponíveis nos tempos anteriores à da janela temporal do estudo. O truncamento à direita é utilizado quando definimos a janela temporal a partir do evento de morte, ou a partir da data de abandono do cliente da carteira ou do fim do contrato (dados prevalentes), o que não é o caso neste trabalho. No presente estudo teremos casos em que o truncamento a esquerda será necessário. Os conceitos de censura e truncamento são detalhadamente discutidos por Hosmer, Lemeshow e May (HOSMER, *et al.*, 2008), (CARVALHO, *et al.*, 2011), entre outros.

A análise de sobrevivência trata de estimar de maneira probabilística o tempo até o indivíduo sofrer o evento de morte, assim sendo o tempo até o cliente deixar de

comprar é uma variável aleatória que segue uma distribuição de probabilidades que usualmente não é conhecida. A distribuição da variável *T*, com *T* sendo o tempo até o cancelamento do cliente é expressa pela função de densidade de probabilidade $f(t)$:

$$f(t) = lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t} \quad (2)$$

Com base nesta $f(t)$ é possível calcular a função de sobrevivência e de risco. A função de sobrevivência nos entrega a probabilidade de um cliente não deixar de comprar antes do tempo t e é expressa por:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^{+\infty} f(u)du \quad (3)$$

A curva de sobrevivência teoricamente seria uma curva suave, mas na prática, sua estimativa, é um gráfico de escada (*stepfunction*), onde os degraus são os momentos onde os clientes deixam de compra (KLEINBAUM e KLEIN, 2005).

A função de risco, igualmente importante, indica o risco instantâneo do cliente deixar de comprar no intervalo de tempo $(t, t + \Delta t)$, dados que não deixou de comprar até o tempo *t* e pode ser escrita como:

$$\lambda(t) = lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)} \quad (4)$$

A função de sobrevivência $S(t)$ e a função risco $\lambda(t)$ podem ser estimadas com abordagens univariadas como as tabelas de vida, que por sua vez fazem uso de estimadores como: atuarial, Kaplan-Meier ou Nelson-Aalen ou de abordagens multivariadas, que fazem a avaliação do impacto simultâneo das diversas variáveis, nesse contexto é usual a utilização do modelo semi-paramétrico de riscos proporcionais de Cox.

3.3.1. Modelos descritivos para estimar sobrevivência

Três métodos descritivos são comumente utilizados para estimar a sobrevivência, são eles: Kaplan-Meier, tábuas de vida e Nelson-Aalen.

Conhecido como estimador limite produto, o método de Kaplan-Meier faz uso de todas as informações disponíveis (com ou sem censura), com os dados observados o método estima a probabilidade condicional de sobrevivência para cada período de observação no tempo para então multiplicar as estimativas e chegar a

função gera de sobrevivência (HOSMER, *et al.*, 2008). Uma limitação natural desse método é a necessidade da categorização de variáveis quantitativas (KLEINBAUM e KLEIN, 2005).

Casos o número de dados disponíveis seja suficientemente, as tábuas de vida são uma alternativa ao método Kaplan-Meier e como esse método estima primeiramente a função de sobrevivência. Diferente destes dois métodos, Nelson-Aalen estima primeiro a função risco acumulado para então estimar a função de sobrevivência acumulada, estimando assim a função de sobrevivência.

As tábuas de vida permitem que o analista escolha o intervalo de tempo já Kaplan-Meier e Nelson-Aalen apresentam os tempos onde os eventos de interesse ocorrem. Quanto ao tamanho de amostra, as tábuas de vida necessitam de um tamanho maior para produzir estimativas não viesadas da função de sobrevivência. Se a amostra for pequena o estimador Kaplan-Meier é o que estima melhor a função de sobrevivência.

Embora sejam mais simples de se implementar, esses três métodos apenas permitem uma variável como sendo preditora e necessitam que as variáveis quantitativas sejam categorizadas. Com essas limitações a construção de modelos mais complexos, quer sejam eles paramétricos ou semi-paramétricos, ganha força, uma vez que permitirá o uso de múltiplos preditores sem a necessidade de categorizar as variáveis quantitativas.

3.3.2. Modelos paramétricos e semi-paramétricos

Modelos paramétricos permitem caracterizar as mudanças que variáveis associadas ao evento de interesse provocam na distribuição de sobrevivência no tempo (HOSMER, *et al.*, 2008). Porém modelos paramétricos necessitam que se saiba *a priori* a distribuição de probabilidades do tempo até ocorrer o evento de interesse. As funções densidade de probabilidade mais utilizadas são Exponencial, Weibull e Lognormal, pois são de relativamente fáceis de adaptar a diversos cenários (CARVALHO, *et al.*, 2011).

Em oposição aos modelos paramétricos, os modelos semi-paramétricos não exigem suposição a priori sobre a distribuição do tempo. Tais modelos permitem determinar como o comportamento das variáveis preditoras modifica a função de sobrevivência e de risco dos clientes observados. Os modelos semi-paramétricos são

utilizados como uma alternativa robusta quando se deseja estimar o impacto das variáveis preditoras na distribuição da sobrevivência ou do risco. O modelo de Cox é o modelo semi-paramétrico mais utilizado para a análise de sobrevivência.

Em 1972 Sir David Cox publicou um artigo onde propunha o modelo de riscos proporcionais, modelo este que ficou conhecido como modelo de Cox, que como o próprio nome diz tem como premissa que o risco do evento acontecer é proporcional ao longo de todo o tempo de observação (CARVALHO, *et al.*, 2011).

A construção do modelo de Cox deixa claro que a função de risco no tempo *t* é o produto entre o risco basal $h_0(t)$ e a exponencial do vetor de covariáveis *x*, conforme descrito a seguir:

$$h_i(t|x) = h_0(t)exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right) \quad (5)$$

Sabendo-se que um cliente *i* apresenta os valores $x_i = \left(x_{1i}, x_{2i}, \ldots, x_{pi}\right)'$ dado a coleção de vaiáveis preditoras $X_1, X_2, \ldots, X_p$ o modelo permite estimar o risco de ocorrência do evento de interesse no tempo *t* (KLEINBAUM e KLEIN, 2005).

A razão entre o risco de ocorrência do evento de dois clientes (*i* e *j*), conhecida como razão de riscos (do inglês *hazard ratio*), com os clientes apresentando os vetores $x_i = \left(x_{1i}, x_{2i}, \ldots, x_{pi}\right)'$ e $x_j = \left(x_{1j}, x_{2j}, \ldots, x_{pj}\right)'$, respectivamente é expressa como:

$$\frac{h_0(t)exp(\beta x_i)}{h_0(t)exp\left(\beta x_j\right)} = exp\left(\beta' x_i - \beta' x_j\right) = exp\left(\beta'\left(x_i - x_j\right)\right) \quad (6)$$

Da fórmula 6 podemos vislumbrar o efeito multiplicativo das variáveis preditoras na função risco o que é diretamente relacionada a suposição de riscos proporcionais que o modelo faz uso. Como o termo não paramétrico $h_0(t)$ acaba sendo cancelado, esta razão se torna constante ao longo do tempo, ou não seja não depende de t (CARVALHO, *et al.*, 2011).

Para estimar $exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right)$ podemos fazer uso de diferentes técnicas: máxima verossimilhança, verossimilhança parcial ou verossimilhança aproximada. Fazendo uso da verossimilhança parcial, temos a vantagem de estimar os coeficientes β_i sem ser preciso especificar o componente não paramétrico do modelo, a função de risco basal $h_0(t)$ (ALISSON, 2010). O risco basal em *t* [$h_0(t)$] é

compreendido como o risco de deixar de comprar comum a todos os clientes com relacionamento ativo no tempo *t*, independente dos valores que as covariáveis $x_1, x_2, \dots, x_p$ assumam neste tempo *t*.

A razão de riscos para o caso de variáveis indicadoras pode ser interpretada como a razão entre o risco estimado para aqueles clientes com variável indicadora=1 em relação àqueles clientes com variável indicadora=0, controlando para demais variáveis preditoras (ALISSON, 2010). Para as variáveis continuas a razão de riscos é interpretada como a razão entre o risco estimado por unidade da variável preditora (CARVALHO, *et al.*, 2011).

O modelo de Cox como descrito na fórmula 6 assume que todas as variáveis foram aferidas em um único momento, geralmente em *t*=0 (início do estudo), e perdurando o mesmo valor durante todo o período de observação. Porém, é desejável que se possa medir algumas variáveis preditoras diversas vezes durante o estudo (variáveis tempo-dependentes). Sob este aspecto a função de risco pode depender mais dos valores observados perto do final do período observacional do que dos valores observados no início do período (HOSMER, *et al.*, 2008).

Esta estrutura tempo-dependente é bem acomodada pelo modelo de Cox estendido. O modelo estendido apresenta uma formulação semelhante à da fórmula 6, com a adição de um termo novo $z(t)$ que irá incorporar as variáveis tempo-dependentes (KLEINBAUM e KLEIN, 2005):

$$h_i(t) = h_0(t) exp\left(\beta' x + \gamma' z(t)\right) \qquad (7)$$

As variáveis medidas repetidas vezes no mesmo cliente são o tipo mais comum de variáveis tempo-dependentes (CARVALHO, *et al.*, 2011).

Os valores dos estimadores $\hat{\beta}$ e $\hat{\gamma}$, obtidos por máxima verossimilhança parcial, vão avaliar a influência das covariáveis na estimativa da curva de risco. Quanto mais distante de 1 estiver o termo exponencial da fórmula 7, mais distante do risco basal estará a curva de risco $h_i(t)$, isto é, os indivíduos deverão ter curvas de risco distintas baseadas nos seus perfis de compra. A partir da estimativa da sobrevivência poderemos então calcular o CLV do cliente *i* no tempo *t* (fórmula 1).

3.4. Estrutura do banco para variáveis tempo-dependentes

Quando existem variáveis disponíveis para o estudo que podem assumir valores distintos em cada instante do tempo, varáveis tempo-dependentes, é comum que o banco de dados seja estruturado na forma de um processo de contagem. Um banco nessa forma irá apresentar diversas linhas, registros, para cada cliente.

No processo de contagem a divisão dos intervalos tempo se dá nos tempos em que são observadas cada uma das compras dos clientes, e nesse momento os valores de cada variável são registrados. Como cada cliente tem um comportamento diferente, os intervalos de tempo tem início e fim distintos, todavia o início do intervalo não precisa necessariamente ser zero, o que torna a incorporação de truncamento uma tarefa fácil (HOSMER, *et al.*, 2008). Já variáveis que são medidas apenas no começo do estudo, variáveis fixas, tem seus os valores repetidos em todas as linhas do cliente.

Após terminada a construção do banco de dados, sua estrutura permite a fácil validação de erros oriundos da programação e agregação dos dados (ALISSON, 2010). Para fins de ilustração, a Figura 1 exibe um pequeno recorte do banco de dados final.

Figura 1 - Exemplo do banco de dados

id_tcc	variavel_7	variavel_1	variavel_3	variavel_2	variavel_4	variavel_5	variavel_6	inicio	fim	variavel_8	variavel_11	variavel_9	variavel_10	produto_1	produto_2	produto_3	produto_4	produto_5	produto_6	produto_7	produto_8	produto_9	produto_10	produto_11	produto_12	produto_13	produto_14	produto_15	produto_16	produto_17	produto_18	produto_19	produto_20	produto_21
23921	43	positivo	negativo	negativo	negativo	categoria_1	categoria_4	0	693	0	0	categoria_2	negativo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23921	43	positivo	negativo	negativo	negativo	categoria_1	categoria_4	693	717	0	0	categoria_1	positivo	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
23921	43	positivo	negativo	negativo	negativo	categoria_1	categoria_4	717	731	0	0	categoria_1	positivo	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
27109	42	positivo	negativo	positivo	positivo	categoria_2	categoria_4	0	1	0	0	categoria_2	negativo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27109	42	positivo	negativo	positivo	positivo	categoria_2	categoria_4	1	207	0	0	categoria_2	positivo	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
27109	42	positivo	negativo	positivo	positivo	categoria_2	categoria_4	207	208	0	0	categoria_2	positivo	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
27109	42	positivo	negativo	positivo	positivo	categoria_2	categoria_4	208	305	0	1	categoria_1	positivo	1	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0
27109	42	positivo	negativo	positivo	positivo	categoria_2	categoria_4	305	581	0	0	categoria_2	positivo	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
30982	50	negativo	negativo	negativo	negativo	categoria_1	categoria_4	0	1	0	0	categoria_2	negativo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30982	50	negativo	negativo	negativo	negativo	categoria_1	categoria_4	1	637	0	0	categoria_1	positivo	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

O uso do método de verossimilhança parcial nos permite desconsiderar aqueles clientes que não estavam em risco em um determinado período de tempo e posteriormente quando expostas ao risco passem a ser consideradas, dando suporte então ao truncamento à esquerda (ALISSON, 2010). Com o banco de dados estruturado na forma de contagem os tempos entre as linhas de cada cliente são completamente disjuntos e o cálculo da verossimilhança parcial envolverá apenas uma observação de cada cliente, tornando as parcelas independentes (CARVALHO, *et al.*, 2011).

4. METODOLOGIA

Fazendo uso da literatura revista na seção 3 desenvolveu-se este trabalho usando a metodologia descrita a seguir:

4.1. Estruturaração do banco de dados

Transformação dos arquivos flat files para que as variáveis tempo-dependentes sejam acomodadas numa estrutura de processo de contagem e que sejam incorporados os truncamentos e censuras. Todos os clientes que realizaram a sua primeira compra no período anterior ao final da coleta de dados serão desconsiderados das análises. Os clientes que realizaram duas ou mais compras no mesmo dia serão retirados das análises, assim como aqueles clientes sem a informação de sexo. A base transacional será transformada para que apresente a forma de contagens.

Para a construção do banco de dados será utilizado o software *postgresql*. Podemos descrever o processo de obtenção do banco final nas seguintes etapas: (*i*) partindo-se do banco cadastral, que possui as variáveis fixas e que devem ser repetidas, une-se o banco transacional, onde cada linha deve representar somente uma compra do cliente; (*ii*) para o processo de contagem, um intervalo de tempo é criado a cada compra observada, assim se faz necessária a inclusão de uma linha para cada cliente, pois existe um intervalo sem compras no início do relacionamento de cada cliente; (*iii*) as variáveis início e fim que denotam o começo e final de cada intervalo de tempo são criadas, sendo que o intervalo é fechado abaixo e aberto acima.

4.2. Modelagem da sobrevivência via regressão de Cox

O ajuste do modelo Cox descrito na fórmula 7 será executado no software R, através da rotina *coxph*. A significância de cada variável será analisada através do teste de Wald ao nível de significância de 1%. As variáveis que não contribuam estatisticamente para o modelo serão retiradas. O processo será executado repetidamente até que apenas variáveis significativas estejam no modelo.

4.2.1. Qualidade do ajuste

A qualidade do ajuste do modelo será avaliada através do Coeficiente de Concordância. Para que o ajuste do modelo seja considerado satisfatório, o valor do coeficiente deve ser de pelo menos 0,6 (CARVALHO, *et al.*, 2011).

Após a avaliação da concordância iremos avaliar a adequação do modelo em relação a premissa da proporcionalidade dos resíduos, existência de pontos mal ajustados e de alavancagem.

O Resíduo de Schoenfeld será usado para medir a proporcionalidade dos riscos, computando a correlação linear entre o tempo de sobrevivência e os resíduos, sendo a hipótese nula a existência de proporcionalidade (CARVALHO, *et al.*, 2011). Devido ao tamanho amostral grande, nesse trabalho a avaliação se dará na análise da estimativa pontual da correlação, ao invés do valor *p* associado. Uma boa discussão de procedimentos para avaliar o ajuste de modelos de regressão diante de tamanhos de amostras grandes pode ser encontrada em (YAO, *et al.*, 2009) e (GHOSE, *et al.*, 2006).

A análise dos resíduos Martingale e Deviance busca encontrar casos mal ajustados (observações discrepantes) pelo modelo (CARVALHO, *et al.*, 2011). Na análise do gráfico as linhas vermelhas evidenciarão os casos com resíduos fora dos limites desejados, indicando assim casos discrepantes ao ajuste do modelo. As linhas vermelhas ocorrerão para resíduos Deviance inferiores a -2 ou superiores a 2; e resíduos Martingale inferiores a -2, pois estes resíduos tem limite superior igual a 1.

Os resíduos Escore nos permitem identificar pontos de alavancagem no modelo, ou seja pontos que tem uma importância grande na estimação dos parâmetros do modelo (CARVALHO, *et al.*, 2011). Como o resultado do resíduo é padronizado, devemos nos ater aos pontos que estão a mais de 2 desvios da média dos resíduos.

4.3. Estimação do CLV:

Para o cálculo do CLV será usada a fórmula proposta por Ferreira, descrita na fórmula 1. A contribuição financeira $[M_i(t)]$ será estimada através da contribuição diária, que por sua vez é o resultado da divisão do total gasto durante o relacionamento pelo total de dias em relacionamento.

O modelo ajustado utilizando a regressão de Cox permitirá obter uma estimativa da probabilidade de sobrevivência até o tempo t para cada cliente, sendo portanto a estimativa do termo $S_i(t)$.

Fazendo uso de um recorte temporal irá se verificar a assertividade da categorização realizada sobre os resultados do modelo. Os clientes serão classificados em dois grupos, conforme a estimativa obtida para o CLV, nos 20% de maior CLV (melhores) e os 80% restantes. Em seguida os clientes serão reordenados em 20%/80% de acordo com o CLV real obtido no período de análise. Finalmente a análise de assertividade será realizada computando o percentual de classificação correta para ambos os grupos (20% melhores e 80% restantes).

5. ESTUDO DE CASO

5.1. Construção do banco para Análise

Para o presente estudo foram utilizadas duas bases de dados: uma com informações de clientes, uma com informação de todas as compras realizadas. Os dados utilizados foram obtidos com o consentimento da empresa meio, que utilizará os resultados do estudo como insumo para a empresa fim.

A base cadastral, com informações dos clientes, tem dados dos clientes que se cadastraram entre 2010-07 e 2017-09; já a base transacional, com as compras realizadas, tem informações do período 2014-01 a 2017-09. Para analise foram considerados todos os clientes que tiveram ao menos uma compra no período 2014-01 a 2017-09. Cada mercado considera um tempo diferente para declarar que o cliente deixou de se relacionar com a empresa. Por exemplo, um supermercado pode dizer que um cliente que não compra com ele a mais de 33 dias é perdido, já a padaria do bairro talvez considere 7 dias ao passo que uma montadora de carros talvez considere 3 anos para aqueles clientes mais identificados com a marca. Para a empresa em estudo o cliente que tenha ficado 366 ou mais dias sem realizar uma compra é definido como sendo inativo ou perdido.

Durante o período em estudo 1.383.031 clientes realizaram ao menos uma compra e o banco estruturado na forma de contagens apresenta 4.325.137 linhas. Nesse período 90.907 experimentaram o evento, ou seja, deixaram de realizar compras, o que representa 6.5% da base de clientes analisados.

5.1.1. Variáveis do estudo

O banco de dados inclui variáveis transacionais e cadastrais. A Tabela 1 apresenta a relação de variáveis cadastrais.

Tabela 1 - Variáveis cadastrais

Variável	Significado	Codificação	Tipo
variavel_1	informação de contato	categoria_1=positivo categoria_2=negativo	categórica
variavel_2	informação de contato	categoria_1=positivo categoria_2=negativo	categórica

variavel_3	qualidade do contato	categoria_1=positivo categoria_2=negativo	categórica
variavel_4	qualidade do contato	categoria_1=positivo categoria_2=negativo	categórica
variavel_5	informação do cliente	categoria_1 categoria_2	categórica
variavel_6	informação do cliente	categoria_1 categoria_2 categoria_3 categoria_4 categoria_5 categoria_6	categórica
variavel_7	informação do cliente		contínua

A Tabela 2 apresenta o banco transacional composto por variáveis tempo – dependentes. Estas variáveis são do tipo categóricas ou indicadoras, isto é, denotam se a característica foi observada (variável recebe 1) ou não (variável recebe 0) em diferentes períodos.

Tabela 2 - Variáveis transacionais

Variável	**Significado**	**Codificação**	**Tipo**
variavel_8	denota que o cliente tornou-se inativo após a compra específica	0=ativo 1=inativo	categórica
variavel_9	informação da compra	categoria_1 categoria_2	categórica
variavel_10	informação da compra	categoria_1=positivo categoria_2=negativo	categórica
variavel_11	informação da compra	categoria_1=positivo categoria_2=negativo	indicadora
produto_1	compra produto tipo 1		indicadora
produto_2	compra produto tipo 2		indicadora
produto_3	compra produto tipo 3		indicadora
produto_4	compra produto tipo 4		indicadora
produto_5	compra produto tipo 5		indicadora
produto_6	compra produto tipo 6		indicadora
produto_7	compra produto tipo 7		indicadora
produto_8	compra produto tipo 8		indicadora

produto_9	compra produto tipo 9	indicadora
produto_10	compra produto tipo 10	indicadora
produto_11	compra produto tipo 11	indicadora
produto_12	compra produto tipo 12	indicadora
produto_13	compra produto tipo 13	indicadora
produto_14	compra produto tipo 14	indicadora
produto_15	compra produto tipo 15	indicadora
produto_16	compra produto tipo 16	indicadora
produto_17	compra produto tipo 17	indicadora
produto_18	compra produto tipo 18	indicadora
produto_19	compra produto tipo 19	indicadora
produto_20	compra produto tipo 20	indicadora
produto_21	compra produto tipo 21	indicadora

5.2. Análise descritiva dos dados

A seguir são apresentadas as tabelas de frequência simples de cada variável disponíveis no banco de dados:

Tabela 3 - Variável 5

	Frequência	**%**
categoria_2	1053487	76,20%
categoria_1	329544	23,80%

Tabela 4 - Variável 6

	Frequência	**%**
categoria_1	53308	3,90%
categoria_2	119921	8,70%
categoria_3	2203	0,20%
categoria_4	1104945	79,90%
categoria_5	1717	0,10%
categoria_6	100937	7,30%

Tabela 5 - Variável 7

1º Quartil	**Mediana**	**Média**	**3º Quartil**
25	32	33,4	40

Tabela 6 - Informação de contato do cliente

	Valor	Frequência	%
variavel_1	positivo	948698	68,60%
	negativo	434333	31,40%
variavel_2	positivo	204833	14,80%
	negativo	1178198	85,20%

Tabela 7 - Qualidade do contato

	Valor	Frequência	%
variavel_3	positivo	1253098	90,60%
	negativo	129933	9,40%
variavel_4	positivo	210929	15,30%
	negativo	1172102	84,70%

Tabela 8 - Informação da compra

	Frequência	%
categoria_2	1999062	46,22%
categoria_1	2326075	53,78%

Tabela 9 - Informação da compra

	Frequência	%
negativo	2309163	53,39%
positivo	2015974	46,61%

Tabela 10 - Informação da compra

	Frequência	%
negativo	4143253	95,79%
positivo	181884	4,21%

As Tabelas 3 a 10 trazem algumas informações que caracterizaram a base de clientes. Podemos observar uma grande concentração da categoria 2 da variavel_5 (75,20%) e que a categoria_4 da variavel_6 está presente em 79% de todos clientes. A maioria dos clientes apresenta um variavel_3 positiva (90,6%), porem apenas 68,60% dos clientes variavel_1 positiva. A variavel_4 positiva é pouco difundida, apresentando apenas 15,30% dos clientes com essa informação. A divisão entre as categorias da

variavel_9 é equilibrada (53,78% e 47,22% respectivamente), equilíbrio também observado na proporção das categorias da variavel_10, onde 46,61% dos clientes são positivo. Por fim podemos verificar que apenas 4,21% das compras foram positivo para a variavel_11.

Tabela 11 - Categorias compradas

	Quantidade de compras com a categoria	**% das compras**	**% dos clientes**
produto_1	781258	18,06%	35,72%
produto_2	2310985	53,43%	79,90%
produto_3	303568	7,02%	12,62%
produto_4	287029	6,64%	14,88%
produto_5	144146	3,33%	8,14%
produto_6	46265	1,07%	2,62%
produto_7	25730	0,59%	1,48%
produto_8	33299	0,77%	2,08%
produto_9	5033	0,12%	0,31%
produto_10	511974	11,84%	25,76%
produto_11	500058	11,56%	26,11%
produto_12	711609	16,45%	32,72%
produto_13	161974	3,74%	9,09%
produto_14	320125	7,40%	16,29%
produto_15	163878	3,79%	8,71%
produto_16	32784	0,76%	2,02%
produto_17	570613	13,19%	30,51%
produto_18	130223	3,01%	7,60%
produto_19	72617	1,68%	4,40%
produto_20	44254	1,02%	2,70%
produto_21	62573	1,45%	3,82%

A Tabela 11 exibe a quantidade total de compras que tiveram dentre seus produtos ao menos 1 da categoria em questão, também apresenta a participação da categoria no total de compras realizadas (4.325.137), e o percentual de clientes que realizaram ao menos uma compra que tivesse a categoria, relativo ao total de clientes (1.383.031). É possível ver a importância de produto_2, pois 53,43% de todas as compras tem algum produto dessa categoria ao passo que 18,06% das compras tem produtos do produto_1, produto_12 com 16,45% é outra categoria que está relativamente presente nas compras.

Ao analisar quais categorias mais impactam os clientes (aqui impactar é o fato de ter realizado ao menos uma compra da categoria) observamos novamente que produto_2 tem o melhor desempenho impactando 79,90% dos clientes, seguida de produto_1 com 35,72%, produto_12 32,72%, produto_17 30,51%, produto_11 26,11% e produto_10 25,76%.

5.3. Ajuste do modelo

Após duas rodadas de seleção de varáveis o modelo apresentou apenas variáveis significativas, resultando no coeficiente de concordância de 0,632 o que, indica um ajuste satisfatório.

O modelo final apresentou 6 variáveis cadastrais e 18 varáveis transacionais (tempo-dependentes). As estimativas do risco relativo podem ser observadas na Tabela 12.

Tabela 12 - Estimativa do risco relativo de deixar de comprar

Variável	**Relação com o tempo**	**Risco Relativo**		
		Pontual	**Intervalo de confiança (95%)**	
variavel_5 categoria_1	Fixas	97,08%	95,32%	98,89%
variavel_7		100,23%	100,17%	100,29%
variavel_6 (ref=categoria_1)				
categoria_2		71,92%	69,53%	74,40%
categoria_3		44,93%	38,43%	52,52%
categoria_4		47,97%	46,30%	49,70%
categoria_5		35,85%	28,79%	44,63%
categoria_6		144,40%	138,60%	150,44%
variavel_1 negativo		88,51%	87,15%	89,89%
variavel_3 negativo		86,36%	84,40%	88,35%
variavel_4 negativo		52,99%	51,85%	54,16%
variavel_11 1	Tempo dependentes	104,18%	101,00%	107,46%
variavel_9 categoria_1		159,87%	155,91%	163,92%
variavel_10 positivo		117,84%	116,06%	119,65%
produto_1 1		146,57%	143,30%	149,93%
produto_2 1		181,26%	177,11%	185,51%
produto_3 1		90,61%	88,24%	93,04%
produto_5 1		107,21%	104,05%	110,47%
produto_7 1		30,76%	25,08%	37,74%
produto_8 1		75,40%	69,42%	81,90%

produto_10 1	115,89%	113,84%	117,98%
produto_11 1	109,89%	107,88%	111,94%
produto_14 1	114,63%	112,25%	117,05%
produto_15 1	90,18%	87,37%	93,08%
produto_16 1	108,23%	102,04%	114,80%
produto_17 1	109,03%	107,12%	110,98%
produto_18 1	107,23%	103,54%	111,05%
produto_19 1	53,27%	49,71%	57,08%
produto_20 1	115,42%	109,07%	122,14%

As estimativas apresentadas na Tabela 12 permitem interpretar a influência de cada uma das varáveis do modelo no risco de deixar de comprar. Importante destacar que devido ao tamanho grande da amostra, a análise sobre o impacto das variáveis no risco de deixar de comprar deve se concentrar em quão distantes de 100% (tanto para mais quanto para menos) estão as estimativas por intervalo dos riscos (LIN, *et al.*, 2013). Por exemplo, em relação à variável quantitativa variavel_7, percebe-se que a cada incremento de uma unidade, o risco de cancelamento cresce 0,23% (IC 95%: 0,17%-0,29%), tendo uma influência relativamente pequena no risco de cancelamento.

No caso da variavel_6, apenas a categoria_6 apresenta estimativa de risco relativo maiores do que 100% em relação à categoria de referência categoria_1. Por exemplo, clientes com categoria_2, apresentam em média risco de deixar de comprar 28,08%(IC 95%: 25,60%-30,47%) inferior aos clientes da categoria_1, ajustando para as demais variáveis. Talvez seja importante realizar uma investigação sobre a qualidade dos contatos que tem categoria_6, já que a o risco relativo aumenta 44,40%(IC 95%: 38,60%-50,44%) em relação a categoria_1. No caso da variavel_5 sendo o cliente da categoria_1, o risco de deixar de comprar diminui em 2,92% (IC 95%: 1,11%-4,68%).

O cliente comprar variavel_9 categoria_1 aumenta em 59,87% (IC 95%: 55,91%-63,92%) o risco de deixar de comprar relativo a comprar variavel_9 categoria_2, esse fato, pode indicar que a experiência de variavel_9 categoria_2 traz um relacionamento mais duradouro com a empresa. Quando o pagamento é realizado variavel_10 positivo o risco aumenta 17,84% (IC 95%: 16,06%-19,65%) frente as outras formas de pagamento.

Das variáveis tempo-dependentes, as que mais aumentam o risco relativo são: ter comprado produto_2 aumenta o risco relativo em 81,26% (IC 95%: 77,11%-85,51%),

produto_1 aumentam o risco em 46,57% (IC 95%: 43,30%-49,93%), produto_10 15,89% (IC 95%: 13,84%-17,98%) e produto_20 15,42% (IC 95%: 9,07%-22,14%).

As variáveis tempo-dependentes que mais diminuem o risco de deixar de compra são: compra produto da categoria produto_7, fato que diminui o risco relativo em 69,24% (IC 95%: 62,26%-74,92%), produto_19 diminuindo o risco 46,76% (IC 95%: 42,92%-50,29%), produto_8 24,60% (IC 95%: 18,10%-30,58%), produto_15 9,82% (IC 95%: 6,92%-12,63%) e produto_3 9,39% (IC 95%: 6,96%-11,76%).

Para as variáveis tempo-dependentes é válido uma análise cruzada com a Tabela 11 visto que algumas categorias tem uma baixa participação sobre a base de clientes, assim produto_2 e produto_1 devem ter uma atenção pois impactam 79,90% e 35,72% da base e causam os maiores aumentos no risco de deixar de comprar, produtos produto_10 tem atingem 25,76% e também deveriam ter uma atenção maior; já produtos da categoria produto_20 impactam apenas 2,70% dos clientes. Das variáveis que diminuem o risco relativo, produtos produto_3 (12,62%) e produtos produto_15 (8,71%) são aquelas que mais impactam clientes, e assim poderiam ter seu comportamento analisado mais a fundo; já produtos produto_19, esporte e produto_7 por terem um baixo impacto (4,40%, 2,08% e 1,48% respectivamente) não necessitariam de maiores cuidados.

5.4. Análise de Resíduos

Observamos na Tabela 13 que embora boa parte das correlações sejam significativas a 1%, podemos observar que a maior correlação amostral encontrada foi de -0,037 (variavel_5 categoria_1), considerada relativamente baixa. Dessa forma, consideramos que nenhuma das variáveis viola a premissa de proporcionalidade dos riscos.

Tabela 13 - Teste da correlação linear dos resíduos Schoenfeld

Variável	***rho***	**p**
variavel_5 categoria_1	-0,037679	$0{,}00\text{x}10^{+0}$
variavel_7	0,018394	$3{,}22\text{x}10^{-8}$
variavel_6 categoria_2	0,02092	$2{,}65\text{x}10^{-10}$
variavel_6 categoria_3	0,016746	$4{,}32\text{x}10^{-7}$
variavel_6 categoria_4	0,066798	$0{,}00\text{x}10^{+0}$
variavel_6 categoria_5	0,005643	$8{,}86\text{x}10^{-2}$
variavel_6 categoria_6	0,031215	$0{,}00\text{x}10^{+0}$

variavel_1 negativo	-0,006571	$4,67x10^{-2}$
variavel_3 negativo	-0,013341	$6,97x10^{-5}$
variavel_4 negativo	0,004241	$2,05x10^{-1}$
variavel_11 positivo	0,029591	$0,00x10^{+0}$
variavel_9 categoria_1	-0,006205	$4,69x10^{-2}$
variavel_10 positivo	0,009721	$2,00x10^{-3}$
produto_1 1	0,010101	$4,10x10^{-3}$
produto_2 1	0,005511	$1,04x10^{-1}$
produto_3 1	0,036549	$0,00x10^{+0}$
produto_5 1	0,000187	$9,54x10^{-1}$
produto_7 1	-0,005945	$7,28x10^{-2}$
produto_8 1	-0,007943	$1,48x10^{-2}$
produto_10 1	-0,004639	$1,63x10^{-1}$
produto_11 1	-0,033085	$0,00x10^{+0}$
produto_14 1	0,014343	$1,71x10^{-5}$
produto_15 1	0,023526	$1,40x10^{-12}$
produto_16 1	0,018106	$4,68x10^{-8}$

Os gráficos dos resíduos Schoenfeld de cada variável estão disponíveis na Figura 4 do anexo. Observa-se que os betas estimados não estão incluídos nos intervalos de confiança durante todos os instantes de tempo. Entretanto, devido ao tamanho da amostra acentuado, os intervalos de confiança são bastante estreitos, tornando a análise visual sobre a proporcionalidade inconclusiva. (CARVALHO, *et al.*, 2011).

Nas Figuras 2 e 3, observamos que apesar da analise visual sugerir um grande número de pontos influentes ou mal ajustados, o total de pontos com resíduo Deviance superior à 2 é de 48.514, representando apenas 1,12% da amostra. Dessa forma, consideramos não haver influência significativa destas observações no modelo ajustado.

Figura 2 - Resíduo Martingale

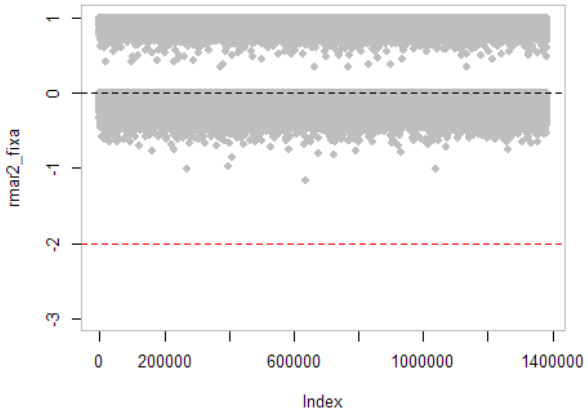


Figura 3 - Resíduo Deviance

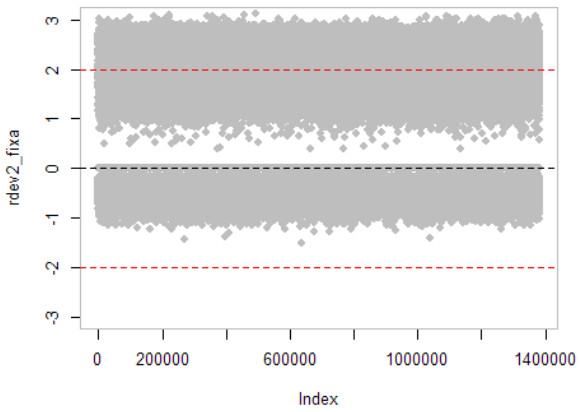


Tabela 14 - Resíduos Escore: valores mínimo e máximo observado

Variável	Valores do resíduo	
	Mínimo	Máximo
variavel_5 categoria_1	0,01060	-0,00683
variável_7	0,01758	-0,01178
variavel_6 categoria_	0,00670	-0,01284
variavel_6 categoria_3	0,07493	-0,01542
variavel_6 categoria_4	0,00987	-0,01899
variavel_6 categoria_5	0,10935	-0,01803
variavel_6 categoria_6	0,01245	-0,01665
variavel_1 negativo	0,00905	-0,00427
variavel_3 negativo	0,01187	-0,00539
variavel_4 negativo	0,01143	-0,01318
variavel_11 positivo	0,01510	-0,01058
variavel_9 categoria_1	0,01643	-0,01529
variavel_10 positivo	0,00602	-0,00818
produto_1 1	0,01480	-0,01386
produto_2 1	0,01479	-0,01693
produto_3 1	0,01544	-0,01304
produto_5 1	0,01688	-0,01103
produto_7 1	0,10364	-0,01411
produto_8 1	0,04226	-0,01507
produto_10 1	0,01215	-0,00954
produto_11 1	0,01232	-0,00899
produto_14 1	0,01194	-0,00806
produto_15 1	0,01664	-0,01299
produto_16 1	0,03023	-0,01877

A Tabela 14 nos mostra os valores mínimo e máximo do resíduo escore, para cada variável do modelo. Ao analisarmos a tabela é possível ver que nenhuma variável apresenta resíduo escore máximo maior do que 0,5, assim como nenhuma apresenta resíduo escore mínimo menor do que 0,5, fato considerado satisfatório (CARVALHO, *et al.*, 2011). O maior resíduo positivo é 0,10935 de variavel_6 categoria_5, já o maior valor negativo é -0,01899 para variavel_6 categoria_4.

5.5. Avaliação do CLV via Modelo

Neste estudo para construção do CLV optou-se por usar a fórmula 1, considerando o conceito de máximo valor de cliente de Berger e Nasr, conforme descrito na metodologia. A estimativa da contribuição financeira do cliente foi obtida através da média de gasto diário realizado até o momento da última compra observada. A curva de sobrevivência $S(t)$ foi então estimada a partir da regressão Cox ajustada aos dados.

Como no banco de dados cada cliente tem um momento específico no tempo para a sua última compra, o valor do tempo que é utilizado para a estimativa de sobrevivência até a data 2017-09-01 é diferente em cada cliente. Como o modelo utilizava um período de 24 meses e o cliente é considerado perdido após 366 dias sem comprar, as estimativas para a data em foco somente serão realizadas para aqueles clientes que terminaram o período de construção do modelo ativos e que o seu tempo até 2017-09-01 é inferior a 731.

Para estimar os valores do CLV utilizou-se como valor das covariáveis os dados da última compra realizada pelo cliente no período de 2014-01 à 2015-12; e como intervalo de tempo a distância, em dias, entre a data da última compra do cliente e a data de interesse 2017-09-01.

Para a avaliação da assertividade do modelo optou-se por classificar os clientes em dois grupos, conforme a estimativa obtida para o CLV, nos 20% de maior CLV (melhores) e os 80% restantes. Em seguida os clientes foram reordenados em 20%/80% de acordo com o CLV real obtido no período de análise (2017-09-01). A Tabela 15 mostra a analise cruzada das classificações obtidas com as duas ordenações.

Tabela 15 - Assertividade do modelo

		Resultado real		
		Botton 80%	Top 20%	Total
Predição do modelo	Botton 80%	276078 (90,41%)	29273 (38,35%)	305351
	Top 20%	29277 (9,59%)	47061 (61,65%)	76338
	Total	305355	76334	381689

Observamos que o modelo classifica corretamente 90,41% dos clientes Bottom 80% e 61,65% dos clientes Top 20% com uma classificação correta global de 84,66% dos clientes. É natural que a empresa invista mais (aloque mais recursos) nos clientes Top 20% visto que eles tem um maior perfil de consumo. Destaca-se que apenas 9,59% dos clientes Bottom 80% seriam tratados como Top 20%, isto é, receberiam mais investimentos do que o desejado. Por outro lado, 38,35% dos clientes Top 20% seriam subvalorizados, recebendo menos investimentos do que deveriam. Os resultados ilustram a boa capacidade preditiva para o CLV da abordagem proposta no trabalho, sendo inclusive, superior em relação à outros estudos similares descritos na literatura, como por exemplo em MALTHOUSE e BLATTERBER (2005).

6. CONSIDERAÇÕES FINAIS

Este trabalho apresentou a modelagem do CLV dos clientes de uma empresa com base na regressão de Cox ajustada a partir de uma base de dados apresentando variáveis cadastrais e transacionais. Através da abordagem proposta, pode-se identificar as variáveis associadas ao tempo de sobrevivência do cliente na base e, dessa forma, realizar a estimativa dos seus gastos esperados. Verificamos através de um recorte temporal no banco de clientes a boa capacidade preditiva do modelo ajustado na determinação do CLV de cada cliente.

Os resultados obtidos neste trabalho nos dão insumos para elaborar relatórios que demonstrem o potencial CLV de um grupo de clientes ou ainda o potencial CLV dos novos clientes adquiridos, bem como identificar os clientes que são mais importantes financeiramente para empresa, nos quais possíveis investimentos devem ser realizados ou os quais devam ter uma régua especial anti-atrito.

Em trabalhos futuros seria interessante o desenvolvimento mais apurado de um modelo para estimar o gasto de cada cliente, considerando interações entre variáveis transacionais no ajuste da regressão de Cox, bem como considerando no cálculo do CLV o valor investido em cada cliente, tanto na sua aquisição, como durante o período de relacionamento com a empresa.

REFERÊNCIAS

ALISSON, P. **Survival analysis using SAS:** a practical guide. [S.l.]: SAS Institute, 2010.

BERGER, P. D.; NASR, N. I. Customer Lifetime Value: Marketing Models and Applications. **Journal of Interactive Marketing**, p. 17-30, 1998.

BLATTBERG, R. C.; GETZ, G.; THOMAS, J. S. **Customer Equity – Building and Managing Relationships as Value Assets**. 1. ed. [S.l.]: Havard Business School Press, 2001.

BLATTERBER, R. C.; DEIGHTON, J. Manage Marketing by the Customer Equity Test. **Harvard Business Review**, p. 136-144, Julho-Agosto 1996.

CARVALHO, M. S.; ANDREOZZI, V. L.; CODEÇO, C. T.; CAMPOS, D. P.; BARBOSA, M. T. S.; EMIKO, S. **Análise de sobrevivência:** teoria e aplicações em saúde. Rio de Janeiro: Editora Fiocruz, 2011.

DWYER, F. R. Customer Lifetime Valuation to Suport Marketing Decision Making. **Journal of Direct Marketing**, p. 6-13, 1997.

FERREIRA, E. C. **Um modelo quantitativo para o valor do cliente**. Escola de Administração de Empresas de São Paulo. São Paulo. 2007.

GHOSE, A.; SMITH, M.; TELANG, R. Internet exchanges for used books: an empirical analysis of product cannibalization and welfare impacat. **Information System Research**, 1, 2006. 3-19.

HOSMER, D. W.; LEMESHOW, S.; MAY, S. **Applied survival analysis.** Hoboken: Wiley-Interscience, 2008.

JAIN, D.; SINGH, S. S. Customer Lifetime Value Research. **Journal of interactive marketing**, v. 16, 2002. ISSN 1.

KLEINBAUM, D. G.; KLEIN, M. **Survival analysis. A self-learning approach.** Nova Iorque: Springer, 2005.

KOTLER, P.; ARMSTRONG, G. **Principles of Marketing**. 7th. ed. [S.l.]: Prentice-Hall, 1996.

KOTLER, P.; KELLER, K. **Administração de Marketing**. 10. ed. São Paulo: Prentice Hall, 2010.

LIN, M.; LUCAS, H. C. J.; SHMUELI, G. To big to fail: large samples and the p-value problem. **Information Systems Research**, 12 abr. 2013. 1-12.

MALTHOUSE, E. C.; BLATTBERG, R. C. Can we predict customer lifetime value. **Journal of Interactive Marketing**, p. 2-16, 2005.

PEPPERS, D.; ROGERS, M. **Return on customer**. [S.l.]: Random House, 2005.

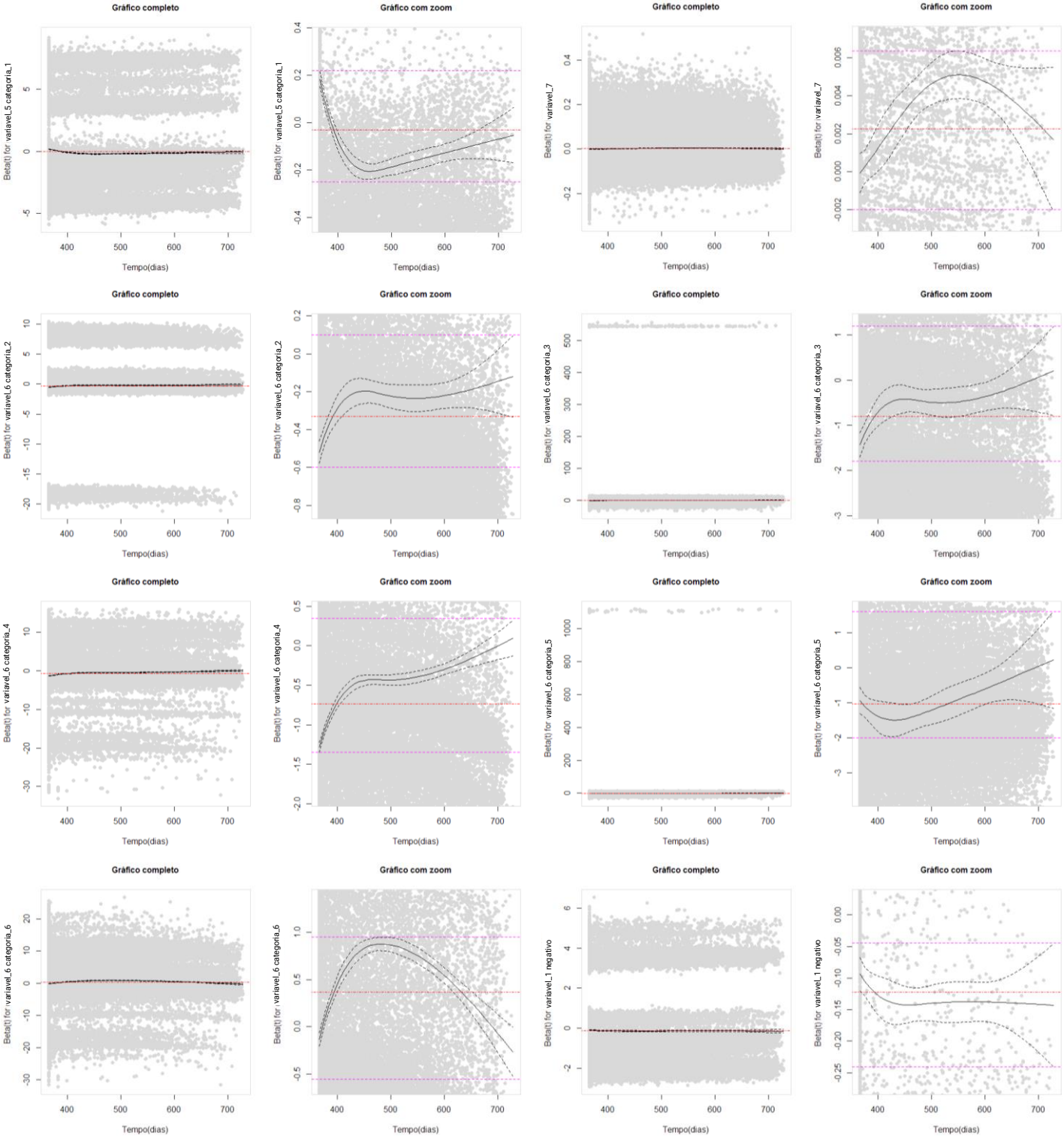
PFEIFER, P. E.; CARRAWAY, R. L. Modeling customer relationships as Markov chains. **Journal of Interactive Marketing**, v. 14, n. 2, p. 43-55, 2000.

POEL, D.; LARIVIÈRE, B. Customer attrition analysis for financial services using proportional hazard models. **European Journal of Operational Research**, n. 157, p. 196-217, 2004.

YAO, Y.; DRESNER, M.; PALMER, J. Private network EDI vs. Internet electronic markets: A direct comparison of fulfillment performance. **Management Science**, 55, n. 5, 2009. 843-852.

APÊNDICE

Figura 4 - Resíduos Schoenfeld



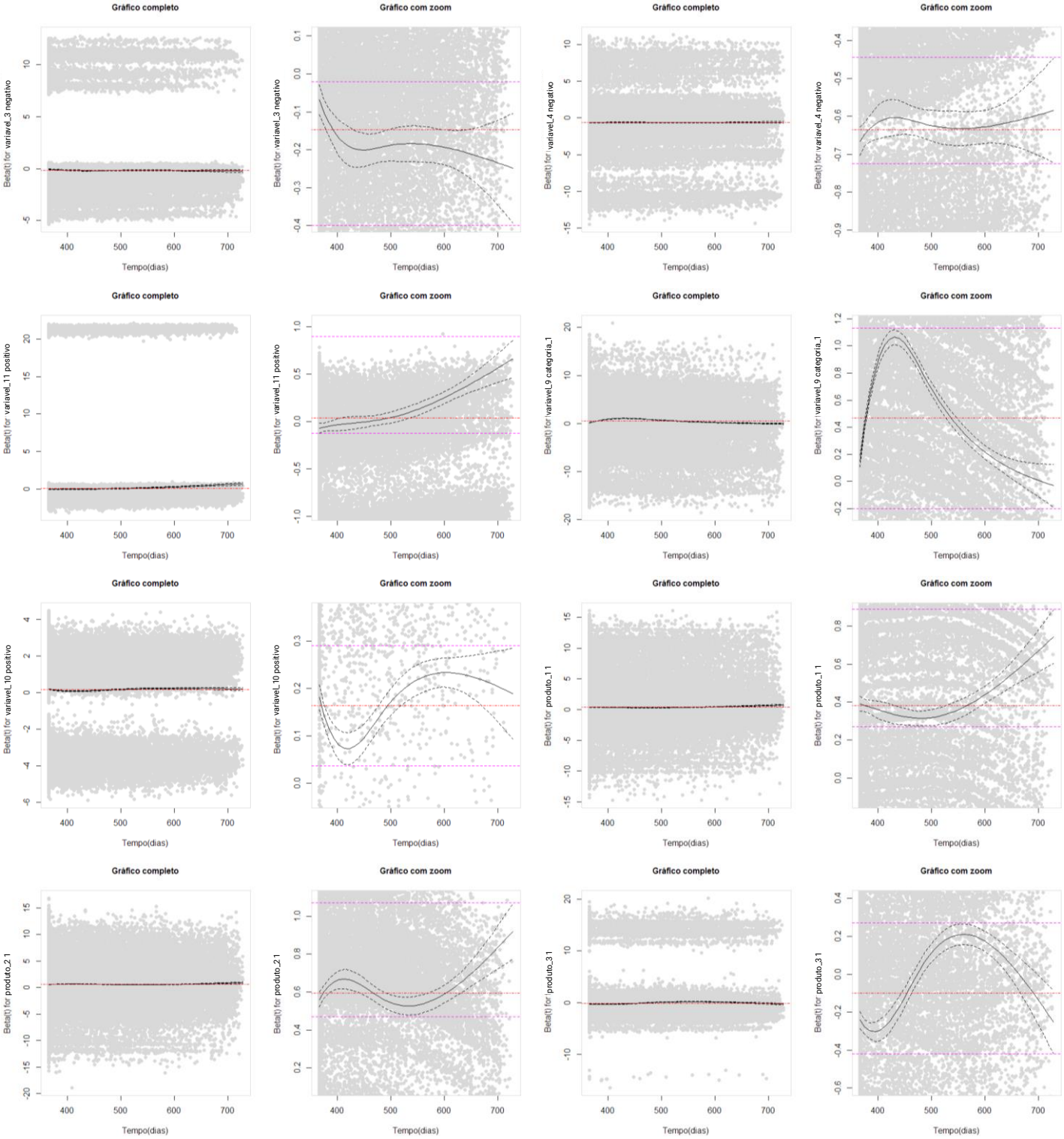


Gráfico completo
Beta(t) for variavel_3 negativo
Tempo(dias)
Gráfico com zoom
Beta(t) for variavel_3 negativo
Tempo(dias)
Gráfico completo
Beta(t) for variavel_4 negativo
Tempo(dias)
Gráfico com zoom
Beta(t) for variavel_4 negativo
Tempo(dias)
Gráfico completo
Beta(t) for variavel_11 positivo
Tempo(dias)
Gráfico com zoom
Beta(t) for variavel_11 positivo
Tempo(dias)
Gráfico completo
Beta(t) for variavel_9 categoria_1
Tempo(dias)
Gráfico com zoom
Beta(t) for variavel_9 categoria_1
Tempo(dias)
Gráfico completo
Beta(t) for variavel_10 positivo
Tempo(dias)
Gráfico com zoom
Beta(t) for variavel_10 positivo
Tempo(dias)
Gráfico completo
Beta(t) for produto_1 1
Tempo(dias)
Gráfico com zoom
Beta(t) for produto_1 1
Tempo(dias)
Gráfico completo
Beta(t) for produto_2 1
Tempo(dias)
Gráfico com zoom
Beta(t) for produto_2 1
Tempo(dias)
Gráfico completo
Beta(t) for produto_3 1
Tempo(dias)
Gráfico com zoom
Beta(t) for produto_3 1
Tempo(dias)

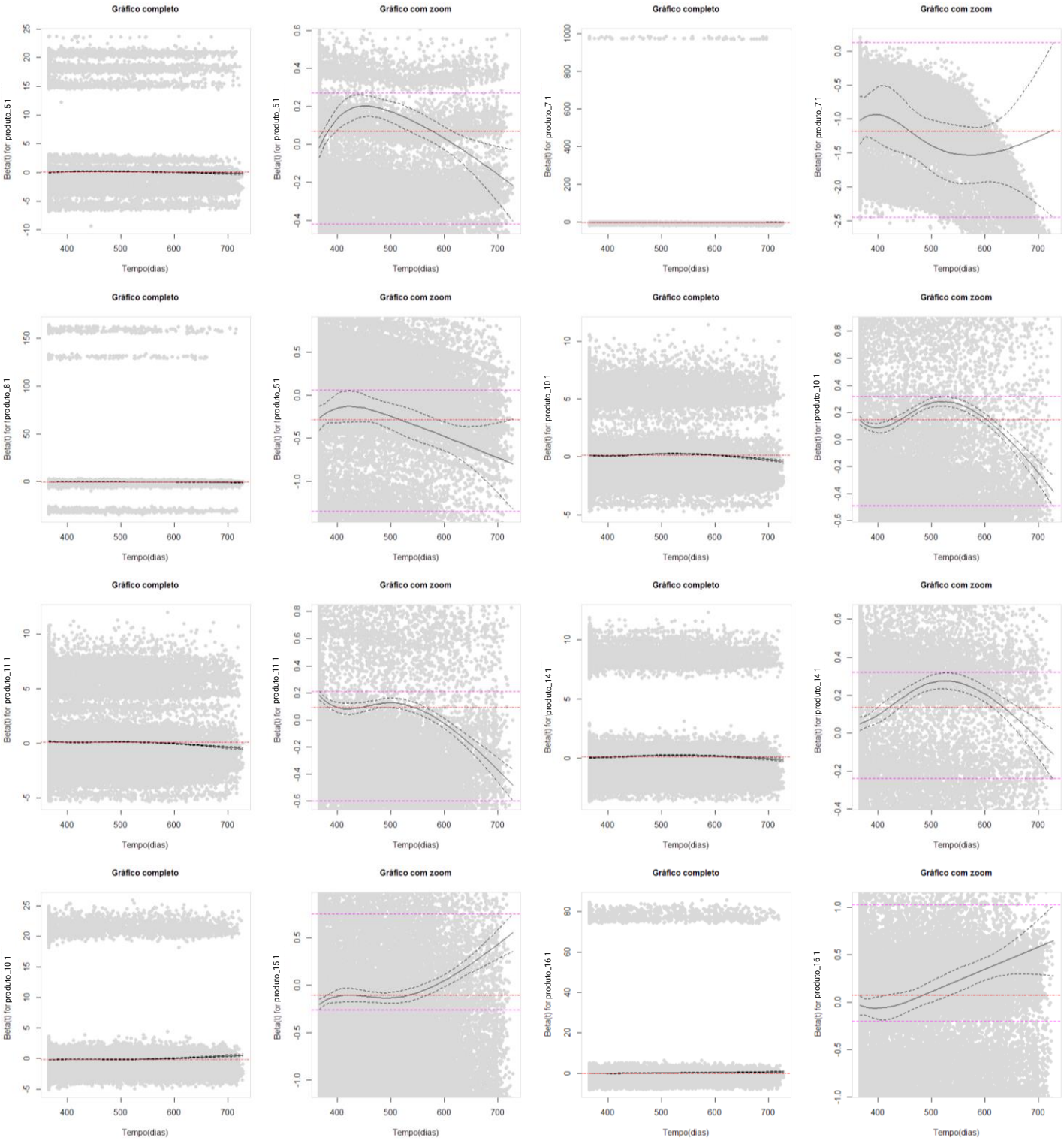


Gráfico completo
Gráfico com zoom
Beta(t) for produto_5 1
Beta(t) for produto_7 1
Beta(t) for produto_8 1
Beta(t) for produto_10 1
Beta(t) for produto_11 1
Beta(t) for produto_14 1
Beta(t) for produto_15 1
Beta(t) for produto_16 1
Tempo(dias)

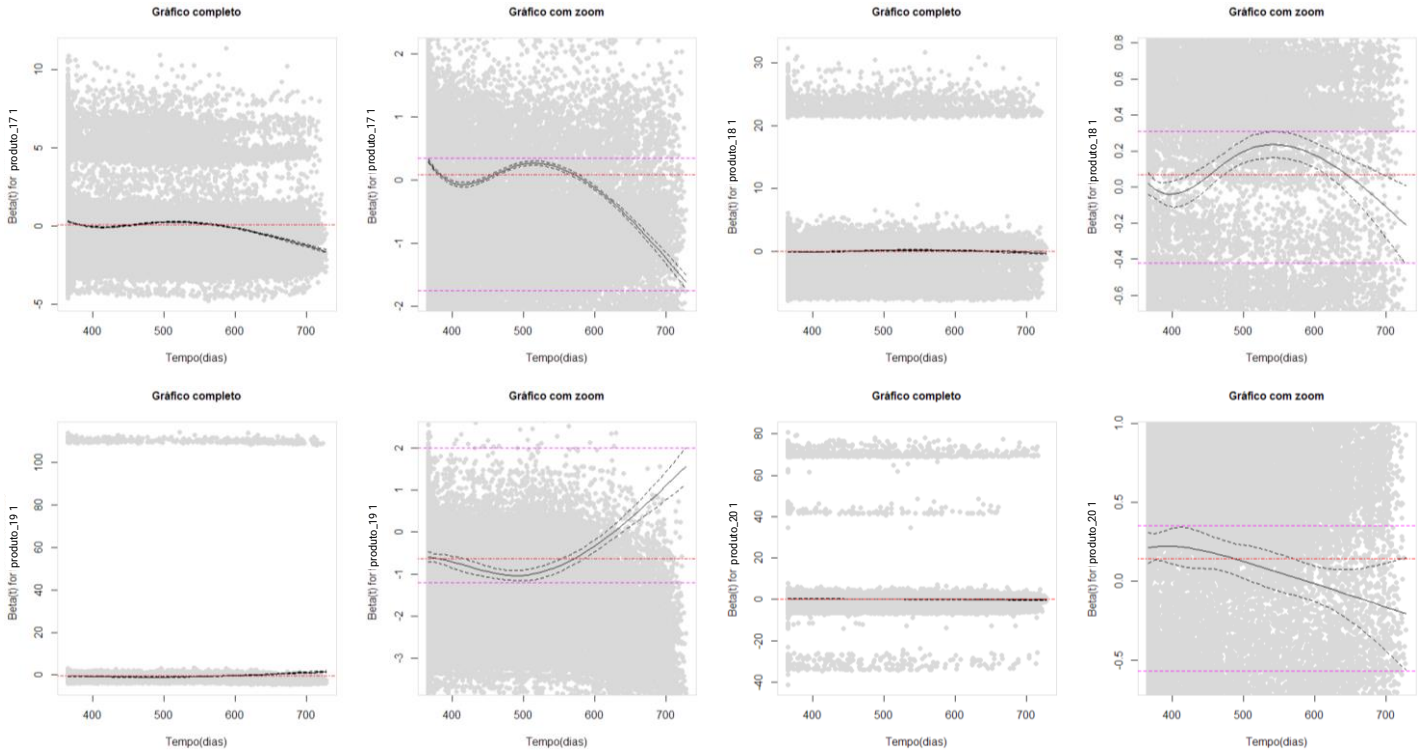


Figura 5 - Resíduos escore: variáveis fixas

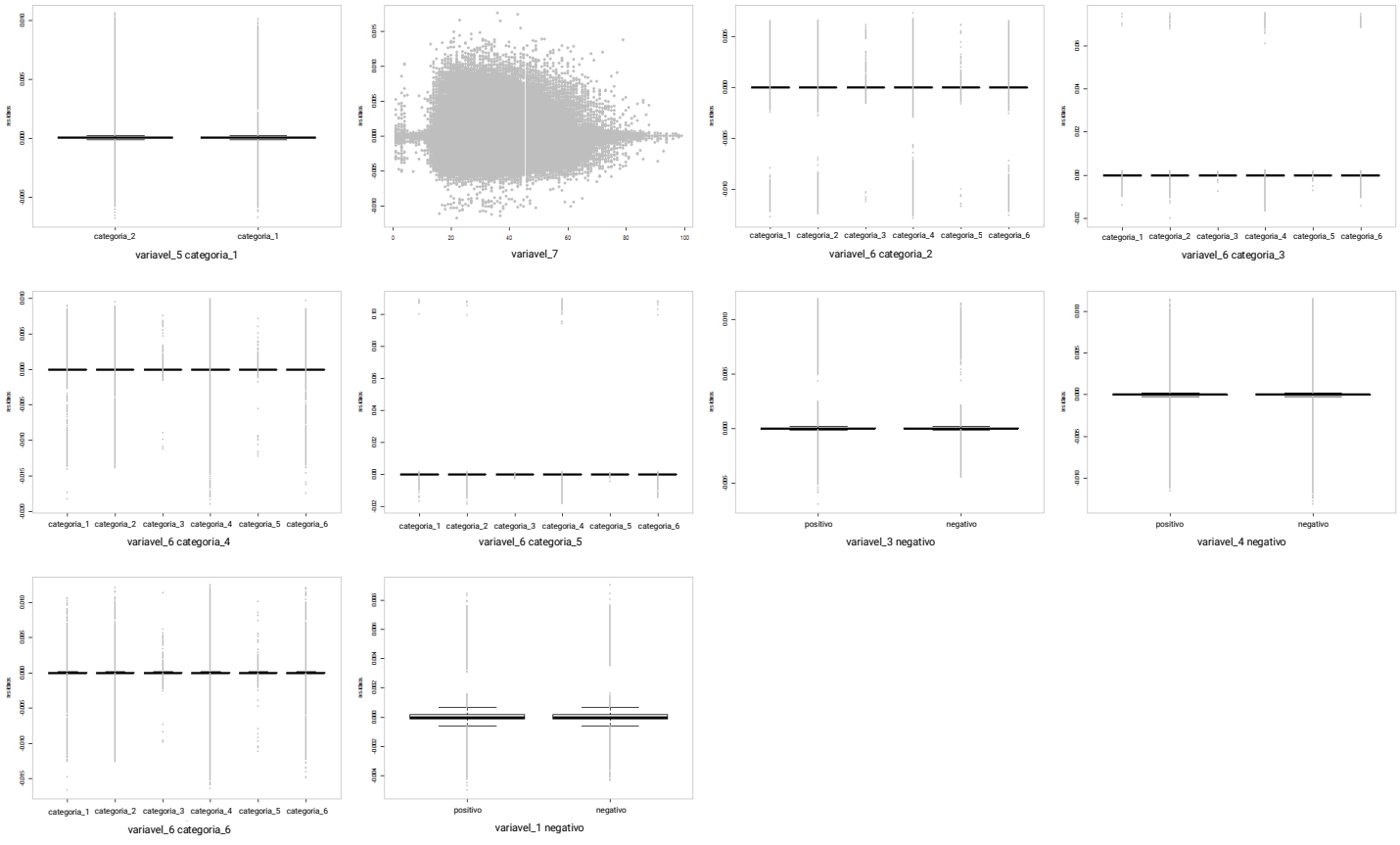


Figura 6 - Resíduos escore: variáveis tempo-dependentes