Universidade Federal de São Carlos
Centro de Ciências Exatas e de tecnologia
Departamento de Estatística

Análise de sobrevivência em marketing, considerando o modelo exponencial com fragilidade compartilhada, na predição de *churn*.

Ana Carolaine Hipollito Cintra

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Análise de sobrevivência em marketing, considerando o modelo exponencial com fragilidade compartilhada, na predição de *churn*.

Ana Carolaine Hipollito Cintra
Orientador: José Carlos Fogo

Trabalho de Conclusão de Curso a ser apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
6 de Julho de 2021

Ana Carolaine Hipollito Cintra

Utilização de análise de sobrevivência na predição de *churn*, em *marketing*, considerando o modelo de regressão exponencial com termo de fragilidade.

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Ana Carolaine Hipollito Cintra e aprovado pela banca examinadora.

São Carlos, 6 de Julho de 2021.

Banca Examinadora

- José Carlos Fogo (Orientador)
- Daiane Aparecida Zuanetti
- Estela Maris Pereira Bereta

Dedicatória

Dedico este trabalho aos meus pais, que edificaram a minha vida através do amor e da educação.

Agradecimentos

À Deus principalmente, por me dar forças e sabedoria mesmo nos momentos mais adversos da vida.

À todos os meus familiares que se fizeram presentes e me apoiaram durante a minha trajetória.

Aos meus amigos Vinícius, Thaina, Amanda e Tatiana, que levarei para vida, agradeço pelos momentos de diversão e companheirismo.

À todos os professores que fizeram parte da minha formação, em particular o meu orientador José Carlos, pela dedicação e paciência no auxílio da elaboração deste estudo.

Ao meu namorado Raul, que esteve sempre ao meu lado durante essa caminhada e me ajudou a enfrentar as dificuldades sempre com muito carinho.

Em especial à memoria dos meus pais, Conceição e Devanir, que sonharam com esse momento junto comigo e fizeram tudo para que isso se concretizasse.

Resumo

As empresas que oferecem serviços por assinatura, como telefonia, planos de saúde, internet, entre outros, tem percebido o quão importante é redirecionar suas estratégias de marketing para retenção e fidelização de seus clientes. Ao manter um relacionamento comercial com seus consumidores, as empresas garantem a arrecadação de capital e também diminuem os gastos na prospecção de novos usuários.

O estudo da ocorrência do *churn* é fundamental na retenção de clientes, pois ao analisar aqueles que romperam o vínculo, as empresas passam a conhecer o perfil dos seus consumidores que provavelmente gerarão o abandono de assinatura.

Esse estudo tem como objetivo predizer a probabilidade de ocorrência do *churn* com base no tempo de relacionamento entre cliente e empresa, nas características do serviço oferecido e no perfil do consumidor, utilizando de técnicas de sobrevivência e modelos de regressão.

A fim de tornar a predição mais aplicável, estimamos um modelo para cada grupo de clientes com características semelhantes. E para tornar os resultados dos grupos comparáveis é acrescido um termo de fragilidade, que descreve o risco comum compartilhado pelos indivíduos da mesma categoria.

O estudo também conta com uma aplicação da metodologia utilizada em um banco de dados da companhia de telecomunicações Telecom, no qual a análise foi realizada utilizando-se duas segmentações de grupo. O modelo desenvolvido para esses segmentos conseguiu, através da fragilidade, ordenar de forma satisfatória o risco dos grupos.

Palavras-chave: *análise de sobrevivência, churn, modelo exponencial, fragilidade, marketing, modelo de regressão.*

Sumário

1 Introdução **1**

2 Metodologia **5**

2.1 Análise de sobrevivência 5

2.1.1 Censura 6

2.1.2 Função de sobrevivência 8

2.1.3 Função de risco 10

2.2 A função de verossimilhança 11

2.3 Modelos de regressão 12

2.3.1 Modelo exponencial 13

2.4 Modelo de fragilidade 14

2.4.1 A distribuição para a fragilidade 14

2.5 O modelo de regressão exponencial com termo de fragilidade 15

2.6 Critério de informação de Akaike (AIC) e critério Bayesiano de Schwarz (BIC) 19

3 Aplicação **21**

3.1 Estudo da covariável categoria do cliente como segmento de grupo 25

3.2 Estudo da covariável escolaridade como segmento de grupo 28

3.3 Comparação dos modelos de segmentos de grupo 33

4 Conclusão **35**

Referências Bibliográficas **37**

Apêndice **39**

A Código **39**

Lista de Tabelas

3.1 Descrição das variáveis do conjunto Telecom. 21
3.2 O *Churn* nas categorias de clientes. 25
3.3 AIC e BIC dos modelos com categoria do cliente como segmento. 26
3.4 O *Churn* nos novos níveis de Escolaridade. 29
3.5 AIC e BIC dos modelos com escolaridade como segmento. 30
3.6 Termo de fragilidade predito dos segmentos por grupo. 33

Lista de Figuras

2.1 Ilustração do mecanismo de censura tipo I com seis indivíduos, em que o tempo limite é igual a 18. • representa falha e ∘ representa censura. 6

2.2 Ilustração do mecanismo de censura tipo II com seis indivíduos, em que r é igual a 4. • representa falha e ∘ representa censura. 7

2.3 Os indivíduos 1 e 3 ilustram o mecanismo de censura aleatória. O estudo é observado até tempo limite 18 e • representa falha e ∘ representa censura. 8

3.1 Box-Plot das variáveis "Tempo como cliente", "Idade", "Tempo de residência", "Renda", "Tempo de emprego" e "Residentes", de acordo com a ocorrência do *churn*. 23

3.2 Box-Plot das variáveis de Receita dos serviços de acordo com a ocorrência do *churn*. 23

3.3 Gráfico de barras empilhadas da proporção de *churn* de acordo com as variáveis "Região", "Estado Civil", "Nível de Escolaridade", "Gênero", "Aposentado" e "Categoria do Cliente". 24

3.4 Gráfico de barras empilhadas da proporção de *churn* de acordo com o serviço. 24

3.5 Gráfico do Estimador de Kaplan-Meier por categoria do plano de serviço. . 25

3.6 Gráfico do estimador de Kaplan-Meier para cada categoria do cliente. . . . 28

3.7 Gráfico de barras empilhadas da proporção de *churn* dos clientes, de acordo com o nível de escolaridade. 29

3.8 Gráfico do Estimador de Kaplan-Meier por nível de escolaridade. 30

3.9 Gráfico do estimador de Kaplan-Meier para cada nível escolar. 31

3.10 Gráfico do estimador de Kaplan-Meier para cada nível escolar. 32

3.11 Gráfico do estimador de Kaplan-Meier para cada modelo. 33

Lista de Quadros

3.1 Saída do ajuste do modelo 4. 27
3.2 Saída do ajuste do modelo 2 32

Capítulo 1

Introdução

A globalização caracterizada pela evolução dos meios de comunicação e transporte possibilitou a intensificação no processo de fusão econômica, política e cultural das sociedades.

Esse processo trouxe inúmeros benefícios, como a possibilidade da produção e da venda de diversos produtos e serviços em toda parte do mundo. Com isso, os consumidores ganharam uma infinidade de opções, enquanto que as empresas precisaram aprender a lidar com o aumento da concorrência.

A concorrência traz ao consumidor a oportunidade de escolha entre produtos de diferentes designers, qualidades e consequentemente diferentes preços. As empresas que antes dominavam o mercado começaram a enfrentar o desafio de competir com preços mais baixos que os seus.

Com tantas opções tornou-se comum a migração de clientes entre as empresas, principalmente entre aquelas que ofertam serviço por assinatura. O rompimento da fidelização deu origem ao termo *churn*, esse por sua vez caracteriza o rompimento do relacionamento entre o consumidor e a organização. Esse evento pode ser causado por vários fatores entre eles:

- O cliente não pode mais arcar com a mensalidade do produto;
- O cliente preferiu o produto da concorrência;
- Problemas de relacionamento entre a empresa e o cliente;
- Período de crise financeira.

A prevenção do *churn* é essencial para o crescimento sustentável da empresa, uma vez que ele possui um efeito negativo na receita, por isso entender o que é essa medida é algo vital para o sucesso de muitos negócios.

É interessante também que se saiba quais clientes ocasionarão o *churn*, para prevenir e solucionar diversas situações. Com esse conhecimento é possível elaborar estratégias de *marketing* que possibilitam a retenção desses clientes ou mesmo investir em consumidores que manterão o vínculo por mais tempo.

Custa cerca de 5 a 10 vezes mais recrutar um novo cliente do que fidelizar um já existente (Lu e Park, 2003). Portanto para que os investimentos sejam especificamente direcionados é relevante identificar quais clientes poderão vir a se tornar possíveis *churn* num futuro próximo.

As estratégias de marketing, como podemos observar, são cruciais no momento da venda. Entretanto, elas também vem sendo muito utilizadas na fidelização de clientes, principalmente pelas empresas que trabalham com assinatura de serviço. E para isso tem se utilizado informações a respeito do *churn* para implementar medidas que melhorem a experiência do usuário, como exemplo investir em equipes de suporte de atendimento, desenvolver ferramentas competitivas com o mercado e até remodelar a precificação.

Na prática todos os clientes, em algum momento, deixarão de utilizar os serviços de uma empresa, seja por cancelamento, troca ou mesmo por falecimento. Por isso, a ocorrência deve ser medida e contabilizada dentro de um espaço de tempo não muito longo.

A análise realizada nesse estudo será aplicada em um banco de dados, no qual serão utilizadas técnicas de sobrevivência, e por isso, dentro desse contexto, o tempo até a ocorrência do *churn* será chamado de tempo de vida.

O termo tempo de vida, em sobrevivência, é utilizado mais especificamente na medicina, no qual há o interesse em observar o tempo de um paciente até a ocorrência de um determinado evento. Entretanto essa medida também é estudada em outros ambientes como na área industrial, porém nesse caso é mais comumente chamada de tempo de falha.

Com essa análise será possível compreender quais características possuem maior influência na fidelidade dos clientes com relação a empresa, e para isso utilizaremos modelos de regressão. Portanto, características dos clientes como idade, estado civil, gênero, renda, educação, tipo de cliente e a região de residência, além de informações acerca do tempo e do tipo de prestação de serviço serão utilizadas para predizer a probabilidade de ocorrer

o *churn*.

Uma vez que a predição é obtida através das características, podemos dizer que indivíduos semelhantes terão uma predição semelhante. Portanto utilizando-se dessa premissa e da tentativa de tornar o processo menos individualizado e mais aplicável, a análise será feita para grupos.

Nesse caso podemos supor que os clientes da mesma categoria possuem tratamentos similar e por isso, do ponto de vista estatístico, existe uma correlação intragrupo, que significa que há certa semelhança no risco de ocasionar o *churn*. Esse fator pode ser corrigido através da adição de um efeito aleatório, a fragilidade (Carvalho *et al.*, 2011).

Portanto, esse estudo tem como objetivo obter um modelo de sobrevivência para grupos de clientes, segmentados conforme algum critério preestabelecido, a fim de promover uma retenção menos individualizada, utilizando uma componente de fragilidade para predizer a probabilidade de ocorrência de *churn* em tais grupos. A predição de *churn* seria, então, aplicada ao grupo como um todo, de tal forma que, os indivíduos com aquele perfil sejam afetados de maneira coletiva.

Capítulo 2

Metodologia

Neste capítulo será abordado todas as metodologias estatísticas utilizadas na análise de sobrevivência para a predição de *churn*, em *marketing*, considerando um modelo de regressão com um termo de fragilidade.

2.1 Análise de sobrevivência

A análise de sobrevivência é uma das áreas que mais cresceu nas últimas décadas do século passado. E esse crescimento é devido ao aprimoramento de técnicas estatísticas combinado ao avanço e rapidez dos computadores (Colosimo e Giolo, 2006).

Esta análise é composta por técnicas estatísticas que visam estudar o tempo de vida dos indivíduos durante um experimento. Ela também pode ser aplicada na duração de componentes até o tempo de falha ou, de maneira geral, em estudos nos quais o tempo de vida é observado até a ocorrência de um evento de interesse (Fogo, 2007).

O evento de interesse, na maioria dos casos, são indesejáveis e comumente são chamados de falha. Nesse estudo, a falha é a ocorrência do rompimento do vínculo entre cliente e empresa.

O tempo de vida t, por sua vez, pode ser considerado como uma observação de uma variável aleatória T, contínua e não negativa. Ele possui um início bem definido e é demarcado a partir do instante em que o indivíduo começa a fazer parte do estudo até a ocorrência do evento de interesse.

Portanto, o tempo de vida será contabilizado até o *churn* ou até o encerramento do experimento. No caso da não ocorrência do *churn*, o tempo de vida é dito como censura e será melhor abordado na Seção 2.1.1.

As medidas de interesse em análise de sobrevivência podem ser resumidas por meio de duas funções: a função de sobrevivência, $S(t)$, e a função de risco, $h(t)$ (Collett, 2015). Ambas as funções serão melhor detalhadas nas Seções 2.1.2 e 2.1.3, respectivamente.

2.1.1 Censura

A censura é uma característica associada aos dados de sobrevivência. Ela é atribuída em geral aos indivíduos que por algum motivo não foram acompanhados até o fim do experimento, ou seja, que possuem tempo de falha superior aquele observado (Lawless, 2011).

Dados sem a presença de censura podem facilmente ser analisados por técnicas estatísticas clássicas como análise de regressão ou planejamento de experimentos, utilizando-se de transformação na variável resposta. Entretanto, se houver censuras, essas técnicas não podem ser utilizadas, pois não se tem todos os tempos de falha. Logo é necessário o uso de análise de sobrevivência (Colosimo e Giolo, 2006).

Atualmente existem três categorias que classificam a censura:

i) Censura de Tipo I: o estudo é conduzido até um tempo limite L, já preestabelecido, no qual é registrado um número r de ocorrências do evento. Indivíduos que ultrapassaram o tempo limite sem ocorrências têm seus tempos censurados. A Figura 2.1 ilustra o mecanismo de censura do tipo I.

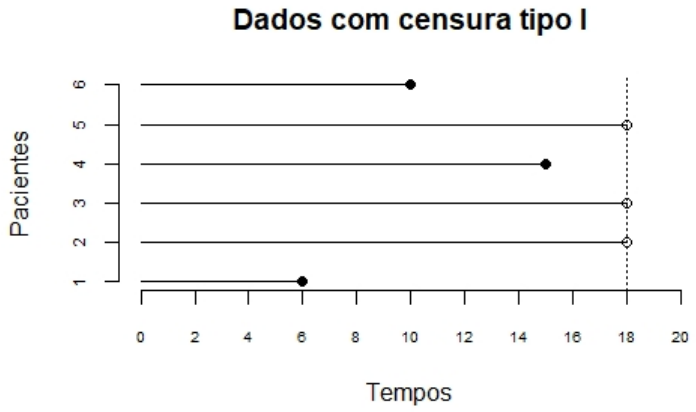

Figura 2.1: Ilustração do mecanismo de censura tipo I com seis indivíduos, em que o tempo limite é igual a 18. • representa falha e ∘ representa censura.

ii) Censura de Tipo II: neste caso, o estudo é conduzido até que se tenha r ocorrências, ou seja, defini-se um número r máximo e aqueles indivíduos que não apresentarem falha até essa contagem são censurados. A Figura 2.2 exemplifica o mecanismo de censura do tipo II.

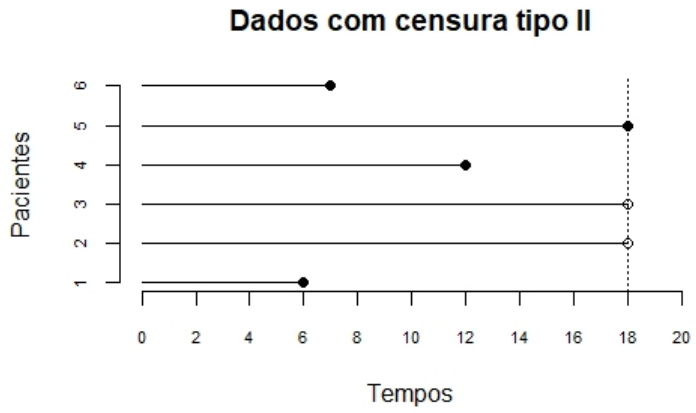

Figura 2.2: Ilustração do mecanismo de censura tipo II com seis indivíduos, em que r é igual a 4. • representa falha e ∘ representa censura.

iii) Censura Aleatória: Na área médica, por exemplo, os pacientes podem ser inseridos no experimento de acordo com a data do diagnóstico. Se o estudo termina numa data pré-estabelecida, então, os tempos de censuras daqueles pacientes que ainda permanecem vivos, ou sem terem experimentado o evento de interesse, serão aleatórios. A representação é feita alinhando todos os indivíduos no tempo inicial, como pode ser visto na Figura 2.3.

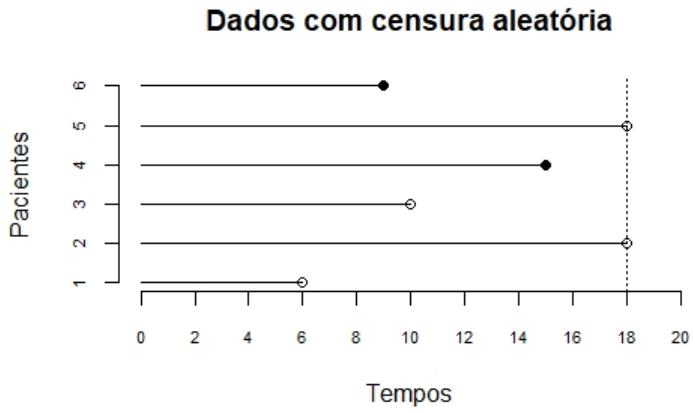

Figura 2.3: Os indivíduos 1 e 3 ilustram o mecanismo de censura aleatória. O estudo é observado até tempo limite 18 e • representa falha e ∘ representa censura.

2.1.2 Função de sobrevivência

Os dados de sobrevivência para o indivíduo i $(i = 1, ..., n)$ sob estudo, são representados, em geral, por t_i, o tempo de falha ou de censura, e δ_i, a variável indicadora de falha ou censura (Colosimo e Giolo, 2006), ou seja,

$$\delta_i = \begin{cases} 1, & \text{se } t_i \text{ é tempo de falha} \\ 0, & \text{se } t_i \text{ é tempo censurado.} \end{cases}$$

A função de sobrevivência é uma das principais funções usadas para descrever estudos de sobrevivência. Ela define a probabilidade de um indivíduo não ocasionar a falha em um tempo t, ou seja, a probabilidade dele sobreviver a esse tempo sem que o evento de interesse aconteça. Em termos probabilísticos (Colosimo e Giolo, 2006),

$$S(t) = P\left(T \geq t\right), \tag{2.1}$$

ou ainda,

$$S(t) = \int_t^{\infty} f(u)\ du, \tag{2.2}$$

em que $f(.)$ é a função densidade assumida para a variável aleatória tempo de vida T.

Por outro lado, a função de distribuição acumulada de T descreve a probabilidade de uma observação não sobreviver ao tempo estipulado, e é definida como

$$F(t) = P\left(T \leq t\right) = 1 - S(t). \tag{2.3}$$

Da teoria da probabilidade, a função densidade de probabilidade é definida por (Casella e Berger, 2002)

$$f(t) = \frac{d}{dt} F(t). \tag{2.4}$$

Estimador de Kaplan-Meier

A função de sobrevivência é a principal componente para realizar a análise descritiva para dados de tempo de vida com censura. Para isso utilizamos o método de estimação desenvolvido por Kaplan e Meier (1958) para estimar a função de sobrevivência $S(t)$.

Esse estimador é não paramétrico, ou seja, não é preciso assumir nenhuma distribuição de probabilidade para a variável aleatória T. A probabilidade de sobrevida até o instante t é estimada considerando a independência entre os tempos. Para isso intervalos temporais são gerados a partir da ordenação dos tempos de falha, de modo que o número de intervalos seja igual ao número de falhas distintas (Colosimo e Giolo, 2006).

A probabilidade do indivíduo não ocasionar o evento de interesse até o tempo t é dada pelo produto das probabilidades de sobreviver a cada um dos intervalos anteriores, pois para sobreviver até certo instante o indivíduo deve sobreviver a todos os instantes passados. Logo,

$$S(t_a) = (1 - q_1)\ (1 - q_2) \ldots (1 - q_a), \tag{2.5}$$

em que q_a é a probabilidade de um indivíduo ocasionar a falha no intervalo $[t_{a-1}, t_a)$, no qual t_a são os tempos de falhas e $a = 1, \ldots, b$, sendo b o número total de eventos de interesse ocasionados.

O estimado de Kaplan-Meier então é basicamente resumido pela estimativa q_a, que é dado por (Colosimo e Giolo, 2006)

$$\hat{q_a} = \frac{\text{número de falhas em } t_a}{\text{número de observações sob risco em } t_{a-1}} \quad a = 1, \ldots, b. \tag{2.6}$$

Portanto, dado que $n(t_a)$ é o total de pessoas em risco no tempo t_a e d_a é o número de falhas, o estimador da probabilidade de sobrevida é expresso por

$$\hat{S}(t) = \prod_{t_a < t} \frac{n(t_a) - d(t_a)}{n(t_a)} = \prod_{t_a < t} 1 - \frac{d(t_a)}{n(t_a)}. \tag{2.7}$$

2.1.3 Função de risco

A função de risco, $h(t)$, é uma outra medida importante na análise de sobrevivência. Ela representa a probabilidade de ocorrência de um determinado evento, nesse caso o *churn*, dentro de um intervalo de tempo.

Fixando o intervalo de tempo $[t, t + dt)$ e condicionando a sobrevivência no instante t, temos por definição que a função de risco é

$$h(t) = \lim_{dt \to 0} \frac{P\left(t < T \leq t + dt \mid T > t\right)}{dt}. \tag{2.8}$$

Portanto definiremos, no contexto de *marketing*, que $h(t)$ é o risco de perder um clientes entre o tempo t e $t + dt$ (Berry e Linoff, 2004). Essa função é muito utilizada para determinar a distribuição dos tempos de sobrevivência e para descrever a maneira com que a chance de ocorrência de um evento muda com o tempo (Klein e Moeschberger, 2006).

A partir da definição de probabilidade condicional podemos obter a seguinte relação:

$$h(t) = \lim_{dt \to 0} \frac{P\left(t \leq T \leq t + dt\right)}{dt\, P\left(T \geq t\right)} = \frac{1}{P\left(T \geq t\right)} \lim_{dt \to 0} \frac{P\left(t \leq T \leq t + dt\right)}{dt} = \frac{f(t)}{S(t)}. \tag{2.9}$$

Como $-f(t) = \dfrac{dS(t)}{dt}$, pode-se ainda escrever

$$h(t) = \frac{-dS(t)}{dt}.\frac{1}{S(t)} = -\frac{d}{dt}\log[S(t)]. \tag{2.10}$$

Ou, integrando ambos os lados em relação a t de (2.9), obtemos a função de risco acumulada, definida por

$$H(t) = -log[S(t)], \tag{2.11}$$

de onde podemos escrever

$$S(t) = \exp\{-H(t)\}, \tag{2.12}$$

ou ainda,

$$S(t) = \exp\left\{-\int_0^t h(u)du\right\}. \tag{2.13}$$

Das Equações (2.9) e (2.13), pode-se obter uma expressão para representar a função densidade de probabilidade da variável aletória T

$$f(t) = h(t)\exp\left\{-\int_0^t h(u)du\right\}. \tag{2.14}$$

2.2 A função de verossimilhança

Suponhamos um conjunto de dados $\mathcal{D}$ com n indivíduos, em que o tempo de vida do i-ésimo indivíduo seja representado pela varável aleatória T_i, com $i = 1, \ldots, n$, e que associado a ele existe uma variável indicadora de falha δ_i.

Os tempos de sobrevivência $T_1, T_2, ..., T_n$ serão assumidos iid (independentes e identicamente distribuídos), com função densidade de probabilidade $f(t_i|\boldsymbol{\theta})$, em que $\boldsymbol{\theta}$ é um vetor de parâmetros. Se $\delta_i = 1$, então, T_i representa o tempo de falha do indivíduo i e se $\delta_i = 0$, T_i representa o seu tempo de censura. Como (Fogo, 2007)

$$P(T_i, \delta_i = 1) = P(T_i|\delta_i = 1)P(\delta_i = 1) = f(t_i|\boldsymbol{\theta}) \tag{2.15}$$

e

$$P(T_i, \delta_i = 0) = P(T_i|\delta_i = 0)P(\delta_i = 0) = S(t_i|\boldsymbol{\theta}), \tag{2.16}$$

a função densidade de probabilidade conjunta de T_i e δ_i é dada por

$$f(t_i, \delta_i|\boldsymbol{\theta}) = [f(t_i|\boldsymbol{\theta})]^{\delta_i}\ [S(t_i|\boldsymbol{\theta})]^{1-\delta_i}. \tag{2.17}$$

Assim, a função de verossimilhança para dados sujeitos a censura é obtida por

$$L(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{n} f(t_i, \delta_i|\boldsymbol{\theta}) = \prod_{i=1}^{n}[f(t_i|\boldsymbol{\theta})]^{\delta_i}[S(t_i|\boldsymbol{\theta})]^{1-\delta_i}, \tag{2.18}$$

em que o conjunto de dados observado é tido por $\mathcal{D} = \{t_i, \delta_i\}$.

Utilizando as relações (2.13) e (2.14), obtem-se uma representação da função de verossimilhança expressa em termos da função de risco, dada por

$$L(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{n} \; [h(t_i|\boldsymbol{\theta})]^{\delta_i} \; \exp\left\{ - \int_0^{t_i} h(u|\boldsymbol{\theta}) du \right\}. \tag{2.19}$$

2.3 Modelos de regressão

Os modelos de regressão são usados para expressar a relação entre duas ou mais variáveis. Nesse contexto as variáveis para as quais se faz a predição são comumente chamadas de resposta ou dependentes e as variáveis que afetam a resposta são ditas como preditoras, explicativas, independentes ou covariáveis.

A influência que as características dos clientes exercem nos tempos de vida dos indivíduos é estudada usualmente por meio de modelos de regressão e, por isso, denominamos essas características como covariáveis.

Essa influência será representada através da modelagem da função de risco $h(t)$, ou seja, essas covariáveis serão utilizadas para predizer a probabilidade de ocorrer o abandono do serviço, ou melhor, o *churn*, em determinado espaço de tempo.

Considere um conjunto de dados de tempos de vida com p covariáveis e n indivíduos. Sendo T uma variável aleatória que representa o tempo de ocorrência de um evento, o modelo de regressão que estabelece a função de risco de um indivíduo, no tempo t, considerando um vetor de covariáveis $\mathbf{z} = (z_1, z_2, \ldots, z_p)$, é

$$h\left(t_i|\mathbf{z}_i\right) = h_0\left(t_i\right) g\left(\mathbf{z}_i|\boldsymbol{\beta}\right), \quad i = 1, 2, \ldots, n, \tag{2.20}$$

sendo $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ o vetor de parâmetros a serem estimados, $g(\cdot)$ uma função positiva das covariáveis e $h_0(t)$ a função de risco de base.

Tendo em conta que $g(\mathbf{z}_i|\boldsymbol{\beta})$ deve ser uma função não negativa, uma escolha apropriada consiste em fazer $g(\mathbf{z}_i|\boldsymbol{\beta}) = \exp\{\boldsymbol{\beta}'\mathbf{z}_i\}$. Dessa forma, a função de risco para o indivíduo i, pode ser reescrita por (Wienke, 2007)

$$h\left(t_i|\mathbf{z}_i\right) = h_0\left(t_i\right) \exp\left\{\boldsymbol{\beta}'\mathbf{z}_i\right\} = h_0\left(t_i\right) \exp\left\{\boldsymbol{\beta}_1 z_{i1} + \boldsymbol{\beta}_2 z_{i2} + \ldots + \boldsymbol{\beta}_p z_{ip}\right\}, \tag{2.21}$$

em que $\boldsymbol{\beta}_0$ é incorporado na função de risco de base.

Utilizando (2.13) e (2.20), podemos obter sua função de sobrevivência como

$$S(t_i|\mathbf{z}_i) = \exp\left\{ - \int_0^{t_i} h_0(u)\, g(\mathbf{z}_i|\boldsymbol{\beta})\, du \right\}, \quad i = 1, 2, \ldots, n. \tag{2.22}$$

2.3.1 Modelo exponencial

A literatura estatística apresenta uma infinidade de modelagens para tempo de vida, podendo ser paramétricas ou não paramétricas. Contudo, o modelo exponencial foi escolhido para ser utilizado devido a sua fácil interpretação e aplicação. O modelo exponencial assume que o risco é constante, por isso, ao o utilizarmos admitimos que o risco do *churn* é o mesmo com o passar dos meses.

Sabemos que o modelo exponencial não descreve perfeitamente o comportamento da probabilidade de ocorrência do *churn*. O risco que mais se assemelharia a realidade é um que tenha a função densidade de probabilidade em formato de banheira. Pois no início do contrato o risco tende a ser alto, uma vez que pode haver uma quebra de expectativa da parte do contratante, e com o passar do tempo ele diminui, evidenciando a satisfação e a comodidade do cliente. Entretanto, após um período a tendência é que o risco de ocasionar o *churn* volte a crescer, como consequência do desgaste da relação entre cliente e empresa que aumenta ao longo dos anos.

Apesar do modelo exponencial não descrever com precisão o comportamento do risco, ele descreve bem o período intermediário descrito acima, em que o risco é constante, portanto mais um motivo para utiliza-lo neste estudo.

Utilizando (2.14) e (2.20), podemos escrever a função densidade de probabilidade de t como

$$f(t_i) = h(t_i|\mathbf{z}_i)\ \mathrm{e}^{-H(t_i|\mathbf{z}_i)} = h_0\left(t_i\right) g\left(\mathbf{z}_i|\boldsymbol{\beta}\right)\ \mathrm{e}^{-H(t_i|\mathbf{z}_i)}, \quad i = 1, 2, \ldots, n. \tag{2.23}$$

Como o risco no modelo exponencial é constante e igual a λ então, $h(t_i|\mathbf{z}_i) = \lambda\ g\left(\mathbf{z}_i|\boldsymbol{\beta}\right)$. Logo,

$$H(t_i|\mathbf{z}_i) = \int_0^{t_i} \lambda\ g\left(\mathbf{z}_i|\boldsymbol{\beta}\right)\ d(u) \quad i = 1, 2, \ldots, n. \tag{2.24}$$

Considerando a relação descrita em (2.24), temos que

$$H(t_i|\mathbf{z}_i) = \lambda\ g\left(\mathbf{z}_i|\boldsymbol{\beta}\right)\ t_i, \quad i = 1, 2, \ldots, n, \tag{2.25}$$

o que nos leva a função de densidade de probabilidade do tempo de vida com distribuição exponencial como

$$f(t_i) = \lambda\ g\left(\mathbf{z}_i|\boldsymbol{\beta}\right)\ \mathrm{e}^{\lambda t_i\ g(\mathbf{z}_i|\boldsymbol{\beta})}, \quad i = 1, 2, \ldots, n. \tag{2.26}$$

2.4 Modelo de fragilidade

Existem algumas situações nas quais as associações entre os indivíduos não podem ser observadas. Por exemplo, sujeitos da mesma família possuem uma relação genética entre eles, assim como observações de uma mesma região tendem a ter uma associação no quesito variáveis ambientais (Fogo, 2007). O estudo em questão apresenta essa relação quando comparamos indivíduos consumidores de um mesmo tipo de serviço.

Os serviços distribuídos pelas empresas de telefonia, em geral, são criados para atender todos os tipos de públicos. E é por isso que existem planos de serviços simples e por isso mais baratos e planos mais completos e mais caros. Isso faz com que os usuários de cada plano possuam características semelhantes, por exemplo, usuários de um serviço mais básico tendem a ter renda aproximada e inferior as de clientes que contratam um serviço mais amplo.

Desta forma, é necessário incluir um peso que torne comparável esses grupos de clientes e iremos fazer isso através da inclusão de um termo de fragilidade, que é definido basicamente como um efeito aleatório que descreve o risco comum dentro do grupo.

Neste contexto, cada grupo de clientes assume uma fragilidade única, compartilhada por todos os sujeitos, esta, por sua vez, representada na função de risco por

$$h\left(t_{ki}|w_k, \mathbf{z}_{ki}\right) = w_k\, h_0\left(t_{ki}\right) \exp\left\{\boldsymbol{\beta}'\mathbf{z}_{ki}\right\}, \quad i = 1, \ldots, n_k, \quad k = 1, \ldots, g, \tag{2.27}$$

em que w_k é a fragilidade não observada, associada ao k-ésimo grupo, n_k o número de indivíduos em cada grupo, $\mathbf{z}_{ki}$ o vetor de covariáveis do i-ésimo indivíduo do k-ésimo grupo e g o número de grupos.

2.4.1 A distribuição para a fragilidade

Diferentes distribuições de probabilidade têm sido sugeridas para o termo de fragilidade, dentre elas, a lognormal, a uniforme, a *Weibull*, porém, a mais utilizada tem sido a distribuição gama (Tomazella, 2003).

A estimação do modelo de fragilidade pode ser feita por meio da construção da função de verossimilhança e sua otimização. Por isso, sejam n_k indivíduos de um grupo de dados $\mathcal{D} = \{t_{ki}, z_{ki}, w_k\}$ sem censura, de acordo com (2.27), a função de verossimilhança para o

k-ésimo grupo, condicionada à fragilidade w_k, pode ser escrita por

$$L_k(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathcal{D}) = \prod_{i=1}^{n_k} w_k \, h_0(t_{ki}) \, \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{ki}} \exp\left\{ - \int_0^{t_{ki}} w_k \, h_0(u) \, \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{ki}} \, du \right\}. \tag{2.28}$$

Dada a verossimilhança em (2.28) e uma densidade $f(w)$ para a fragilidade, tem-se que

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathcal{D}) = \prod_{k=1}^{g} L_k(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathcal{D}) \, f\,(w_k)$$

$$= \prod_{k=1}^{g} w_k^{n_k} \left[\prod_{i=1}^{n_k} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{ki}} \, h_0(t_{ki}) \exp\left\{ - \int_0^{t_{ki}} w_k \, h_0(u) \, \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{ki}} \, du \right\} \right] f(w_k). \tag{2.29}$$

Considerando, ainda, a presença de censuras, ou seja, $\mathcal{D} = \{t_{ki}, z_{ki}, w_k, \delta_{ki}\}$, a função de verossimilhança em (2.29), passa a ser expressa na forma

$$L_k(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathcal{D}) = \prod_{k=1}^{g} \left[\prod_{i=1}^{n_k} w_k^{\delta_{ki}} \, \mathrm{e}^{\delta_{ki}\boldsymbol{\beta}'\mathbf{z}_{ki}} \, [h_0(t_{ki})]^{\delta_{ki}} \times \right.$$

$$\left. \times \exp\left\{ - \int_0^{t_{ki}} w_k \, h_0(u) \, \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{ki}} \, du \right\} \right] f(w_k). \tag{2.30}$$

2.5 O modelo de regressão exponencial com termo de fragilidade

Considerando o modelo exponencial definido em (2.26) para os tempos de vida, temos que a função de risco de base é do tipo $h_0(t) = \lambda$, constante.

Assim sendo, de (2.28), a função de verossimilhança para o k-ésimo grupo, sem a presença de censuras e sem considerar o termo de fragilidade, ou seja, $\mathcal{D} = \{t_{ki}, z_{ki}\}$, é dada por:

$$L_k(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathcal{D}) = \prod_{i=1}^{n_k} \lambda \, \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{ki}} \exp\left\{ -\lambda \, t_{ki} \, \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{ki}} \right\}. \tag{2.31}$$

Assumindo dados com censuras ($\mathcal{D} = \{t_{ki}, z_{ki}, \delta_{ki}\}$), em que δ_{ki} é o indicador de falhas

do i-ésimo sujeito do k-ésimo grupo, temos que (2.31) é escrita como

$$L_k(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathcal{D}) = \prod_{i=1}^{n_k} \left(\lambda \, \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{ki}}\right)^{\delta_{ki}} \exp\left\{-\lambda \, t_{ki} \, \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{ki}}\right\}$$

$$= \lambda^{\delta_k} \, \mathrm{e}^{-\lambda \, T_k(\boldsymbol{\beta})} \left[\prod_{i=1}^{n_k} \mathrm{e}^{\delta_{ki}\boldsymbol{\beta}'\mathbf{z}_{ki}}\right], \tag{2.32}$$

em que $\delta_k = \sum_{i=1}^{n_k} \delta_{ki}$ é o número de falhas referentes ao grupo k e $T_k(\boldsymbol{\beta}) = \sum_{i=1}^{n_k} t_{ki} \, \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{ki}}$, $k = 1, 2, \ldots, g$.

Incluindo o termo de fragilidade $w_k > 0$ no modelo, para o $k-$ésimo grupo, considerando $\mathcal{D} = \{t_{ki}, z_{ki}, w_k, \delta_{ki}\}$, temos

$$L_k(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathcal{D}) = \prod_{i=1}^{n_k} \left(w_k \, \lambda \, \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{ki}}\right)^{\delta_{ki}} \exp\left\{-w_k \, \lambda \, t_{ki} \, \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{ki}}\right\}$$

$$= \lambda^{\delta_k} \, w_k^{\delta_k} \, \exp\Big\{ - \lambda \, w_k \, T_k(\boldsymbol{\beta})\Big\} \left[\prod_{i=1}^{n_k} \mathrm{e}^{\delta_{ki}\boldsymbol{\beta}'\mathbf{z}_{ki}}\right]. \tag{2.33}$$

A função de verossimilhança global, considerando os g grupos e uma densidade $f(w_k)$ para W_k, é dada por:

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathcal{D}) = \prod_{k=1}^{g} L_k(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathcal{D}) \, f(w_k)$$

$$= \prod_{k=1}^{g} \lambda^{\delta_k} \, w_k^{\delta_k} \, \exp\Big\{ - \lambda \, w_k \, T_k(\boldsymbol{\beta})\Big\} \left[\prod_{i=1}^{n_k} \mathrm{e}^{\delta_{ki}\boldsymbol{\beta}'\mathbf{z}_{ki}}\right] \, f(w_k). \tag{2.34}$$

O termo de fragilidade em (2.34) pode ser eliminado integrando-se a verossimilhança em relação a w_k, ou seja,

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathcal{D}) = \int_0^{\infty} \prod_{k=1}^{g} \lambda^{\delta_k} \, w_k^{\delta_k} \, \exp\Big\{ - \lambda \, w_k \, T_k(\boldsymbol{\beta})\Big\} \, \left[\prod_{i=1}^{n_k} \mathrm{e}^{\delta_{ki}\boldsymbol{\beta}'\mathbf{z}_{ki}}\right] \, f(w_k) \, dw_k$$

$$= \prod_{k=1}^{g} \left[\prod_{i=1}^{n_k} \mathrm{e}^{\delta_{ki}\boldsymbol{\beta}'\mathbf{z}_{ki}}\right] \, \lambda^{\delta_k} \int_0^{\infty} w_k^{\delta_k} \, \exp\Big\{ - \lambda \, w_k \, T_k(\boldsymbol{\beta})\Big\} \, f(w_k) \, dw_k. \tag{2.35}$$

Assumindo uma densidade $gama(\alpha,\, \alpha)$ para W_k, temos $E(W_k) = 1$ e $Var(W_k) = 1/\alpha$. Assim, a expressão em (2.35) é escrita por:

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathcal{D}) = \prod_{k=1}^{g} \left[\prod_{i=1}^{n_k} \mathrm{e}^{\delta_{ki}\boldsymbol{\beta}'\mathbf{z}_{ki}}\right] \lambda^{\delta_k} \int_0^\infty w_k^{\delta_k} \exp\Big\{ -\lambda\, w_k\, T_k(\boldsymbol{\beta})\Big\} \frac{\alpha^\alpha}{\Gamma(\alpha)}\, w_k^{\alpha-1}\, \mathrm{e}^{-\alpha w_k}\, dw_k$$

$$= \prod_{k=1}^{g} \left[\prod_{i=1}^{n_k} \mathrm{e}^{\delta_{ki}\boldsymbol{\beta}'\mathbf{z}_{ki}}\right] \lambda^{\delta_k} \frac{\alpha^\alpha}{\Gamma(\alpha)} \int_0^\infty w_k^{\delta_k+\alpha-1} \exp\Big\{ -w_k\big[\lambda\, T_k(\boldsymbol{\beta}) + \alpha\big]\Big\}\, dw_k. \tag{2.36}$$

A integral em (2.36) é, de fato, o núcleo de uma gama com parâmetros

$$\alpha_k^* = \alpha + \delta_k \quad \mathrm{e} \quad \beta_k^* = \alpha + \lambda T_k(\boldsymbol{\beta}),$$

em que a esperança posteriori da fragilidade é dada por

$$E(W_k) = \frac{\alpha_k^*}{\beta_k^*} \;=\; \frac{\alpha + \delta_k}{\alpha + \lambda\; T_k(\beta)}. \tag{2.37}$$

Portanto, temos que

$$\int_0^\infty w_k^{\alpha+\delta_k-1} \exp\Big\{-w_k\, \Big[\alpha + \lambda\, T_k(\boldsymbol{\beta})\Big]\Big\}\, dw_k = \frac{\Gamma(\alpha+\delta_k)}{\Big[\alpha + \lambda\, T_k(\boldsymbol{\beta})\Big]^{\alpha+\delta_k}}. \tag{2.38}$$

Substituindo o resultado da integral, dado em (2.38) em (2.36), temos a função de verossimilhança independente do termo de fragilidade w_k, ou seja,

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathcal{D}) = \prod_{k=1}^{g} \lambda^{\delta_k} \left[\prod_{i=1}^{n_k} \mathrm{e}^{\delta_{ki}\boldsymbol{\beta}'\mathbf{z}_{ki}}\right] \frac{\alpha^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+\delta_k)}{\Big[\alpha + \lambda\, T_k(\boldsymbol{\beta})\Big]^{\alpha+\delta_k}}$$

$$= \prod_{k=1}^{g} \frac{\lambda^{\delta_k}}{\alpha^{\delta_k}\, \Gamma(\alpha)} \left[\prod_{i=1}^{n_k} \mathrm{e}^{\delta_{ki}\boldsymbol{\beta}'\mathbf{z}_{ki}}\right] \frac{\Gamma(\alpha+\delta_k)}{\Big[1 + \frac{\lambda}{\alpha}\, T_k(\boldsymbol{\beta})\Big]^{\alpha+\delta_k}}$$

$$= \frac{\lambda^\delta}{\alpha^\delta\; [\,\Gamma(\alpha)\,]^g} \left\{\prod_{k=1}^{g} \frac{\Gamma(\alpha+\delta_k)}{\Big[1 + \frac{\lambda}{\alpha}\, T_k(\boldsymbol{\beta})\Big]^{\alpha+\delta_k}}\right\} \left[\prod_{k=1}^{g}\prod_{i=1}^{n_k} \mathrm{e}^{\delta_{ki}\boldsymbol{\beta}'\mathbf{z}_{ki}}\right]. \tag{2.39}$$

em que $\delta = \sum_{k=1}^{g} \delta_k$, é o número total de falhas.

Desta forma, podemos utilizar o método da Máxima Verossimilhança para estimar os parâmetros do modelo utilizando as estimativas que tornam máximo o valor da função. Portanto, aplicando-se o logaritmo na Equação (2.39), temos

$$\ell(\boldsymbol{\beta}, \alpha \,|\mathcal{D}) = \log\Big[L(\boldsymbol{\beta}, \alpha \,|\mathcal{D})\Big]$$

$$= \delta \log(\lambda) - \delta \log(\alpha) - g \log\Big[\Gamma(\alpha)\Big] + \sum_{k=1}^{g} \log\Big[\,\Gamma(\alpha + \delta_k)\,\Big] -$$

$$- \sum_{k=1}^{g} (\alpha + \delta_k)\ \log\Big[1 + \frac{\lambda}{\alpha}\, T_k(\boldsymbol{\beta})\Big] + \sum_{k=1}^{g}\sum_{i=1}^{n_k} \delta_{ki} \boldsymbol{\beta}' \mathbf{z}_{ki}\,. \tag{2.40}$$

As derivadas do logaritmo da função de verossimilhança são dadas por

$$\frac{\partial \ell}{\partial \alpha} = \frac{-\delta}{\alpha} - g\ \psi(\alpha) + \sum_{k=1}^{g} \psi(\alpha + \delta_k) - \sum_{k=1}^{g} \log\left[1 + \frac{\lambda}{\alpha}\ T_k(\boldsymbol{\beta})\right]\ + \sum_{k=1}^{g} \frac{\lambda\ T_k(\boldsymbol{\beta})\ (\alpha + \delta_k)}{\alpha^2 + \alpha\ \lambda\ T_k(\boldsymbol{\beta})}$$

$$= \frac{-\delta}{\alpha} - g\ \psi(\alpha) + \sum_{k=1}^{g} \psi(\alpha + \delta_k)\ -\ \sum_{k=1}^{g} \log\Big[1 + \frac{\lambda}{\alpha}\ T_k(\boldsymbol{\beta})\Big]\ + \frac{\lambda}{\alpha}\ \sum_{k=1}^{g}\ \frac{\big(\alpha + \delta_k\big)\ T_k(\boldsymbol{\beta})}{\alpha + \lambda\ T_k(\boldsymbol{\beta})}\,; \tag{2.41}$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{\delta}{\lambda} - \sum_{k=1}^{g}\ \frac{\big(\alpha + \delta_k\big)\ T_k(\boldsymbol{\beta})}{\alpha + \lambda\ T_k(\boldsymbol{\beta})}\,; \tag{2.42}$$

$$\frac{\partial \ell}{\partial \beta_j} = -\sum_{k=1}^{g}\ \frac{\lambda\ (\alpha + \delta_k)}{\alpha\ \Big[1 + \dfrac{\lambda}{\alpha}\ T_k(\boldsymbol{\beta})\Big]}\ \frac{\partial}{\partial \beta_j}\Big[T_k(\boldsymbol{\beta})\Big] + \sum_{k=1}^{g}\ \sum_{i=1}^{n_k} \delta_{ki}\ z_{kij}$$

$$= -\sum_{k=1}^{g}\ \frac{\lambda\ (\alpha + \delta_k)}{\Big[\alpha + \lambda\ T_k(\boldsymbol{\beta})\Big]} \left(\sum_{i=1}^{n_k}\ z_{kij}\ t_{ki}\ \mathrm{e}^{\delta_{ki} \boldsymbol{\beta}' \mathbf{z}_{ki}}\right)\ + \sum_{k=1}^{g}\ \sum_{i=1}^{n_k}\ \delta_{ki}\ z_{kij}, \tag{2.43}$$

em que $\psi(x) = d\ \log[\Gamma(x)]\ /dx$ é a função digama (Abramowitz e Stegun, 1965) e $j = 1, ..., p$.

A estimação dos parâmetros α, λ e $\boldsymbol{\beta}$ será feita através de métodos numéricos, uma vez que o processo de estimação não pode ser resolvido algebricamente. Após essa estimação

podemos predizer o valor de w_k através da equação (2.37), como

$$\hat{w_k} = \frac{\hat{\alpha} + \delta_k}{\hat{\alpha} + \hat{\lambda}\ T_k(\hat{\boldsymbol{\beta}})} \ = \ \frac{\hat{\alpha} + \delta_k}{\hat{\alpha} + \hat{\lambda}\ \sum_{i=1}^{n_k} t_{ki}\ \mathrm{e}^{\hat{\boldsymbol{\beta}}'\mathbf{z}_{ki}}}, \tag{2.44}$$

que corresponde ao valor esperado de uma variável aleatória com distribuição de probabilidade gama$\Big(\ \hat{\alpha} + \delta_k\ ,\ \hat{\alpha} + \hat{\lambda}\ T_k(\hat{\boldsymbol{\beta}})\Big)$ (Tomazella, 2003).

2.6 Critério de informação de Akaike (AIC) e critério Bayesiano de Schwarz (BIC)

Escolher um modelo estatístico apropriado é extremamente importante em uma análise de dados, uma vez que buscamos um modelo que tenha o mínimo de parâmetros possível e que ao mesmo tempo explique de forma satisfatória o comportamento da variável resposta. Para isso, exitem diversos critérios para seleção de modelos, dentre eles os mais utilizados são o teste da razão de verossimilhança (TRV), o critério de informação de Akaike (AIC) e o critério Bayesiano de Schwarz (BIC). Neste estudo utilizaremos o AIC e o BIC.

Visto que o viés é a distância entre a média do conjunto de estimativas e o verdadeiro valor do parâmetro $\boldsymbol{\theta}$ e que, o método da máxima verossimilhança estima os parâmetros do modelo utilizando os mesmos dados para estimar essa média, temos um viés que varia de acordo com a dimensão do vetor de parâmetros.

Akaike (1974), mostrou que o viés é dado por

$$AIC = -2\log(L(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}|\mathcal{D})) + 2r,$$

em que r é o número de parâmetros a serem estimados no modelo. Por isso, ao nos depararmos com dois ou mais modelos devemos optar pelo que apresentar o menor AIC.

Já o BIC, proposto por Schwarz a partir de métodos Bayesianos, é dado por

$$BIC = -2\log(L(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}|\mathcal{D})) + r\log(n),$$

em que n é o número de observações. Assim como o AIC, o melhor modelo eleito pelo BIC é aquele que apresenta o menor valor entre os modelos.

Capítulo 3

Aplicação

Para a aplicação da metodologia e estimação da probabilidade de *churn* através da análise de sobrevivência, considerando o modelo de regressão exponencial com o termo de fragilidade utilizaremos um conjunto de dados extraídos do software IBM SPSS (2010) que é proveniente de uma empresa de telecomunicação, a Telecom.

A amostra é composta por 1000 clientes, dentre os quais 274 ocasionaram o *churn* e 726 foram censurados, sendo que o experimento teve duração máxima de 72 meses.

Na Tabela 3.1 podemos observar com detalhes as 20 covariáveis, como idade, renda, estado civil, sexo e entre outras, seguida da variável resposta *churn*.

Tabela 3.1: Descrição das variáveis do conjunto Telecom.

Variável	**Valores**	**Informações**
Regiao	Região 1 Região 2 Região 3 Região 4 Região 5	Região de Residência
ClienteDuracao		Tempo como cliente (mês)
Idade		Idade (anos)
EstadoCivil	Não-casado Casado	Estado civil
ResidenciaDuracao		Tempo na residência (ano)
Renda		Renda familiar anual (milhar)

continua na página seguinte

continuação da página anterior

Variável	**Valores**	**Informações**
Educacao	Ensino médio incompleto Ensino médio completo Ensino superior incompleto Ensino superior completo Pós-graduação completa	Nível educacional
EmpregoDuracao		Tempo de trabalho no emprego atual (ano)
Aposentado	Não Sim	Aposentadoria
Sexo	Masculino Feminino	Gênero
Residentes		Quantidade de pessoas na residência
Servico1	Não Sim	Possui serviço 1
Servico2	Não Sim	Possui serviço 2
Servico3	Não Sim	Possui serviço 3
Servico4	Não Sim	Possui serviço 4
ReceitaS1		Receita no último mês com serviço 1
ReceitaS2		Receita no último mês com serviço 2
ReceitaS3		Receita no último mês com serviço 3
ReceitaS4		Receita no último mês com serviço 4
ClienteCategoria	Básico Eletrônico Plus Total	Categoria do cliente
Churn	Não Sim	Abandonou o serviço no mês anterior

Com a finalidade de analisarmos a proporção de *churn* com relação a cada variável, fizemos os seguintes gráficos:

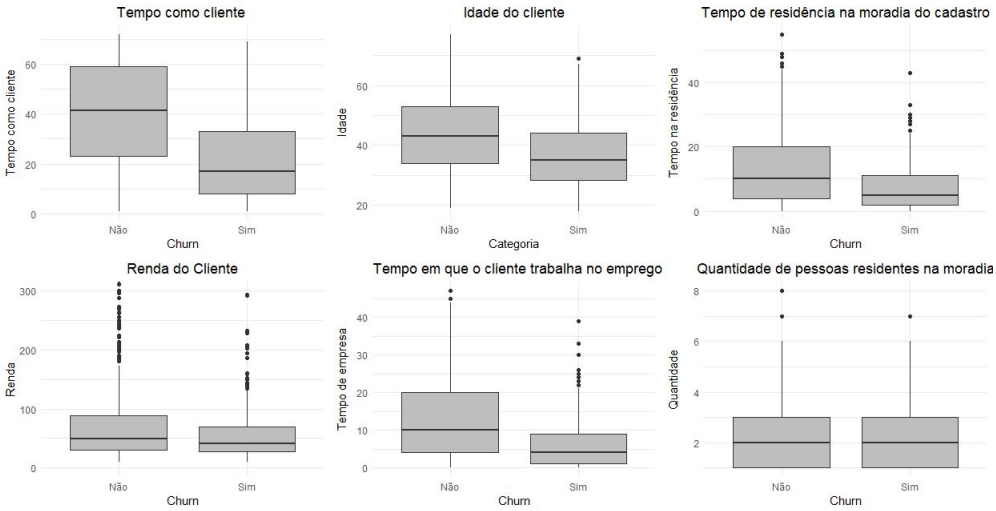

Figura 3.1: Box-Plot das variáveis "Tempo como cliente", "Idade", "Tempo de residência", "Renda", "Tempo de emprego" e "Residentes", de acordo com a ocorrência do *churn*.

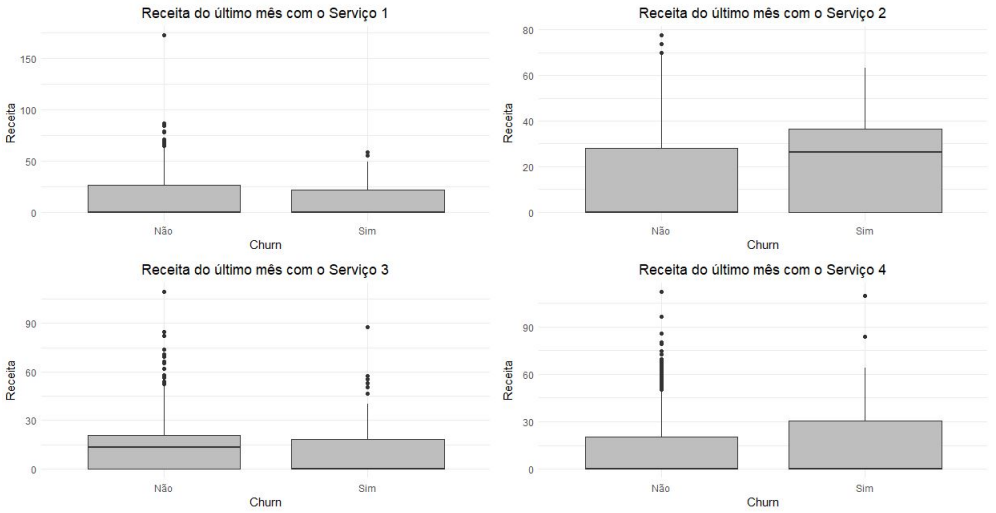

Figura 3.2: Box-Plot das variáveis de Receita dos serviços de acordo com a ocorrência do *churn*.

Ao analisarmos a ocorrência do *churn* de acordo com as variáveis numéricas, vistas nas Figuras 3.1 e 3.2, notamos que pode haver uma diferença entre os contratantes que abandonaram ou não o serviço com relação às variáveis que representam o tempo como cliente, a idade, o tempo de moradia na residência de cadastro, o tempo em que o cliente está em seu emprego e a receita do último mês com o serviço 2.

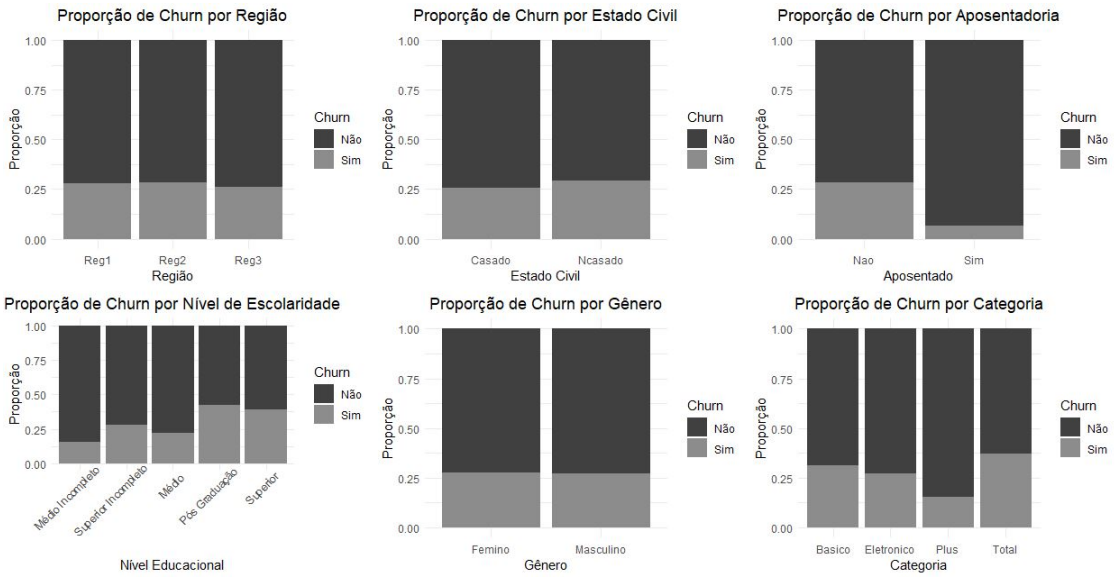

Figura 3.3: Gráfico de barras empilhadas da proporção de *churn* de acordo com as variáveis "Região", "Estado Civil", "Nível de Escolaridade", "Gênero", "Aposentado" e "Categoria do Cliente".

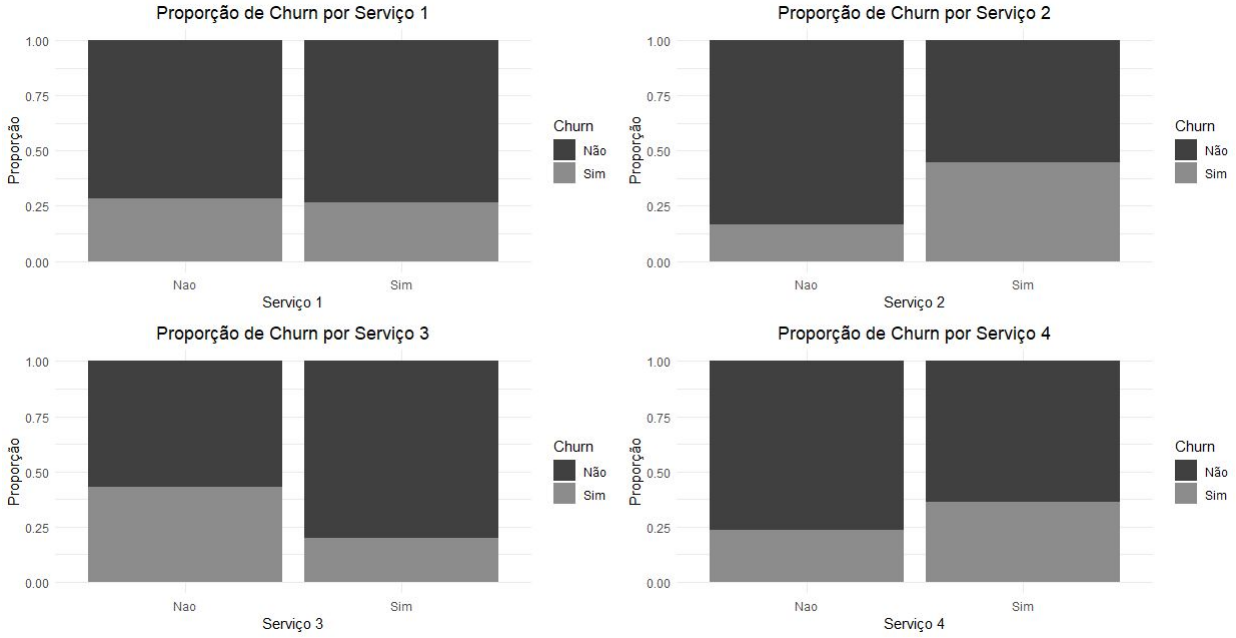

Figura 3.4: Gráfico de barras empilhadas da proporção de *churn* de acordo com o serviço.

Nas Figuras 3.3 e 3.4 vemos a proporção de ocorrência do *churn* conforme as variáveis categóricas. Por meio delas, é possível notar que possivelmente as variáveis que indicam aposentadoria, escolaridade, categoria do cliente e utilização do serviço 2 e 3 são significativamente diferentes para clientes que ocasionam ou não o *churn*.

3.1 Estudo da covariável categoria do cliente como segmento de grupo

Com a finalidade de tornar o estudo mais aplicável, as análises serão realizadas para grupos de clientes que possuem similaridade. Por isso, a covariável categoria do cliente, a princípio, será utilizada como segmentação de grupo. Logo os clientes serão divididos de acordo com o plano assinado na empresa Telecom.

Tabela 3.2: O *Churn* nas categorias de clientes.

Categoria	Número de clientes	Quantidade de churn	% *churn*
Básico	266	83	30,3%
Eletrônico	217	59	21,5%
Plus	281	44	16,1%
Total	236	88	32,1%

Como podemos observar na Tabela 3.2, os clientes estão distribuídos em quatro categorias: Básico, Eletrônico, Plus e Total, esses por sua vez são referentes a planos contratados. Cada plano possui uma abrangência e uma qualidade diferente dos demais, por isso entendemos que cada um possui consumidores de perfil semelhante. Também é possível ver que o número de clientes é bem distribuído entre as categorias, entretanto podemos concluir que o Plus possui uma porcentagem de *churn* menor entre os planos.

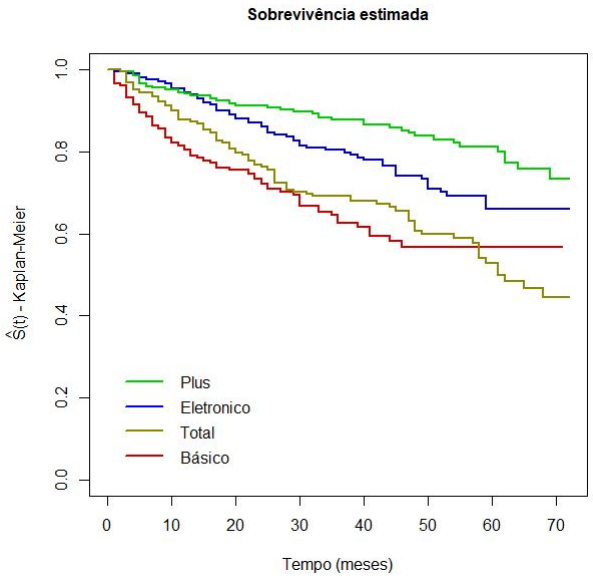

Figura 3.5: Gráfico do Estimador de Kaplan-Meier por categoria do plano de serviço.

Calculamos o estimador de Kaplan-Meier para as quatro categorias. Na Figura 3.5, podemos ver que o plano Plus, segundo o estimador, possui a maior sobrevivência entre as categorias, seguido do plano Eletrônico. As categorias Total e Básico, possuem os menores tempos de sobrevida, ou seja, os clientes contratantes desses planos tendem a ocasionar o *churn* em menor tempo. A partir desse gráfico também é possível notar que as categorias representados pelas cores azul e vermelho tem sua sobrevivência estabilizada após um certo tempo decorrido. Com essa análise, esperamos que os contratantes dos planos Plus e Eletrônico mantenham mais tempo de vínculo com a empresa.

Para estimar o modelo de sobrevivência compartilhada, considerado neste estudo, utilizamos os pacotes *survival* e *parfm* presentes no software R (R Core Team, 2019), com eles é possível estimar os parâmetros do modelo, predizer a fragilidade de cada grupo e também a sobrevivência de cada clientes utilizando as covariáveis individualmente e a fragilidade do grupo a qual ele pertence.

Uma vez que tínhamos como premissa resumir as covariáveis dos indivíduos de forma a representar o segmento a qual ele faz parte, para assim tornar a análise mais aplicável, concluímos que seria prudente somente utilizar as variáveis preditoras quantitativas para a modelagem, isso possibilitou a utilização da medida de resumo mais usual, a média. As demais variáveis são boas opções para segmentação de grupo, principalmente aquelas que observamos na análise descritiva como significativas na discriminalização do *churn*.

No modelo completo, algumas variáveis não apresentaram significância no ajuste, e com o intuito de melhorar o modelo, analisamos esse com e sem a presença dessas covariáveis. Consideramos o melhor ajuste aquele que possui uma quantidade menor de variáveis explicativas e que tenha o menor AIC e BIC comparado aos demais.

Tabela 3.3: AIC e BIC dos modelos com categoria do cliente como segmento.

Variáveis	Modelo 1	Modelo 2	Modelo 3	Modelo 4
Idade	X			
ResidenciaDuracao	X	X	X	X
Renda	X	X	X	X
EmpregoDuracao	X	X	X	X
Residentes	X	X	X	X
ReceitaS1	X	X		
ReceitaS2	X	X	X	X
ReceitaS3	X	X	X	X
ReceitaS4	X	X	X	
AIC	2941	2939	2937	2937
BIC	2995	2988	2982	2976

Os modelos descritos na Tabela 3.3 foram obtidos através da extração uma a uma das covariáveis que não foram significativas no modelo completo (Modelo 1).

O modelo 4 foi ajustado com seis variáveis, todas significativas para o ajuste como pode ser visto no Quadro 3.1 a partir do valor-p emitido pelo comando *parfm* do *software* R. Três dessas variáveis já apresentavam indícios que seriam boas preditoras pela na análise descritiva vista através da Figura 3.1. Esse modelo também foi o que resultou nos melhores valores de AIC e BIC, por isso ele foi escolhido como o modelo mais adequado dentre esses descritos na Tabela 3.3.

Quadro 3.1: Saída do ajuste do modelo 4.

```
        ESTIMATE SE      p-val
theta    0.054    0.049
lambda   0.037    0.008
z2      -0.054    0.009  <.001 ***
z3       0.001    0.001  0.015 *
z4      -0.072    0.010  <.001 ***
z5      -0.115    0.045  0.01  *
z7       0.014    0.004  <.001 ***
z8      -0.028    0.006  <.001 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dessa forma, a função de risco e de sobrevivência compartilhada ajustada podem ser escritas como

$$\hat{h}_k\left(t_k|\hat{w}_k, z_k\right) = \hat{w}_k\ 0,037\ \exp\{-0.054\ z_{k2} + 0.001\ z_{k3} - 0.072\ z_{k4} - 0.115\ z_{k5} + 0.014\ z_{k7} - 0.028\ z_{k8}\}, \tag{3.1}$$

$$\hat{S}_k\left(t_k|\hat{w}_k, z_k\right) = \exp\left\{ - \hat{w}_k\, 0.037\, t_k \exp\left[- 0.054\ z_{k2} + 0.001\ z_{k3} - 0.072\ z_{k4} - 0.115\ z_{k5} + 0.014\ z_{k7} - 0.028\ z_{k8}\right] \right\}, \tag{3.2}$$

em que $\hat{w}_k$ é a fragilidade predita, associada ao k-ésimo grupo, $\mathbf{z}_k$ o vetor de covariável do grupo k sendo z_2 o "Tempo de residência", z_3 a "Renda", z_4 o "Tempo de emprego",

z_5 o "Número de Residentes", z_7 a "Receita do Serviço 2" e z_8 a "Receita do Serviço 3" e t_k a média dos tempos de vida do k-ésimo grupo.

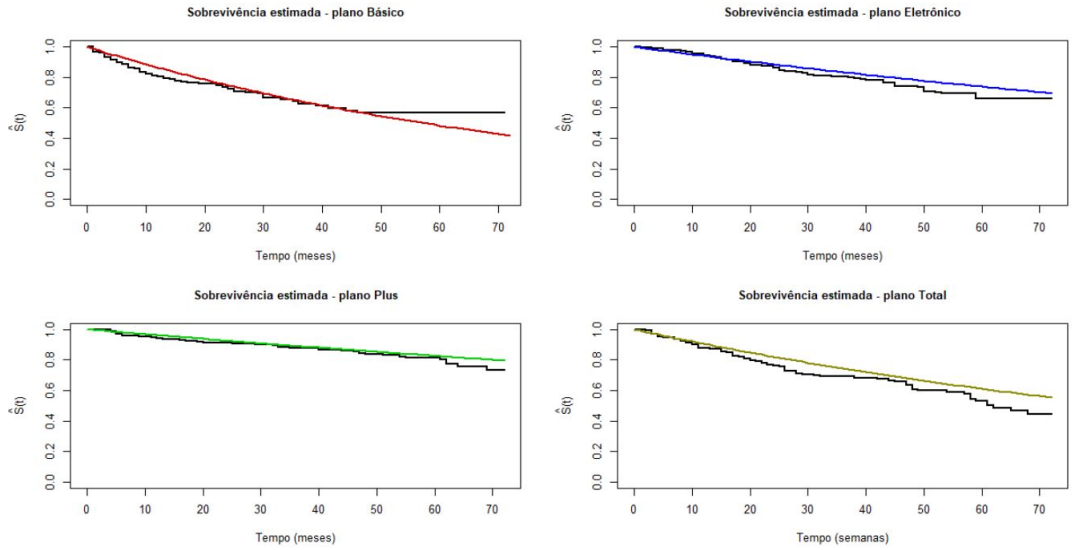

Figura 3.6: Gráfico do estimador de Kaplan-Meier para cada categoria do cliente.

Na Figura 3.6 podemos observar as mesmas curvas coloridas vistas na Figura 3.5, agora em preto, que representam o ajuste pelo estimador de Kaplan-Meier. Já as curvas coloridas representam as funções de sobrevivência ajustadas pelo modelo 4 para cada um dos grupos. Curvas próximas são evidências que o modelo está bem ajustado e faz boas predições.

A fragilidade predita por esse modelo é de $1,303$ para o plano Básico, $0,759$ para o plano Eletrônico, $0,866$ para o Plus e 1.072 para o Total. Quanto mais próximo a 1 está esse valor mais próximo o grupo está do risco neutro, ou seja, segundo a fragilidade o grupo não tem risco nem baixo nem alto de ocasionar o *churn*. Números abaixo desse marco, indicam menor risco e acima maior, por isso podemos concluir que o plano Básico, segundo o modelo, é o grupo com mais risco de ocasionar o *churn* e o plano Eletrônico o menos arriscado.

3.2 Estudo da covariável escolaridade como segmento de grupo

Além da divisão pelo plano de assinatura, uma segunda segmentação foi utilizada levando-se em conta o nível de escolaridade do cliente, com a qual foram obtidos resultados

muito bons que serão descritos nesta seção.

A princípio a variável Escolaridade contava com 5 níveis como descrito na Tabela 3.1, entretanto, por meio de tentativas, percebemos que os resultados ficaram mais satisfatórios ao se utilizar quatro níveis. Por isso, unimos "Ensino superior completo" e "Pós-graduação completa"em um nível e o nomeamos como "Ensino Superior+".

Tabela 3.4: O *Churn* nos novos níveis de Escolaridade.

Categoria	Número de clientes	Quantidade de churn	% *churn*
Ensino Médio Incompleto	204	32	15,7%
Ensino Médio Completo	287	59	20,6%
Ensino Superior Incompleto	209	63	30,1%
Ensino Superior+	300	120	40,0%

Observamos na Tabela 3.4 os clientes distribuídos nas quatro categorias de acordo com a escolaridade. Entendemos que essa categoria classifica o público em vários aspectos, como por exemplo idade, faixa salarial, local de moradia, e por isso concluímos que cada uma possui consumidores de perfis semelhantes.

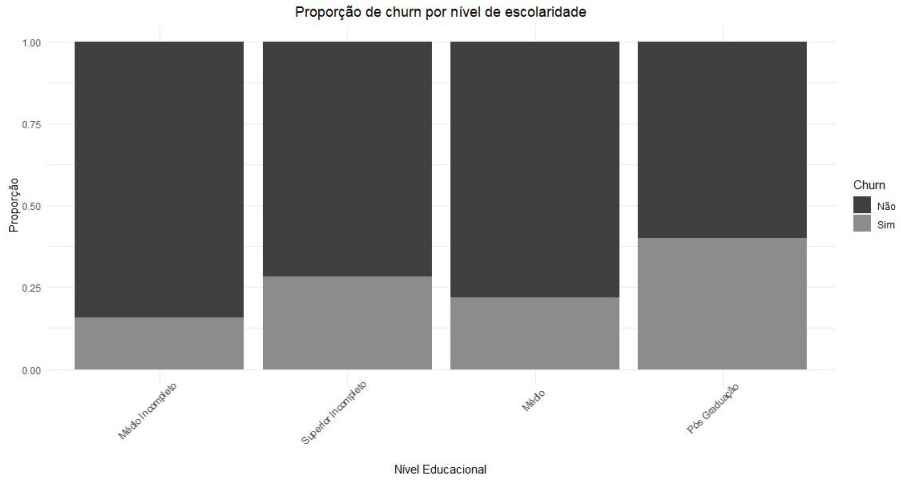

Figura 3.7: Gráfico de barras empilhadas da proporção de *churn* dos clientes, de acordo com o nível de escolaridade.

Por meio da Tabela 3.2 e da Figura 3.7 vemos que o nível de escolaridade ordena o porcentual de *churn*, ou seja, quanto maior a formação acadêmica maior a proporção de pessoas que abandonam o serviço.

Na Figura 3.8 vemos o estimador de Kaplan-Meier para as quatro categorias. Podemos ver que indivíduos com o ensino médio incompleto, segundo o estimador, possui a

maior sobrevivência entre as categorias, seguido de quem tem o ensino médio completo. As categorias formadas por clientes com ensino superior incompleto e completo, possuem o menor tempo de sobrevida, ou seja, os clientes contratantes desses planos tendem a ocasionar o *churn* em menor tempo. Por meio desse gráfico também é possível notar que as categorias representados pelas cores azul, vermelho e dourado tem sua sobrevivência estabilizada após um certo tempo decorrido. Com essa análise, esperamos que os contratantes com formação acadêmica até o segundo grau utilizem o serviço disponibilizado pela empresa por mais tempo.

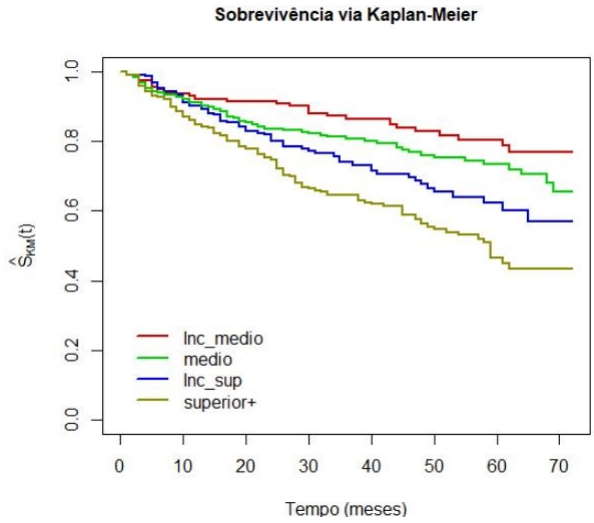

Figura 3.8: Gráfico do Estimador de Kaplan-Meier por nível de escolaridade.

Utilizando as mesmas variáveis preditoras do modelo com segmento categoria do cliente, ajustamos um modelo com a segmentação educação, ou seja, agrupamos os clientes segundo o nível de escolaridade. Inicialmente ajustamos um modelo com todas as variáveis, entretanto o algoritmo não convergiu, por isso fomos adicionando as variáveis uma a uma e acompanhando o AIC e o BIC.

Tabela 3.5: AIC e BIC dos modelos com escolaridade como segmento.

Variáveis	Mod. 1	Mod. 2	Mod. 3	Mod. 4	Mod. 5	Mod. 6	Mod. 7	Mod. 8	Mod. 9	Mod. 10
Idade	X	X	X	X				X	X	X
ResidenciaDuracao		X	X	X	X	X	X	X	X	X
Renda			X							
EmpregoDuracao				X	X	X	X			
Residentes					X	X	X	X	X	X
ReceitaS1						X				
ReceitaS2							X	X	X	X
ReceitaS3									X	X
ReceitaS4										X
AIC	3061	3035	3035	2995	2986	2986	Não Converge	3014	2980	2982
BIC	3076	3055	3060	3019	3019	3015	Não converge	3044	3015	3021

Segundo os valores de AIC e BIC presentes na Tabela 3.5, o melhor modelo que utiliza a segmentação escolaridade é o Modelo 9, pois ele apresenta os menores valores para ambos os critérios de seleção, além de possuir menos variáveis preditoras que o Modelo 10. O modelo 9 é formado pelas covariáveis "Idade", "Tempo de residência", "Residentes", "Receita do Serviço 2" e "Receita do Serviço 3".

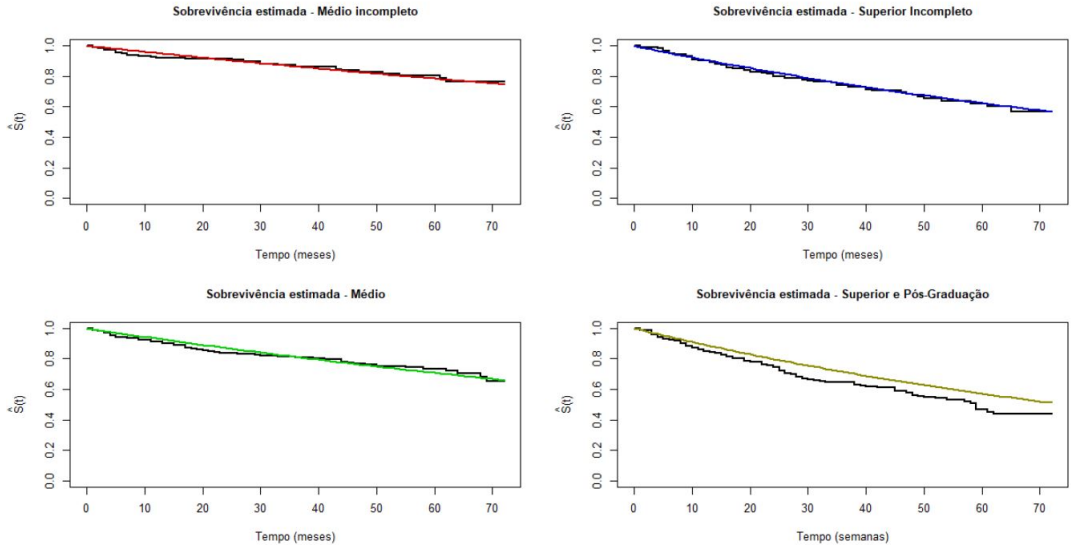

Figura 3.9: Gráfico do estimador de Kaplan-Meier para cada nível escolar.

Na Figura 3.9 temos em colorido a curva de sobrevivência ajustada com o modelo 9 e o ajuste por Kaplan-Meier em preto. Para os níveis médio incompleto, médio e superior incompleto podemos dizer que o modelo é assertivo, pois as curvas se sobrepõe, entretanto para o nível que engloba o superior e a pós-graduação o mesmo não acontece.

A fragilidade predita pelo modelo 9 é $0,99$ para o grupo do ensino médio incompleto, $0,995$ para o ensino médio, $0,993$ para o ensino superior incompleto e $1,022$ para o superior e pós graduação. Nesse modelo as fragilidades de todos os grupos se aproximaram de 1, o que indica que o modelo não discriminou o risco de forma satisfatória.

Dado que a predição da fragilidade através do modelo 9 não diferenciou os riscos dos níveis de escolaridade, resolvemos predize-la por meio dos outros modelos, e concluímos que o modelo 2, apesar de apresentar um AIC e BIC maior, predisse uma fragilidade de $0,765$ para o grupo do ensino médio incompleto, $0,862$ para o médio, $1,013$ para o superior incompleto e $1,360$ para o ensino superior completo e pós graduação.

As variáveis desse modelo, "Idade" e "Tempo de residência" são significativas para o ajuste segundo o valor-p visto no Quadro 3.2, dado pelo comando *parfm* do *software* R. Elas já se mostravam ser boas preditoras na análise descritiva descrita na Figura 3.1.

Quadro 3.2: Saída do ajuste do modelo 2

```
         ESTIMATE SE      p-val
theta     0.061    0.054
lambda    0.056    0.015
z1       -0.036    0.007  <.001  ***
z2       -0.053    0.010  <.001  ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
\label{quadro2}
```

Sendo assim, a função de risco e de sobrevivência compartilhada ajustada podem ser escritas como

$$\hat{h}_k\left(t_k|w_k, z_k\right) = \hat{w}_k \; 0,056 \; \exp\{-0,036 \; z_{k1} - 0,053 \; z_{k2}\}, \tag{3.3}$$

$$\hat{S}_k\left(t_k|\hat{w}_k, z_k\right) = \exp\left\{ - \hat{w}_k \, 0.037 \, t_k \exp\left[- 0,036 \; z_{k1} - 0,053 \; z_{k2}\right]\right\}, \tag{3.4}$$

em que $\hat{w}_k$ é a fragilidade predita, associada ao k-ésimo grupo, e $\mathbf{z}_k$ o vetor de covariável do grupo k sendo z_1 a "Idade" e z_2 o "Tempo de residência" e t_k a média dos tempos de vida do k-ésimo grupo.

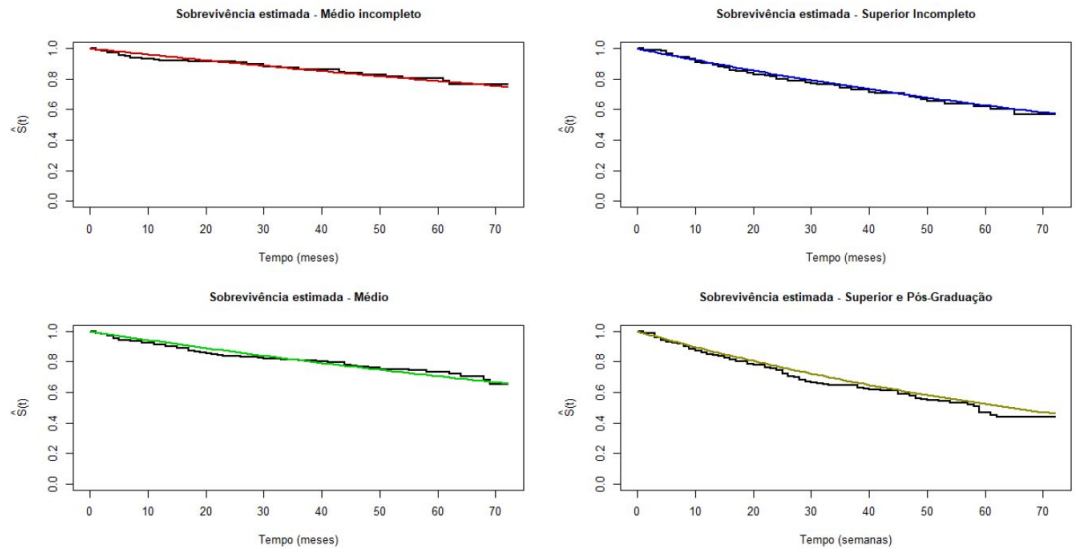

Figura 3.10: Gráfico do estimador de Kaplan-Meier para cada nível escolar.

Temos, na Figura 3.10, em colorido a curva de sobrevivência ajustada com o modelo 2

e o ajuste por Kaplan-Meier em preto. A sobreposição das curvas nos indica que o modelo faz uma boa predição do tempo, logo acreditamos que ele seja o melhor modelo para o segmento de nível de escolaridade.

3.3 Comparação dos modelos de segmentos de grupo

Apesar das duas segmentações, a categoria do cliente e a escolaridade, não fazerem a mesma divisão do público e por isso serem implementadas em âmbitos diferentes, gostaríamos de nessa seção fazer um comparativo entre os dois melhores modelos de cada segmento.

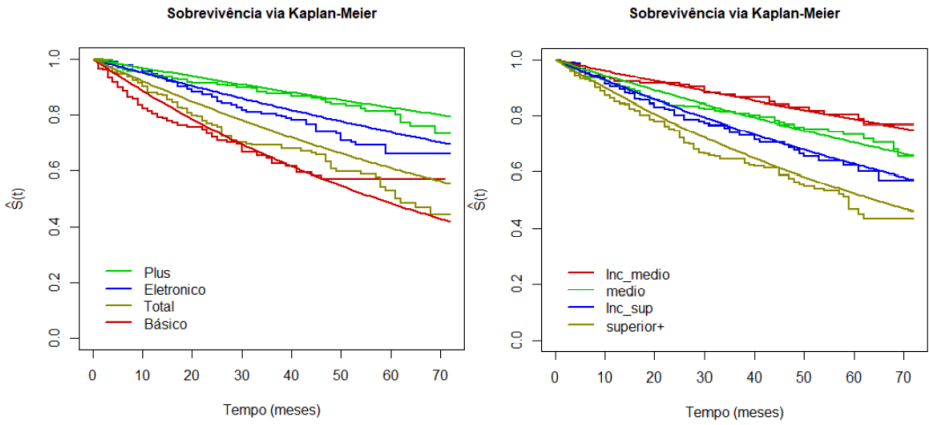

Figura 3.11: Gráfico do estimador de Kaplan-Meier para cada modelo.

Tabela 3.6: Termo de fragilidade predito dos segmentos por grupo.

Segmentação	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Categoria do cliente (modelo 4)	1,303	0,759	0,866	1,072
Nível de escolaridade (modelo 2)	0,765	0,862	1,013	1,360

Como é possível observar na Figura 3.11 o modelo que tem como segmento a educação produz uma predição melhor, pois a sobrevivência se assemelha muito a curva de Kaplan-Meier, comparado ao modelo que tem a categoria do cliente como segmentação. Entretanto, dado que o modelo 2 apresentou um valor maior para os critérios de seleção, podemos dizer que ele apresenta um maior viés.

Contudo, como podemos ver na Tabela 3.6 que ambos os modelos discriminam bem os grupos a partir da fragilidade predita, logo podemos concluir que os dois modelos

atingiram a expectativa esperada e podem ser usados de acordo com o objetivo e estratégia da empresa contratante.

Capítulo 4

Conclusão

Com o desenvolvimento global tornou-se comum a concorrência entre empresas e a variabilidade de serviços oferecidos aos consumidores. Essa situação acarretou no aumento da migração de clientes entre as empresas e por consequência o *churn*. A fim de fidelizar e obter uma maior retenção de consumidores surgiu a necessidade de estudar e retardar a quebra de vínculo com os clientes.

O estudo teve como objetivo obter um modelo de sobrevivência para grupos de clientes, segmentados conforme algum critério preestabelecido, utilizando uma componente de fragilidade para predizer a probabilidade de ocorrência do *churn* e aplicar a predição para o grupo como um todo.

Toda a metodologia foi desenvolvida utilizando um um distribuição exponencial para os tempos de vida. Esse modelo foi utilizado mesmo sabendo que ele não descreve com precisão o comportamento do risco, entretanto o utilizamos devido a sua fácil aplicação e interpretação.

Nesse estudo sugerimos uma distribuição gama para a componente de fragilidade, devido a sua popularidade. Na construção do modelo de sobrevivência utilizamos as variáveis preditoras quantitativas, uma vez que precisávamos criar medidas resumo para os grupos do segmento e utilizamos a média como medida. Os dados da Telecom apresentavam muitas variáveis qualitativas e uma forma de utilizá-las nesse estudo foi como segmentação, com isso criamos dois modelos com segmentações diferentes, plano do serviço e nível de escolaridade.

A princípio, os modelos foram selecionados através dos valores do AIC e do BIC, entretanto percebemos, que nesse caso, nem sempre o modelo que possui a melhor predição, segundo esses critérios, é o modelo que melhor discrimina os grupos segundo a fragilidade.

Os ajustes foram verificados graficamente por meio de curvas de sobrevivências e do estimador de Kaplan-Meier dos grupos de clientes. Concluímos que o modelo com segmento escolaridade foi mais assertivo, entretanto apresenta maior viés, enquanto que o modelo com segmentação por categoria de cliente não possui um ajuste tão preciso, mas possui um viés menor.

Esse tipo de estudo é essencial para identificar grupos de clientes mais propensos ao *churn* e utilizar estratégias de retenção direcionada a esse conjunto de consumidores ou investir no marketing voltado a grupos que tenham maior tempo de sobrevivência e por isso tardarão a abandonar o serviço.

Essa análise possibilita retrabalhar futuramente alguns pontos como a implementação de um modelo que melhor descreva os tempos de vida, por exemplo o modelo *Weibull*, a validação do componente de fragilidade através de um intervalo e a utilização das variáveis qualitativas como subgrupos. Em vez de só dividirmos os clientes por nível de escolaridade podemos subdividi-los por estado civil, de modo a termos uma categoria de contratantes com ensino superior incompleto e casados, por exemplo.

Referências Bibliográficas

Abramowitz, M. e Stegun, I. A. (1965). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government Printing Office.

Berry, M. J. e Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

Carvalho, M. S., Andreozzi, V. L., Codeço, C. T., Campos, D. P., Barbosa, M. T. S. e Shimakura, S. E. (2011). *Análise de sobrevivência: teoria e aplicações em saúde*. SciELO-Editora FIOCRUZ.

Casella, G. e Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Collett, D. (2015). *Modelling survival data in medical research*. CRC press.

Colosimo, E. A. e Giolo, S. R. (2006). *Análise de sobrevivência aplicada*. Editora Blucher.

Fogo, J. C. (2007). *Modelo de regressao para um processo de renovaçao Weibull com termo de fragilidade*. Ph.D. thesis, Universidade de São Paulo.

Klein, J. P. e Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data*, volume 362. John Wiley & Sons.

Lu, J. e Park, O. (2003). Modeling customer lifetime value using survival analysis—an application in the telecommunications industry. *Data Mining Techniques*, pages 120–128.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Tomazella, V. L. D. (2003). *Modelagem de dados de eventos recorrentes via processo de Poisson com termo de fragilidade.*. Ph.D. thesis, Universidade de São Paulo.

Wienke, A. (2007). *Frailty models in survival analysis*. Universitäts-und Landesbibliothek Sachsen-Anhalt.

Apêndice A

Código

```
library(MASS)
library(stats4)
library(survival)
library(parfm)
library(readr)
library(ggplot2)
library(ggridges)
library(tidyr)
library(tidyverse)
library(ggthemes)
library(cowplot)
library(magrittr)

## Lendo os dados
telecom <- read_delim("C:/Users/Acer/Desktop/TG/Material/Dados_telecom.csv",
                      ";", escape_double = FALSE, trim_ws = TRUE)
attach(telecom)


##############################Análise Exploratória#########################

summary(telecom)
```

```
#Região
tregiao = data.frame(xtabs(~regiao + churn, data = telecom))
Barra_1=ggplot(tregiao, aes(x = regiao, y = Freq, fill = churn)) +
  geom_bar(position = "fill", stat="identity")+
  scale_fill_manual(values = c("gray25","gray55"))+
  labs(x = "Região", y = "Proporção",
       title = "Proporção de Churn por Região",
       fill="Churn")+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

#Duração como cliente
Box_1=ggplot(telecom) +
  geom_boxplot(aes(x = churn, y = clienduracao),fill="grey")+
  labs(x = "Churn", y = "Tempo como cliente",
       title = "Tempo como cliente ")+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

#Idade
Box_2=ggplot(telecom) +
  geom_boxplot(aes(x = churn, y = idade),fill="grey")+
  labs(x = "Categoria", y = "Idade",
       title = "Idade do cliente") +
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

#Estado civil
tcivil = data.frame(xtabs(~estadocivil + churn, data = telecom))
Barra_2=ggplot(tcivil, aes(x = estadocivil, y = Freq, fill = churn)) +
  geom_bar(position = "fill", stat="identity")+
  scale_fill_manual(values = c("gray25","gray55"))+
```

```
  labs(x = "Estado Civil", y = "Proporção",
       title = "Proporção de Churn por Estado Civil",
       fill="Churn")+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

#Duração na residência
Box_3=ggplot(telecom) +
  geom_boxplot(aes(x = churn, y = residduracao),fill="grey")+
  labs(x = "Churn", y = "Tempo na residência",
       title = "Tempo de residência na moradia do cadastro") +
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

#Renda
Box_4=ggplot(telecom) +
  geom_boxplot(aes(x = churn, y = renda),fill="grey")+
  labs(x = "Churn", y = "Renda",title = "Renda do Cliente") +
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

#Educação
teducacao = data.frame(xtabs(~educacao + churn, data = telecom))
Barra_3=ggplot(teducacao, aes(x = educacao, y = Freq, fill = churn)) +
  geom_bar(position = "fill", stat="identity")+
  scale_fill_manual(values = c("gray25","gray55"))+
  scale_x_discrete(labels=c("Médio Incompleto",
                            "Superior Incompleto","Médio",
                            "Pós Graduação","Superior"))+
  labs(x = "Nível Educacional", y = "Proporção",
       title = "Proporção de Churn por Nível de Escolaridade",
       fill="Churn")+theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5),
```

```
        axis.text.x = element_text(angle = 45,hjust = 0.8))

#Duração no emprego
Box_5=ggplot(telecom) +
  geom_boxplot(aes(x = churn, y = emprduracao),fill="grey")+
  labs(x = "Churn", y = "Tempo de empresa",
  title = "Tempo em que o cliente trabalha no emprego") +
  theme(plot.title = element_text(hjust = 0.5))+theme_minimal()

#Aposentado
taposentado = data.frame(xtabs(~aposentado + churn, data = telecom))
Barra_4=ggplot(taposentado, aes(x = aposentado, y = Freq, fill = churn)) +
  geom_bar(position = "fill", stat="identity")+
  scale_fill_manual(values = c("gray25","gray55"))+
  labs(x = "Aposentado", y = "Proporção",
       title = "Proporção de Churn por Aposentadoria",
       fill="Churn")+theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

#Sexo
tsexo = data.frame(xtabs(~sexo + churn, data = telecom))
Barra_5=ggplot(tsexo, aes(x = sexo, y = Freq, fill = churn)) +
  geom_bar(position = "fill", stat="identity")+
  scale_fill_manual(values = c("gray25","gray55"))+
  scale_x_discrete(labels=c("Femino","Masculino"))+
  labs(x = "Gênero", y = "Proporção",
       title = "Proporção de Churn por Gênero",
       fill="Churn")+theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

#Residentes
Box_6=ggplot(telecom) +
  geom_boxplot(aes(x = churn, y = residentes),fill="grey")+
```

```
  labs(x = "Churn", y = "Quantidade",
       title = "Quantidade de pessoas residentes na moradia") +
  theme_minimal()+theme(plot.title = element_text(hjust = 0.5))

plot_grid(Box_1,Box_2,Box_3,Box_4,Box_5,Box_6,align="hv")

#Serviço1
ts1 = data.frame(xtabs(~servico1 + churn, data = telecom))
s1=ggplot(ts1, aes(x = servico1, y = Freq, fill = churn)) +
  geom_bar(position = "fill", stat="identity")+
  scale_fill_manual(values = c("gray25","gray55"))+
  labs(x = "Serviço 1", y = "Proporção",
       title = "Proporção de Churn por Serviço 1",
       fill="Churn")+theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

#Serviço2
ts2 = data.frame(xtabs(~servico2 + churn, data = telecom))
s2=ggplot(ts2, aes(x = servico2, y = Freq, fill = churn)) +
  geom_bar(position = "fill", stat="identity")+
  scale_fill_manual(values = c("gray25","gray55"))+
  labs(x = "Serviço 2", y = "Proporção",
       title = "Proporção de Churn por Serviço 2",
       fill="Churn")+theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

#Serviço3
ts3 = data.frame(xtabs(~servico3 + churn, data = telecom))
s3=ggplot(ts3, aes(x = servico3, y = Freq, fill = churn)) +
  geom_bar(position = "fill", stat="identity")+
  scale_fill_manual(values = c("gray25","gray55"))+
  labs(x = "Serviço 3", y = "Proporção",
       title = "Proporção de Churn por Serviço 3",
```

```
       fill="Churn")+theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

#Serviço4
ts4 = data.frame(xtabs(~servico4 + churn, data = telecom))
s4=ggplot(ts4, aes(x = servico4, y = Freq, fill = churn)) +
  geom_bar(position = "fill", stat="identity")+
  scale_fill_manual(values = c("gray25","gray55"))+
  labs(x = "Serviço 4", y = "Proporção",
       title = "Proporção de Churn por Serviço 4",
       fill="Churn")+theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

plot_grid(s1,s2,s3,s4,align="hv")

#Receita1
r1=ggplot(telecom) +
  geom_boxplot(aes(x = churn, y = receitas1),fill="grey")+
  labs(x = "Churn", y = "Receita",
       title = "Receita do último mês com o Serviço 1") +
  theme_minimal()+theme(plot.title = element_text(hjust = 0.5))

#Receita2
r2=ggplot(telecom) +
  geom_boxplot(aes(x = churn, y = receitas2),fill="grey")+
  labs(x = "Churn", y = "Receita",
       title = "Receita do último mês com o Serviço 2") +
  theme_minimal()+theme(plot.title = element_text(hjust = 0.5))

#Receita3
r3=ggplot(telecom) +
  geom_boxplot(aes(x = churn, y = receitas3),fill="grey")+
  labs(x = "Churn", y = "Receita",
```

```
       title = "Receita do último mês com o Serviço 3") +
  theme_minimal()+theme(plot.title = element_text(hjust = 0.5))

#Receita4
r4=ggplot(telecom) +
  geom_boxplot(aes(x = churn, y = receitas4),fill="grey")+
  labs(x = "Churn", y = "Receita",
       title = "Receita do último mês com o Serviço 4") +
  theme_minimal()+theme(plot.title = element_text(hjust = 0.5))

#Churn
tchurn = data.frame(xtabs(~churn + clientecateg, data = telecom))
Barra_6=ggplot(tchurn, aes(x = clientecateg, y = Freq, fill = churn)) +
  geom_bar(position = "fill", stat="identity")+
  scale_fill_manual(values = c("gray25","gray55"))+
  labs(x = "Categoria", y = "Proporção",
       title = "Proporção de Churn por Categoria",
       fill="Churn")+theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

plot_grid(Barra_1,Barra_2,Barra_4,Barra_3,Barra_5,Barra_6,align="v")

##############função para alguns cálculos################
somas <-function(grupo, tempo, h.bt, vz, cens, j){
  grp <- names(table(grupo))
  g <- dim(table(grupo))
  d.k <- numeric(g)
  T.k <- numeric(g)
  t.k <- numeric(g)
  z1 <- numeric(g)
  z2 <- numeric(g)
  z3 <- numeric(g)
  z4 <- numeric(g)
```

```
  z5 <- numeric(g)
  z6 <- numeric(g)
  for(k in 1:g){
    linhas <- which(grupo==grp[k])
    d.k[k] <- sum(cens[linhas])
    t.k[k] <- sum(tempo[linhas])
    z1[k] <- mean(vz[linhas,1])
    z2[k] <- mean(vz[linhas,2])
    z3[k] <- mean(vz[linhas,3])
    z4[k] <- mean(vz[linhas,4])
    z5[k] <- mean(vz[linhas,5])
    z6[k] <- mean(vz[linhas,6])
    T.k[k] <- exp(h.bt%*%t(vz[linhas,1:j]))%*%tempo[linhas]
  }
  list(z1=z1, z2=z2, z3=z3, z4=z4, z5=z5, z6=z6, T.k=T.k, d.k=d.k, g=g, t.k=t.k)
}

#############################################################################
## Caso 1: Grupos = Tipo de Plano
#############################################################################
# valores iniciais
o <- order(clientecateg)
grupox <- c(rep(1,266),rep(2,217),rep(3,281),rep(4,236))
vz1 <- idade[o]
vz2 <- residduracao[o]
vz3 <- renda[o]
vz4 <- emprduracao[o]
vz5 <- residentes[o]
vz6 <- receitas1[o]
vz7 <- receitas2[o]
vz8 <- receitas3[o]
vz9 <- receitas4[o]
vz <- cbind(vz1, vz2, vz3, vz4, vz5, vz6, vz7, vz8, vz9)
```

```
tempo <- clienduracao[o]
cens <- churn01[o]

dados <- data.frame(tempo,grupox,vz,cens)
attach(dados)

## a) modelo com as covariáveis:
##    idade, residduracao, renda, emprduracao, residentes,
##    receitas1, receitas2, receita3, receita4

ajuste1 <- parfm(Surv(tempo,cens)~vz1+vz2+vz3+vz4+vz5+vz6+vz7+vz8+vz9,
                 cluster="grupox",
                 data=dados, dist="exponential", frailty="gamma")
ajuste1
AIC(ajuste1)
#[1] 2941.193
BIC(ajuste1)
#[1] 2995.178

## b) modelo com as covariáveis:
##    residduracao, renda, emprduracao, residentes,
##    receitas1, receitas2, receita3, receita4

ajuste2 <- parfm(Surv(tempo,cens)~vz2+vz3+vz4+vz5+vz6+vz7+vz8+vz9,
                 cluster="grupox",
                 data=dados, dist="exponential", frailty="gamma")
ajuste2
AIC(ajuste2)
#[1] 2939.245
BIC(ajuste2)
#[1] 22988.323

## c) modelo com as covariáveis:
```

```
##    residduracao, renda, emprduracao, residentes,
##    receitas2, receita3, receita4
ajuste3 <- parfm(Surv(tempo,cens)~vz2+vz3+vz4+vz5+vz7+vz8+vz9,
                 cluster="grupox",
                 data=dados, dist="exponential", frailty="gamma")
ajuste3
AIC(ajuste3)
#[1] 2937.861
BIC(ajuste3)
#[1] 2982.031

## d) modelo com as covariáveis:
##    residduracao, renda, emprduracao, residentes,
##    receitas2, receita3
ajuste4 <- parfm(Surv(tempo,cens)~vz2+vz3+vz4+vz5+vz7+vz8,
                 cluster="grupox",
                 data=dados, dist="exponential", frailty="gamma")
ajuste4
AIC(ajuste4)
#[1] 2937.314
BIC(ajuste4)
#[1] 2976.576

## Parâmetros ajustados
h.alf <- 1/ajuste4[1]
h.lam <- ajuste4[2]
h.bt <- c(ajuste4[3], ajuste4[4], ajuste4[5],
          ajuste4[6], ajuste4[7], ajuste4[8])

vz <- cbind(vz2,vz3,vz4,vz5,vz7,vz8)
sm <- somas(grupox, tempo, h.bt, vz, cens, j=6)
ex <- exp(cbind(sm$z1,sm$z2,sm$z3,sm$z4,sm$z5,sm$z6)%*%h.bt)
wk <- round((h.alf+sm$d.k)/(h.alf+h.lam*sm$T.k),3)
```

```
## Valores preditos dados pelo pacote
predict(ajuste4)

#grupox frailty
#1      1.303
#2      0.759
#3      0.866
#4      1.072

## Cálculos para os gráficos
t <- seq(0,72,by=1)
S1 <- exp(-wk[1]*h.lam*ex[1]*t)
S2 <- exp(-wk[2]*h.lam*ex[2]*t)
S3 <- exp(-wk[3]*h.lam*ex[3]*t)
S4 <- exp(-wk[4]*h.lam*ex[4]*t)
ekm <- survfit(Surv(tempo,cens)~grupox)
ekm.g1 <- survfit(Surv(tempo[grupox==1],cens[grupox==1])~1)
ekm.g2 <- survfit(Surv(tempo[grupox==2],cens[grupox==2])~1)
ekm.g3 <- survfit(Surv(tempo[grupox==3],cens[grupox==3])~1)
ekm.g4 <- survfit(Surv(tempo[grupox==4],cens[grupox==4])~1)

## Gráficos das sobrevivências estimadas
par(mar=c(4,4.5,4,2)+0.2)
par(mfrow=c(1,1))
plot(c(0,72), c(0,1), type ="n", main="Sobrevivência estimada", cex.main=1,
     xlab="Tempo (meses)", ylab=expression(paste(hat(S),'(t)')))
lines(t,S1, col="red3", lwd=2)
lines(t,S2, col="blue", lwd=2)
lines(t,S3, col="green3", lwd=2)
lines(t,S4, col="yellow4", lwd=2)
legend(0,0.3, legend=c("Plus","Eletronico","Total","Básico"),lty=1,lwd=2,
       col=c("green3","blue","yellow4","red3"), bty="n")
```

```
## Gráficos Kaplan-Meier - estimados
plot(ekm, main="Sobrevivência via Kaplan-Meier", cex.main=1, xlab="Tempo (meses)",
     ylab=expression(paste(hat(S),'(t)')), conf.int=F, lwd=2,
     col=c("red3","blue","green3","yellow4"))
lines(t,S1, col="red3", lwd=2)
lines(t,S2, col="blue", lwd=2)
lines(t,S3, col="green3", lwd=2)
lines(t,S4, col="yellow4", lwd=2)
legend(0,0.3, legend=c("Plus","Eletronico","Total","Básico"),lty=1,lwd=2,
       col=c("green3","blue","yellow4","red3"), bty="n")
par(mfrow=c(1,1))

## Gráficos individuais
par(mfrow=c(2,2))
plot(ekm.g1, main="Sobrevivência estimada - plano Básico", xlab="Tempo (meses)",
     ylab=expression(paste(hat(S),'(t)')), conf.int=F, lwd=2, cex.main=1)
lines(t,S1, col="red3", lwd=2)

plot(ekm.g2, main="Sobrevivência estimada - plano Eletrônico", xlab="Tempo (meses)",
     ylab=expression(paste(hat(S),'(t)')), conf.int=F, lwd=2, cex.main=1)
lines(t,S2, col="blue", lwd=2)

plot(ekm.g3, main="Sobrevivência estimada - plano Plus", xlab="Tempo (meses)",
     ylab=expression(paste(hat(S),'(t)')), conf.int=F, lwd=2, cex.main=1)
lines(t,S3, col="green3", lwd=2)

plot(ekm.g4, main="Sobrevivência estimada - plano Total", xlab="Tempo (semanas)",
     ylab=expression(paste(hat(S),'(t)')), conf.int=F, lwd=2, cex.main=1)
lines(t,S4, col="yellow4", lwd=2)
par(mfrow=c(1,1))

#########################################################################
```

```
## Caso 2: Grupo = Escolaridade
#########################################################################
# valores iniciais
educ <- educacao
educ[which(educacao=="superior")] <- "superior+"
educ[which(educacao=="posgrad")] <- "superior+"
table(educ)
o <- order(educ)
grupox <- c(rep(1,204),rep(2,209),rep(3,287),rep(4,300))
vz1 <- idade[o]
vz2 <- residduracao[o]
vz3 <- renda[o]
vz4 <- emprduracao[o]
vz5 <- residentes[o]
vz6 <- receitas1[o]
vz7 <- receitas2[o]
vz8 <- receitas3[o]
vz9 <- receitas4[o]
vz <- cbind(vz1, vz2, vz3, vz4, vz5, vz6, vz7, vz8, vz9)
tempo <- clienduracao[o]
cens <- churn01[o]

dados <- data.frame(tempo,grupox,vz,cens)
attach(dados)

teducacao = data.frame(xtabs(~grupox + cens, data = dados))
ggplot(teducacao, aes(x = educacao, y = Freq, fill = churn)) +
  geom_bar(position = "fill", stat="identity")+
  scale_fill_manual(values = c("gray25","gray55"))+
  scale_x_discrete(labels=c("Médio Incompleto",
                            "Superior Incompleto","Médio",
                            "Pós Graduação","Superior"))+
  labs(x = "Nível Educacional", y = "Proporção",
```

```
       title = "Proporção de Churn por Nível de Escolaridade",
       fill="Churn")+theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45,hjust = 0.8))


## a) modelo com as covariáveis:
##    idade, residduracao, renda, emprduracao, residentes,
##    receitas1, receitas2, receita3, receita4
ajuste1 <- parfm(Surv(tempo,cens)~vz1+vz2+vz3+vz4+vz5+vz6+vz7+vz8+vz9,
                 cluster="grupox",
                 data=dados, dist="exponential", frailty="gamma")
ajuste1
AIC(ajuste1)
BIC(ajuste1)
#Não converge

## b) modelo com as covariáveis:
##    idade
ajuste2 <- parfm(Surv(tempo,cens)~vz1,
                 cluster="grupox",
                 data=dados, dist="exponential", frailty="gamma")
ajuste2
AIC(ajuste2)
#[1] 3061.846
BIC(ajuste2)
#[1] 3076.569

## c) modelo com as covariáveis:
##    residduracao, residduracao
ajuste3 <- parfm(Surv(tempo,cens)~vz1+vz2,
                 cluster="grupox",
                 data=dados, dist="exponential", frailty="gamma")
```

```
ajuste3
AIC(ajuste3)
#[1] 3035.945
BIC(ajuste3)
#[1] 3055.576

## d) modelo com as covariáveis:
##    idade,
##    residduracao, renda
ajuste4 <- parfm(Surv(tempo,cens)~vz1+vz2+vz3,
                 cluster="grupox",
                 data=dados, dist="exponential", frailty="gamma")
ajuste4
AIC(ajuste4)
#[1] 3035.89
BIC(ajuste4)
#[1] 3060.429

## e) modelo com as covariáveis:
##    idade,
##    residduracao, emprduracao
ajuste5 <- parfm(Surv(tempo,cens)~vz1+vz2+vz4,
                 cluster="grupox",
                 data=dados, dist="exponential", frailty="gamma")
ajuste5
AIC(ajuste5)
#[1] 2995.358
BIC(ajuste5)
#[1] 3019.896

## f) modelo com as covariáveis:
##    residduracao, emprduracao
##    residentes
```

```
ajuste6 <- parfm(Surv(tempo,cens)~vz2+vz4+vz5,
                 cluster="grupox",
                 data=dados, dist="exponential", frailty="gamma")
ajuste6
AIC(ajuste6)
#[1] 2986.641
BIC(ajuste5)
#[1] 3019.896

## g) modelo com as covariáveis:
##    residduracao, emprduracao
##    residentes, receitas1
ajuste7 <- parfm(Surv(tempo,cens)~vz2+vz4+vz5+vz6,
                 cluster="grupox",
                 data=dados, dist="exponential", frailty="gamma")
ajuste7
AIC(ajuste7)
#[1] 2986.262
BIC(ajuste7)
#[1] 3015.709

## h) modelo com as covariáveis:
##    idade,renda, residentes
##    receitas2
ajuste8 <- parfm(Surv(tempo,cens)~vz2+vz4+vz5+vz7,
                 cluster="grupox",
                 data=dados, dist="exponential", frailty="gamma")
ajuste8
AIC(ajuste8)
BIC(ajuste8)
#Não converge

## i) modelo com as covariáveis:
```

```
##    idade,renda, residentes
##    receitas2
ajuste9 <- parfm(Surv(tempo,cens)~vz2+vz1+vz5+vz7,
                 cluster="grupox",
                 data=dados, dist="exponential", frailty="gamma")
ajuste9
AIC(ajuste9)
#[1] 3014.86
BIC(ajuste9)
#[1] 3044.307

## j) modelo com as covariáveis:
##    idade,renda, residentes
##    receitas2, receitas3
ajuste10 <- parfm(Surv(tempo,cens)~vz1+vz2+vz5+vz7+vz8,
                 cluster="grupox",
                 data=dados, dist="exponential", frailty="gamma")
ajuste10
AIC(ajuste10)
#[1] 2980.708
BIC(ajuste10)
#[1] 3015.062

## k) modelo com as covariáveis:
##    idade,renda, residentes
##    receitas2, receitas3, recitas4
ajuste11 <- parfm(Surv(tempo,cens)~vz1+vz2+vz5+vz7+vz8+vz9,
                  cluster="grupox",
                  data=dados, dist="exponential", frailty="gamma")
ajuste11
AIC(ajuste11)
#[1] 2982.651
BIC(ajuste11)
```

```
#[1] 3021.913

## Parâmetros ajustados
h.alf <- 1/ajuste10[1]
h.lam <- ajuste10[2]
h.bt <- c(ajuste10[3], ajuste10[4], ajuste10[5],
          ajuste10[6], ajuste10[7])

## função para alguns cálculos
somas <-function(grupo, tempo, h.bt, vz, cens, j){
  grp <- names(table(grupo))
  g <- dim(table(grupo))
  d.k <- numeric(g)
  T.k <- numeric(g)
  t.k <- numeric(g)
  z1 <- numeric(g)
  z2 <- numeric(g)
  z3 <- numeric(g)
  z4 <- numeric(g)
  z5 <- numeric(g)
  for(k in 1:g){
    linhas <- which(grupo==grp[k])
    d.k[k] <- sum(cens[linhas])
    t.k[k] <- sum(tempo[linhas])
    z1[k] <- mean(vz[linhas,1])
    z2[k] <- mean(vz[linhas,2])
    z3[k] <- mean(vz[linhas,3])
    z4[k] <- mean(vz[linhas,4])
    z5[k] <- mean(vz[linhas,5])
    T.k[k] <- exp(h.bt%*%t(vz[linhas,1:j]))%*%tempo[linhas]
  }
  list(z1=z1, z2=z2, z3=z3, z4=z4, z5=z5, T.k=T.k, d.k=d.k, g=g, t.k=t.k)
}
```

```
vz <- cbind(vz1,vz2,vz5,vz7,vz8)
sm <- somas(grupox, tempo, h.bt, vz, cens, j=5)
ex <- exp(cbind(sm$z1,sm$z2,sm$z3,sm$z4,sm$z5)%*%h.bt)
wk <- round((h.alf+sm$d.k)/(h.alf+h.lam*sm$T.k),3)

## Valores preditos dados pelo pacote
predict(ajuste10)

#grupox frailty
#1      0.99
#2      0.993
#3      0.995
#4      1.022

## Cálculos para os gráficos
t <- seq(0,72,by=1)
S1 <- exp(-wk[1]*h.lam*ex[1]*t)
S2 <- exp(-wk[2]*h.lam*ex[2]*t)
S3 <- exp(-wk[3]*h.lam*ex[3]*t)
S4 <- exp(-wk[4]*h.lam*ex[4]*t)
ekm <- survfit(Surv(tempo,cens)~grupox)
ekm.g1 <- survfit(Surv(tempo[grupox==1],cens[grupox==1])~1)
ekm.g2 <- survfit(Surv(tempo[grupox==2],cens[grupox==2])~1)
ekm.g3 <- survfit(Surv(tempo[grupox==3],cens[grupox==3])~1)
ekm.g4 <- survfit(Surv(tempo[grupox==4],cens[grupox==4])~1)

## Gráficos das sobrevivências estimadas
par(mar=c(4,4.5,4,2)+0.2)
par(mfrow=c(1,2))
plot(c(0,72), c(0,1), type ="n", main="Sobrevivência estimada", cex.main=1,
     xlab="Tempo (meses)", ylab=expression(paste(hat(S),'(t)')))
lines(t,S1, col="red3", lwd=2)
```

```
lines(t,S2, col="blue", lwd=2)
lines(t,S3, col="green3", lwd=2)
lines(t,S4, col="yellow4", lwd=2)
legend(0,0.3, legend=c("Inc_medio","medio","Inc_sup","superior+"),lty=1,lwd=2,
       col=c("red3","green3","blue","yellow4"), bty="n")

## Gráficos Kaplan-Meier - estimados
plot(ekm, main="Sobrevivência via Kaplan-Meier", cex.main=1, xlab="Tempo (meses)",
     ylab=expression(paste(hat(S),'(t)')), conf.int=F, lwd=2,
     col=c("red3","blue3","green3","yellow4"))
lines(t,S1, col="red3", lwd=2)
lines(t,S2, col="blue", lwd=2)
lines(t,S3, col="green3", lwd=2)
lines(t,S4, col="yellow4", lwd=2)
legend(0,0.3, legend=c("Inc_medio","medio","Inc_sup","superior+"),lty=1,lwd=2,
       col=c("red3","green3","blue","yellow4"), bty="n")
par(mfrow=c(1,1))

## Gráficos individuais
par(mfrow=c(2,2))
plot(ekm.g1, main="Sobrevivência estimada - Médio incompleto", xlab="Tempo (meses)",
     ylab=expression(paste(hat(S),'(t)')), conf.int=F, lwd=2, cex.main=1)
lines(t,S1, col="red3", lwd=2)

plot(ekm.g2, main="Sobrevivência estimada - Superior Incompleto", xlab="Tempo (meses)",
     ylab=expression(paste(hat(S),'(t)')), conf.int=F, lwd=2, cex.main=1)
lines(t,S2, col="blue3", lwd=2)

plot(ekm.g3, main="Sobrevivência estimada - Médio", xlab="Tempo (meses)",
     ylab=expression(paste(hat(S),'(t)')), conf.int=F, lwd=2, cex.main=1)
lines(t,S3, col="green3", lwd=2)

plot(ekm.g4, main="Sobrevivência estimada - Superior e Pós-Graduação", xlab="Tempo (sema
```

```
     ylab=expression(paste(hat(S),'(t)')), conf.int=F, lwd=2, cex.main=1)
lines(t,S4, col="yellow4", lwd=2)
par(mfrow=c(1,1))
```