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

Estimação do risco de cancelamento de clientes no setor de gerenciamento de frotas por meio de modelo de Cox com variáveis tempo-dependentes

Autor: Mariana Wink Hohgraefe
Orientador: Professor Dr. Álvaro Vigo

Porto Alegre, 10 de Julho de 2015.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Estimação do risco de cancelamento de clientes no setor de gerenciamento de frotas por meio de modelo de Cox com variáveis tempo-dependentes

Autor: Mariana Wink Hohgraefe

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

Banca Examinadora:
Professor Dr. Álvaro Vigo (orientador)
Professora Dra. Suzi Alves Camey
Bruna Trierweiler Ribas

Porto Alegre, 10 de Julho de 2015.

“All models are wrong.
Some are useful”
(George Box)

AGRADECIMENTOS

Aos meus pais, Gilberto e Sandra. São eles os grandes responsáveis pela pessoa que sou e pelo caminho que trilhei, pois sempre prezaram e me demonstraram que o maior bem que alguém pode ter é o conhecimento. Obrigada por estarem ao meu lado sempre, por me incentivarem e vibrarem comigo a cada nova conquista.

Ao professor Álvaro Vigo, que pacientemente me orientou no processo de construção deste trabalho, transmitindo seu conhecimento, dedicando seu tempo e prezando pelo meu desenvolvimento ao longo de toda a minha trajetória acadêmica. Obrigada por compartilhar comigo a tua sabedoria e experiência como profissional de estatística.

Ao meu namorado, Charles, que com muito carinho sempre me apoiou e incentivou na busca pelos meus objetivos.

Aos meus queridos amigos "*outliers*", Alessandra, Yasmine, Gabriel e Juliana, pelo companheirismo durante todo o curso.

Aos meus amigos e professores de Cornell, que dividiram comigo um ano repleto de aprendizados, oportunidades e realizações, que tenho certeza serão fundamentais para o meu futuro.

Finalmente, agradeço à Empresa, em especial a minha chefe, Bruna, que confiou no meu trabalho e permitiu que eu pudesse aplicar a estatística em um projeto relevante para o negócio. Obrigada pelos aprendizados constantes e por te preocupares, diariamente, com o meu desenvolvimento e de toda a equipe.

RESUMO

Se bem gerenciada, a retenção de clientes pode reduzir custos e incrementar margens de lucro das empresas, demonstrando até mesmo vantagens frente aos dispendiosos esforços de atrair novos consumidores. Nesse cenário, modelos de cancelamento de clientes vêm atraindo a atenção dos pesquisadores de marketing, que através da aplicação de técnicas estatísticas como regressão logística, análise discriminante e análise de sobrevivência conseguem identificar fatores associados com a perda de clientes. Sendo assim, este trabalho se propôs a modelar o cancelamento de clientes por meio de um modelo de Cox considerando variáveis tempo-dependentes, sob uma abordagem de processo de contagem. O ajuste do modelo se deu em uma base de dados de uma empresa do setor de gerenciamento de frotas, que compreendia mais de 10 mil clientes. Com os resultados obtidos, tornou-se possível identificar fatores associados ao cancelamento e estimar o risco de cancelamento para clientes de acordo com suas características modeladas. O modelo final construído neste estudo configura um primeiro passo para prever a perda de clientes que futuramente pode servir de base para a adoção de uma estratégia de marketing orientada ao Valor Vitalício do Cliente.

Palavras-chave: modelo de Cox, variáveis tempo-dependentes, processo de contagem, retenção de clientes, modelo de cancelamento de clientes.

SUMÁRIO

1. INTRODUÇÃO

O presente trabalho surgiu de uma necessidade estratégica da área de Marketing de uma conceituada empresa de Gestão de Frotas, presente no mercado a nível nacional que, por motivos de confidencialidade, ao longo do trabalho será referida como Empresa X. Valente (2008) define gestão de frotas como a atividade de administrar ou gerenciar um conjunto de veículos pertencentes a uma mesma empresa. Canhoto et al. (2005) acrescenta que através do acompanhamento de informações de forma ágil a empresa pode fazer um controle detalhado de custos da frota. A Empresa X opera neste cenário oferecendo um sistema de gestão de frota para empresas clientes (portanto, os clientes são pessoas jurídicas) que abrange desde o pagamento das despesas com frota (abastecimento/manutenção) através de um "cartão combustível", até a coleta de informações, elaboração de relatórios de acompanhamento de performance da frota para, finalmente, gerar subsídio para que os gestores das empresas clientes posam tomar decisões estratégicas em relação à sua frota, usualmente visando a redução de custos e controle.

Em termos de dimensionamento da representatividade do setor de gestão de frotas no Brasil, destaca-se que o país possui uma frota de mais de 87,3 milhões de veículos (DENATRAN, 2015). Destes, estima-se que 5 milhões de veículos pertencem a frota corporativa, mercado alvo da Empresa X (SILVA, 2014). O grande potencial de mercado verificado no Brasil vem atraindo empresas internacionais especializadas em Gestão de Frotas, impulsionando, portanto, a concorrência pelo mercado de empresas com frotas corporativas. Atrelado ao aumento da concorrência, as ofertas do mercado atual estão mais padronizadas, semelhantes, enquanto os clientes estão cada vez mais conscientes em relação aos preços e mais exigentes (Kotler, 2010).

No caso do mercado *Business to Business* (*B2B*) em que está inserida a Empresa X, as operações de compra e venda ocorrem entre duas empresas, não existindo, portanto, consumidor final constituído por uma pessoa física, e sim um representante dos interesses de uma pessoa jurídica. Neste sentido, qualquer decisão de compra de um produto ou serviço por uma empresa pode impactar em toda a sua operação, demandando um processo parcimonioso e analítico de compra (Kotler, 2010). Em virtude destas características, a empresa que busca um fornecedor de serviços de gestão de frotas não se preocupa apenas com a operacionalização do pagamento das suas despesas com frota (abastecimento/manutenção), mas também com os indicadores de desempenho que o fornecedor pode entregar. Neste

sentido, as plataformas de serviços disponibilizadas pela Empresa X diferem positivamente da oferta dos concorrentes. Neste cenário complexo de atração de novos clientes, as empresas passam a reconhecer a crescente importância de satisfazer e reter seus clientes atuais (Reichheld, 1996).

A retenção de clientes que já estão na carteira das empresas costuma ser mais barato do que conquistar novos, enquanto que o lucro de reter clientes pode ser significativamente maior do que o resultado obtido através de novos fechamentos de negócios (Kotler, 2010). Para reter clientes, todavia, é necessário, primeiramente, entender porque a empresa os está perdendo, identificando tendências no comportamento daqueles clientes propensos ao cancelamento. Por meio da identificação de fatores associados à um comportamento de cancelamento por parte do cliente, torna-se possível monitorar tais fatores e agir sobre eles, de forma a evitar a perda e incrementar a retenção.

Uma das abordagens que vem sendo utilizada na identificação de padrões comportamentais de cancelamento de clientes são os modelos para previsão de cancelamento de clientes através da utilização de técnicas estatísticas como regressão logística, análise discriminante e análise de sobrevivência. Este trabalho irá se dedicar à aplicação de um modelo de análise de sobrevivência para investigar o cancelamento de clientes numa empresa do setor de gestão de frotas. Para isso, o trabalho está estruturado a partir de uma breve revisão da literatura de marketing sobre o viés de retenção de clientes e dos métodos que podem ser utilizados para identificar tendências ao cancelamento entre clientes. Num segundo momento, são descritas as principais abordagens de análise de sobrevivência que podem ser utilizadas a fim de predizer o cancelamento e identificar fatores associados a este evento.

A partir do delineamento da fundamentação teórica, os objetivos principais do trabalho consistem em modelar o tempo de relacionamento entre empresa e cliente até o cancelamento para estimar risco e identificar fatores associados (1) e iniciar discussão sobre a potencialidade do modelo para estimação do valor vitalício do cliente (2). Para tanto, uma etapa fundamental, intermediária, é a construção do banco de dados para utilização do modelo de Cox com variáveis tempo-dependentes na abordagem de processo de contagem. O entendimento da estrutura e o algoritmo para construção desta base de dados de análise são considerados como um objetivo específico. Tais objetivos serão endereçados no capítulo 4, de estudo de caso. Neste estudo, um modelo de Cox estendido será ajustado a uma base de dados reais de clientes da Empresa X de forma a identificar os principais fatores associados ao cancelamento dos clientes e mensurar seu impacto no risco de cancelamento. Serão discutidas as etapas de preparação da base de dados numa estrutura de processo de contagem para

análise, descritas as variáveis envolvidas nesta e finalmente ajustado o modelo, cujos resultados serão apresentados para que seja discutido na empresa seu potencial em termos estratégicos para que se possa aprimorá-lo e futuramente implementá-lo, com objetivo de reduzir a perda de clientes, e consequentemente, de ativos financeiros pela organização.

2. REVISÃO DA LITERATURA E MÉTODOS

2.1. RETENÇÃO DE CLIENTES E MODELOS DE CANCELAMENTO

Tanto a atração, quanto a satisfação e retenção de clientes fazem parte de um processo amplo de Marketing conhecido como CRM (*Customer Relationship Management*), abordado na literatura brasileira através dos termos "Gerenciamento do Relacionamento com Clientes" e "Gestão do Relacionamento com o Cliente". Segundo o especialista em Marketing Kotler (2010), o Gerenciamento do Relacionamento com Clientes compreende o processo de atrair e reter clientes rentáveis construindo relações de longo prazo por meio da entrega de valor e de satisfação aos clientes. Este trabalho aborda um dos aspectos do processo de Gerenciamento do Relacionamento com Clientes, referente a retenção destes, ou seja, a manutenção de clientes já conquistados pela empresa em sua base de clientes.

A retenção de clientes tem como objetivo diminuir a perda de clientes pelas empresas, reduzindo o número de clientes que cancelam seu contrato com a empresa e evitando migração para a concorrência. A falha na retenção ou perda de clientes já existentes é comumente abordada na literatura de marketing pelos termos em inglês "*Churn*" e "*Customer Attrition*", podendo ser traduzidos como "o ato de mudar de um fornecedor de um determinado produto/serviço para outro" e "desgaste de clientes".

Kotler (2010) afirma que esforços para atração de novos clientes podem gerar um custo até cinco vezes maior à empresa em relação aos investimentos para satisfação e retenção de clientes já existentes. O autor acrescenta ainda que dependendo do segmento de atuação de cada empresa, uma redução de 5% no índice de perda de clientes pode gerar um incremento de até 25 a 85% em seus lucros, uma vez que conforme se estende o período de relacionamento entre empresa e cliente a lucratividade por cliente tende a aumentar. Sendo assim, é imprescindível para a gestão do relacionamento com o cliente aprender a reter clientes, sendo que para isso é necessário descobrir por que os clientes partem, identificando padrões entre os clientes perdidos (Van den Poel e Larivière, 2003).

Uma das abordagens que vem sendo utilizada na identificação de padrões comportamentais de cancelamento de clientes são os modelos para previsão de cancelamento de clientes. Modelos preditivos exploram padrões históricos de comportamento encontrados em bases de dados ao longo do tempo para identificação de possíveis riscos e oportunidades de negócio. Neste âmbito, modelos preditivos permitem analisar entre diversos fatores àqueles

que estejam de fato associados a determinadas condições (no caso deste trabalho o cancelamento de clientes da Empresa X), de forma a guiar a tomada de decisões de negócio.

A literatura atual acerca de modelos de cancelamento de clientes encontra-se, em sua maioria, aplicada às áreas de serviços financeiros, telecomunicações (televisão por assinatura) e seguros, em virtude de suas taxas elevadas de perda de clientes e/ou fácil possibilidade de migração entre fornecedores. Em todos estes casos, o consumidor é uma pessoa física, portanto os fatores usualmente observados ao longo do tempo para construção de modelos preditivos de cancelamento de cliente poderiam ser agrupados, segundo Van den Poel e Larivière (2003) em indicadores de comportamento do cliente; percepções dos clientes, característica demográficas e variáveis macro-ambientais (que incluíram indicadores de prosperidade econômica, por exemplo).

Em um levantamento realizado por Van den Poel e Larivière (2003) acerca dos estudos em cancelamento de clientes publicados entre 1994 e 2002 fica evidente que as técnicas estatísticas mais utilizadas na modelagem do cancelamento de clientes são regressão logística, análise discriminante e análise de sobrevivência. Importante destacar que as duas primeiras técnicas permitem predizer apenas o cancelamento ou não dos clientes, enquanto a técnica de análise de sobrevivência amplia os resultados permitindo predizer ainda o tempo até o cancelamento. Os autores também revelam que a maioria dos estudos anteriormente realizados sobre o tópico teve suas análises baseadas em dados obtidos através de questionários de pesquisa (*survey*), podendo, portanto, estarem sujeitos a viés de seleção.

Para Van den Poel e Larivière (2003), os modelos preditivos de cancelamento de clientes podem ser divididos em modelos estáticos ou dinâmicos, dependendo do período de tempo em que os dados foram observados. Nos modelos estáticos, o comportamento dos clientes é observado em um momento específico no tempo, enquanto nos modelos dinâmicos o comportamento dos clientes é acompanhado ao longo do tempo, permitindo, portanto, que sejam estimados riscos que evoluem com o passar do tempo. Modelos dinâmicos, que permitem que a mesma variável assuma valores diferentes ao longo do tempo, produzem predições mais precisas do que os estáticos, devem ser preferidos.

Além de oferecerem melhores predições, modelos que levam em consideração não somente o evento do cancelamento, mas também o tempo até que este ocorra, são mais interessantes no contexto de Marketing, pois permitem que ações estratégicas sejam planejadas e implementadas no momento adequado, de forma a maximizar o potencial de reverter um possível cancelamento. Finalmente, modelos para predição de cancelamento de clientes utilizando análise de sobrevivência permitem que seja estimada a probabilidade de

cancelamento para cada instante no tempo, informação esta que pode ser utilizada no cálculo de outra métrica que vêm ganhando a atenção dos especialistas em marketing, o Valor Vitalício dos Clientes (VVC).

O Valor Vitalício do Cliente é uma medida do retorno financeiro que o cliente pode gerar para a empresa durante o período em que existir vínculos contratuais entre as partes (Rust, Lemon, Narayandas, 2005). Para que se possa calcular o Valor Vitalício do Cliente faz-se necessário modelar a margem de contribuição financeira gerada pelos clientes, estimar a probabilidade dos clientes permanecerem na carteira da empresa até um determinado período e determinar a taxa de desconto apropriada pela empresa para investimentos em marketing (Ferreira, 2007).

A equação 1, proposta por Ferreira (2007) com base nas discussões de Blattberg et. al (2000), define o Valor Vitalício para o cliente i, até o período T, em função da sua margem de contribuição financeira M(t) , da probabilidade S(t) de permanência na carteira de clientes até o período T, e de uma taxa de desconto (R) apropriada para os investimentos em marketing

$$VVC_i(T) = \sum_{t=0}^{T} S_i(t) M_i(t) (1 + R)^{-t}. \quad (1)$$

Sendo assim, este trabalho tem como objetivo a modelagem de um dos componentes necessários para o cálculo do VVC, a probabilidade de um cliente permanecer na base de clientes da empresa até um determinado tempo T, justificando a escolha da técnica de modelagem do cancelamento através de análise de sobrevivência, conforme será descrito nas próximas seções.

2.2. MÉTODOS DE ANÁLISE DE SOBREVIVÊNCIA

A análise de sobrevivência compreende uma classe de métodos estatísticos que permite modelar o tempo até a ocorrência de um determinado evento, sendo o evento de interesse neste caso, o cancelamento do cliente. Através da análise de sobrevivência é possível construir um modelo para predizer o risco de um evento ocorrer em função de diversas variáveis preditoras. Conforme destacado anteriormente, uma das principais vantagens da análise de sobrevivência frente à outras técnicas estatísticas como regressão logística e análise de discriminante está justamente na possibilidade de estimar o tempo até que um evento de

interesse ocorra, e não apenas a ocorrência ou não deste, limitação encontrada nas duas últimas técnicas (Kleinbaum e Klein, 2005).

Outra vantagem do método de análise de sobrevivência consiste em permitir o uso de informações de clientes mesmo que estes ainda não tenham experimentado o evento de interesse. Trata-se do conceito de dados censurados, em que não se tem a informação completa da observação para as análises, mas a parte da informação a qual se tem acesso é utilizada para as estimativas. Existem basicamente três tipos de censura: censura a direita, à esquerda e intervalar. A censura à direita ocorre quando o evento de interesse vem a acontecer em período posterior ao acompanhamento dos sujeitos (seria o caso de um cliente que cancelou contrato após o fim do levantamento de dados para análise e por isso seu cancelamento não foi contabilizado no estudo) ou quando algum evento e/ou circunstância fora do controle do pesquisador acaba eliminando este sujeito do conjunto de risco sem que se possa mais acompanhar a ocorrência do evento. Já a censura à esquerda ocorre quando a data de início é desconhecida, não ocorrendo no contexto deste trabalho, uma vez que se tem conhecimento da data de início de relacionamento com a empresa para todos os clientes. A censura intervalar ocorre quando ambas as censuras, à esquerda e à direita, estão presentes. As definições de censura devem ser cuidadosamente examinadas caso a caso, de acordo com o tipo de evento que está sendo modelado. Informações detalhadas sobre censura podem ser encontradas na literatura específica (Hosmer, Lemeshow e May, 2008).

Por ocupar-se em estimar o tempo até um evento, a abordagem de análise de sobrevivência pode ser classificada como probabilística ou estocástica. Os momentos no tempo em que os eventos de interesse ocorrem são considerados realizações de um processo aleatório e, portanto, um tempo t específico em que ocorreu o evento de cancelamento de um determinado cliente, é considerado uma variável aleatória com uma distribuição de probabilidade específica, muitas vezes desconhecida. A distribuição da variável aleatória T, onde T descreve o tempo até o cancelamento para um cliente específico pode ser expressa pela função densidade de probabilidade $f(t)$

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

A função densidade de probabilidade $f(t)$ serve de base para o cálculo de outras duas funções de suma importância para a análise de sobrevivência: a função de sobrevivência e a função risco. A função de sobrevivência denota a probabilidade de um cliente não cancelar o contrato antes do tempo t, podendo ser expressa como

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

em que a probabilidade de um cliente cancelar até um tempo t é representada pela função de distribuição acumulada $F(t)$. Teoricamente a curva de sobrevivência é representada por uma curva suave, porém na prática o seu gráfico possui uma aparência de escada (*stepfunction*), com degraus nos tempos em que ocorrem as falhas, neste caso, os cancelamentos (Kleinbaum e Klein, 2005).

Finalmente, a função risco representa o risco instantâneo de um cliente cancelar o contrato no intervalo de tempo $(t, t + \Delta t)$, dado que não cancelou o contrato até tempo t, definida 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 diferentes abordagens: tabelas de vida (estimador atuarial, Kaplan-Meier, Nelson-Aalen). No entanto, o cancelamento (evento de interesse para a análise de sobrevivência deste trabalho) depende de múltiplos fatores, e estes métodos geralmente não são adequados. Para avaliar o impacto simultâneo desses fatores é usual utilizar modelos semiparamétricos, como o modelo de risco proporcionais proposto por Cox. A seguir, serão descritos estes diferentes métodos e em virtude do objetivo da presente análise será justificada a escolha do modelo mais apropriado.

Modelos descritivos da função de sobrevivência: Kaplan-Meier, Tábuas de vida e Nelson-Aalen

Kaplan-Meier: também conhecido como estimador limite produto, incorpora as informações de todas as observações da base de dados (com censura ou não), considerando a sobrevivência até qualquer momento t ordenado no tempo como uma série de passos definidos até a ocorrência do evento ou censura. Com os dados observados, é estimada a probabilidade condicional de sobrevivência para cada período de observação no tempo, para então multiplicar-se as estimativas chegando a uma função geral de sobrevivência. (Hosmer, Lemeshow e May, 2008). A função de sobrevivência estimada é usualmente representada graficamente por uma curva (na prática uma função escada), permitindo a comparação das curvas de sobrevivência de grupos de interesse diferentes. Em um contexto de modelagem,

em que diversos preditores estão sendo avaliados, estes métodos podem ser utilizados para uma descrição inicial, por meio dos gráficos das funções de sobrevivência, pois a finalidade é descritiva e não inferencial. Além disso, uma limitação natural desse método encontra-se na necessidade de categorizar preditores quantitativos. ((Kleinbaum e Klein, 2005).

Uma alternativa ao método de Kaplan-Meier, especialmente para estudos com grandes bases de dados e eventos apresentados ao longo do calendário, consiste na utilização de estimadores de tabelas de vida. Este estimador é usado há mais de 100 anos para descrever a mortalidade humana e está entre os primeiros exemplos de aplicação de métodos estatísticos (Hosmer, Lemeshow e May, 2008). Tanto o método de Kaplan-Meier quanto as tabelas de vida, focam-se em estimar primeiramente a função de sobrevivência S(t). O método proposto por Nelson e Aalen, estima a função de sobrevivência acumulada e, posteriormente, S(t). Conhecido como estimador de Nelson-Aalen, o método é reconhecido na área pelo pioneirismo na abordagem da análise de sobrevivência sob uma abordagem de processo de contagem para derivação do estimador de risco H(t).

A diferença fundamental do método da tabela de vida (estimador atuarial) em relação ao Kaplan-Meier e Nelson-Aalen encontra-se na especificação dos intervalos. No método atuarial os intervalos são definidos pelo usuário, ao passo que nos dois últimos são definidos pelos tempos distintos nos quais se registra o evento de interesse (o cancelamento, no contexto deste trabalho). O método atuarial exige tamanhos de amostra grandes para produzir estimativas não viesadas da função de sobrevivência. Para tamanhos de amostra menores, os métodos de Kaplan-Meier e Nelson-Aalen têm um comportamento melhor, no sentido de gerar estimativas não viesadas da função de sobrevivência, com alguma vantagem para o estimador Kaplan-Meier.

Embora as técnicas descritivas de análise de sobrevivência sejam úteis em muitos contextos, todos os métodos previamente apresentados possuem limitações: não permitem a incorporação de múltiplas variáveis preditoras, as variáveis quantitativas precisam ser categorizadas, o que acarreta perda de informação, e os métodos descritivos de Kaplan-Meier e tábuas de vida não produzem estimativas de risco de cancelamento. Sendo assim, a utilização de modelos mais sofisticados e versáteis (como paramétricos ou semi-paramétricos) se torna muito atraente, pois permite a incorporação de múltiplas variáveis, inclusive quantitativas sem categorização, além de possibilitar a estimativa de risco de ocorrência do evento sob a perspectiva de múltiplas variáveis preditoras.

Modelos paramétricos e semi-paramétricos

Os modelos paramétricos permitem que sejam alcançados dois objetivos da análise de sobrevivência: descrever a distribuição de sobrevivência no tempo (componente do erro) e caracterizar como esta distribuição varia no tempo como função de outras variáveis associadas ao evento de interesse (componente sistemático do modelo). Em alguns casos é importante ter um modelo que cumpra ambos os objetivos, mas em outros cenários é necessário somente caracterizar as mudanças que variáveis associadas ao evento de interesse provocam na distribuição de sobrevivência no tempo (Hosmer, Lemeshow e May, 2008).

A grande dificuldade dos modelos paramétricos consiste justamente em o pesquisador saber, a priori, a distribuição de probabilidades do tempo até a ocorrência do evento de interesse. A forma da distribuição deve ser conhecida, com exceção dos valores dos parâmetros desconhecidos (Kleinbaum e Klein, 2005). Neste caso, as funções de densidade de probabilidade mais utilizadas são exponencial, Weibull e lognormal, pois apresentam flexibilidade para adaptação a uma variedade de situações (Carvalho et. al, 2005).

Todavia, quando não se conhece a priori a distribuição do tempo até o evento de interesse, mas se busca caracterizar as mudanças que variáveis associadas ao evento de interesse provocam nesta distribuição, podem ser utilizados os modelos semiparamétricos, uma alternativa robusta para estimar razão de azar e curvas de sobrevivência ajustadas.

Os modelos semiparamétricos apresentam a mesma estrutura de regressão encontrada nos modelos paramétricos, mas permitem que a distribuição do tempo de sobrevida não seja especificada. Nestes casos a determinação da distribuição do tempo até o evento de interesse pode vir a ser secundária, como no presente estudo de caso, pois o objetivo principal pode concentrar-se em determinar como variáveis associadas a tal evento modificam a função de sobrevivência e de risco das unidades observadas. O modelo semi-paramétrico mais popular na análise de sobrevivência é o Modelo de Cox.

Também conhecido por modelo de riscos proporcionais, este modelo semiparamétrico acabou conhecido como Modelo de Cox, pois sua proposição, em 1972, se deu por meio de um artigo do estatístico britânico Sir David Cox. A partir de então, o modelo passou a ser o mais utilizado na análise de dados de sobrevivência. Conforme o próprio nome indica, a suposição básica do modelo para estimação do efeito de variáveis no risco de um determinado evento ocorrer está na proporcionalidade dos riscos ao longo de todo o tempo de observação (Carvalho et al., 2005).

A formulação do modelo de Cox explicita que a função de risco no tempo t é um produto de dois termos: o risco basal $h_0(t)$ e um termo que depende de um vetor de covariáveis $\boldsymbol{x}$

$$h_i(t\,|\boldsymbol{x}) = h_0(t) exp\big(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\big). \quad (5)$$

Sendo assim, o modelo permite estimar para cada sujeito i o risco de ocorrência do evento de interesse no tempo t, dado que o sujeito apresenta os valores $\boldsymbol{x_i} = (x_{1i}, x_{2i}, \dots, x_{pi})'$ uma coleção de variáveis preditoras $X_1, X_2, \dots, X_p$ (Kleinbaum e Klein, 2005).

Conhecida como razão de azares (ou *hazard ratio* na literatura em língua inglesa), a razão entre o risco de ocorrência do evento de interesse de dois indivíduos (i e j) com variáveis preditoras $\boldsymbol{x_i} = (x_{1i}, x_{2i}, \dots, x_{pi})'$ e $\boldsymbol{x_j} = (x_{1j}, x_{2j}, \dots, x_{pj})'$ é expressa como

$$\frac{h_0(t)\exp(\boldsymbol{\beta x_1})}{h_0(t)\exp(\boldsymbol{\beta x_2})} = exp\,(\boldsymbol{\beta' x_1} - \boldsymbol{\beta' x_2}) = exp\big(\boldsymbol{\beta'}(\boldsymbol{x_1} - \boldsymbol{x_2})\big). \quad (6)$$

Essa formulação evidencia um efeito multiplicativo das variáveis preditoras na função de risco, o que está diretamente relacionado à suposição de riscos proporcionais inerente ao modelo. A razão é constante ao longo do tempo, uma vez que a parcela não-paramétrica da função de risco que depende do tempo, o risco basal $h_0(t)$, acaba sendo cancelada na operação de divisão do risco dos dois indivíduos (Carvalho et al., 2005).

A estimação do componente paramétrico $exp(\boldsymbol{\beta' x}) = exp\big(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\big)$ do modelo, descrito na equação 5, pode se dar de diversas maneiras, por máxima verossimilhança, verossimilhança parcial ou verossimilhança aproximada. A vantagem do método de verossimilhança parcial, amplamente utilizado no processo de estimação, consiste em estimar os coeficientes β_i do modelo de risco proporcional sem que haja necessidade de especificação do componente não paramétrico do modelo, a função de risco basal $h_0(t)$ (Allisson,2010).

A razão de azares pode ser interpretada de forma similar a razão de chances dos modelos de regressão logística. Para o caso de variáveis indicadoras (assumindo valores 0 ou 1 apenas), a razão de azares pode ser interpretada como a razão entre o risco estimado para aqueles indivíduos com variável indicadora=1 em relação àqueles indivíduos com variável indicadora=0, controlando para demais variáveis preditoras (Allisson,2010).

A formulação do modelo de Cox na equação 6 assume que o valor de todas as variáveis preditoras tenha sido medido em um único momento no tempo, em geral no início

do estudo ($t = 0$) permanecendo este valor durante todo o período de observação de um sujeito/unidade observacional. Todavia, existem situações em que os valores de uma ou mais variáveis preditoras são medidos ao longo do tempo de estudo, não apenas uma única vez, permitindo que a cada medição observe-se um valor diferente para cada uma destas variáveis. Neste tipo de cenário, o valor estimado para a função de risco pode depender mais dos valores recentemente observados para cada variável preditora do que do valor observado no início do estudo (Hosmer, Lemeshow e May, 2008).

Para permitir a incorporação no modelo de Cox de variáveis preditoras cujos valores mudam ao longo do tempo de estudo (variáveis conhecidas como tempo-dependentes) pode-se utilizar uma abordagem de modelo de Cox estendido. A formulação básica do modelo exemplificada na equação 6 permanece a mesma, mas é incorporado um novo vetor $\boldsymbol{z}(t)$ que abrange apenas aquelas variáveis tempo-dependentes:

$$h_i(t) = h_0(t) exp(\boldsymbol{\beta}'\boldsymbol{x} + \boldsymbol{\gamma}'\boldsymbol{z}(t)). \quad (7)$$

O vetor de variáveis $\boldsymbol{x}$ continua representado os valores observados para o i-ésimo sujeito naquelas variáveis que assumem apenas um único valor ao longo do tempo, enquanto a inclusão do vetor $\boldsymbol{z}(t)$ permite que sejam incorporadas informações sobre variáveis cujos valores para diferentes tempos no estudo variam (Kleinbaum e Klein, 2005). Os tipos mais comuns de variáveis tempo-dependentes são as medidas repetidas num mesmo indivíduo sob estudo (Carvalho et al., 2005).

Embora a modificação do modelo de Cox para permitir a inclusão de variáveis tempos-dependentes seja simples, o processo de estimação por verossimilhança parcial torna-se mais dispendioso, sendo a implementação prática do procedimento mais complexa (Allisson,2010).

Sendo assim, o ajuste de um modelo de estendido de Cox depende, muitas vezes, da disponibilidade de *software* que permita a organização do banco de dados para análise num processo de contagem, o que possibilita que as estimativas dos parâmetros do modelo sejam obtidas de forma simples.

2.3. ABORDAGEM DE PROCESSO DE CONTAGEM PARA ACOMODAR VARIÁVEIS TEMPO-DEPENDENTES

Em situações em que se deseja estudar o efeito de múltiplas variáveis tempo-dependentes no risco de ocorrência de um determinado evento é conveniente construir o banco de dados sob uma abordagem de processo de contagem. Um banco de dados organizado em processo de contagem possui múltiplos registros para um mesmo indivíduo, sendo cada registro correspondente a um intervalo de tempo em que as variáveis observadas permaneceram constantes. Embora este método de estruturação do banco de dados exija, algumas vezes, um nível considerável de conhecimento de programação, uma vez finalizado permite uma checagem fácil e direta de quaisquer erros na manipulação de dados (Allisson,2010).

As variáveis medidas no início do estudo, ou cujos valores não se alteram ao longo do tempo, são repetidas nos múltiplos registros de intervalos de tempo. O início e fim de cada um destes intervalos de tempo representam um período em que a observação continuava sob-risco de ocorrência do evento de interesse, sendo que tal evento só pode ocorrer no fim de cada intervalo (Allisson,2010). Um exemplo hipotético de transformação do banco de dados original para o formato de processo de contagem pode ser verificado na Figura 1.

Estrutura do Banco de Dados original

id cliente	Tempo estudo	Cancela	Canal de Venda	Filial Operadora	Prazo Pgto	Cartões mês 1	Cartões mês 2	Cartões mês 3
1	52	1	Venda Direta	PA	12	85	90	57
2	16	1	Venda Remota	Venda Remota	15	23	30	10
3	90	0	Venda Direta	BA	10	235	300	285

Estrutura do Banco de Dados em Processo de Contagem

id cliente	inicio	fim	Cancela	Canal de Venda	Filial Operadora	Prazo Pgto	Cartões
1	0	1	0	Venda Direta	PA	12	85
1	1	2	0	Venda Direta	PA	12	90
1	...		0	Venda Direta	PA	12	57
1	52	53	1	Venda Direta	PA	12	...
2	0	1	0	Venda Remota	Venda Remota	15	23
2	1	2	0	Venda Remota	Venda Remota	15	30
2	...		0	Venda Remota	Venda Remota	15	10
2	16	17	1	Venda Remota	Venda Remota	15	...
3	0	1	0	Venda Direta	BA	10	235
3	1	2	0	Venda Direta	BA	10	300
3	...		0	Venda Direta	BA	10	285
3	90	91	0	Venda Direta	BA	10	...

FIGURA 1 – Exemplos de estruturas de banco de dados para a análise de sobrevivência: no topo, estrutura usual e, abaixo, estrutura para a abordagem de processo de contagem.

Para um mesmo cliente, as múltiplas linhas do banco de dados referem-se à mesma observação, podendo surgir preocupação em relação à independência das observações, usualmente necessária para o funcionamento adequado dos processos de estimação dos parâmetros do modelo. Todavia, como os intervalos de tempo definidos para cada sujeito do banco de dados são disjuntos, os resultados da estimação são completamente válidos. A garantia desses resultados pode ser explicada pelo fato do cálculo da verossimilhança parcial considerar no máximo uma observação de cada sujeito do banco de dados em qualquer momento (Carvalho et al., 2005).

Após a construção do banco de dados na estrutura de processo de contagem, as variáveis tempo-dependentes passam a ser tratadas normalmente no procedimento de ajuste do modelo. A diferença no procedimento de ajuste do modelo restringe-se a modificações na variável dependente (tempo até desfecho de interesse), que neste caso passa a ser intervalar, uma vez que para cada indivíduo existem múltiplos intervalos de tempo em que foram observadas as variáveis. No caso do *software* para análises estatísticas SAS (*Statistical Analysis System*), utilizado neste trabalho, em que o ajuste de modelo de Cox se da por meio da utilização do procedimento PROC PHREG, a única diferença na especificação do modelo a ser ajustado consiste na utilização do início e fim de cada faixa de intervalo de tempo definida como variáveis dependentes, ao invés da especificação de uma única variável dependente de tempo até ocorrência do evento ou censura (Allisson,2010).

2.4. TRUNCAMENTO

Além de permitir que o modelo de Cox estendido seja ajustado de forma simples, a abordagem de processo de contagem permite ainda a fácil incorporação de truncamento à esquerda nas análises, recurso este que será utilizado no estudo de caso do presente trabalho. Conforme descrito anteriormente, o processo de contagem pressupõe a divisão do tempo de estudo em intervalos de tempo em que são registrados os valores observados para cada variável de um determinado sujeito. Sendo assim, cada um desses intervalos tem um tempo de início e fim, sendo que o início do intervalo não precisa necessariamente ser zero, o que torna a incorporação de truncamento uma tarefa fácil (Hosmer, Lemeshow e May, 2008).

Diferentemente da censura à esquerda, que ocorre aleatoriamente no nível de cada unidade observada no estudo, o truncamento à esquerda envolve um processo definido de seleção que permeia todas as unidades de observação do estudo. Este processo de seleção leva em consideração o período de exposição ao risco de ocorrência do evento de interesse para

cada unidade de observação. No caso do truncamento à esquerda, as unidades observadas no estudo não estão expostas ao risco até que certo tempo tenha decorrido desde sua origem.

Numa abordagem padrão de verossimilhança parcial pressupõe-se que qualquer sujeito/unidade de observação esteja exposta ao risco de ocorrência do evento de interesse desde o tempo zero e continue exposta ao risco até que este evento ocorra ou o estudo termine (caso de censura à direita). A partir deste ponto, o sujeito/unidade de observação seria removido do conjunto de risco, não podendo mais retornar. Todavia, a pressuposição de que qualquer sujeito esteja exposto ao risco desde o tempo zero até o registro do evento ou censura não se faz necessária. O método de verossimilhança parcial permite que unidades de observação que não estavam no conjunto de risco em um determinado período de tempo sejam desconsideradas para aquele período e passem a ser consideradas posteriormente, em outro período de tempo, acomodando, portanto, o truncamento à esquerda (Allisson,2010). Nestes casos, sabe-se quanto tempo se passou desde a origem da observação, mas se utiliza informações sobre as variáveis observadas somente durante o tempo de estudo.

O conceito de truncamento e a estruturação dos dados em um processo de contagem serviram de base para a construção da base de dados que será analisada por meio de um modelo de Cox estendido no estudo de caso apresentado no Capítulo 4.

3. OBJETIVOS

Gerais:

- ✓ Modelar o tempo até o cancelamento do contrato dos clientes para estimar risco e identificar fatores associados ao cancelamento; e,
- ✓ A partir do modelo estimado, iniciar discussão na empresa sobre a potencialidade do modelo ajustado para a estimação do valor vitalício do cliente.

Específico:

- ✓ Desenvolver um algoritmo para criar computacionalmente o banco de dados para utilização do modelo de Cox com preditores com medidas repetidas ao longo do tempo (tempo-dependentes) na abordagem de processo de contagem.

4. ESTUDO DE CASO

O modelo de Cox Estendido descrito anteriormente será aplicado a uma base de dados de clientes da Empresa X de forma a idendificar os fatores associados ao cancelamento de clientes (desfecho de interesse deste trabalho). Para o estudo foi considerado um único produto da Empresa X, o cartão para serviços de gestão de frota. Optou-se por segmentar o estudo apenas a este produto devido a sua importância estratégica para empresa e em virtude deste exigir um maior número de informações do cliente no momento das transações, fornecendo, consequentemente, um maior número de informações sobre comportamento de compra e de gestão por parte da empresa cliente.

4.1. OBTENÇÃO DA BASE DE DADOS

A base de dados foi obtida, com o consentimento da empresa mediante assinatura de termo de confidencialidade, com suporte da área de Tecnologia da Informação. A autora do trabalho desenvolveu as consultas necessárias para extração da base de dados no software de *Business Intelligence* próprio da empresa, construindo a estrutura do banco de dados a ser extraído: listando as variáveis necessárias e aplicando filtros. Uma vez construída a estrutura do banco de dados, o processamento se deu com o auxílio da equipe de TI (Tecnologia da Informação) em virtude do grande volume de dados a ser extraído.

A extração da base de dados gerou arquivos em Excel (formato xls), agrupados por ano de transação. Com as bases de dados transacionais e a planilha de variáveis fixas ao longo do tempo em mãos, criou-se uma rotina computacional na linguagem do programa SAS (macro) de importação, de forma a facilitar o agrupamento dos dados por cliente. Além de importar todos os bancos de dados originalmente em Excel para o software SAS, estas rotinas computacionais permitiram que todos os dados fossem consolidados em um único arquivo de banco de dados respeitando a unidade de cada cliente.

4.2. DESCRIÇÃO DAS VARIÁVEIS DO ESTUDO

Para garantir confidencialidade das informações, as variáveis utilizadas nas análises deste trabalho foram mascaradas nesta versão pública do trabalho. Para as variáveis quantitativas foram realizadas transformações. A base de dados analisada compreende 9

variáveis, sendo 4 destas fixas no tempo (identificadas na Tabela 1 pelo rótulo X), ou seja, não variam de acordo com o tempo e 5 variáveis tempo-dependentes (identificadas pelo rótulo Z), ou seja, que variam ao longo do tempo.

Tabela 1 – Descrição das variáveis do estudo,

Rótulo	**Descrição**	**Tempo Dependente**
X1	Variável categórica: 6 classes	não
X2	Variável categórica: 15 níveis	não
X3	Variável quantitativa discreta	não
X4	Variável categórica: 10 níveis	não
Z1	Variável quantitativa discreta	sim
Z2	Variável quantitativa contínua	sim
Z3	Variável quantitativa contínua	sim
Z4	Variável quantitativa contínua	sim
Z5	Variável quantitativa contínua	sim

As variáveis tempo-dependentes foram consolidadas de forma mensal, pois esta é uma prática comum nas análises da Empresa X, que apresenta seus indicadores consolidados mensalmente, o que garante ao trabalho interpretabilidade e comparabilidade no contexto do negócio. Vale ressaltar ainda que a variável X4 era originalmente quantitativa contínua, mas em virtude de inconsistência de dados acabou sendo agrupada em classes (não especificadas nesta versão do trabalho, por motivo de confidencialidade).

4.3. DEFINIÇÕES DE ANÁLISE DE SOBREVIVÊNCIA APLICADAS AO ESTUDO

Além das variáveis anteriormente descritas na Tabela 1, o banco de dados conta com duas variáveis chaves para a análise de sobrevivência: a data de início de relacionamento de cada cliente com a empresa e sua respectiva data de cancelamento. A data de cancelamento configura o desfecho de interesse do presente estudo, trata-se do evento de interesse sob o ponto de vista de análise de sobrevivência, uma vez que marca o fim do relacionamento entre a empresa e o cliente em virtude da rescisão contratual. Para preservar a confidencialidade das

informações, na versão pública deste trabalho a data de parada da observação dos dados foi omitida, sendo genericamente designada como "data final" (denominada t_5 na Figura 2). Assim, para todos os clientes com contrato ativo além desta data, o tempo desde a primeira transação é definido como censurado à direita.

Embora a base de dados extraída do sistema interno (*data warehouse)* da empresa compreenda todo o período de histórico cadastral e de transações desde o início da operação da Empresa X, em virtude do cenário e portfólio de produtos da companhia no início de sua operação ser bastante diferente do que se tem hoje, optou-se por restringir a base de dados para análise ao período transacional posterior à "data de início" (denominada t_2 na Figura 2 por motivos de confidencialidade), de forma a garantir um cenário homogêneo de oferta do produto em questão, evitando que diferenças oriundas do período de consolidação da empresa afetassem as análises.

O processo de seleção aplicado em que se considera apenas a exposição ao risco de cancelamento a partir de um dado momento após a origem (t_2) configura truncamento à esquerda para aqueles clientes que ingressaram na carteira da empresa antes de t_2, conforme discutido anteriormente na seção de revisão da literatura. Desta forma, para cada cliente na base de dados para análise existem dois registros de tempo: o tempo total de relacionamento com a empresa (medido em meses) e o tempo de estudo (também medido em meses). O tempo total consiste na diferença entre a data de entrada de cada cliente e sua respectiva data de cancelamento ou fim do estudo (t_5). Já o tempo de estudo configura apenas o período em que foram observados os dados do cliente, estando, portanto, sujeito ao corte pré-definido de truncamento dos dados anteriores a t_2. No caso de clientes cuja data de ativação é posterior a t_2, o tempo total e de estudo será o mesmo, pois os mesmos não estão sujeitos a truncamento. Para clientes cuja data de ativação é anterior a esta data de truncamento, todavia, o tempo de estudo será menor do que o tempo total, pois serão reduzidos deste último os meses em que o cliente já estava ativo, porém seus dados não foram incluídos na análise. A Figura 2 permite entender melhor a dinâmica de registro do tempo até o cancelamento aplicado ao estudo:

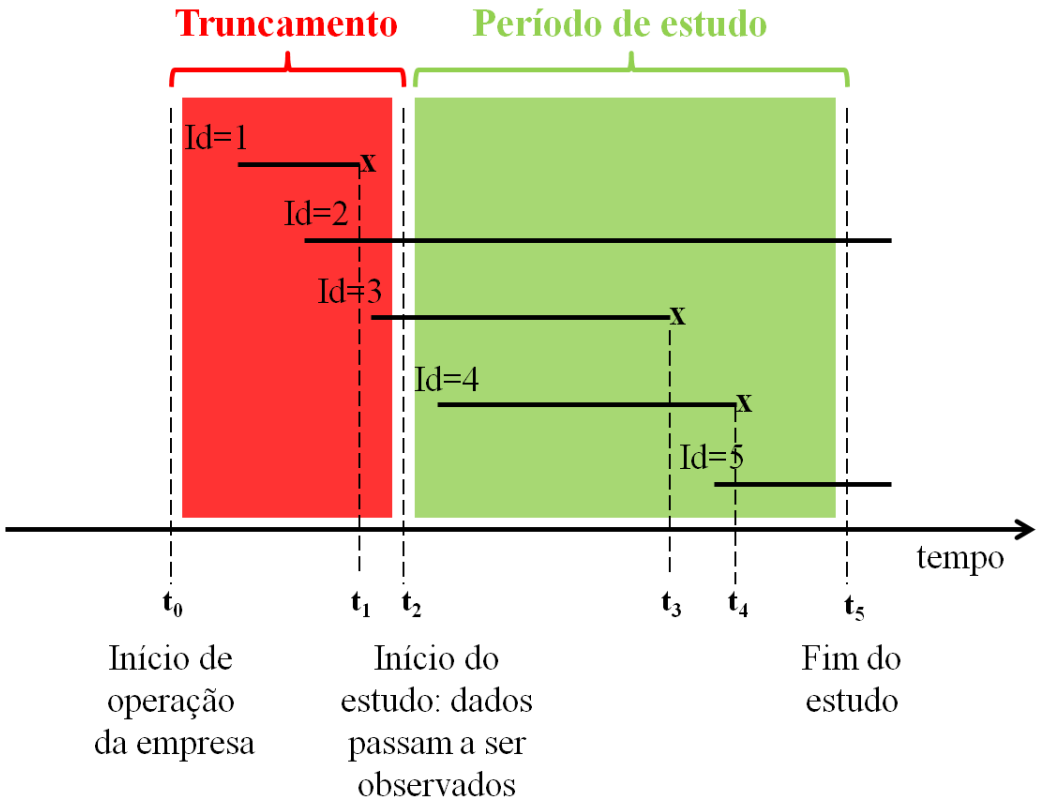


Figura 2 – Exemplificação do tempo total e tempo de estudo

Na Figura 2 são exemplificadas cinco possibilidades diferentes de disposição do tempo de relacionamento com a Empresa X encontradas na base de dados.

- Cliente hipoteticamente identificado com id=1: o tempo total de relacionamento entre cliente e empresa foi de 15 meses (t_1-t_0), mas o tempo de estudo foi de 0 meses, pois o cancelamento ocorreu ainda no período de truncamento (anterior a t_2, demarcado em vermelho na figura). Neste caso, o cliente não foi incorporado às análises.
- Cliente hipoteticamente identificado com id=2: o tempo total entre entrada e censura (fim do estudo) deste cliente corresponde a 136 meses (entrada até t_5), todavia seu tempo de estudo é de apenas 119 meses (área demarcada em verde na figura, que corresponde a t_5-t_2), pois os primeiros 17 meses do relacionamento do cliente com a empresa sofreram truncamento (área demarcada em vermelho na figura). Sabe-se, portanto, que estes 17 meses ocorreram, mas não se analisa os dados transacionais referentes a este período.
- Cliente hipoteticamente identificado com id=3: o tempo total entre a entrada do cliente e o cancelamento foi de 62 meses (entrada até t_3), todavia seu tempo de

estudo é de 55 meses (área demarcada em verde na figura, correspondente a t_3-t_2), pois os primeiros 7 meses do relacionamento do cliente com a empresa antecedem t_2 e portanto estão sujeitos ao truncamento (área demarcada em vermelho na figura).

- Cliente hipoteticamente identificado com id=4: tempo total decorrido entre a entrada na carteira e o cancelamento (t_4) foi de 61 meses e como sua data de ativação já ocorreu após a data de truncamento (t_2) não existe período de truncamento, portanto seu tempo de estudo também é de 61 meses, igual ao seu tempo total.
- Cliente hipoteticamente identificado com id=5: o tempo total é igual ao tempo de estudo, 55 meses, pois todos os registros do cliente estão na área demarcada em verde na figura, pois este não era cliente durante o período de truncamento (t_0 até t_2).

4.4. EXCLUSÃO DE CASOS

Além dos casos de censura e truncamento descritos na subseção anterior, foram excluídos do banco de dados clientes sem registro de data de entrada na carteira e clientes com data de entrada a partir do último mês considerado no estudo t_3 (omitida por motivo de confidencialidade), garantindo assim, que todos os clientes tenham um histórico de pelo menos um mês de relação empresa-cliente, pois caso contrário o relacionamento destes clientes com a empresa seria praticamente nulo, não contribuindo para o modelo de análise de sobrevivência.

Foram excluídos também clientes cuja data de cancelamento dependia de fatores contratuais pré-determinados, pois o relacionamento entre a Empresa X e o cliente estaria sujeito ao fim (cancelamento) em virtude do período de vigência de um termo, e não por iniciativa do cliente, o que do ponto de vista estratégico não agregaria informações ao modelo estatístico. Visando a confidencialidade das informações, o tamanho de amostra final, utilizado nas análises, também foi omitido nesta versão pública do trabalho.

4.5. ANÁLISE DESCRITIVA DOS DADOS

Uma vez concluídas as exclusões anteriormente citadas, realizou-se uma análise das variáveis que não dependem do tempo, de forma a detectar possíveis problemas nos dados e entender a distribuição dos clientes conforme algumas variáveis cadastrais. As variáveis tempo-dependentes, em virtude da sua estrutura de medidas repetidas (organizadas num

processo de contagem) não terão suas medidas resumo apresentadas, pois o interesse recai sobre a distribuição destas para cada cliente, não havendo, portanto, sentido prático na apresentação destas medidas resumo de forma global.

Tabela 2 – Distribuição de frequência da variável X1

X1	**(%)**
Categoria 1	69,4
Categoria 2	16,8
Categoria 3	9,3
Categoria 4	4,2
Outros	0,3

Verifica-se na Tabela 2 que a categoria de clientes 1 agrupa a maior parcela de clientes (69,4%) do estudo. A categoria "outros" foi criada com o objetivo de agrupar clientes cuja classificação de acordo com o critério X1 refletia inconsistência de dados.

Tabela 3 – Distribuição de frequência da variável X2

X2	**(%)**
Categoria 1	16,7
Categoria 2	13,0
Categoria 3	11,1
Categoria 4	9,3
Categoria 5	7,8
Categoria 6	6,9
Categoria 7	5,8
Categoria 8	5,7
Categoria 9	5,5
Categoria 10	5,3
Categoria 11	5,1
Categoria 12	4,2
Categoria 13	2,8
Categoria 14	0,5
Outros	0,3

A partir da Tabela 3 tem-se um detalhamento das categorias de X2, anteriormente representadas pela sua variável resumo (X1) na Tabela 2. Verifica-se que as categorias 2 e 3 são as grandes responsáveis pela representatividade da categoria 1 de X1, contabilizando 13,0% e 11,1% dos clientes da empresa respectivamente. Novamente foi criada a categoria "outros" para agrupamento daqueles clientes cujos registros pareciam inconsistentes. Embora

tenha-se verificado os clientes com estes registros em outras bases de dados da empresa não foi possível melhorar o dado.

A variável X3 possui natureza quantitativa discreta, e através de análise descritiva, verificou-se que o mínimo de X3 para os clientes da base de dados é de 2 unidades, enquanto o máximo consiste em 30 unidades. A mediana de X3 é de 14 unidades, mas vale ressaltar esta acabou influenciada pela manipulação de dados feita para todos os casos de clientes que não possuíam valor registrado para X3 na base de dados (11,7% dos clientes da base), pois estes casos receberam imputação da moda, que era justamente 14 unidades. Finalmente, na Tabela 4 é apresentada a descrição da variável X4.

Tabela 4 – Distribuição de frequência da variável X4

X4	**(%)**
Categoria 1	22,4
Categoria 2	20,6
Categoria 3	16,9
Categoria 4	11,2
Categoria 5	10,9
Categoria 6	10,5
Categoria 7	5,1
Categoria 8	1,6
Categoria 9	0,7
Categoria 10	0,1

Embora a variável X4 seja por natureza quantitativa contínua, optou-se por agrupá-la em faixas em virtude das inconsistências de base dados (muitos clientes apresentavam valores que não eram plausíveis). A Tabela 4 consolida a descrição destas faixas deixando claro que as faixas 1 e 2 agrupam o maior número de clientes (responsáveis por aproximadamente 43% dos clientes).

4.6. AJUSTE MODELO

O primeiro passo executado para que fosse possível ajustar o modelo de Cox estendido foi a estruturação da base de dados em processo de contagem. Para isso, foi implementada uma rotina computacional específica escrita na linguagem do programa SAS. Com o banco de dados devidamente montado no formato de processo de contagem, procedeu-se com o ajuste do modelo. A rotina computacional utilizada para o ajuste do modelo final é mostrada abaixo.

```
proc phreg data=BaseFINAL;
     CLASS X2 (ref='Categoria 4') / param=ref;
     model (START, STOP)*CANCELA(0) =  X2 X3 Z2 Z4 Z5 / rl ties=efron;
     output out=residuos xbeta=xbeta resmart=mart dfbeta=dbetax2_1-
dbetax2_14 dbetax3 dbetaZ2 dbetaZ4 dbetaZ5;
     id id TempoEstudo Cancela;
run;
```

O processo para seleção das variáveis iniciou com a modelagem de todas as variáveis disponíveis no banco de dados e estas foram sendo retiradas do modelo, sucessivamente, à medida que não contribuíam significativamente na estimação do risco de cancelamento. Tal contribuição para o ajuste do modelo foi medida através da estatística de Wald, ao nível de significância de 5%.

Finalizada a seleção das variáveis, o modelo final ajustado, contendo apenas àquelas variáveis significativas a um nível de 5%, considerou 5 variáveis, sendo duas destas fixas no tempo (X2 e X3) e três tempo-dependentes (Z2, Z4 e Z5). As estimativas de risco relativo de cancelamento (*hazard ratio* – HR) com base nestas 5 variáveis preditoras podem ser observadas na Tabela 5 a seguir.

Tabela 5 – Estimativas de risco relativo de cancelamento

Variável	**Risco Relativo**		
	Estimativa Pontual	**Intervalo Confiança (95%)**	
X2 (ref= Categoria 4)			
Categoria 1	2,30	1,62	3,26
Categoria 2	4,22	3,07	5,80
Categoria 3	3,84	2,79	5,28
Categoria 5	3,18	2,19	4,61
Categoria 6	3,47	2,53	4,76
Categoria 7	5,92	3,63	9,66
Categoria 8	3,02	2,18	4,19
Categoria 9	3,57	2,62	4,88
Categoria 10	3,67	2,69	5,01
Categoria 11	4,43	3,23	6,08
Categoria 12	4,15	3,06	5,64
Categoria 13	3,84	2,83	5,23
Categoria 14	2,29	1,69	3,10
Outros	9,63	6,39	14,53
X3	1,03	1,02	1,04
Z2	0,46	0,39	0,54
Z4	0,34	0,31	0,37
Z5	0,24	0,16	0,34

As estimativas apresentadas na Tabela 5 permitem interpretar a influência de cada uma das varáveis do modelo no risco de cancelamento. No caso da variável X2, percebe-se que todas as categorias apresentam estimativa de risco relativo maiores do que 1 em relação à categoria de referência (4). Por exemplo, clientes categoria 1 apresentam em média risco de cancelamento 130% [(2,30-1) x 100] (IC 95%: 60,x%-225,9%) superior aos clientes da categoria 4 ajustando para as demais variáveis. As demais interpretações são feitas de maneira análoga.

Já em relação à variável quantitativa X3, percebe-se que a cada incremento de 1 unidade nesta variável, o risco de cancelamento cresce 3% (IC 95%: 2,3%-4,0%) . No caso da variável Z2 verifica-se que para cada incremento de 1 unidade o risco de cancelamento diminui em 54,5% (IC 95%: 46,5%-61,4%). Observa-se a mesma tendência para as variáveis Z4 e Z5. Para cada incremento de 1 unidade em Z4 por mês estima-se que o risco de cancelamento reduza 66,4% (IC 95%: 63,0%-69,5%) , enquanto que para um incremento de 1 unidade em Z5 estima-se que o risco de cancelamento recue 76,4% (IC 95%: 65,7%-83,7%).

4.7. AVALIAÇÃO DO MODELO

De forma a avaliar a qualidade de ajuste do modelo final procedeu-se com uma análise gráfica dos resíduos de Martingale e dos *dfbetas*. Os gráficos de análise dos resíduos e da medida *dfbeta* em relação aos clientes (ids do estudo) e ao tempo podem ser consultados na seção de apêndice. De maneira geral, parece não haver discrepâncias importantes no modelo, exceto para alguns clientes que merecem investigação por apresentarem comportamento diferenciado, podendo sinalizar a presença de *outliers*. Há um aparente aumento na variabilidade dos resíduos martingales com a evolução do tempo, que poderia estar relacionado com o aumento da capilaridade da rede credenciada. Ainda, como passar do tempo, foram sendo estruturados outros canais de venda, trazendo um perfil de clientes mais heterogêneo à carteira. Estes aspectos, juntamente com o refinamento do modelo, merecem uma discussão mais profunda.

5. CONSIDERAÇÕES FINAIS

A abordagem utilizada para modelar o tempo até o cancelamento mostrou resultados satisfatórios e com grande potencial para utilização após refinamento pela empresa. O que se esperaria do modelo é que fosse capaz de identificar clientes que estão "prestes" a cancelar, expectativa esta alcançada neste estudo conforme resultados previamente apresentados acerca da qualidade de discriminação do modelo.

Não foram apenas os resultados do modelo que contribuíram para a compreensão dos fatores associados à perda de clientes pela empresa. O processo como um todo, desde a etapa de obtenção dos dados, limpeza e de verificação de inconsistências nas variáveis no banco de dados também foram importantes e peças fundamentais para entender quais tipos de informações estão disponíveis nos sistemas da empresa e de que forma elas estão organizadas. Neste sentido, foram identificadas possibilidades de melhoria na coleta e organização dos dados que podem ser implementadas internamente de forma a garantir análises mais confiáveis e seguras no futuro.

Embora os resultados preliminares deste trabalho já forneçam insumos interessantes sobre o cancelamento de clientes para a Empresa X, algumas limitações de estudo foram observadas, podendo ser trabalhadas e transpostas num segundo momento de refinamento do modelo. A primeira destas seria a possibilidade de existência de correlação entre códigos de clientes de um mesmo grupo econômico. Todavia, como o cancelamento ocorre no nível de código de cliente, não foi levada em consideração neste momento uma estrutura de covariância, que poderia ser investigada futuramente. Além disso, seria conveniente uma análise mais aprofundada daquelas observações cujos *dfbetas* e resíduos de Martingale apresentaram valores relativamente discrepantes de forma a validar a consistência das estimativas antes de fazer inferências. Finalmente, como parte do refinamento do modelo, especialmente para predição de cancelamento, uma abordagem de validação do mesmo precisa ser considerada, por exemplo, por meio técnicas como *split sampling* ou *bootstraping*.

Uma vez investigadas e corrigidas as possíveis limitações, pode-se focar esforços no potencial refinamento do modelo. Nesse sentido, sugere-se a verificação da suposição de linearidade para as variáveis tempo-dependentes e um estudo dos métodos disponíveis para contornar essa situação. Além disso, o modelo poderia ser ampliado para permitir a inclusão de novas variáveis comportamentais dos clientes a fim de potencializar um melhor ajuste e poder de discriminação. Uma das sugestões seria, inclusive, explorar no modelo variáveis sob

uma abordagem de defasagem no tempo (*lags*), para predizer de forma mais acurada e prática o risco de cancelamento, possibilitando até mesmo uma abordagem de controle da influência no cancelamento do crescimento/queda nos valores observados para certas variáveis entre unidades de tempo específicas.

Os aspectos aqui mencionados para melhoria do modelo não foram realizados em parte pelo período restrito de tempo para estruturação da base de dados, bem como pelo grande volume de dados a ser extraído e das etapas computacionais de manipulação, com checagem constante dos dados para estruturação da base de dados em processo de contagem. Sendo assim, talvez uma das principais contribuições deste trabalho seja a apropriação dos conceitos e do como fazer o algoritmo para manipulação da base de dados e ajuste do modelo. As potenciais limitações descritas anteriormente podem ser superadas com o refinamento do modelo de forma relativamente simples. Para isso, será necessária uma discussão profunda acerca dos aspectos a serem trabalhados junto à empresa para um posterior desenvolvimento e implementação de um sistema de monitoramento dos clientes com respeito ao risco de cancelamento.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ALLISON, P. D. **Survival analysis using SAS: a practical guide**. Sas Institute, 2010.

BLATTBERG, R. C., GETZ, G., THOMAS J. **Customer Equity – Building and Managing Relationships as Value Assets**. Boston, Massachusetts: Harvard Business School Press, 2001.

CANHOTO, P.; JESUS, M.; RAMOS, Célia MQ. **Sistema de informação para a gestão de uma frota**. 2005. Disponível em: < http://hdl.handle.net/10400.1/1091>. Acesso em: 20 mai. 2015.

CARVALHO, M. **Análise de sobrevida: teoria e aplicações em saúde.** Rio de Janeiro: Editora Fiocruz, 2005.

DENATRAN. Anuário Estatístico Frota de Veículos – Frota Nacional (Abril de 2015). Brasília, 2015. Disponível em: < http://www.denatran.gov.br/frota2015.htm>. Acesso em: 20 mai. 2015.

FERREIRA, E. C. **Um modelo quantitativo para o valor do cliente**. Tese de Doutorado, 2007.

HOSMER, D. W.; LEMESHOW, S.; M., Susanne. **Applied survival analysis**. Hoboken. 2008.

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

KOTLER, P. **Administração de marketing**. 10. ed. São Paulo: Prentice Hall, 2000.

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

REICHHELD, F. F. **The Loyalty Effect: The Hidden Force Behind Growth**, Profits and Lasting Value. Boston: Harvard Business School Press, 1996.

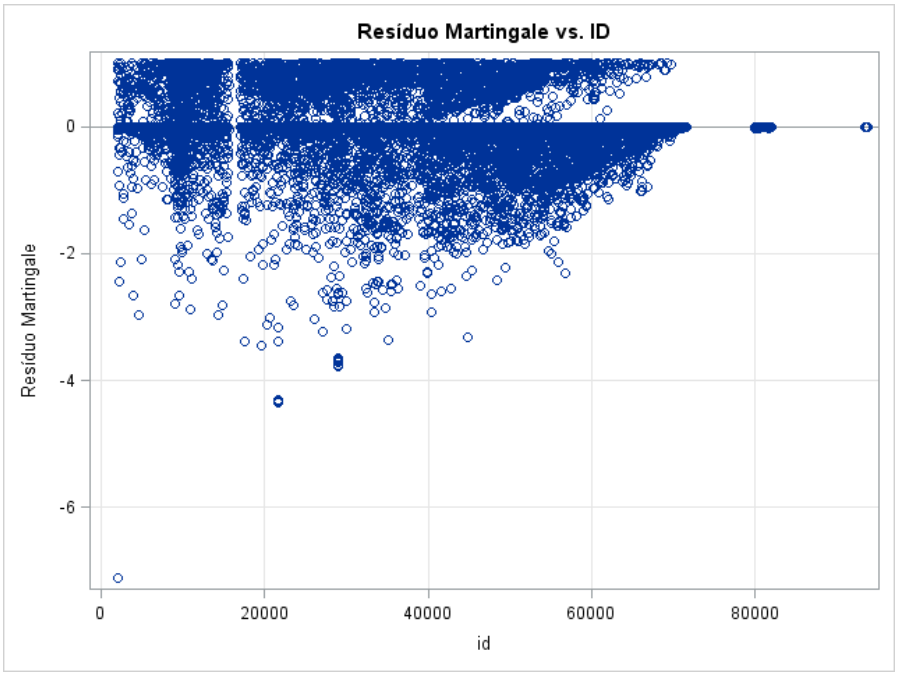
RUST, R., LEMON K., NARAYANDAS, D. **Customer Equity Management**. Upper Saddle River, NJ: Pearson/Prentice Hall, 2005.

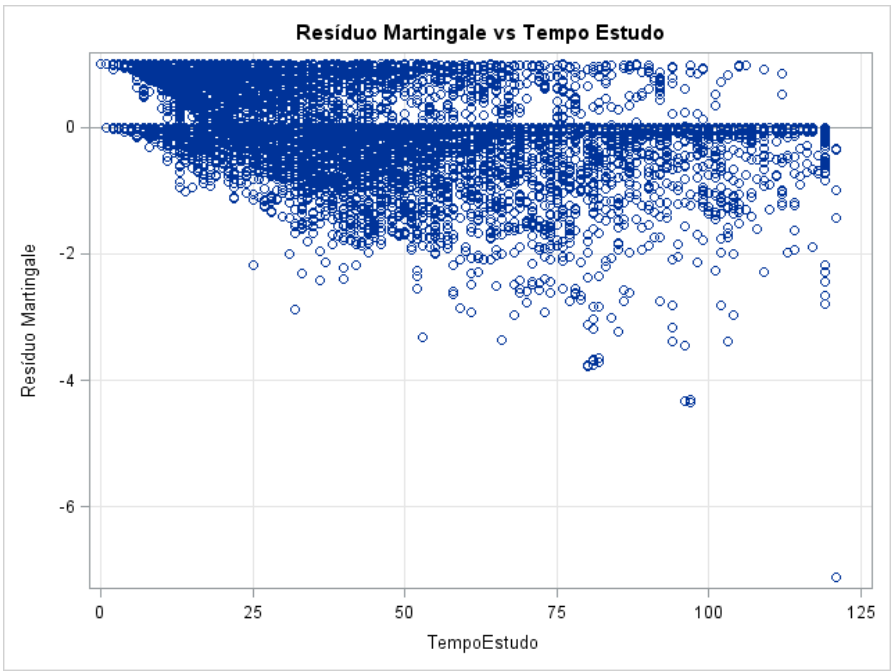
SILVA, M.. **O que está acontecendo no segmento de terceirização de frotas leve no Brasil?** Portal Administradores, 2014. Disponível em: < http://www.administradores.com.br/artigos/negocios/o-que-esta-acontecendo-no-segmento-de-terceirizacao-de-frotas-leve-no-brasil/76658/>. Acesso em: 20 mai. 2015.

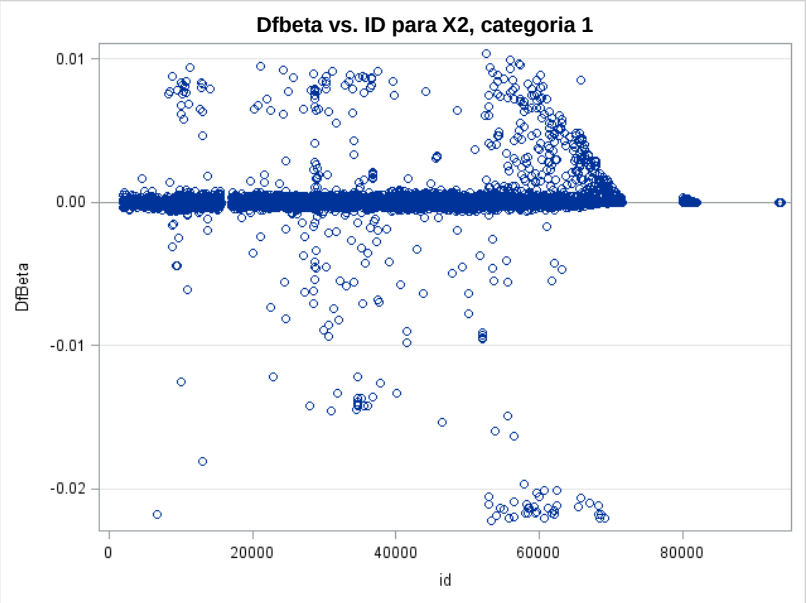
VALENTE, A. M. **Gerenciamento de transportes e frotas**. 2 ed. Editora Cengage Learning, 2008.

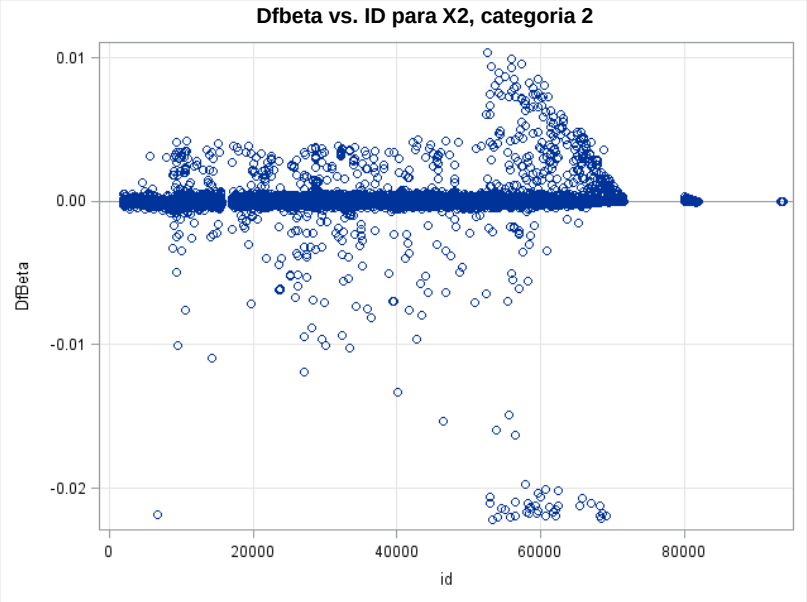
7. APÊNDICE

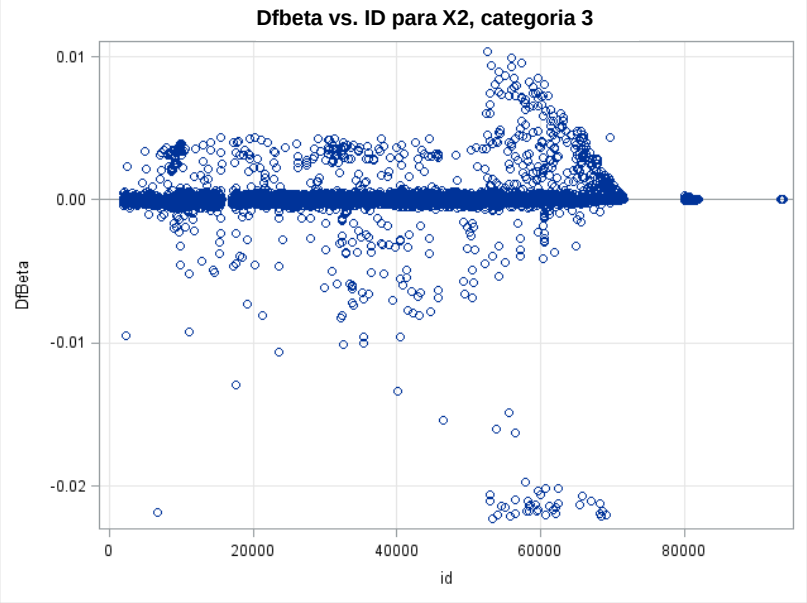
Gráficos dos resíduos de martingale e *dfbetas*

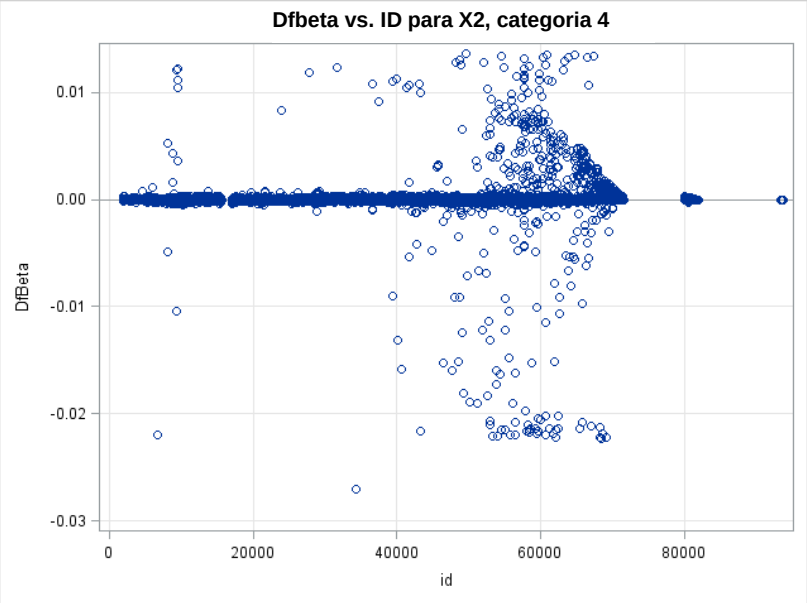


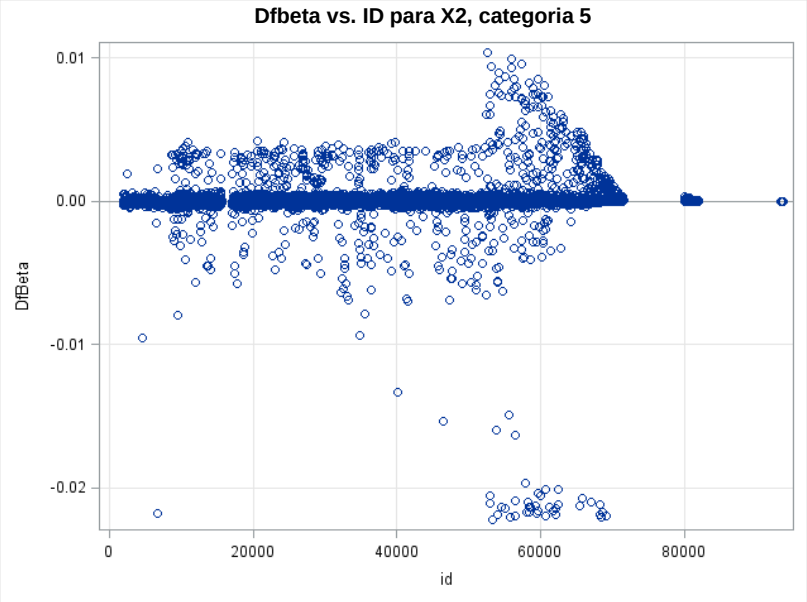


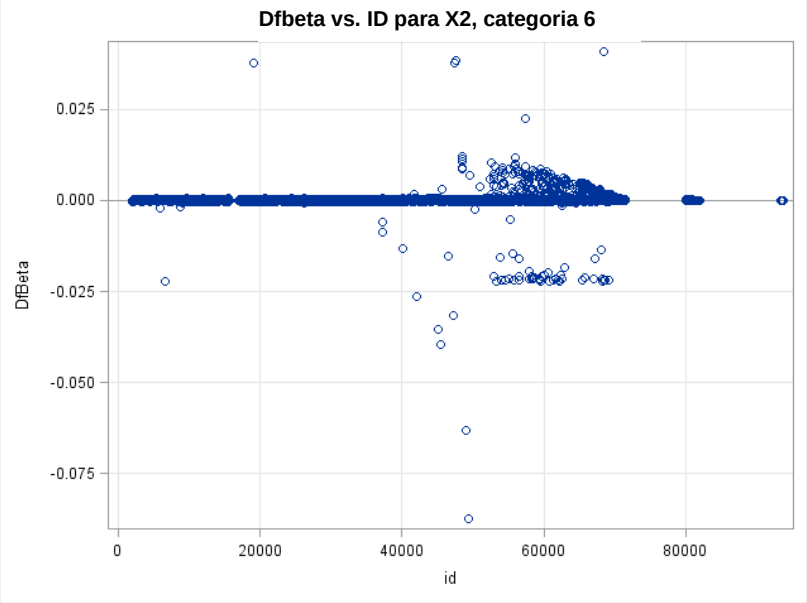

Dfbeta vs. ID para X2, categoria 1
DfBeta
0.01
0.00
-0.01
-0.02
0
20000
40000
60000
80000
id

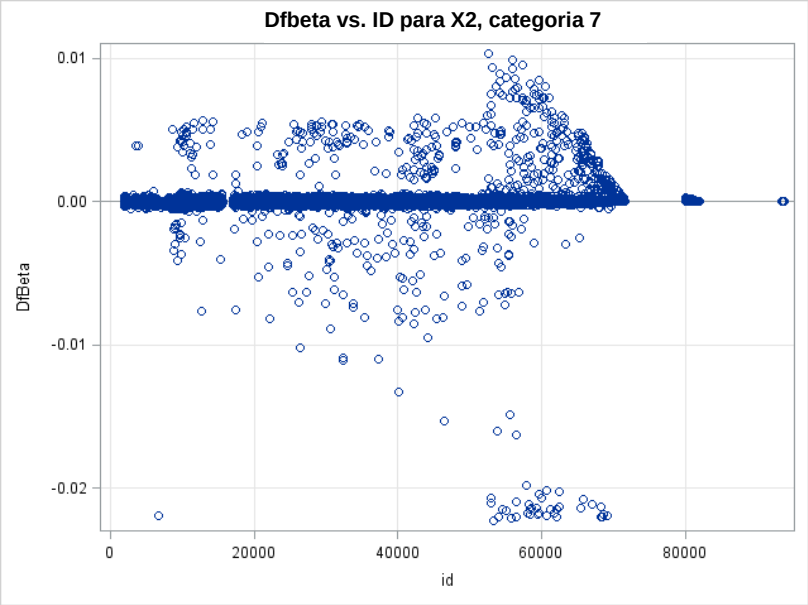

Dfbeta vs. ID para X2, categoria 2
DfBeta
0.01
0.00
-0.01
-0.02
0
20000
40000
60000
80000
id

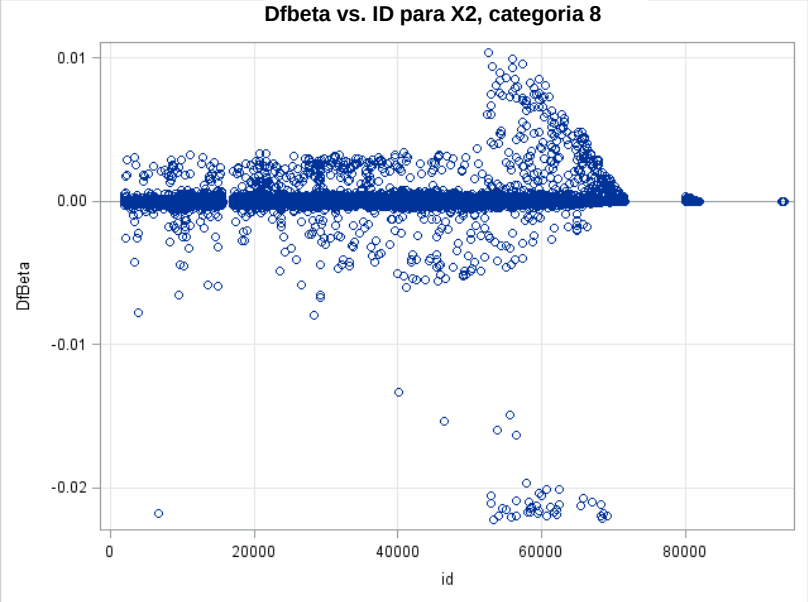

Dfbeta vs. ID para X2, categoria 3
DfBeta
0.01
0.00
-0.01
-0.02
0
20000
40000
60000
80000
id

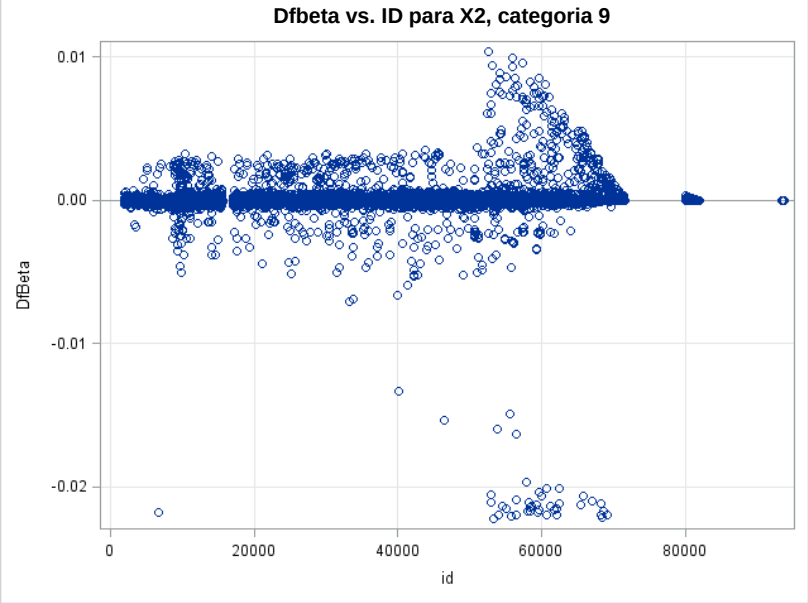

Dfbeta vs. ID para X2, categoria 4
DfBeta
0.01
0.00
-0.01
-0.02
-0.03
0
20000
40000
60000
80000
id

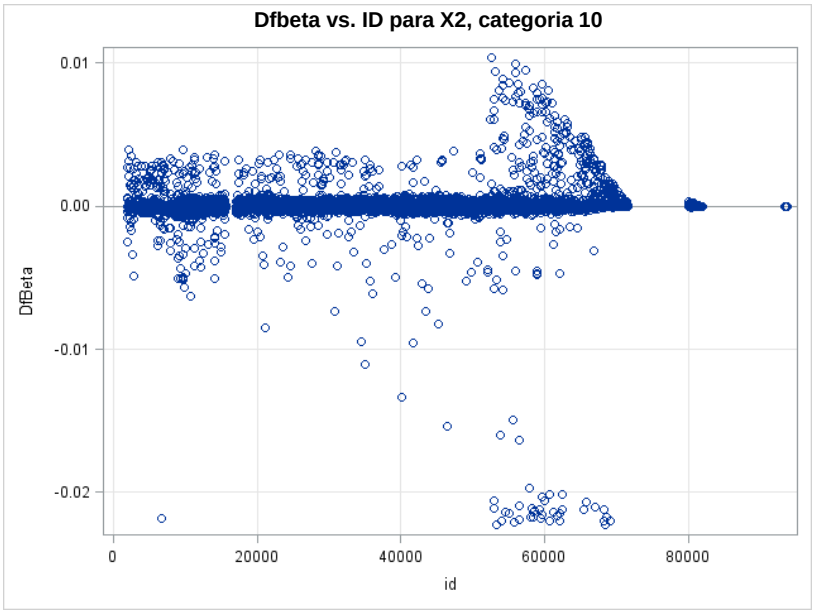

Dfbeta vs. ID para X2, categoria 5
DfBeta
0.01
0.00
-0.01
-0.02
0
20000
40000
60000
80000
id

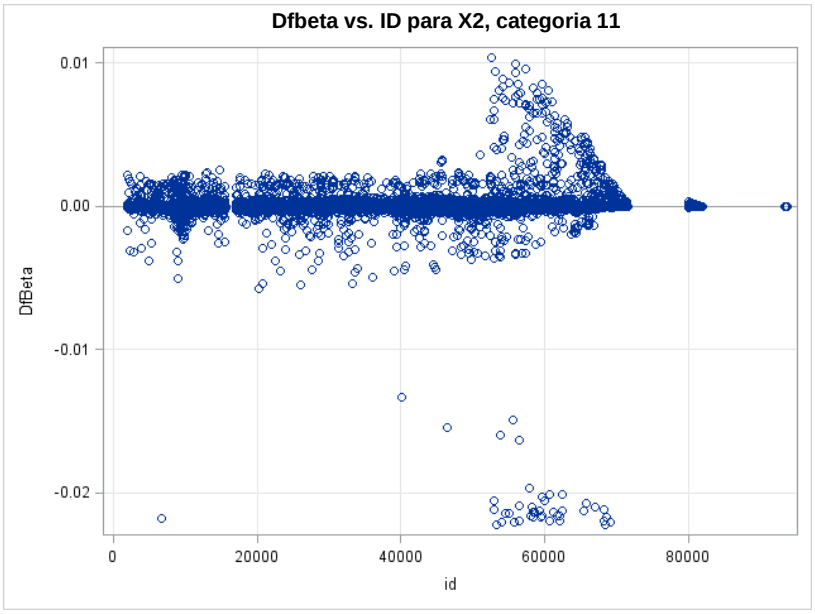

Dfbeta vs. ID para X2, categoria 6
DfBeta
0.025
0.000
-0.025
-0.050
-0.075
0
20000
40000
60000
80000
id

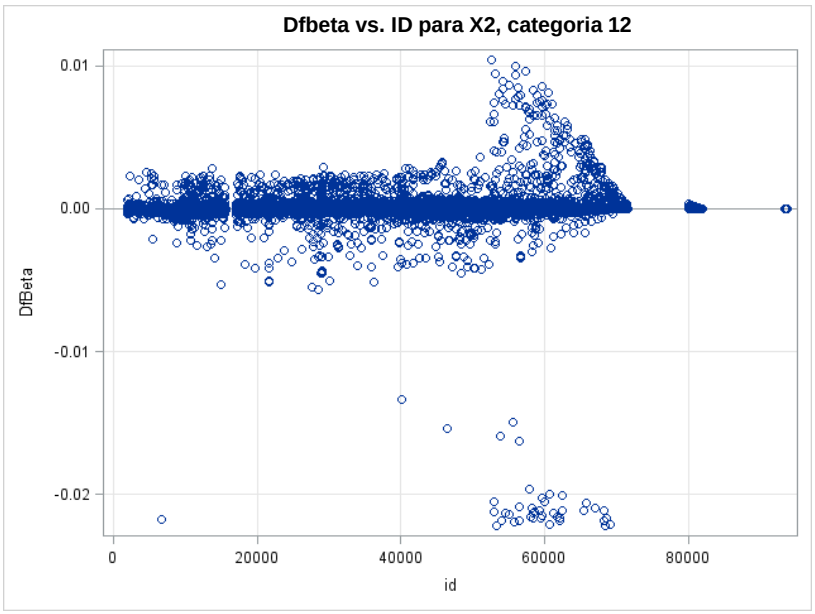

Dfbeta vs. ID para X2, categoria 7
DfBeta
0.01
0.00
-0.01
-0.02
0
20000
40000
60000
80000
id

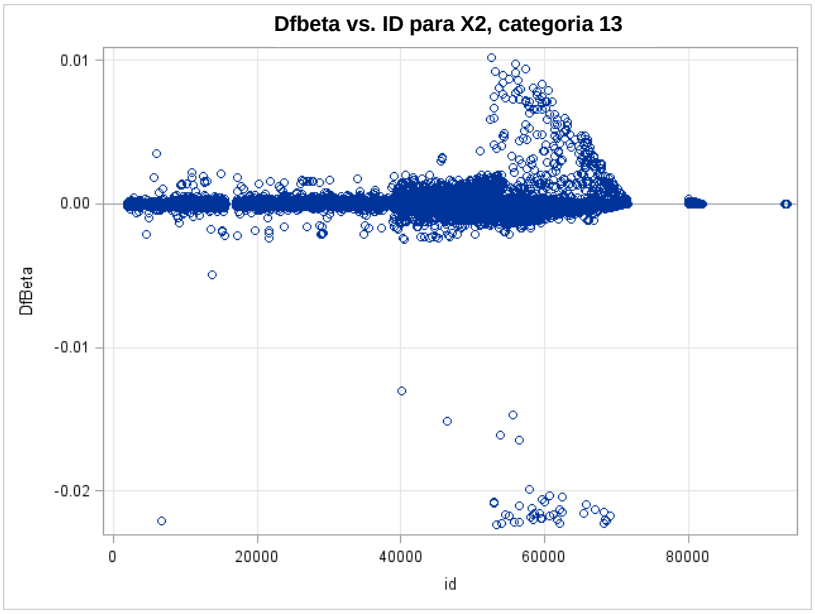

Dfbeta vs. ID para X2, categoria 8
DfBeta
0.01
0.00
-0.01
-0.02
0
20000
40000
60000
80000
id

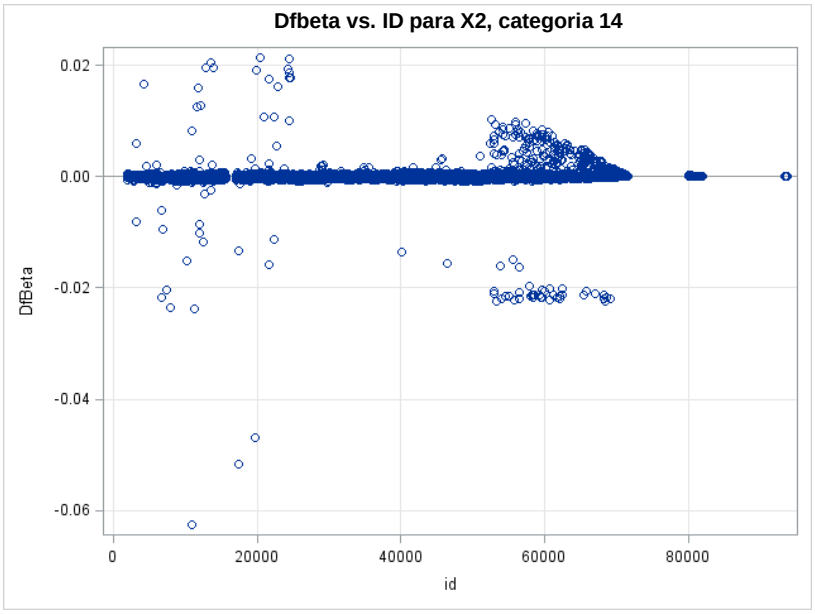

Dfbeta vs. ID para X2, categoria 9
DfBeta
0.01
0.00
-0.01
-0.02
0
20000
40000
60000
80000
id

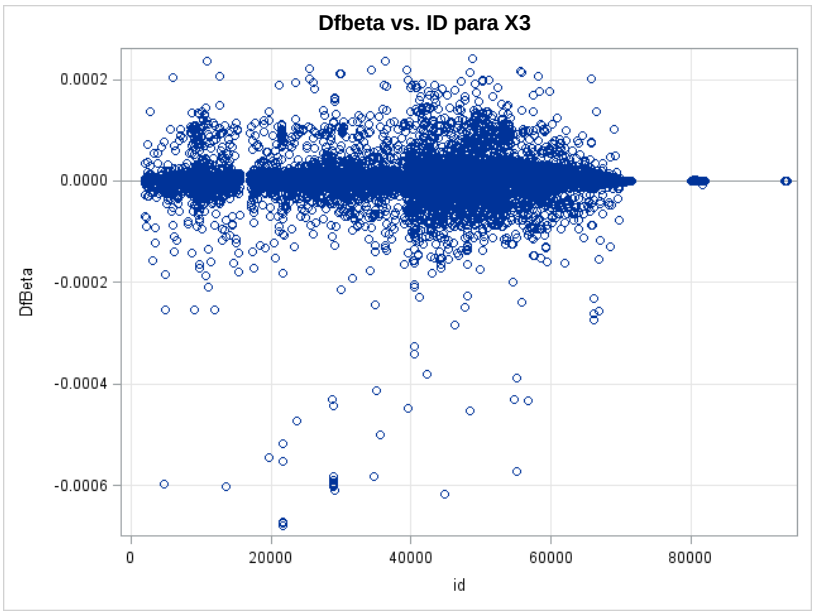

Dfbeta vs. ID para X2, categoria 10
0.01
0.00
-0.01
-0.02
DfBeta
0
20000
40000
60000
80000
id

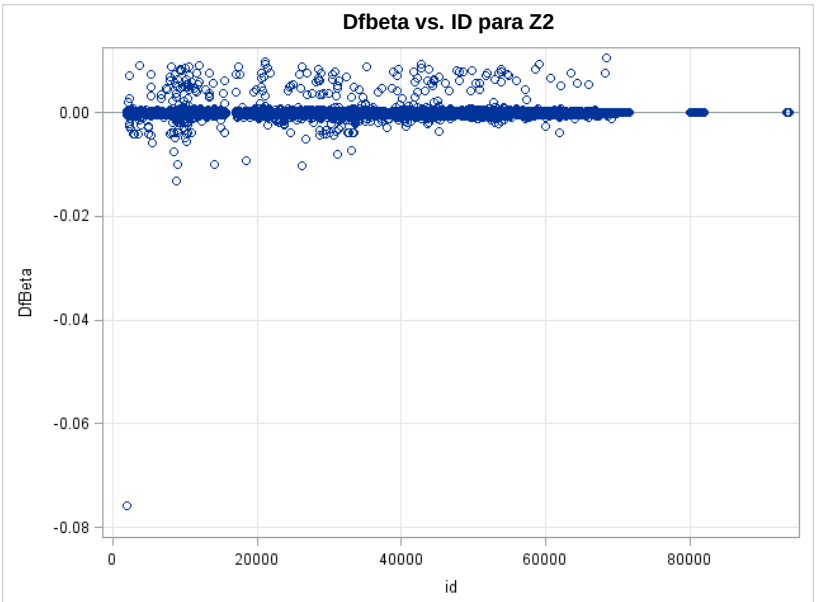

Dfbeta vs. ID para X2, categoria 11
0.01
0.00
-0.01
-0.02
DfBeta
0
20000
40000
60000
80000
id

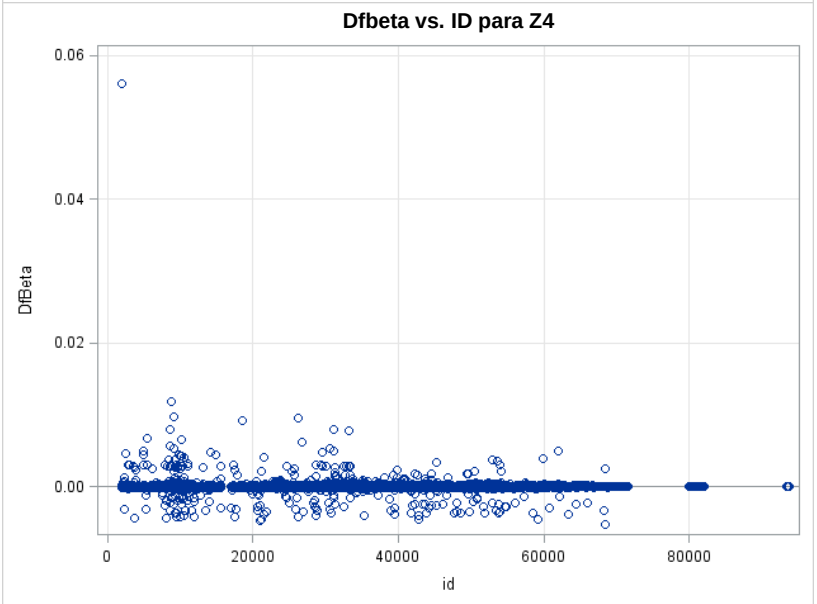

Dfbeta vs. ID para X2, categoria 12
0.01
0.00
-0.01
-0.02
DfBeta
0
20000
40000
60000
80000
id

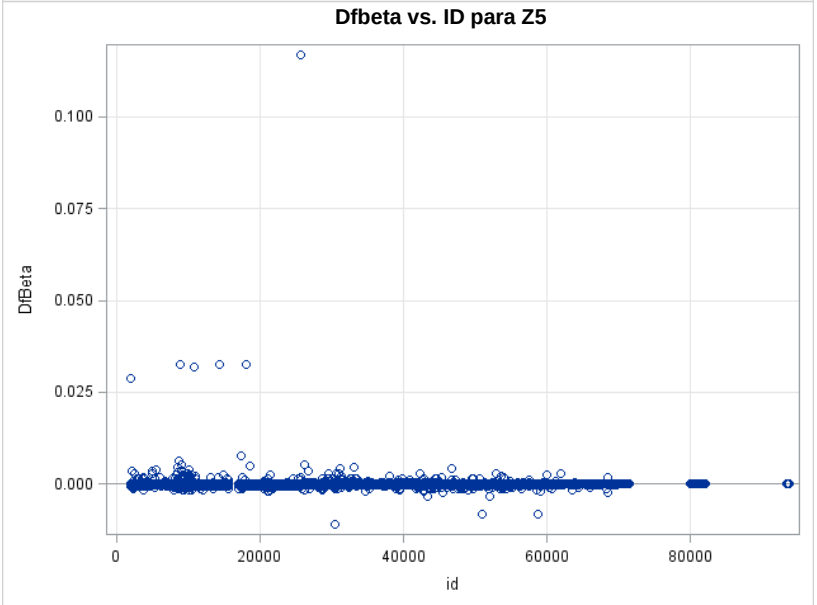

Dfbeta vs. ID para X2, categoria 13
0.01
0.00
-0.01
-0.02
DfBeta
0
20000
40000
60000
80000
id


Dfbeta vs. ID para X2, categoria 14
0.02
0.00
-0.02
-0.04
-0.06
DfBeta
0
20000
40000
60000
80000
id


Dfbeta vs. ID para X3
0.0002
0.0000
-0.0002
-0.0004
-0.0006
DfBeta
0
20000
40000
60000
80000
id


Dfbeta vs. ID para Z2
DfBeta
0.00
-0.02
-0.04
-0.06
-0.08
0
20000
40000
60000
80000
id


Dfbeta vs. ID para Z4
DfBeta
0.06
0.04
0.02
0.00
0
20000
40000
60000
80000
id


Dfbeta vs. ID para Z5
DfBeta
0.100
0.075
0.050
0.025
0.000
0
20000
40000
60000
80000
id