Gean Carlo Gomes
Jéssica Jabczenski Roslindo

Análise de Sobrevivência como ferramenta auxiliar na originação e manutenção do ciclo de crédito

Curitiba – PR

2008

Gean Carlo Gomes
Jéssica Jabczenski Roslindo

Análise de Sobrevivência como ferramenta auxiliar na originação e manutenção do ciclo de crédito

Trabalho de Conclusão de Curso apresentado à Universidade Federal do Paraná para obtenção do título de Bacharel em Estatística.

Orientadora:
Prof. Ph.D Silvia Emiko Shimakura

UNIVERSIDADE FEDERAL DO PARANÁ
DEPARTAMENTO DE ESTATÍSTICA

Curitiba – PR

2008

Trabalho de Conclusão de Curso sob o título *"Análise de Sobrevivência como ferramenta auxiliar na originação e manutenção do ciclo de crédito"*, defendido por Gean Carlo Gomes e Jéssica Jabczenski Roslindo e aprovado em 26 de junho de 2008, em Curitiba, Paraná, pela banca examinadora constituída pelos doutores:

Prof. Ph.D Silvia Emiko Shimakura
Departamento de Estatística - UFPR
Orientador

Prof. Dr. Joel Maurício Corrêa da Rosa
Departamento de Estatística - UFPR

À Deus, a nossa família e a todos que nos ajudaram.

Agradecimentos

Agradecemos inicialmente a Professora Silvia Emiko Shimakura, pela orientação e incentivo. Por estar sempre pronta a nos ajudar e aconselhar, pois só assim foi possível a execução deste trabalho acadêmico.

Agradecemos também nossas famílias, que desde o começo do curso nos deram forças para continuar e principalmente pelo apoio e compreensão despendidos nessa última fase.

Não esqueceremos de agradecer aos nossos colegas que em muitos momentos se fizeram professores, sendo de grande importância para nossa formação e construção deste.

Agradecimento especial a Fabiane Oliveira pelo incentivo e auxilio no desenvolvimento deste.

Resumo

Um dos grandes interesses das instituições financeiras é encontrar maneiras de reter e até mesmo fidelizar seus clientes com campanhas promocionais, descontos e novos serviços afim de mantê-los o maior tempo possível em suas carteiras, recuperando assim seu investimento inicial e tornando o negócio rentável. Já existem diversos métodos quantitativos que auxiliam essas instituições na originação e manutenção do ciclo de crédito dos seus clientes. Desse modo, o objetivo desse trabalho é ajudar uma instituição financeira a identificar os fatores que possam indicar previamente eventuais perdas, para isto serão apresentadas técnicas de análise de sobrevivência. Serão analisados os principais fatores que interferem na originação e manutenção do ciclo de crédito, e ainda encontrar modelos paramétricos com base nas técnicas de análise de sobrevivência que modelem as duas variáveis de interesse, cancelamento voluntário e cancelamento por inadimplência.

Palavras-chave: Análise de Sobrevivência, originação e ciclo de crédito, estimação paramátrica.

Sumário

1 Introdução p. 1

2 Revisão de Literatura p. 3

2.1 Conceitos de Análise de Sobrevivência . . . . . p. 3

2.1.1 Tempo de Falha . . . . . p. 4

2.1.2 Censura . . . . . p. 5

2.1.3 Representação dos Dados de Sobrevivência . . . . . p. 7

2.1.4 Função de Sobrevivência . . . . . p. 7

2.1.5 Função de Taxa de Falha ou de Risco . . . . . p. 8

2.1.6 Função de Taxa de Falha Acumulada . . . . . p. 9

2.1.7 Tempo Médio e Vida Média Residual . . . . . p. 9

2.1.8 Relações entre Funções . . . . . p. 10

2.2 Estimação Não-Paramétrica . . . . . p. 11

2.2.1 O Estimador de Kaplan-Meier . . . . . p. 11

2.2.2 Estimação de Quantidades Básicas . . . . . p. 16

2.2.3 Teste log-rank para Comparação de Curvas de Sobrevivência p. 17

2.3 Estimação Paramétrica . . . . . p. 19

2.3.1 Distribuição Exponencial . . . . . p. 19

2.3.2 Distribuição de Weibull . . . . . p. 20

2.3.3 Distribuição Log-normal . . . . . p. 22
2.3.4 Estimação dos Parâmetros pelo Método da Máxima Verossimilhança . . . . . p. 24
2.3.5 Intervalos de Confiança para os Parâmetros Estimados . . . p. 27
2.3.6 Testes de Hipóteses . . . . . p. 28
2.3.7 Seleção de Modelos . . . . . p. 29
2.4 Modelos de Regressão Paramétricos . . . . . p. 33
2.4.1 Modelagem Paramétrica . . . . . p. 34
2.4.2 Adequação do Modelo Ajustado . . . . . p. 36
2.4.3 Interpretação dos Coeficientes Estimados . . . . . p. 39

3 Análise dos Dados de Originação do Ciclo de Crédito p. 40
3.1 Descrição do Estudo e das Variáveis . . . . . p. 40
3.2 Análise Descritiva e Exploratória . . . . . p. 42
3.3 Ajuste dos Modelos Probabilísticos . . . . . p. 57
3.4 Ajuste dos Modelos de Regressão . . . . . p. 64
3.5 Interpretação dos Coeficientes Estimados . . . . . p. 75

4 Conclusões p. 77

Referências Bibliográficas p. 79

Apêndice A – Comparativo $S(t)$ p. 80
A.1 Cancelamento Voluntário . . . . . p. 80
A.2 Cancelamento por Inadimplência . . . . . p. 82

Apêndice B – Modelos ajustados p. 83

B.1 Cancelamento Voluntário . . . . . p. 83

B.2 Cancelamento por Inadimplência . . . . . p. 86

Apêndice C – Summary p. 89

C.1 Cancelamento Voluntário . . . . . p. 89

C.2 Cancelamento por Inadimplência . . . . . p. 91

1 Introdução

No Brasil, até a metade da década de 90 as análises de crédito não eram sustentadas por políticas tão desenvolvidas como as que temos hoje, pois as correções inflacionárias compensavam as perdas de crédito através da ciranda financeira e das remarcações de preços, apenas a experiência do analista de crédito se fazia suficiente para originação de clientes.

Com a criação do plano Real e a constância da moeda as perdas se tornaram evidentes, com o reaquecimento da economia ocorreu o crescimento e popularização do mercado de crédito.

Não havendo uma cultura de crédito, as instituições financeiras não conseguiram manter a qualidade frente a forte demanda, tornou-se então inevitável elevar a importância do ciclo de crédito, desenvolver e aperfeiçoar ferramentas que garantissem maior objetividade à análise, aumentando a rapidez e criando uma padronização no momento da concessão e manutenção do crédito.

Diante deste cenário de crescente concorrência, aliado ao alto custo de aquisição de novos clientes, cada vez mais as instituições procuram formas de reter e até mesmo fidelizar os clientes com campanhas promocionais, descontos e novos serviços para mantê-lo o maior tempo possível, recuperando assim seu investimento inicial e tornando o negócio rentável.

Dentre os diversos métodos quantitativos utilizados hoje, os mais conhecidos são os modelos de escoragem, conhecidos nas Instituições como *Credit Score* para originação e *Behaviour Score*, *Score* de Fraude, *Score* Anti-*Atrition*, entre outros inúmeros modelos para manutenção do ciclo de crédito.

Portanto, o objetivo das instituições é constantemente melhorar a qualidade de entrada do cliente e também identificar o momento ideal para agir de maneira a reter cada tipo de cliente.

Com este estudo, pretendemos auxiliar uma Instituição Financeira atuante no mercado varejista de crédito, de agora em diante denominada de FINANCEIRA que se utiliza apenas de modelos de escoragem para originação e manutenção de seus clientes, a identificar os fatores que possam indicar previamente eventuais perdas, para isso, nos utilizaremos de metodologias e técnicas da análise de sobrevivência, para obtermos não apenas a probabilidade de ocorrência do evento, mas também quando ele irá ocorrer, possibilitando uma ação na originação do crédito bem como em qualquer outro momento do ciclo. Para isso, serão observados dois eventos relevantes para instituição: cancelamento voluntário e cancelamento por inadimplência.

Escolheu-se este tipo de análise pela natureza longitudinal dos dados, pois não temos a necessidade de fixar um tempo exato, podemos estudar a probabilidade para todos os instantes até o período máximo observado na amostra. Outra questão que defende o uso da análise de sobrevivência é a existência de tempos de falha e censura cujo os conceitos e definições serão apresentadas no capítulo 2 (Revisão Literária).

No capítulo 2 são descritas as metodologias utilizadas nos estudos de análise de sobrevivência, poderemos conhecer melhor os conceitos e métodos utilizados neste tipo de estudo, veremos como são estimados os modelos não-paramétricos, paramétricos e como escolher um modelo probabilístico que se ajuste bem aos dados.

No capítulo 3 (Análise dos dados de Originação de crédito) será apresentado o conjunto de dados, os principais fatores que influenciam na ocorrência dos eventos, os critérios de seleção das distribuições escolhidas e os modelos escolhidos para explicar cada um dos eventos observados, veremos também os testes realizados para verificar graficamente a qualidade dos ajustes.

2 *Revisão de Literatura*

2.1 Conceitos de Análise de Sobrevivência

Muitas vezes se tem o interesse em analisar os tempos de vida de pessoas ou até mesmo de produtos. Estudos com esses interesses, têm crescido muito nos últimos anos, conjuntamente com o desenvolvimento, e aprimoramento da informática, e das técnicas estatísticas. A área da estatística que estuda dados de tempo de vida é a análise de sobrevivência, também chamada de análise de sobrevida, estes termos referem-se basicamente a situações médicas, entretanto situações onde se possa utilizar esse tipo de análise ocorrem em diversas áreas, como acontece na área financeira que é o foco desse trabalho.

Como já foi mencionado, a análise de sobrevivência consiste em estudar os tempos de vida de objetos ou pessoas, sendo que a variável resposta é por natureza longitudinal, ou seja, é o tempo até a ocorrência de um evento de interesse. O objetivo é encontrar modelos estatísticos para estimar o tempo de vida. Em algumas situações modelos paramétricos podem ser usados para representar as distribuições desses dados. Em outras situações, esses modelos não se ajustam bem, nesses casos pode-se tentar uma modelagem através de modelos não-paramétricos. Dados de sobrevivência são caracterizados pelos tempos de falha e, muito freqüentemente, pelas censuras ([3]). Em alguns casos pode ser observadas covariáveis para cada indivíduo. Estes componentes constituem a resposta.

Tempo de falha é o tempo até a ocorrência do evento de interesse, e censura é a observação parcial ou incompleta da resposta. Sem a presença de censura, outras técnicas estatísticas poderiam ser utilizadas para a análise desses dados. Ressalta-

mos que mesmo censurados, todos os resultados provenientes de um estudo de sobrevivência devem ser usados na análise estatística, pois a omissão das censuras no cálculo das estatísticas de interesse pode acarretar conclusões viciadas.

Esta seção tem o objetivo de esclarecer conceitos básicos e definir funções da área de análise de sobrevivência que serão utilizados na análise dos dados. Nas próximas seções serão definidos os conceitos de tempo de falha e de dados censurados e ainda serão apresentados os vários tipos de censura. Nas últimas seções, será explicado como os dados de sobrevivência são caracterizados e serão apresentadas as principais funções que ajudam a explorar e analisar os dados estudados, bem como algumas relações matemáticas entre essas funções.

2.1.1 Tempo de Falha

O tempo de falha é constituído pelos seguintes elementos: o tempo inicial, a escala de medida e o evento de interesse (falha). O termo falha surgiu no contexto de análise de confiabilidade, na qual se busca modelar o tempo até a falha de algum equipamento ou componente ([6]). O tempo inicial é o tempo de início do estudo, e deve ser precisamente definido. Os indivíduos devem ser comparáveis na origem do estudo, com exceção de diferenças medidas pelas covariáveis ([3]).

A escala de medida é o tempo real ou "de relógio". O evento de interesse é na maioria dos casos indesejáveis, chamados de falha. É muito importante definir o que é o tempo de falha ao começar um estudo de análise de sobrevivência. Em algumas situações, a definição de falha já é clara, tais como morte em estudos clínicos.

A falha pode ainda ocorrer devido a uma única causa ou devido a duas ou mais causas. Situações em que causas de falha competem entre si são denominadas na literatura de riscos competitivos.

Definimos que o tempo de falha de um indivíduo i é uma variável aleatória chamada T_i, para $i = 1, .., n$.

2.1.2 Censura

Estudos que envolvem a análise de pessoas ou objetos, por longos períodos de tempo, trabalham com censuras, ou seja, durante o estudo perde-se a informação sobre o indivíduo, ou ao chegar ao término do estudo o paciente não falhou, então dizemos que ele censurou. Por exemplo: suponhamos que estamos estudando uma amostra de pacientes, cujo o objetivo é que os pacientes falhem, ou seja, que o evento de interesse aconteça, mas devido à longa duração do estudo, pode ocorrer a perda do acompanhamento de alguns pacientes, por causas diferentes da falha, ou ainda ao chegar ao término do estudo o paciente não falhou. Nesses casos temos observações incompletas ou parciais. A única informação que temos desses pacientes que censuraram, é que o tempo até a ocorrência do evento de interesse, para cada um deles, é superior ao tempo registrado até o último acompanhamento.

As censuras podem ser classificadas em três tipos diferentes: censuras tipo I, censuras do tipo II e censuras aleatórias. A seguir cada uma delas será definida.

- **Censura Tipo I**

  É aquela em que o estudo será terminado após um período pré-estabelecido de tempo. Suponhamos que temos um experimento com n indivíduos com os seguintes tempos de falha: $T_1, T_2, ..., T_n$, e que esses indivíduos foram observados por um período de tempo τ, que foi fixado no início do experimento. Se o indivíduo vier a falhar no tempo t, tal que $t < \tau$, então dizemos que t é o tempo de falha; por outro lado, se o indivíduo não falhou ao término do experimento, então $t = \tau$ é o tempo de censura.

- **Censura Tipo II**

  É aquela em que o estudo será terminado após ter ocorrido o evento de interesse (falha) em um número pré-estabelecido de indivíduos. Em experimentos com esse tipo de censura o tempo de término é aleatório e o número de falhas é fixo. Se todos os indivíduos do experimento são colocados em teste ao mesmo tempo, teremos as censuras simples do tipo II. Já se os indivíduos

entram no experimento em tempos diferentes, essas censuras são chamadas censuras múltiplas do tipo II.

- **Censura Aleatória**

  Nos tipos de censura apresentados acima, temos que o número de falhas é aleatório no tipo I, enquanto que nas censuras do tipo II o número de falhas é fixo. Podemos observar que nas censuras do tipo I, t é uma variável aleatória mista com um componente contínuo e outro discreto. Em ambos os casos os tempo de censuras são determinados pelo planejamento do experimento, sendo o término fixo ou o número de falhas fixo.

  Entretanto, os tempos de censura são freqüentemente aleatórios. Por exemplo: em um ensaio clínico um indivíduo pode ser retirado no decorrer do estudo sem ter ocorrido à falha, ou também, o paciente pode morrer por uma razão diferente da estudada, ou ainda os pacientes podem entrar no estudo de modo aleatório, de acordo com o tempo de diagnóstico.

  Como já mencionado na seção 2.1.1 iremos denominar o tempo de falha de um indivíduo como sendo uma variável aleatória, chamada aqui de T. O tempo de censura também é uma variável aleatória independente de T, representado por C. Quando observamos um indivíduo em um estudo de análise de sobrevivência temos que,

$$t = min(T, C)$$

e

$$\delta = \begin{cases} 1 & \text{se } T \leq C \\ 0 & \text{se } T > C. \end{cases}$$

As censuras podem ser ainda classificadas em censuras à direita, ou à esquerda, ou intervalares. Temos censura à direita quando o tempo de ocorrência do evento de

interesse está à direita do tempo registrado. Já a censura à esquerda ocorre quando o tempo registrado é maior do que o tempo de falha, ou seja, o evento de interesse já aconteceu quando o indivíduo foi observado. Censura intervalar acontece quando, por exemplo, em um certo experimento os indivíduos são acompanhados periodicamente, e é conhecido somente que o evento de interesse aconteceu, mas não se sabe o tempo exato. Assim sabemos que o tempo de falha pertence a um intervalo, isto é, $T \in (L,U]$. Sendo que L é o tempo da última vez em que o individuo foi observado e U é o tempo "agora", a falha já aconteceu. Censuras à direita são mais comuns em dados de sobrevivência.

2.1.3 Representação dos Dados de Sobrevivência

O indivíduo i $(i = 1,...,n)$ em um estudo de análise de sobrevivência é representado, em geral, pelo par $(t_i,\ \delta_i)$, sendo t_i o tempo de falha ou de censura e δ_i a variável indicadora de falha ou censura, isto é,

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

Em casos em que foram observadas covariáveis (x_i), como: idade, sexo, grau de instrução, dentre outras, os dados ficam representados por (t_i, δ_i, x_i), para cada indivíduo.

2.1.4 Função de Sobrevivência

Essa função é definida como a probabilidade de uma observação não falhar até um certo tempo t, ou seja, a probabilidade de uma observação sobreviver ao tempo t. Isto é,

$$S(t) = P(T \geq t).$$

Se fizermos 1 - $S(t)$, ou seja, tirar o complementar de $S(t)$ obtém-se a função de distribuição acumulada, $F(t)$ = 1 - $S(t)$, que é a probabilidade de uma observação não sobreviver ao tempo t. $S(t)$ é uma função escada com saltos somente nos tempos de falha.

2.1.5 Função de Taxa de Falha ou de Risco

A função taxa de falha ou de risco é definida da seguinte maneira: é a probabilidade da falha ocorrer em um intervalo de tempo $[t_1, t_2)$, já ([6]) explicam que essa função permite analisar o risco de um indivíduo sofrer um evento em um determinado tempo t, dado que ele já sobreviveu até aquele momento. A função de taxa de falha pode ser expressa em termos da função de sobrevivência como:

$$S(t_1) - S(t_2).$$

A taxa de falha no intervalo $[t_1, t_2)$ considera que a probabilidade de que a falha ocorra esteja dentro desse intervalo, dado que não ocorreu antes de t_1, dividida pelo comprimento do intervalo. Assim, a taxa de falha no intervalo $[t_1, t_2)$ é expressa por:

$$\frac{S(t_1) - S(t_2)}{(t_2 - t_1)S(t_1)}. \tag{2.1}$$

De forma geral, redefinindo o intervalo como $[t, t + \Delta t)$, a expressão (2.1) assume a seguinte forma:

$$\lambda(t) = \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}.$$

Assumindo que Δt é bem pequeno, $\lambda(t)$ representa a taxa de falha instantânea no tempo t condicional à sobrevivência até o tempo t. A taxa de falha assume números positivos, mas sem limite superior. Ela descreve a forma em que a taxa instantânea de falha muda com o tempo. Na prática podemos utilizar essa taxa para descrever o

tempo de vida de um paciente, por exemplo. A função de taxa de falha de T é, então, definida como:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2.2}$$

A função de taxa de falha pode se comportar de três maneiras: pode ser crescente, decrescente ou constante. Uma função crescente indica que a taxa de falha aumenta ao decorrer do tempo. Se for constante indica que ela não muda com o passar do tempo. Se a taxa de falha apresentar uma forma decrescente, mostra que a taxa de falha diminui à medida que o tempo passa. Essa função é mais informativa do que a função de sobrevivência. As funções $f(t)$ e $S(t)$ são representações comuns da distribuição de uma variável aleatória e podem apresentar formas semelhantes, já as respectivas funções de taxa de falha podem ser totalmente diferentes. A modelagem dessa taxa é um importante método para os dados de sobrevivência.

2.1.6 Função de Taxa de Falha Acumulada

A função taxa de falha acumulada é expressa por:

$$\Lambda(t) = \int_0^t \lambda(u) du.$$

Essa função não tem uma interpretação direta, mas pode ser útil na avaliação da função de maior interesse que é a taxa de falha, $\lambda(t)$. Como na estimação não-paramétrica $\lambda(t)$ é difícil de ser estimada, podemos utilizar $\Lambda(t)$ que possui propriedades ótimas.

2.1.7 Tempo Médio e Vida Média Residual

O tempo médio de vida pode ser obtido através da seguinte equação:

$$t_m = \int_0^\infty S(t) dt.$$

A vida média residual é definida condicional a um certo tempo de vida t. Ou seja, para indivíduos com idade t esta quantidade mede o tempo médio restante de vida, e é expressa por:

$$\text{vmr}(t) = \frac{\int_t^\infty (u-t) f(u) du}{S(t)} = \frac{\int_t^\infty S(u) du}{S(t)}.$$

Essa expressão nos mostra a divisão entre a área sob a curva de sobrevivência à direita do tempo t e $S(t)$, sendo $f(\cdot)$ a função de densidade de T. Observe que vmr(0)=t_m.

2.1.8 Relações entre Funções

Em [3] são apresentadas algumas relações imporantes entre as funções abordadas até agora. Consideremos T uma variável aleatória contínua e não-negativa, temos que:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\left(logS(t)\right),$$

$$\Lambda(t) = \int_0^t \lambda(u) du = -logS(t)$$

e

$$S(t) = exp\{-\Lambda(t)\} = exp\left\{-\int_0^t \lambda(u) du\right\}.$$

Tais relações mostram que o conhecimento de uma das funções, por exemplo $S(t)$, implica no conhecimento das demais, isto é, de $F(t)$, $f(t)$, $\lambda(t)$ e $\Delta(t)$. Outras relações envolvendo estas funções são as seguintes:

$$S(t) = \frac{\text{vmr}(0)}{\text{vmr}(t)} exp\left\{-\int_0^t \frac{du}{\text{vmr}(u)}\right\}$$

e

$$\lambda(t) = \left(\frac{d\text{vmr}(t)}{dt} + 1 \right) / \text{vmr}(t).$$

Até agora foram apresentados conceitos e funções que nos auxiliarão a explorar o conjunto de dados desse trabalho. O próximo passo será apresentar técnicas não-paramétricas para uma análise um pouco mais avançada dos dados, uma análise descritiva.

2.2 Estimação Não-Paramétrica

Quando iniciamos um estudo, o primeiro passo é desenvolver uma análise descritiva dos dados, ou seja, extrair as primeiras informações para poder definir qual a melhor técnica a ser usada para alcançar o objetivo final. Em muitas situações as técnicas usadas são estatísticas simples envolvendo medidas de tendência central e variabilidade, como por exemplo: média e desvio-padrão. Em estudos de análise de sobrevivência a análise descritiva deve ser feita com um maior cuidado, pois esses dados possuem censuras, o que requer técnicas estatísticas especializadas para acomodar essa informação. Sendo assim o principal componente da análise descritiva envolvendo dados de tempo de vida é a função de sobrevivência. As técnicas são não-paramétricas, pois não estamos trabalhando com nenhuma função de probabilidade para os dados ainda, dessa maneira não temos parâmetros de funções a serem estimados.

Nesta seção iremos primeiramente encontrar uma estimativa para a função de sobrevivência pelo método de Kaplan-Meier, para a partir dela estimar as estatísticas de interesse, e finalmente iremos fazer comparações entre curvas de sobrevivência usando o teste log-rank.

2.2.1 O Estimador de Kaplan-Meier

O estimador de Kaplan-Meier, proposto por Kaplan e Meier (1958), é o método mais utilizado para estimar a função de sobrevivência, pois ele considera a censura nos dados. Esse método utiliza os conceitos de independência de eventos e de pro-

babilidade condicional para desdobrar a condição sobreviver até o tempo t em uma seqüência de eventos independentes que caracterizam a sobrevivência em cada intervalo de tempo anterior a t e cuja probabilidade é condicional aos que estão em risco em cada período ([6]). A seguir, iremos apresentar passo a passo como encontrar o estimador de Kaplan-Meier.

Inicialmente, consideremos dados sem censuras, a função de sobrevivência para esses dados é definida como:

$$\widehat{S}(t) = \frac{\text{n}^o \text{ de observações que não falharam até o tempo } t}{\text{n}^o \text{ total de observações no estudo}}. \tag{2.3}$$

A função $\hat{S}(t)$ tem forma de escada, onde os degraus representam os tempos observados de falha de tamanho $\frac{1}{n}$, em que n é o tamanho da amostra. Quando existem várias falhas em um certo tempo t, o tamanho do degrau fica multiplicado pelo número de empates. Consideremos $S(t)$ uma função discreta com probabilidade maior que zero somente nos tempos de falha $t_j, j = 1,...,k$, sendo $k(\leq n)$ o número de falhas distintas e os tempos t_j sendo distintos e ordenados. O estimador de Kaplan-Meier é definido como:

$$\widehat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right) \tag{2.4}$$

onde d_j é o número de falhas e n_j o número de indivíduos sob risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j.

A expressão (2.4) surge do fato que $S(t)$ pode ser escrita em função de probabilidades condicionais, como foi mencionado acima. Suponhamos que existam n indivíduos e k falhas distintas nos tempos $t_1 < t_2 < ...t_k.$, temos que:

$$S(t_j) = (1-q_1)(1-q_2)...(1-q_j) \tag{2.5}$$

onde q_j é a probabilidade de um indivíduo morrer no intervalo $[t_{j-1}, t_j)$ sabendo que ele não morreu até t_{j-1} e considerando $t_0 = 0$. Podemos escrever q_j da seguinte maneira:

$$q_j = P(T \in [t_{j-1}, t_j) | T \geq t_{j-1}). \tag{2.6}$$

Temos então que $S(t)$ é escrita em termos de probabilidades condicionais. A função do estimador de Kaplan-Meier é estimar q_j, que adaptado da expressão (2.3), é dado por:

$$\widehat{q}_j = \frac{\text{n}^o \text{ de falhas em } t_{j-1}}{\text{n}^o \text{ de observações sob risco em } t_{j-1}} \tag{2.7}$$

Uma justificativa para o estimador de Kaplan-Meier, que foi apresentado na expressão (2.4), é que esse vem da decomposição de $S(t)$ em termos dos q_j's apresentada em (2.5). O estimador (2.4) é o estimador de máxima verossimilhança de $S(t)$. A seguir será apresentada a prova que o estimador de Kaplan-Meier para $S(t)$ é um estimador de máxima verossimilhança.

Suponha, que d_j sejam as observações que falharam no tempo t_j, para $j = 1, ..., k$ e m_j as observações que censuraram no intervalo de tempo $[t_j, t_{j+1})$, nos tempos $t_{j1}, ..., t_{jm_j}$. A probabilidade de falha no tempo t_j é:

$$S(t_j) - S(t_j+),$$

com $S(t_j+) = \lim_{\Delta t \to 0} S(t_j + \Delta t), j = 1, ..., k$. Por outro lado, a contribuição para a função de verossimilhança de um tempo de sobrevivência censurado em t_{jl}, para $l = 1, ..., m_j$, é:

$$P(T > t_{jl}) = S(t_{jl}+).$$

A função de verossimilhança pode, então, ser escrita como:

$$L(S(\cdot)) = \prod_{j=0}^{k} \left\{ \left[S(t_j) - S(t_j+) \right]^{d_j} \prod_{l=1}^{m_j} S(t_{jl}+) \right\}.$$

Assim temos a função de verossimilhança do estimador Kaplan-Meier que é uma

generalização do conceito usual utilizado em modelos paramétricos em que se tem tantos parâmetros quanto falhas distintas. Este estimador também mantém esta forma em estudos envolvendo os mecanismos de censura do tipo I e tipo II mas não atinge $\hat{S}(t) = 0$, pois as últimas observações são censuradas. As propriedades desse estimador são as seguintes:

1. é não-viciado para amostras grandes,
2. é fracamente consistente,
3. converge assintoticamente para um processo gaussiano e
4. é estimador de máxima verossimilhança de $S(t)$.

Para melhor entendermos o estimador Kaplan-Meier apresentaremos a figura 2.1 onde foram estudados os tempos de sobrevida de dois grupos distintos, cada grupo foi submetido a um certo tipo de tratamento médico. Esta figura foi retirada de [3].

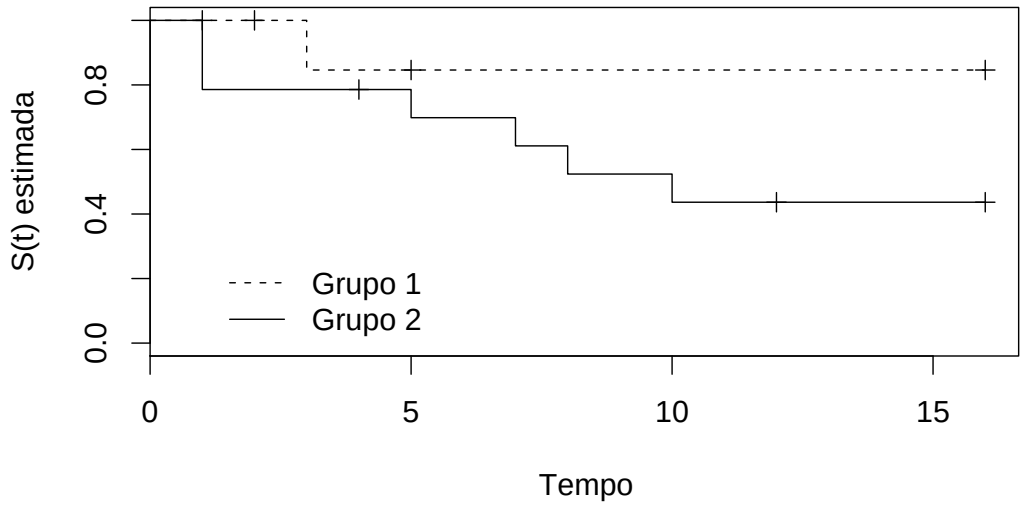


Figura 2.1: Estimativas de Kaplan-Meier para os grupos fictícios G1 e G2. Os tempos representados por + mostram onde ocorreram censuras em cada grupo.

A figura 2.1 foi construída mantendo-se o valor de $\hat{S}(t)$ constante entre os tempos de falha. Podemos observar que o grupo 1 (G1) apresenta estimativas de sobrevivência maiores que o grupo 2 (G2), este apresenta um "curva" que cai rapidamente e se torna constante a partir do tempo igual a dez semanas.

Depois que encontramos o estimador de $S(t)$, devemos encontrar o intervalo de confiança, pois todo estimador está sujeito a variações que podem ser descritas em termos de estimações intervalares. E devemos também testar hipóteses para $S(t)$ para avaliar a precisão do estimador de Kaplan-Meier. Sendo assim, a variância assintótica do estimador de Kaplan-Meier é dada por:

$$\widehat{Var}(\widehat{S}(t)) = \left[\widehat{S}(t)\right]^2 \sum_{j:t_j<t} \frac{d_j}{n_j(n_j-d_j)}.$$

Como $\hat{S}(t)$, para t fixo, tem distribuição assintótica Normal, segue que o intervalo aproximado de $100(1-\alpha)\%$ de confiança para $S(t)$ é dado por:

$$[\widehat{S}(t)]^{exp\left\{\pm z_{\alpha/2}\sqrt{\widehat{Var}(\widehat{U}(t))}\right\}}$$

onde $\hat{U}(t) = log[-log(\hat{S}(t))]$ é uma transformação de $S(t)$, para garantir que o limite inferior e superior do intervalo de confiança de $\hat{S}(t)$ não seja negativo ou maior que um.

É apresentado a nível de ilustração (dados fictícios) a figura 2.2, que contém uma curva de sobrevida com os seus respectivos intervalos com 95% de confiança estimados a partir do estimador de Kaplan-Meier.

Ressaltamos que existem outros estimadores de $S(t)$, eles são: o estimador de Nelson-Aalen e o estimador da tabela de vida ou tábua de vida. Neste trabalho utilizaremos o estimador de Kaplan-Meier por ser o estimador mais utilizado em análise de sobrevivência. Maiores informações sobre o estimador Nelson-Aalen e o estimador da tabela de vida ou tábua de vida podem ser encontrados em [3], [6] e [2].

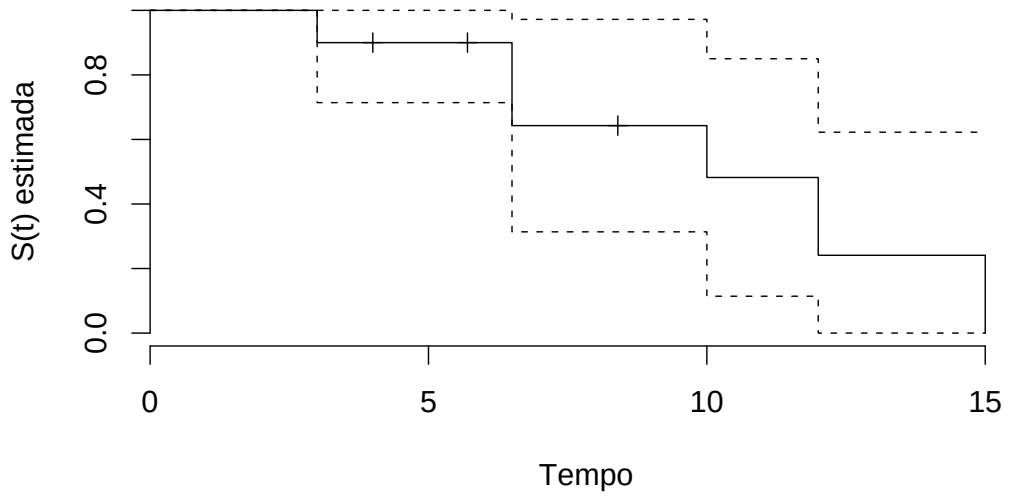


Figura 2.2: Sobrevivência e respectivos intervalos de 95% de confiança estimados a partir do estimador de Kaplan-Meier para dados fictícios.

2.2.2 Estimação de Quantidades Básicas

Quando se tem o interesse em estimar a sobrevida de algum paciente, por exemplo, em um determinado tempo t e este tempo estiver ao longo de um degrau da curva de Kaplan-Meier, é preferível usar a interpolação linear para fazer essa estimativa. Para isso usa-se a seguinte fórmula:

$$\frac{t_j - t_{j-1}}{\hat{S}(t) - (Valor\ obtido\ diretamente\ da\ curva\ de\ Kaplan - Meier)}$$

Esta forma usualmente gera uma melhor representação da distribuição contínua dos tempos de vida.

O tempo médio de vida é estimado por:

$$\widehat{t_m} = t_1 + \sum_{j=1}^{k-1} \widehat{S}(t_j)(t_{j+1} - t_j)$$

que é a área sob a curva de Kaplan-Meier estimada, mas como esta curva é uma função escada, esta integral é simplesmente a soma de áreas de retângulos, onde $t_1 < ... < t_k$ são os k tempos distintos e ordenados de falha. Deve-se tomar cuidado na hora de interpretar tal estimativa, pois quando o maior tempo observado for uma censura a curva de Kaplan-Meier não atinge o valor zero e o valor do tempo médio de vida fica subestimado.

A variância assintótica de $\hat{t}_m$ pode ser estimada por:

$$\widehat{Var}(\widehat{t_m}) = \frac{r}{r-1}\left[\sum_{j=1}^{r-1} \frac{(A_j)^2}{n_j(n_j - d_j)}\right],$$

onde $A_j = \hat{S}(t)(t_{j+1} - t_j) + ... + \hat{S}(t_{r-1})(t_r - t_{r-1})$ e r é o número de falhas.

Podemos estimar também o tempo médio restante de vida daqueles indivíduos que se encontram livres do evento de interesse em um determinado tempo t. Porém esse estimador apresenta as mesma limitações de $\hat{t}_m$.

$$\widehat{vmr}(t) = \frac{\text{área sob a curva } \widehat{S}(t) \text{ à direita de } t}{\widehat{S}(t)}.$$

2.2.3 Teste log-rank para Comparação de Curvas de Sobrevivência

Na figura 2.1 apresentada na seção 2.2.1 mostramos duas estimativas das curvas de $S(t)$ para dois grupos distintos, suponhamos que se tenha interesse em comparar essas duas curvas, para isto existem na literatura diversos teste, aqui iremos abordar o teste log-rank. Sua estatística é:

$$T = \frac{\left[\sum_{j=1}^{k}(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k}(V_j)_2}, \tag{2.8}$$

d_{2j} é o número de falhas do grupo 2 que segue uma distribuição hipergeométrica;

w_{2j} é a média de d_{2j} expressa por: $w_{2j} = n_{2j}d_j{n_j}^{-1}$, onde n_{2j} é o número de indivíduos que estão sob risco em um tempo imediatamente inferior a t_j no grupo 2;

d_j é o número total de falhas nos grupos 1 e 2, e

${n_j}^{-1}$ é $1/$(número total de indivíduos sob risco).

A variância de d_{2j} obtida a partir da distribuição hipergeométrica é:

$$(V_j)_2 = n_{2j}(n_j - n_{2j})d_j(n_j - d_j)n_j^{-2}(n_j - 1)^{-1}.$$

O teste de hipótese log-rank compara a distribuição da ocorrência dos eventos observados em cada estrato com a distribuição que seria esperada se a incidência fosse igual em todos os estratos ([6]), isto é, a estatística (2.8) consiste em testar se as duas curvas $S_1(t)$ e $S_2(t)$ são iguais, ou seja, $H_0 : S_1(t) = S_2(t)$ para todo t no período de acompanhamento, (2.8) tem uma distribuição qui-quadrado com um grau de liberdade.

Consideremos agora que o interesse seja em comparar $r > 2$ funções. Para isso usaremos o seguinte teste:

$$T = v'V^{-1}v$$

onde $v = \sum_{j}^{k} v_j$ é um vetor de dimensão $(r-1)$x1, cujos elementos são as diferenças entre os totais observados e esperados de falha e sua variância é $V = V_1 + ... + V_k$. T tem uma distribuição qui-quadrado com $r-1$ graus de liberdade.

Existem na literatura, como por exemplo em [3], outros testes para comparação de curvas, dois deles são: Wilcoxon e Tarone-Ware.

Na próxima seção falaremos sobre a função de sobrevivência através de técnicas paramétricas, isto é, iremos encontrar uma distribuição de probabilidades para a variável aleatória T, e em seguida encontrar $\hat{S}(t)$.

2.3 Estimação Paramétrica

Até agora usamos o estimador não-paramétrico Kaplan-Meier para estimar a função de sobrevivência, e algumas outras quantidades básicas. Porém podemos realizar essas estimações através de técnicas paramétricas, ou seja, assume-se que o tempo T até a ocorrência do evento segue uma distribuição conhecida de probabilidade, tais distribuições são denominadas de modelos probabilísticos ou paramétricos. Existem várias distribuições de probabilidade na literatura, mas apenas algumas se adequam ao fato de que a variável resposta, neste caso é uma variável aleatória contínua e não-negativa, representando o tempo até o evento de interesse. As distribuições que iremos apresentar aqui são: exponencial, Weibull e log-normal. Em seguida será utilizado o método de Máxima Verossimilhança para estimar os parâmetros de cada distribuição e serão discutidas algumas técnicas para auxiliar na escolha do melhor modelo probabilístico.

2.3.1 Distribuição Exponencial

A distribuição exponencial é historicamente a distribuição mais utilizada. Ela apresenta um único parâmetro e é caracterizada por ter uma função de taxa (ou de risco) constante. A função de densidade dessa distribuição tem a seguinte forma:

$$f(t) = \frac{1}{\alpha}\, exp\Big\{-\Big(\frac{t}{\alpha}\Big)\Big\},\ t \geq 0,$$

onde o parâmetro $\alpha > 0$ é o tempo médio de vida e possui a mesma unidade do tempo de falha t.

A função de sobrevida e a taxa de falha são dadas, respectivamente, por:

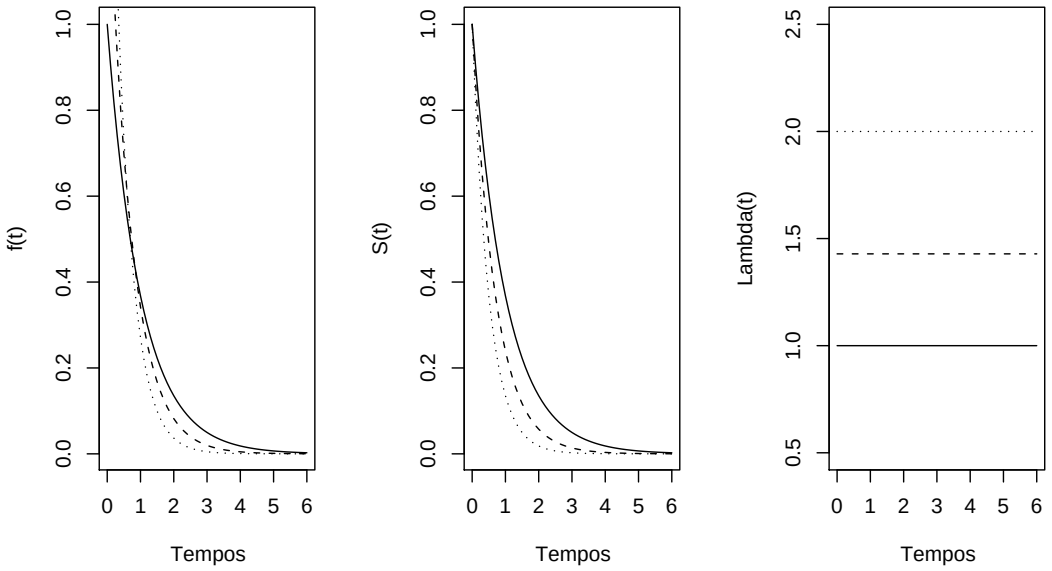
$$S(t) = exp\Big\{-\Big(\frac{t}{\alpha}\Big)\Big\}$$

e

$$\lambda(t) = \frac{1}{\alpha}\ para\ t \geq 0.$$

Na figura abaixo, 2.3, retirada de [3], são ilustradas as formas dessas três funções para

diferentes valores de α.



Podemos observar que quanto menor o valor de α mais rapidamente as funções de densidade de probabilidade $f(t)$ e de sobrevida $S(t)$ caem e que a função de taxa de falha é constante para qualquer α, esta propriedade é chamada de falta de memória.

A média da distribuição exponencial é α, a variância é α^2, e o percentil é $100p\%$ que corresponde ao tempo em que $100p\%$ indivíduos falharam, que pode ser obtido com a seguinte equação:

$$t_p = -\alpha log(1-p).$$

2.3.2 Distribuição de Weibull

A distribuição de Weibull é bastante utilizada pelo fato da sua função de risco ser monótona, isto é, ela é crescente, decrescente ou constante. As funções de probabilidade, de sobrevivência e de risco para a variável aleatória T, são dadas, respectivamente, por:

$$f(t) = \frac{\gamma}{\alpha^{\gamma}}\, t^{\gamma-1} exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\},\ t \geq 0,$$

e

$$S(t) = exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}$$

e

$$\lambda(t) = \frac{\gamma}{\alpha^{\gamma}}\, t^{\gamma-1},$$

onde γ é o parâmetro que determina a forma da função de risco, sendo:

$\gamma < 1$, a função de risco decresce, $\gamma > 1$, a função de risco cresce, $\gamma = 1$, a função de risco é constante.

O parâmetro α, determina a escala da distribuição, sendo que γ e α são positivos. Na figura 2.4, retirada de [3], são apresentados os gráficos das funções de densidade de probabilidade, de sobrevivência e de risco, para alguns valores de α e γ.

As expressões para a média e a variância da Weibull incluem o uso da função gama, isto é,

$$E[T] = \alpha\Gamma\left[1 + \left(\frac{1}{\gamma}\right)\right],$$

$$Var[T] = \alpha^2\left[\Gamma\left[1 + \left(\frac{2}{\gamma}\right)\right] - \Gamma\left[1 + \left(\frac{1}{\gamma}\right)\right]^2\right],$$

sendo a função gama, $\Gamma(k)$, definida por $\Gamma(k) = \int_0^{\infty} x^{k-1} exp\{-x\}\, dx$. Os percentis são dados por:

$$t_p = \alpha\left[-log(1-p)\right]^{\frac{1}{\gamma}}.$$

Ressaltamos que existe uma distribuição que é bastante relacionada à de Weibull, ela é chamada de distribuição do valor extremo ou de Gambel e surge quando se toma

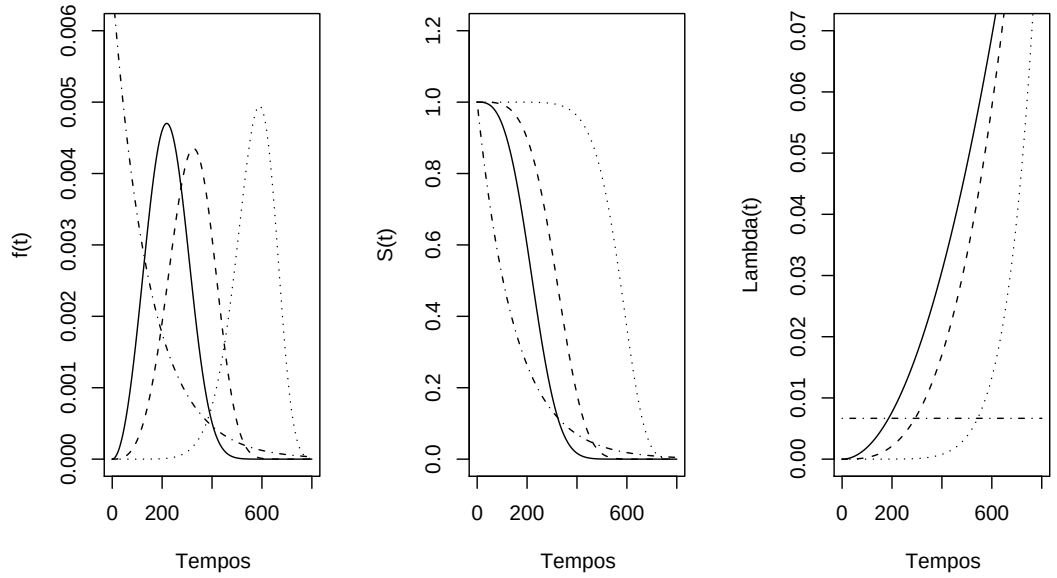


Figura 2.4: Forma típica das funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição de Weibull para os valores: $(-)$ para $(3,0;\ 250)$, $(--)$ para $(4,0;\ 350)$, $(\cdots)$ para $(8,0;\ 600)$ e $(\cdot - \cdot)$ para $(1,0;\ 150)$, dos parâmetros α e γ, respectivamente.

o logaritmo de uma variável com a distribuição de Weibull. Maiores informações sobre a distribuição do valor extremo podem ser obtidas em [3], [6] e [5].

2.3.3 Distribuição Log-normal

A distribuição Log-normal também é muito utilizada para caracterizar tempos de vida de produtos e indivíduos, ela assume que T segue distribuição log-normal, isto é, que o logaritmo de T segue uma distribuição normal com parâmetros μ, que é a média do logaritmo do tempo de falha e σ é o desvio-padrão. Assim, a função de densidade de uma variável aleatória T com distribuição log-normal é dada por:

$$f(t) = \frac{1}{\sqrt{2\pi t\sigma}}\, exp\left\{ -\frac{1}{2}\left(\frac{log(t)-\mu}{\sigma} \right)^2 \right\},\ t > 0,$$

As funções de sobrevivência e de taxa de falha para este modelo não apresentam

uma forma analítica explícita, elas são:

$$S(t) = \Phi\left(\frac{-log(t)+\mu}{\sigma}\right)$$

e

$$\lambda(t) = \frac{f(t)}{S(t)}$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada da normal padrão.

Como fizemos para as distribuições exponencial e Weibull iremos apresentar os gráficos das funções de densidade, sobrevivência e de taxa de falha, vide figura 2.5, retirada de [3].

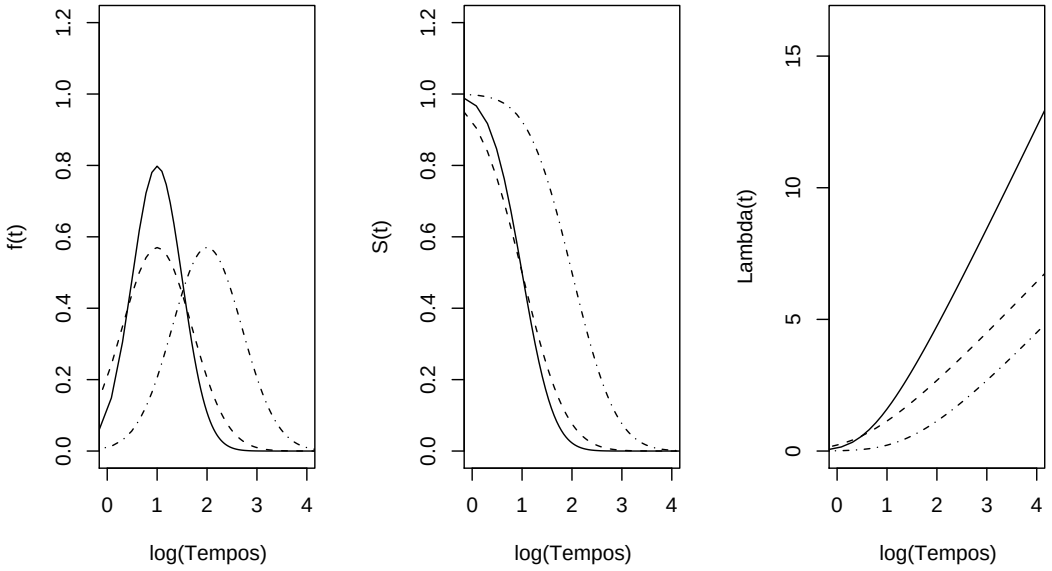


Figura 2.5: Forma típica das funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição log-normal para os valores: $(-)$ para $(0;\ 0,5)$, $(--)$ para $(0;\ 0,7)$ e $(\cdots)$ para $(0;\ 1,5)$, dos parâmetros μ e σ, respectivamente.

Diferente da distribuição Weibull, a função de taxa de falha para a distribuição log-normal não é monótona, elas crescem atingem um valor máximo e depois decrescem.

As expressões da média, da variância e os percentis, que podem ser obtidos a partir da tabela da normal padrão, para essa distribuição são dados a seguir:

$$E[T] = exp\left\{\mu + \frac{\sigma^2}{2}\right\},$$

$$Var[T] = exp\{2\mu + \sigma^2\}\left(exp\{\sigma^2\} - 1\right)$$

e

$$t_p = exp\{z_p\sigma + \mu\}$$

onde z_p o $100p\%$ percentil da distribuição normal padrão.

A distribuição exponencial, Weibull e a log-normal são freqüentemente utilizadas para estimar $\hat{S}(t)$, mas cabe ressaltarmos que existem outras distribuições que também podem ser usadas, desde que essas se ajustem bem aos dados. Algumas delas são: log-logística, gama, gama generalizada, dentre outras. Maiores detalhes sobre essas distribuições podem ser encontrados em: [3], [6] e [5].

O próximo passo agora é estimar os parâmetros dessas distribuições, para isso será usado o método de máxima verossimilhança.

2.3.4 Estimação dos Parâmetros pelo Método da Máxima Verossimilhança

Nas seções anteriores foram apresentadas três distribuições de probabilidades para a variável aleatória T, essas distribuições são caracterizadas por quantidades desconhecidas, denominadas parâmetros. Nesta seção iremos abordar uma técnica específica para estimar esses parâmetros, pois em estudos que envolvem tempos de falha, os parâmetros devem ser estimados a partir das observações amostrais. O método que faz essa estimação de uma forma apropriada é o método de máxima verossimilhança, pois ele incorpora as censuras, é simples de ser entendido e possui propriedades assintóticas ótimas para amostras grandes.

O método de máxima verossimilhança consiste em escolher os melhores valo-

res dos parâmetros das distribuições para a amostra observada. Por exemplo, se a distribuição da variável aleatória T seguir uma distribuição log-normal, o método da máxima verossimilhança irá escolher aquele par μ e σ que melhor explique a amostra observada. Passando para termos matemáticos isso significa encontrar os parâmetros que maximizem a função de verossimilhança que será apresentada a seguir.

Se fizermos a suposição de que temos uma amostra de observações $t_1, ..., t_k$ de uma certa população de interesse e que T tem uma função de densidade $f(t)$, e ainda suponhamos que esses tempos são não censurados, a função de verossimilhança para um vetor de parâmetros θ é expressa por:

$$L(\theta) = \prod_{i=1}^{n} f(t_i; \theta).$$

O parâmetro θ pode estar representando um único parâmetro, como é o caso da distribuição exponencial $\theta = \alpha$, ou um conjunto de parâmetros, como no caso da distribuição log-normal, onde $\theta = (\mu, \sigma)$.

Como já foi mencionado, o método de máxima verossimilhança é o método que melhor incorpora as censuras no seu processo de estimação, então para cada tipo de censura (vide seção 2.1.2) a função de verossimilhança tem uma expressão particular, elas são:

- **Censura tipo I**

  Quando determinamos que o estudo será encerrado após um período pré-estabelecido de tempo, temos dados com censuras do tipo I, assim temos r falhas e $n - r$ censuras observadas ao término do experimento. A função de verossimilhança para dados desse tipo tem a seguinte forma geral:

$$L(\theta) = \prod_{i=1}^{r} f(t_i; \theta) \prod_{i=r+1}^{n} S(t_i; \theta).$$

- **Censura do tipo II**

  Se antes de começar o estudo determinamos que este só acabara quando se

atingir um número fixo de falhas, a função de verossimilhança assume a seguinte forma:

$$L(\theta) = \frac{n!}{(n-r)!} \prod_{i=1}^{r} f(t_i;\theta) \prod_{i=r+1}^{n} S(t_i;\theta),$$

mas podemos notar que $\frac{n!}{(n-r)!}$ é uma constante e que não envolve nenhum parâmetro de interesse. Desse modo a função de verossimilhança fica assim:

$$L(\theta) \propto \prod_{i=1}^{r} f(t_i;\theta) \prod_{i=r+1}^{n} S(t_i;\theta).$$

- **Censura do tipo aleatória**

  Nesta situação, temos que T é o tempo de falha e C o tempo de censura. Consideramos que T e C são independentes e suponhamos que $g(c)$ e $G(c)$ são as funções de densidade e de sobrevivência de C, assim para o i-ésimo indivíduo temos as seguintes expressões, se for observada uma censura e se for observada uma falha, respectivamente:

$$P[t_i = t, \delta_i = 0] = P[C_i = t, T_i > C_i] = P[C_i = t, T_i > t] = g(t)S(t;\theta)$$

e

$$P[t_i = t, \delta_i = 1] = P[T_i = t, T_i \leq C_i] = P[T_i = t, C_i \geq t] = f(t;\theta)G(t).$$

Desta forma,

$$L(\theta) = \prod_{i=1}^{r} f(t_i;\theta)\, G(t_i) \prod_{i=r+1}^{n} g(t_i) S(t_i;\theta).$$

Sob a suposição de que o mecanismo de censura não carrega informações sobre os parâmetros, os termos $G(t)$ e $g(t)$ podem ser desprezados, portanto a função de verossimilhança fica da seguinte maneira:

$$L(\theta) \propto \prod_{i=1}^{r} f(t_i;\theta) \prod_{i=r+1}^{n} S(t_i;\theta).$$

Podemos observar que para todos os tipos de censuras a função de verossimilhança é a mesma, ou seja, os tempos de censuras não ajudam na estimação dos parâmetros das distribuições de probabilidade de T. Assim, para todos os tipos de censuras podemos utilizar a seguinte função de verossimilhança:

$$L(\theta) \propto \prod_{i=1}^{r} f(t_i;\theta) \prod_{i=r+1}^{n} S(t_i;\theta).$$

2.3.5 Intervalos de Confiança para os Parâmetros Estimados

Até o momento utilizamos o método de máxima verossimilhança para fazer estimativas pontuais, porém podemos utilizar esse mesmo método para construir intervalos de confiança.

Antes de construirmos os intervalos de confiança é importante entendermos algumas propriedades, umas delas é a que diz respeito à distribuição assintótica do estimador de máxima verossimilhança $\hat{\theta}$, isto é, para grandes amostras o vetor $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_k)'$ segue uma distribuição Normal multivariada, com média θ e matriz de variância-covariância $Var(\hat{\theta})$, isto é,

$$\widehat{\theta} \sim N_k\left(\theta, Var(\widehat{\theta})\right)$$

onde k é a dimensão de $\hat{\theta}$.

Outra propriedade importante diz respeito à precisão deste estimador e estabelece que, sob certas condições de regularidade,

$$Var(\widehat{\theta}) \approx -\left[E(F(\theta))\right]^{-1}.$$

Ou seja, que a matriz de variância-covariância dos estimadores de máxima verossimilhança é aproximadamente o negativo da inversa da esperança da matriz de derivadas segundas do logaritmo de $L(\theta)$. Maiores informações sobre essas propriedades podem ser encontradas em [3].

Deste modo, temos um intervalo aproximado de $(1-\alpha)100\%$ de confiança para θ:

$$\widehat{\theta} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\widehat{\theta})}.$$

2.3.6 Testes de Hipóteses

Nesta seção iremos ver alguns testes de hipóteses a respeito do vetor de parâmetros $\theta = (\theta_1, ..., \theta_p)'$. Os testes mais utilizados são os testes de Wald, o da Razão de Verossimilhança e o Escore. Considerando-se a hipótese nula:

$H_0 : \theta = \theta_0$, as estatísticas de teste são:

- **Teste de Wald**

$$W = \left(\widehat{\theta} - \theta_0\right)' \left[-F(\theta_0)\right] \left(\widehat{\theta} - \theta_0\right).$$

  Esse teste é baseado na distribuição assintótica de $\hat{\theta}$ e é uma generalização do teste de t de Student. Geralmente, esse teste é usado para testar hipóteses relativas a um único parâmetro θ_j.

- **Teste da Razão de Verossimilhança**

$$TRV = -2\ log\left[\frac{L(\theta_0)}{L(\widehat{\theta})}\right].$$

  O teste acima é baseado na função de verossimilhança e envolve a comparação dos valores do logaritmo da função de verossimilhança maximizada sem restrição e sob H_0.

- **Teste Escore**

$$S = U'(\theta_0)\left[-F(\theta_0)\right]^{-1} U(\theta_0),$$

em que $U(\theta_0)$ é a função escore $U(\theta) = \frac{\partial logL(\theta)}{\partial \theta}$ avaliada em θ_0, e $-F(\theta_0)$ a matriz de variância-covariância de $U(\theta)$ também avaliada em θ_0.

As três estatísticas acima possuem, sob H_0, distribuição χ^2 com p graus de liberdade.

2.3.7 Seleção de Modelos

Até o momento vimos que a variável aleatória T pode assumir uma distribuição de probabilidades, já nesta seção iremos abordar os métodos para a escolha da distribuição mais adequada, pois se fizermos a escolha inadequada toda a análise estatística fica comprometida e conseqüentemente as respostas às perguntas de interesse ficam distorcidas, e também o método de máxima verossimilhança só pode ser aplicado mediante uso de uma distribuição. Uma das maneiras mais eficientes para selecionar o "melhor" modelo, é por meio de técnicas gráficas, mas também existem testes de hipóteses e testes estatísticos como é o caso do teste da razão de verossimilhança. A seguir todas essas técnicas serão apresentadas brevemente.

- **Métodos Gráficos**

  Umas das técnicas gráficas, que iremos chamar de **método 1**, consiste em comparar a função de sobrevivência do modelo proposto (log-normal e de Weibull, por exemplo) com o estimador de Kaplan-Meier. Para fazer essa comparação é feito os ajustes dos modelos propostos ao conjunto de dados e, a partir das estimativas dos parâmetros de cada modelo, estima-se suas respectivas funções de sobrevivência e obtém-se também a estimativa de Kaplan-Meier. Abaixo serão apresentados os gráficos, vide 2.6, de $S(t)$ estimada por Kaplan-Meier *versus* $S(t)$ estimada pelos modelos de probabilidade, os dados que foram utilizados na construção desse gráfico foram retirados do livro [3] com finalidade apenas de ilustração.

  onde $x = \hat{S}(t)$ e $y = \hat{S}_e(t)$, $y = \hat{S}_w(t)$ e $y = \hat{S}_{ln}(t)$, em que $\hat{S}(t)$, $\hat{S}_e(t)$, $\hat{S}_w(t)$ e $\hat{S}_{ln}(t)$ são as curvas estimadas de $S(t)$ por Kaplan-Meier, pela distribuição exponencial, Weibull e log-normal, respectivamente. A distribuição que será

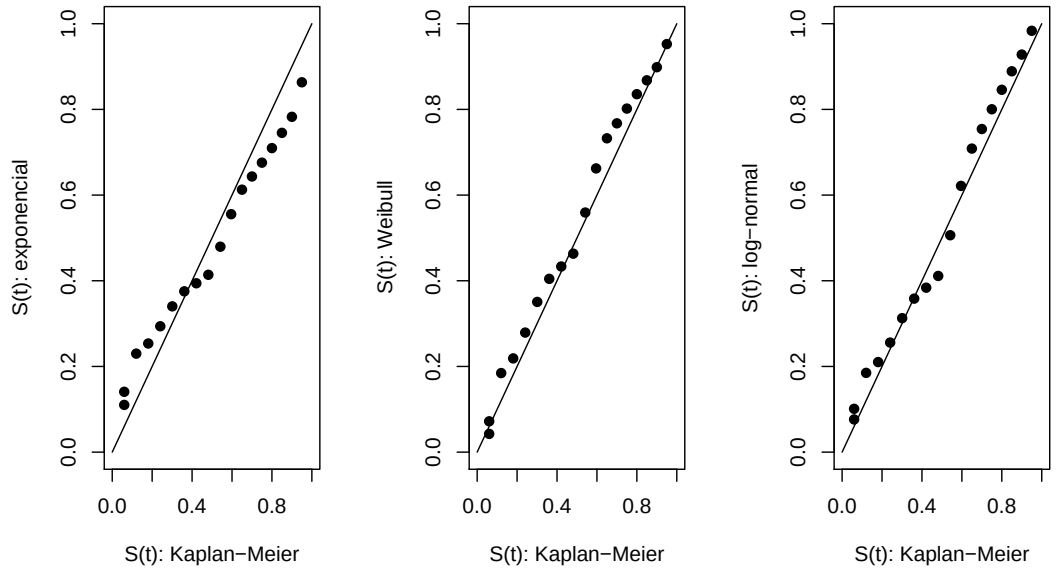


Figura 2.6: Gráficos das sobrevivências estimadas por Kaplan-Meier *versus* as sobrevivências estimadas pelos modelos exponencial, de Weibull e log-normal.

escolhida é aquela cujos os pontos estiverem mais próximos da reta $y = x$. Pela figura 2.6, podemos observar que os modelos de Weibull e log-normal são os que apresentam seus pontos mais próximos da reta para este exemplo.

O **método 2** consiste na linearização da função de sobrevivência tendo como idéia básica a construção de gráficos que sejam aproximadamente lineares, caso o modelo proposto seja apropriado. A seguir iremos apresentar exemplos de linearização para os modelos exponencial, Weibull e log-normal.

1. **Linearização no modelo exponencial**

   Para este modelo a linearização é obtida com a seguinte equação:

   $$-log\,[S(t)] = \frac{t}{\alpha} = \left(\frac{1}{\alpha}\right) t$$

   onde $-log\,[S(t)]$ é uma função linear de t, assim o gráfico de $-log\left[\hat{S}(t)\right]$ *versus* t deve ser aproximadamente linear, passando pela origem para que o modelo seja apropriado aos dados. $\hat{S}(t)$ é o estimador de Kaplan-Meier.

2. **Linearização no modelo de Weibull**

   A equação de linearização para a distribuição de Weibull é a seguinte:

$$log\left[-log\left[S(t)\right]\right] = -\gamma\, log(\alpha) + \gamma\, log(t)$$

   o que mostra que $log\left[-log\left[S(t)\right]\right]$ é uma função linear de $log(t)$. Portanto, o gráfico de $log\left[-log\left[\hat{S}(t)\right]\right]$ *versus* $log(t)$, sendo $\widehat{S}(t)$ o estimador de Kaplan-Meier, neste caso para o modelo de Weibull ser considerado o "melhor" os pontos do gráfico de linearização devem formar uma reta e passar pela origem, mas se a inclinação for igual a um, é uma indicação a favor do modelo exponencial.

3. **Linearização no modelo log-normal**

   Aqui a equação de linearização tem a seguinte forma:

$$\Phi^{-1}\left(S(t)\right) = \frac{-log\, t + \mu}{\sigma}$$

   em que $\Phi^{-1}(\cdot)$ são os percentis da distribuição Normal padrão. Isto significa que o gráfico de $\Phi^{-1}(\widehat{S}(t))$ *versus* $log(t)$ deve ser aproximadamente linear, com intercepto μ/σ e inclinação $-1/\sigma$, se o modelo log-normal for apropriado.

   A figura 2.7, retirada de [3], mostra os gráficos de linearização para as três distribuições apresentadas aqui, com o mesmo conjunto de dados que foi utilizado na construção da figura 2.6.

   Um terceiro gráfico que pode ser construído, é um gráfico das estimativas de $S(t)$, tanto pelo estimador Kaplan-Meier como pelas distribuições de probabilidade, *verus* t. A "melhor" distribuição é aquela em que a curva de sobrevida estimada pelo método paramétrico aproxima-se da curva estimada por Kaplan-Meier. Vide figura 2.8, retirada de [3]:

- **Comparação de Modelos**

  Muitas vezes é difícil escolher o melhor modelo apenas pela análise dos gráficos. Outra forma de discriminar modelos é por meio de testes de hipóteses. As hipóteses são:

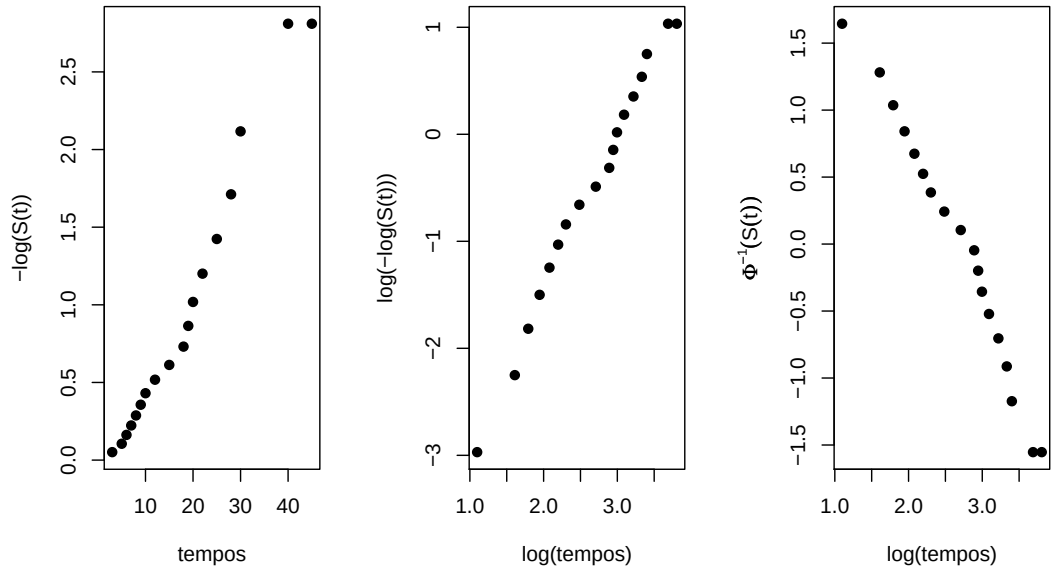


Figura 2.7: Gráficos de t versus $-log(\widehat{S}(t))$, $log(t)$ *versus* $log(-log(\widehat{S}(t)))$ e $log(t)$ *versus* $\Phi^{-1}(\widehat{S}(t))$.

$$H_0 : \text{O modelo de interesse é adequado}$$

versus uma hipótese alternativa de que o modelo não é adequado.

Este teste é usualmente realizado utilizando-se a estatística da razão de verossimilhança. O teste é realizado a partir dos seguintes ajustes: (1) modelo generalizado e obtenção do valor do logaritmo de sua função de verossimilhança $(logL(\hat{\theta}_G))$; (2) modelo de interesse e obtenção do valor do logaritmo de sua função de verossimilhança $(logL(\hat{\theta}_M))$. A partir desses valores, é possível calcular a estatística da razão de verossimilhança, isto é,

$$TRV = -2\, log\left[\frac{L(\widehat{\theta}_M)}{L(\widehat{\theta}_G)}\right] = 2\left[log\, L(\widehat{\theta}_G) - log\, L(\widehat{\theta}_M)\right]$$

que, sob H_0, tem aproximadamente uma distribuição qui-quadrado com graus de liberdade igual a diferença do número de parâmetros ($\hat{\theta}_G$ e $\hat{\theta}_M$) dos modelos sendo comparados. Maiores detalhes podem ser encontrados em [3].

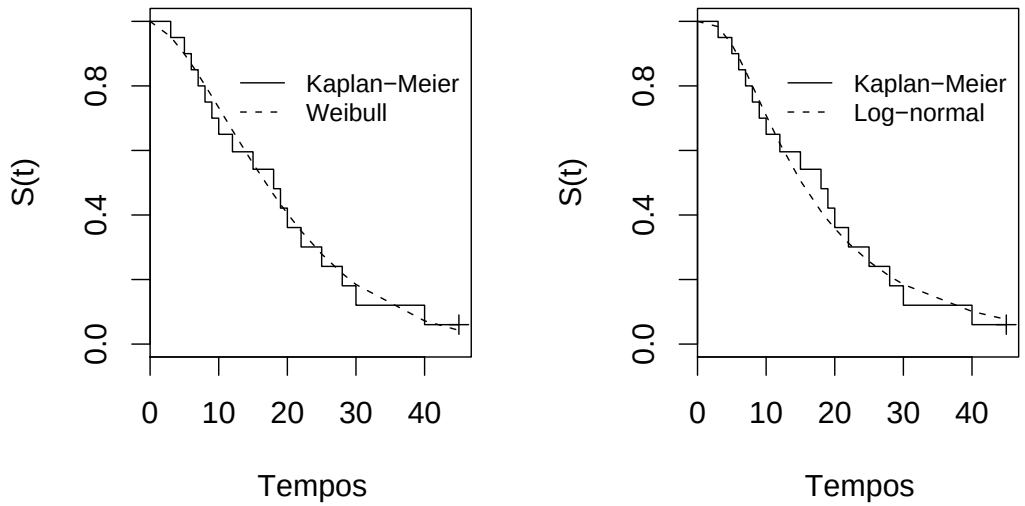


Figura 2.8: Curvas de sobrevivência estimadas pelos modelos de Weibull e log-normal *versus* a curva de sobrevivência estimada por Kaplan-Meier.

A seguir veremos como trabalhar com a presença de covariáveis que podem estar relacionadas com o tempo de sobrevivência.

2.4 Modelos de Regressão Paramétricos

Muitas vezes ao analisar certos conjuntos de dados observamos que algumas covariáveis podem estar relacionadas com o tempo de sobrevivência. As técnicas não-paramétricas não permitem a inclusão direta de covariáveis na análise, mas existem métodos específicos para essas situações, como o ajuste de um modelo de regressão. Nesta seção veremos como modelar a relação entre diversas covariáveis de interesse e o tempo de sobrevivência, assumindo para este alguma das distribuições paramétricas estudadas na seção anterior.

2.4.1 Modelagem Paramétrica

O objetivo de um modelo de regressão é estimar o efeito de covariáveis. Por exemplo, se temos apenas uma covariável, um gráfico de dispersão dessa covariável *versus* a variável resposta (tempo até a ocorrência do evento de interesse) pode detectar uma possível associação entre elas. Se esse gráfico de dispersão sugerir uma relação linear, ou seja, há um nuvem de pontos dispersos em torno de uma reta, podemos modelar essa associação por meio de um modelo de regressão linear. A representação deste modelo é a seguinte:

$$Y = \beta_0 + \beta_1 x + \varepsilon \tag{2.9}$$

em que Y é a resposta, x é a covariável, β_0 e β_1 são os parâmetros a serem estimados e ε é o erro aleatório com distribuição normal.

Para aplicar esse raciocínio para dados de sobrevivência utilizaremos as distribuições paramétricas para o tempo de sobrevivência, incluindo nas observações de cada indivíduo, além dos já descritos tempo de sobrevida e status, (T, δ), o vetor de covariáveis do indivíduo, $x = (x_1, x_2, ..., x_p)$, pois o objetivo é utilizar um modelo de regressão para estudar a relação entre as variáveis (Y e x_p). Como os dados de sobrevivência possuem características especiais, não podemos utilizar diretamente o modelo (2.9), temos que realizar alguns ajustes. Uma dessas características especiais é o fato de que a distribuição da variável resposta tende a ser assimétrica na direção dos maiores tempos de sobrevivência, o que torna inapropriado o uso da distribuição normal para o componente estocástico (variação em torno da reta) do modelo ([3]). Para resolver esse problema iremos utilizar um componente determinístico (equação da reta), não-linear nos parâmetros e uma distribuição assimétrica para o componente estocástico.

- **Modelo de Regressão Exponencial**

  O modelo de regressão exponencial é o mais simples e o mais utilizado na literatura de análise de sobrevivência. Este modelo tem a seguinte expressão:

$$T = exp\{\beta_0 + \beta_1 x\}\,\varepsilon. \tag{2.10}$$

Este modelo admite uma relação linear entre T e x no seu componente determinístico e erro com distribuição assimétrica. O modelo (2.10) é linearizável se for considerado o logaritmo de T. Assim, temos:

$$Y = log(T) = \beta_0 + \beta_1 x + \nu \tag{2.11}$$

onde $\nu = log(\varepsilon)$, que segue uma distribuição do valor extremo padrão ($f(\nu) = exp\{\nu - exp\{\nu\}\}$), essa distribuição caracteriza de forma adequada a distribuição do logaritmo de certos tempos de vida. Em (2.10) e (2.11), x atua linearmente em Y e multiplicativamente em T. Ainda, a função de sobrevivência para Y condicional a x é expressa para este modelo por:

$$S(y \mid x) = exp\{-exp\{y - (\beta_0 + \beta_1 x)\}\}.$$

Para T condicional a x, a função de sobrevivência correspondente é:

$$S(t \mid x) = exp\left\{-\left(\frac{t}{exp\{\beta_0 + \beta_1 x\}}\right)\right\}.$$

O próximo passo é fazer a estimação dos parâmetros, $\theta = (\beta_0, \beta_1)$, pelo método de máxima verossimilhança como apresentada na seção 2.3.4. Para o modelo (2.11), considerando uma amostra de tamanho n, a função de verossimilhança é:

$$L(\theta) = \prod_{i=1}^{n} [f(y_i \mid x_i)]^{\delta_i} [S(y_i, \mid x_i)]^{(1-\delta_i)} \tag{2.12}$$

em que $y_i = log(t_i)$, ou, ainda, para modelos na forma (2.10),por:

$$L(\theta) = \prod_{i=1}^{n} [f(t_i \mid x_i)]^{\delta_i} [S(t_i, \mid x_i)]^{(1-\delta_i)} \tag{2.13}$$

Lembremos que para o i-ésimo indivíduo os dados são representados por $(t_i; \delta_i; x_i)$, onde se $\delta = 1$, temos para t um tempo de falha, mas se $\delta = 0$ temos para t um tempo de censura.

Para a obtenção dos estimadores de máxima verossimilhança, é necessário substituir as funções de densidade e sobrevivência por aquelas da distribuição do

valor extremo em (2.12) ou da exponencial em (2.13). Maiores detalhes podem ser encontrados em [3].

- **Modelo de Regressão Weibull**

  A utilização da distribuição Weibull no contexto da modelagem de sobrevida significa que o tempo T segue uma distribuição de Weibull ([6]). Para isso iremos incluir um parâmetro extra de escala em (2.11), este passa a ter a seguinte forma:

$$Y = log(T) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \sigma\nu \tag{2.14}$$

  como T tem distribuição Weibull, $\log(T)$ segue uma distribuição do valor extremo com parâmetro de escala σ. Sendo assim, a função de sobrevivência para Y condicional a x é expressa por:

$$S(y|\, x) = exp\left\{-exp\left\{\frac{y - x'\beta}{\sigma}\right\}\right\}$$

  em que $x' = (1, x_1, ..., x_p)$ e $\beta = (\beta_0, \beta_1, ..., \beta_p)'$, para T condicional a x, temos:

$$S(t|\, x) = exp\left\{-\left(\frac{t}{exp\{x'\beta\}}\right)^{1/\sigma}\right\}.$$

Vimos os modelos de regressão quando T apresenta um distribuição exponencial ou uma distribuição Weibull, mas podemos ter outras distribuições para ν e para T. Outras distribuições para o tempo podem ser: log-normal, gama ou log-logística, dentre outras. De forma correspondente, a distribuição de ν é normal, log-gama ou logística. Após ser especificada uma distribuição para T, ou de forma equivalente para Y, a função de verossimilhança fica completamente determinada. Sua forma geral é a equação (2.13).

2.4.2 Adequação do Modelo Ajustado

Como no modelo de regressão usual, aqui também temos que avaliar a adequação do modelo ajustado. As técnicas gráficas são bastante utilizadas para examinar dife-

rentes aspectos do modelo. Uma delas é avaliar, por meio dos resíduos, a distribuição dos erros. Estas técnicas devem ser utilizadas como um meio de rejeitar modelos claramente inapropriados. Apresentaremos aqui algumas técnicas de análise de resíduos, com o resíduo de Cox-Snell, os resíduos padronizados, os resíduos de Martingal e os resíduos Deviance.

- **Resíduos de Cox-Snell**

  Essa medida auxilia a examinar o ajuste global do modelo, e é definida por:

$$\widehat{e_i} = \widehat{\Lambda}(t_i|\ x_i)$$

  em que $\hat{\Lambda}(\cdot)$ é a função de risco acumulado obtida do modelo ajustado. Para os modelos de regressão exponencial, Weibull e log-normal, os resíduos de Cox-Snell são dados, respectivamente, por:

$$Exponencial : \widehat{e_i} = \left[t_i \ exp\left\{ -x_i'\widehat{\beta} \right\} \right],$$

$$Weibull : \widehat{e_i} = \left[t_i \ exp\left\{ -x_i'\widehat{\beta} \right\} \right]^{\widehat{\gamma}}$$

  e

$$log - normal : \widehat{e_i} = -log\left[1 - \Phi\left(\frac{log(t_i) - x_i'\widehat{\beta}}{\widehat{\sigma}} \right) \right].$$

  Os resíduos $\hat{e}_i$ vêm de uma população homogênea e devem seguir uma distribuição exponencial padrão se o modelo for adequado. Desse modo, o gráfico $\hat{e}_i$ *versus* $\hat{\Lambda}(\hat{e}_i)$ deve ser aproximadamente uma reta com coeficiente angular igual a 1, quando o modelo exponencial for adequado, uma vez que $\hat{\Lambda}(\hat{e}_i) = -log(\hat{S}(\hat{e}_i))$. Aqui, $\hat{S}(\hat{e}_i)$ é a função de sobrevivência dos $\hat{e}_i$'s obtida pelo estimador de Kaplan-Meier. O gráfico das curvas de sobrevivência desses resíduos, obtidas por Kaplan-Meier e pelo modelo exponencial padrão, também auxiliam na verificação da qualidade do modelo ajustado. Quanto mais próximas elas se apresentam, melhor é considerado o ajuste do modelo aos dados.

Um problema com esta medida é que ela não indica o tipo de violação do modelo quando o gráfico dos resíduos não é linear. Um outro problema ocorre quando não são usados os valores verdadeiros dos parâmetros e sim as estimativas, pois podem ocorrer falhas quanto à distribuição exponencial devido, parcialmente, à incerteza no processo de estimação do vetor de parâmetros β. Essa incerteza é maior na cauda direita da distribuição e para amostras pequenas ([6]).

- **Resíduos Padronizados**

  Os resíduos padronizados são baseados na representação dos modelos log-lineares e são quatidades calculadas por:

$$\widehat{v}_i = \frac{(y_i - x_i'\widehat{\beta})}{\widehat{\sigma}}$$

  com $y_i = log(t_i)$.

  Assim, se o modelo de regressão log-normal for adequado, esses resíduos devem ser uma amostra censurada da distribuição normal padrão. Os resíduos $\hat{v}_i$ são estimativas dos erros que vêm de uma população homogênea. O modelo de regressão log-normal, por exemplo, é considerado adequado se o gráfico de probabilidade normal dos resíduos $\hat{v}_i$ for aproximadamente uma reta.

- **Resíduos Martingal e Resíduos Deviance**

  A definição de resíduos martingal para modelos paramétricos é dada por:

$$\widehat{m}_i = \delta_i - \widehat{e}_i$$

  em que δ_i é a variável indicadora de falha e $\hat{e}_i$, os resíduos de Cox-Snell. Esses resíduos são vistos como uma estimativa do número de falhas em excesso observada nos dados mas não preditos pelo modelo e são um ligeira modificação dos resíduos de Cox-Snell. Eles são utilizados para determinar a forma funcional (linear, quadrática, etc.) de uma covariável, em geral contínua, sendo incluída no modelo de regressão.

  Outros tipos de resíduos, denominados resíduos deviance, são uma tentativa de fazer com que os resíduos martingal sejam mais simétricos em torno de zero e

que possam auxiliar a detectar pontos atípicos (outliers). Eles são definidos por:

$$\widehat{d_i} = sinal(\widehat{m}_i)\left[-2\left(\widehat{m}_i + \delta_i log(\delta_i - \widehat{m}_i)\right)\right]^{1/2}.$$

Gráficos dos resíduos martingal ou deviance contra o tempo ou contra o índice da observação, por exemplo, fornecem uma maneira de verificar globalmente a qualidade do ajuste do modelo. Se o modelo é correto, então essas duas medidas de resíduos não devem apresentar qualquer padrão claro, ou seja, devem se comportar como ruídos aleatórios.

2.4.3 Interpretação dos Coeficientes Estimados

Uma proposta para a interpretação dos coeficientes estimados é a de fazer uso da razão de tempos medianos, ou seja, pode-se mostrar para uma covariável binária que a razão dos tempos medianos é:

$$\frac{t_{0,5}(x=1,\widehat{\beta})}{t_{0,5}(x=0,\widehat{\beta})} = e^{\widehat{\beta}}.$$

Neste caso, a covariável é representada por variáveis indicadoras e a interpretação acima vale para cada uma delas.

3 *Análise dos Dados de Originação do Ciclo de Crédito*

3.1 Descrição do Estudo e das Variáveis

Inicialmente foram selecionados junto a FINANCEIRA 261.115 clientes que nos meses de julho a outubro do ano de 2006 adquiriram determinada linha de crédito, para este estudo foram observados dois eventos de interesse, cancelamento voluntário a pedido do cliente e cancelamento por inadimplência. Os clientes do estudo foram observados por um período de até 22 meses, onde foi registrado para cada cliente a ocorrência ou não de algum dos eventos de interesse. Foram coletadas também seis covariáveis que classificam o cliente no momento da entrada, são elas: Grupo, Produto, Canal, Credit Score, Débito e Limite.

- **Grupo**

  A covariável Grupo representa a classificação de cada cliente para a FINANCEIRA sendo dividida em Correntista, Facility, Não-Correntista e Universitário.

Código	Descrição
C	Correntistas: clientes que possuem conta corrente
F	Facility: clientes provenientes de mailings de mercado
N	Não-Correntistas: clientes que não possuem conta corrente
U	Universitário: clientes com perfil universitária

  Clientes classificados como Correntistas são subdivididos em três grupos: C, H e N.

Código	Descrição
CC	Correntistas com marcação de recebimento de salário na instituição
CH	Correntistas com confirmação de recibento de salário na instituição
CN	Correntistas que não se enquadram nas classificações acima

- **Produto**

A linha de crédito estudada dispõem de diferentes produtos, variando de acordo com o perfil de cada cliente.

Código	Descrição
A	Auto: para clientes com financiamento de automóveis
C	Classic: para clientes de renda baixa a média
G	Golden: para clientes de renda média a alta
O	Open: para clientes de baixa renda
P	Premium: para clientes de alta renda
S	Social: para todos os clientes

- **Canal de Entrada**

O Canal de Entrada representa a origem da venda da linha de crédito em estudo.

Código	Descrição
A	Agência
F	Célula de financiamento de automóveis
I	Internet
O	Outros
T	Telemarketing ativo

- **Credit Score**

A covariável Credit Score, representa a classificação recebida no momento da proposta com base em dados cadastrais, classificação esta representada por níveis de 1 a 5, onde 1 representa um bom pagador e 5 representa um mau

pagador. Clientes com problemas cadastrais são classificados com Credit Score H.

- **Débito Automático**

  A covariável Débito Automático sinaliza clientes que possuem as faturas de sua linha de crédito debitadas diretamente em sua conta corrente. Classificados aqui em dois grupos: Não possuem (N) e possuem (S).

- **Limite**

  Os limites de cada cliente, na linha de crédito em estudo, estão aqui representados em faixas.

Código	Descrição
<500	Até R$500,00
500-999	De R$500,00 a R$999,00
1000-2499	De R$1.000,00 a R$2.499,00
2500-4999	De R$2.500,00 a R$4.999,00
>5000	Acima de R$5.000,00

Antes de se iniciar o estudo quantitativo, foram sorteados e reservados 200 clientes com a seguinte combinação de covariáveis: Grupo = CC , Produto = O, Canal = A, Débito = N, Limite = <500 e Credit Score =3, os dados reservados serão utilizados posteriormente para validar o ajuste, com isso poderemos verificar graficamente se o modelo está se adequando bem a dados fora da amostra utilizada pela modelagem.

3.2 Análise Descritiva e Exploratória

Num primeiro passo em realizaremos uma análise descritiva dos dados. Como foi explicado na seção 2.2, em estudos de análise de sobrevivência são utilizados métodos não-paramétricos, como o método de Kaplan-Meier para realizar essa primeira análise. Primeiramente, apresentaremos os dois gráficos, representados pelas figuras 3.1 e 3.2, das curvas de $S(t)$ estimadas por Kaplan-Meier *versus* os tempos.

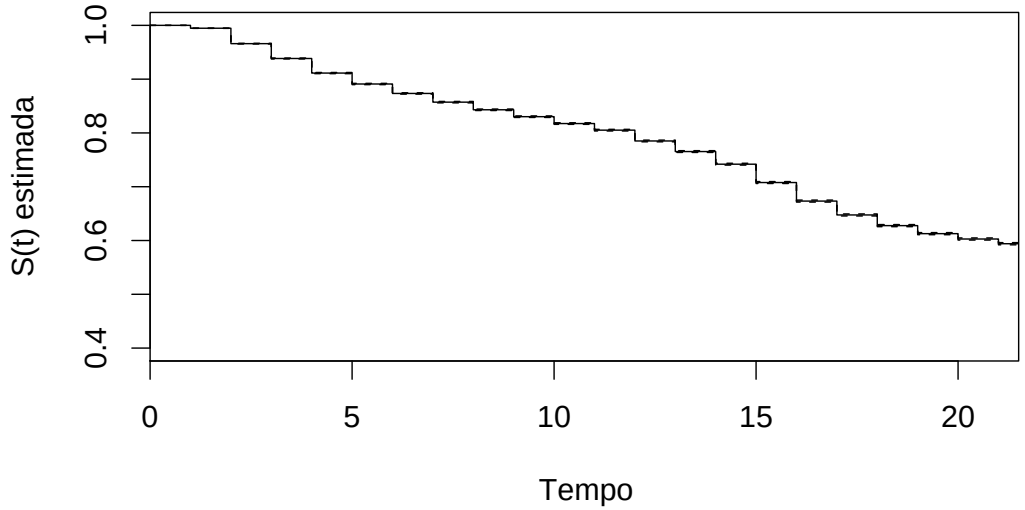


Figura 3.1: Sobrevivência e respectivos intervalos de 95% de confiança estimados a partir do estimador de Kaplan-Meier para os dados de originação de crédito com base na variável resposta tempo até o cancelamento voluntário.

Observando a figura 3.1, onde estão representados os tempos de cancelamento voluntário da linha de crédito, podemos ver que a curva de sobrevida inicia seu caimento no primeiro mês observado, isso sinaliza a existencia de clientes que solicitam o cancelamento da linha de crédito já no primeiro mês. Outro ponto que devemos observar é um caimento acelerado entre o décimo segundo e o décimo quinto mês. Uma justificativa para esta ocorrência é o fato de que apenas no segundo ano, o cliente será cobrado pela anuidade para a manutenção dessa linha de crédito, desse modo, muitos clientes pedem o cancelamento desta.

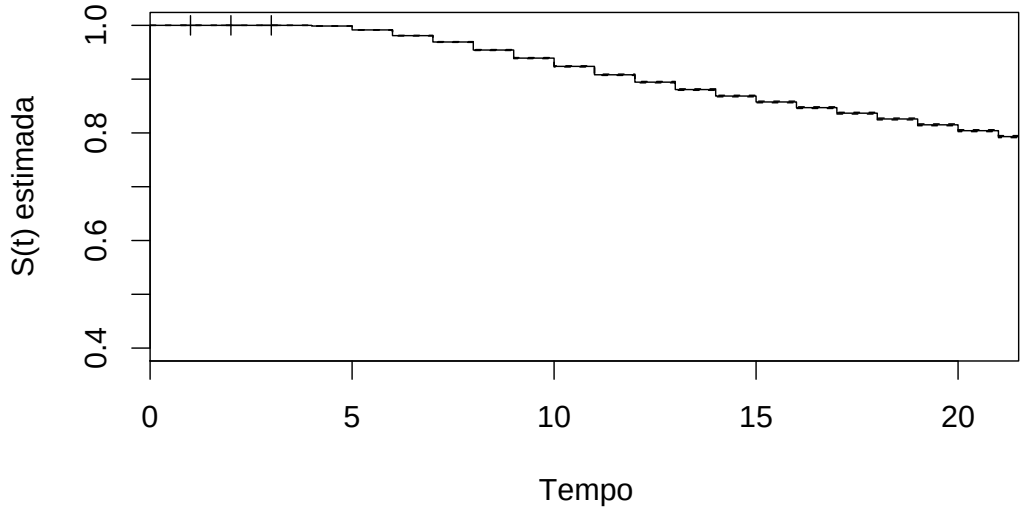


Figura 3.2: Sobrevivência e respectivos intervalos de 95% de confiança estimados a partir do estimador de Kaplan-Meier para os dados de originação de crédito com base na variável resposta tempo até o cancelamento por inadimplência.

Podemos observar na figura 3.2 que a curva de sobrevida inicia seu caimento partir do terceiro mês, como essa curva representa a inadimplência dos clientes, esses são considerados inadimplentes após o terceiro mês de atraso do pagamento dos débitos da linha de crédito. Comparando os dois gráficos no tempo $t = 15$ meses, por exemplo, na figura 3.1 a sobrevida para este tempo é de 0.7, já na figura 3.2 a sobrevida fica em torno de 0.9. Isso deixa claro que as duas curvas comportam-se de maneiras diferentes, pois tratam de tempos com características bem distintas.

Em ambos os gráficos os intervalos de confiança para $\hat{S}(t)$ são bem pequenos, ficando assim, bem próximos das curvas de $\hat{S}(t)$. Isso se deve ao fato de termos uma amostra de tamanho consideravelmente grande.

Nos gráficos que ilustrados acima, foram analisados os comportamentos das variáveis respostas para os eventos cancelamento voluntário e cancelamento por inadimplência. Nos tópicos abaixo iremos analisar o comportamento dos níveis das covariáveis, também

utilizando os gráficos de Kaplan-Meier *versus* os tempos.

- **Grupo**

  A primeira covariável a ser estudada é a covariável denominada Grupo. Esta é dividida em 6 níveis que foram definidos na seção 3.1. Abaixo temos as figuras, 3.3 e 3.4, contendo os gráficos das curvas Kaplan-Meier para esta covariável para os dois eventos em estudo: cancelamento voluntário e inadimplência, respectivamente.

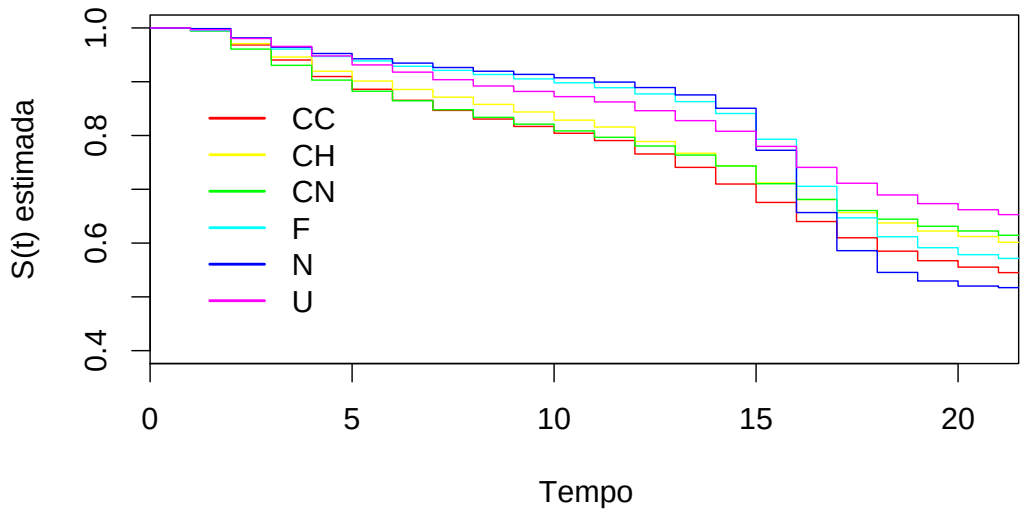


Figura 3.3: Curvas de sobrevivência estimadas a partir do estimador de Kaplan-Meier para a covariável Grupo com base na variável resposta tempo até o cancelamento voluntário.

A figura 3.3 indica que o grupo de Não-Correntistas possui uma curva de sobrevida que cai lentamente até o décimo quinto mês, neste mês a sobrevida desse grupo fica por volta de 0.8, por volta do décimo sétimo mês a sobrevida aproxíma-se de 0.5, após esse período, a curva volta a cair lentamente. Faz-se necessária uma investigação para apurar o que está acontecendo entre esses

meses, que levam a um índice maior de cancelamento dos Não-Correntistas. Podemos observar ainda, que o grupo Facility possui um comportamento parecido com o grupo Não-Correntista. Comparando as figuras 3.3 e 3.4 podemos ver que os grupos dos Correntistas (CC, CH e CN) possuem uma sobrevida menor comparado aos demais grupos na figura 3.3, na figura 3.4 a situação se inverte, os grupos dos Correntistas possuem uma sobrevida maior comparado aos demais grupos. Isso nos leva a entender que esse grupo apresenta uma menor inadimplência, porém solicita mais o cancelamento da linha de crédito.

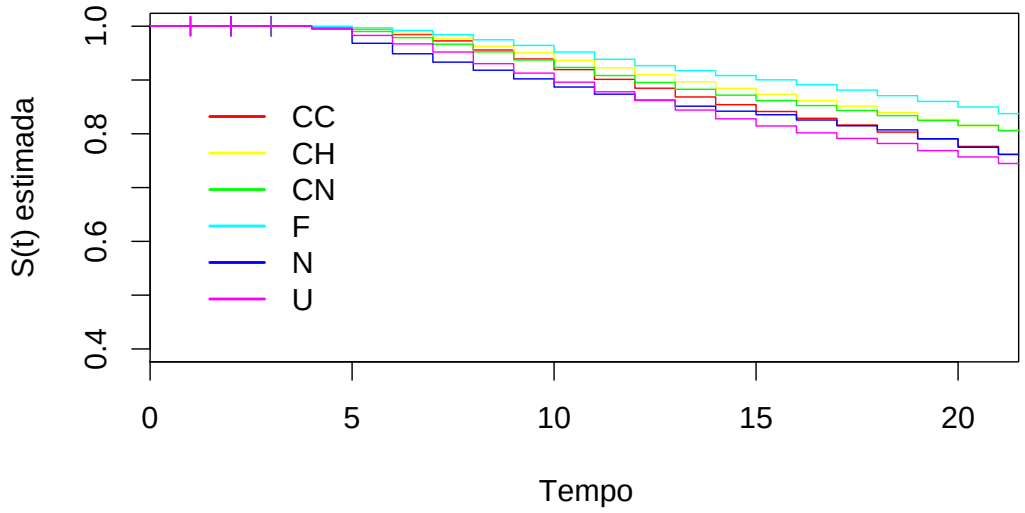


Figura 3.4: Curvas de sobrevivência estimadas a partir do estimador de Kaplan-Meier para a covariável Grupo com base na variável resposta tempo até o cancelamento por inadimplência.

- **Produto**

  As figuras 3.5 e 3.6 são distintas apenas na velocidade de caimento das curvas devido as características de cada uma das variáveis resposta. Mas podemos obsevar que o grupo de clientes classificados com Auto (clientes que adquiriram a linha de crédito através de um finaciamento de automóvel feito pela FINAN-

CEIRA), pois sua curva de sobrevida apresenta um caimento brusco a partir do tempo $t = 15$ meses.

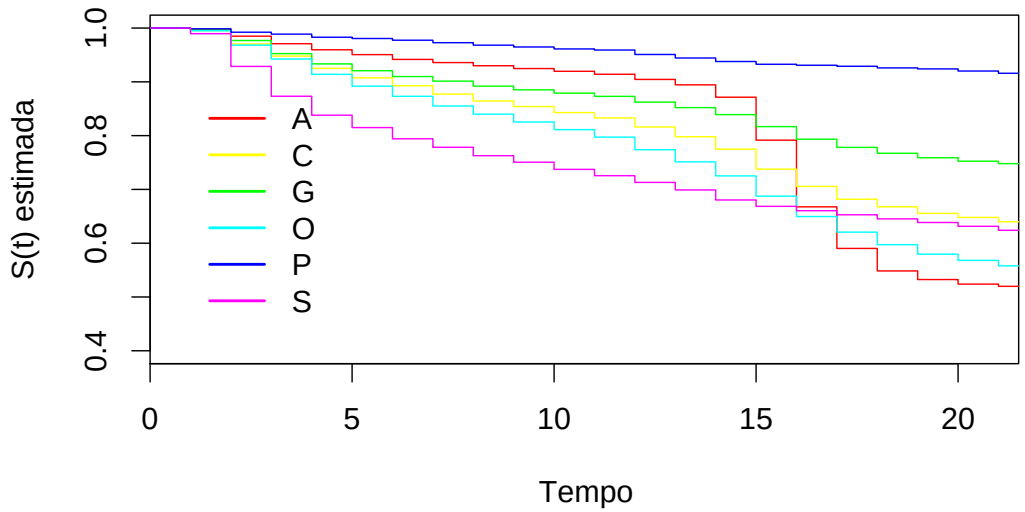


Figura 3.5: Curvas de sobrevivência estimadas a partir do estimador de Kaplan-Meier para a covariável Produto com base na variável resposta tempo até o cancelamento voluntário.

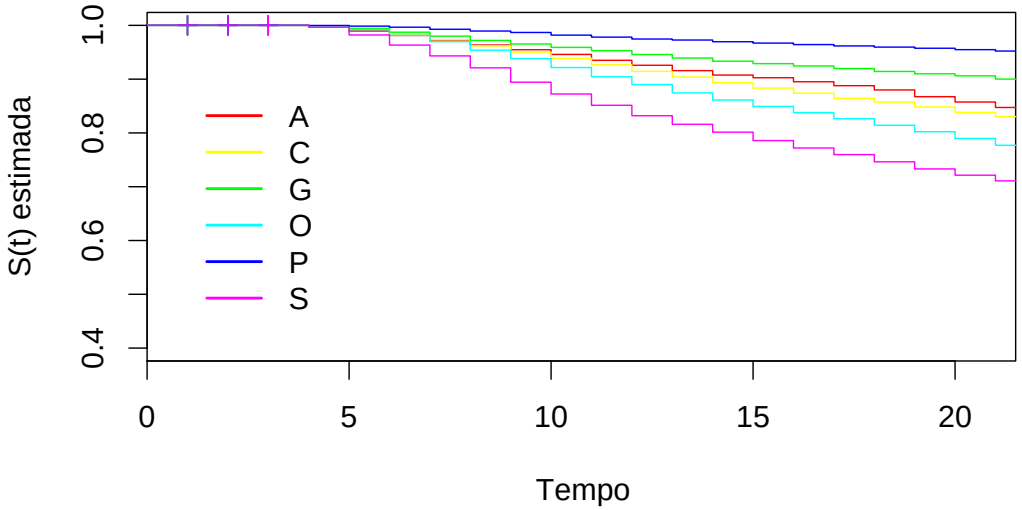


Figura 3.6: Curvas de sobrevivência estimadas a partir do estimador de Kaplan-Meier para a covariável Produto com base na variável resposta tempo até o cancelamento por inadimplência.

- **Canal de Entrada**

  A covariável Canal de Entrada mostra por qual meio o cliente adquiriu a linha de crédito. O primeiro gráfico a ser analisado é o gráfico do evento cancelamento voluntário, e logo em seguida do evento inadimplência.

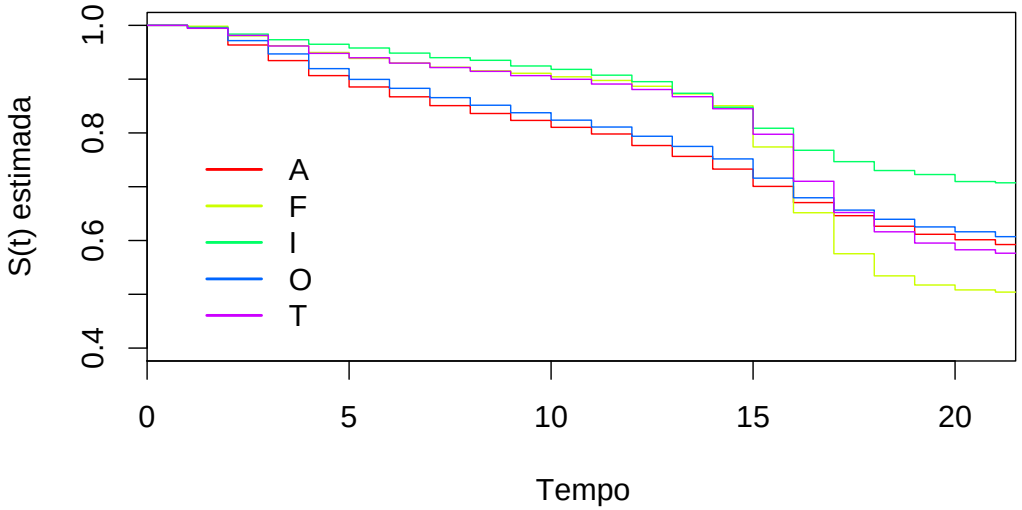


Figura 3.7: Curvas de sobrevivência estimadas a partir do estimador de Kaplan-Meier para a covariável Canal de Entrada com base na variável resposta tempo até o cancelamento voluntário.

A figura 3.7 mostra que os clientes que adquirem a linha de crédito através do Canal Célula de financiamento de automóveis e do Telemarketing Ativo, apresentam uma queda brusca na curva de sobrevida a partir do décimo quinto mês. Fato que acontece também com o produto Auto e com o grupo dos Não-Correntista, todos esses casos aparecem nos gráficos cuja a variável resposta é o tempo até o cancelamento voluntário. Reforçando assim a hipótese de cancelamentos após o recebimento da primeira anuidade para a manutenção da linha de crédito. O grupo que possui uma sobrevida maior quanto ao cancelamento voluntário, é formado pelos clientes que adquirem a linha de crédito pela Internet. Esse comportamento pode ser justificado pelo fato de que quando o cliente solicita sua linha de crédito pela Internet, ele tem mais tempo e/ou facilidade para ler os termos do contrato, assim ele está ciente do serviço que está solicitando, o que não acontece muitas vezes nas vendas pelo Telemarketing, por exemplo, às vezes devido ao fato de que visando o cumprimento de metas, os vendedores

"empurram" a linha de crédito para o cliente, e este acaba cancelando a mesma, pois no ato da venda ele não foi esclarecido sobre todas as taxas e anuidades.

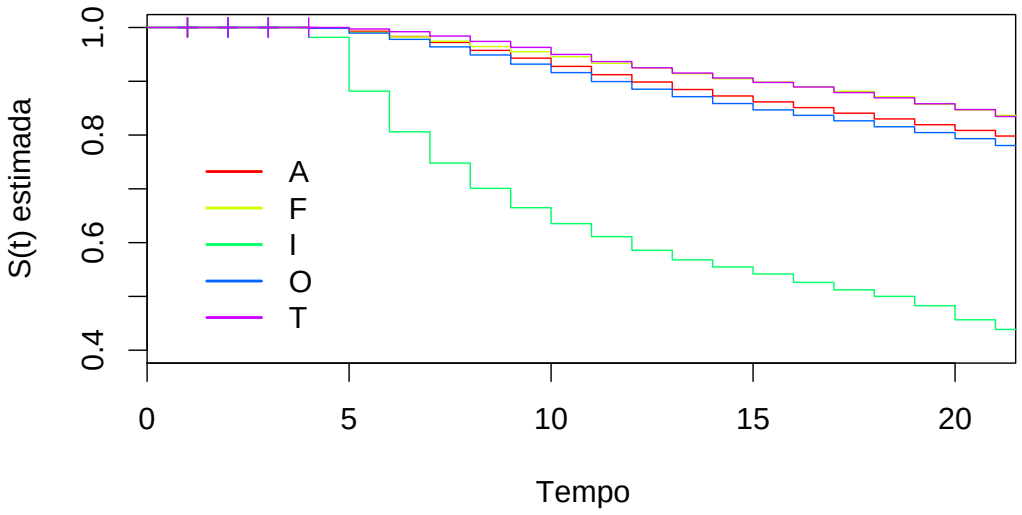


Figura 3.8: Curvas de sobrevivência estimadas a partir do estimador de Kaplan-Meier para a covariável Canal de Entrada com base na variável resposta tempo até o cancelamento por inadimplência.

Na figura 3.8 o Canal Internet possui uma curva de sobrevida com caimento muito acelerado. No tempo $t = 10$ meses, menos de um ano após a contratação da linha de crédito, a sobrevida desse grupo para o evento inadimplência aproxima-se de 0.6. Enquanto a sobrevida dos clientes que "entraram" pelo Canal Telemarketing para esse mesmo tempo $t = 10$, a sobrevida fica em torno de 0.9. Vimos anteriormente que a sobrevida do Canal Internet era alta comparada as sobrevidas dos outros Canais para o evento cancelamento voluntário. Já para a variável inadimplência é a pior curva de sobrevida, fato este que deve ser analisado pela FINANCEIRA.

- **Credit Score**

  Para cada cliente é atribuído um score quando este adquire a linha de crédito,

tentando assim classificar o cliente em níveis de 1 a 5 ou H, esses níveis foram explicados na seção 3.1. Abaixo temos as duas figuras, 3.9 e 3.10, representando essa covariável para as duas variáveis respostas.

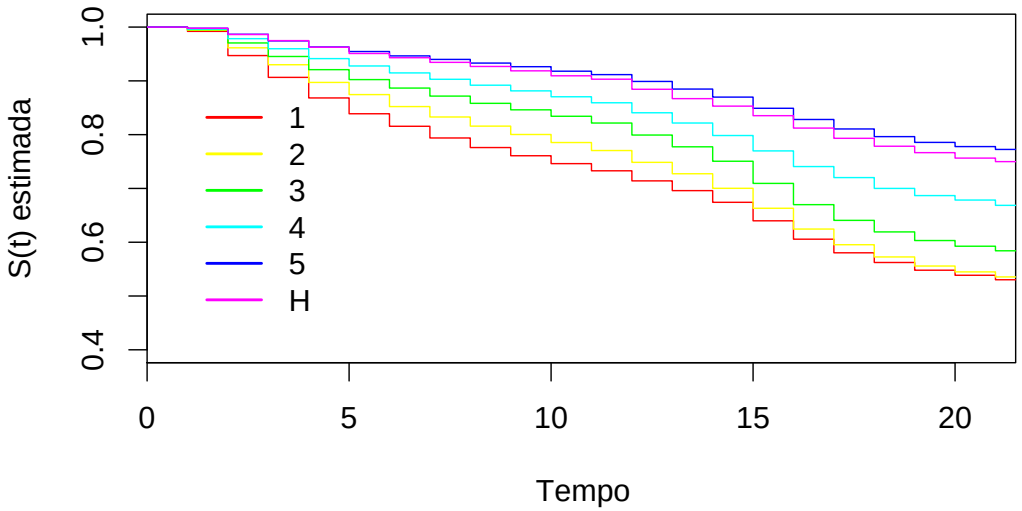


Figura 3.9: Curvas de sobrevivência estimadas a partir do estimador de Kaplan-Meier para a covariável Credit Score com base na variável resposta tempo até o cancelamento voluntário.

Na figura 3.9 os clientes classificados com escores menores, apresentam menor sobrevida em relação aos clientes com escores maiores. Na figura 3.10 observamos com clareza a distinção feita pela covariável Credit Score com relação a inadimplêcia, como já era esperado, clientes com escores menores apresentam maior sobrevida com relaçãoa inadimplência se comparados aos clientes com escores maiores.

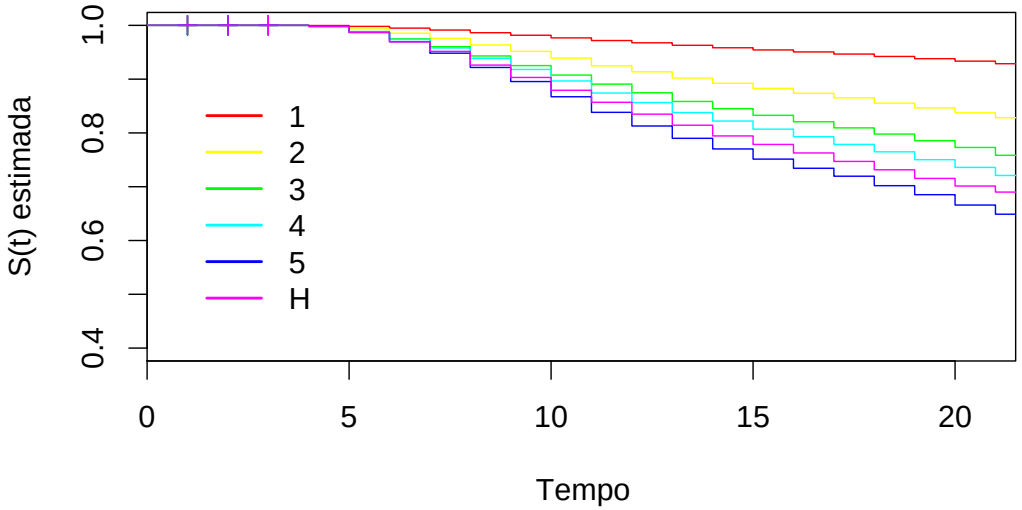


Figura 3.10: Curvas de sobrevivência estimadas a partir do estimador de Kaplan-Meier para a covariável Credit Score com base na variável resposta tempo até o cancelamento por inadimplência.

- **Débito Automático**

  Os gráficos a serem analisados agora, são os gráficos da covariável Débito Automático em Conta Corrente, isto é, se as faturas da linha de crédito são debitadas diretamente da conta corrente do cliente.

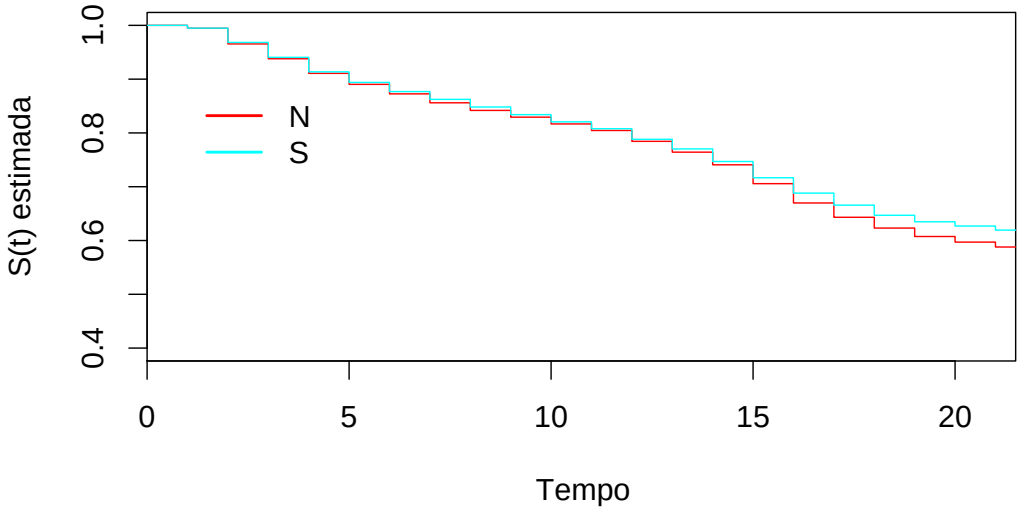


Figura 3.11: Curvas de sobrevivência estimadas a partir do estimador de Kaplan-Meier para a covariável Débito Automático em Conta Corrente com base na variável resposta tempo até o cancelamento voluntário.

A figura 3.11 mostra que a curva de sobrevida dos clientes que possuem Débito Automático em Conta Corrente e a curva dos que não possuem são praticamente iguais, para a variável resposta cancelamento voluntário.

Na figura 3.12 referente ao evento inadimplência, podemos observar que a sobrevida dos clientes que não possuem Débito Automático em Conta Corrente é menor do que a sobrevida dos clientes que possuem. Uma hipótese para esse fato é que a inadimplência é menor para os que possuem débito automático, pois como o pagamento da fatura é efetuado diretamente na conta torna-se mais díficil o cliente deixar de pagar a fatura por problemas financeiros.

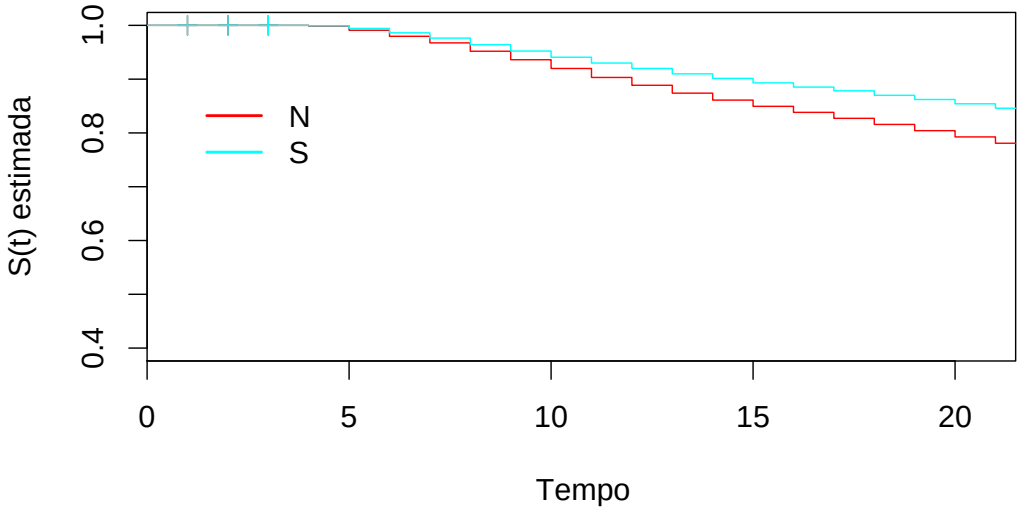


Figura 3.12: Curvas de sobrevivência estimadas a partir do estimador de Kaplan-Meier para a covariável Débito Automático em Conta Corrente com base na variável resposta tempo até o cancelamento por inadimplência.

- **Limite**

  Para terminar, iremos analisar os gráficos da covariável Limite.

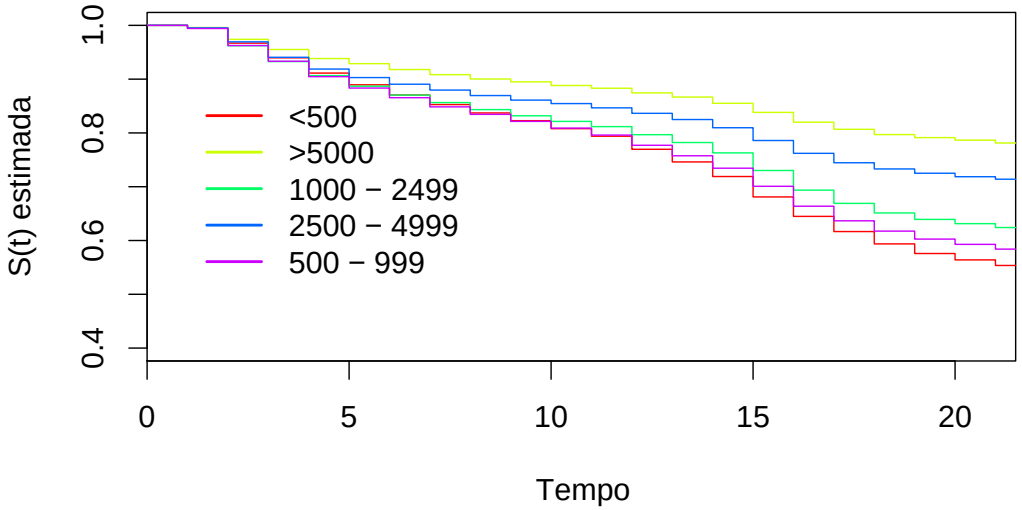


Figura 3.13: Curvas de sobrevivência estimadas a partir do estimador de Kaplan-Meier para a covariável Limite com base na variável resposta tempo até o cancelamento voluntário.

Analisando as figuras 3.13 e 3.14, podemos ver que quanto maior o limite que o cliente possui na sua linha de crédito, maior é a sobrevida para a variável cancelamento voluntário e para a variável inadimplência. Já a "pior" curva, tanto para a variável cancelamento voluntário quanto para inadimplência, é a curva dos clientes que possuem limite de até R$500,00. Este fato já é esperado, pois clientes que possuem um alto limite geralmente não querem perder essa vantagem e acabam não cancelando a linha de crédito voluntariamente. Com relação a inadimplência, como já era esperado, clientes com limites baixos apresentam uma sobrevida baixa se comparados a as faixas de limites mais elevados.

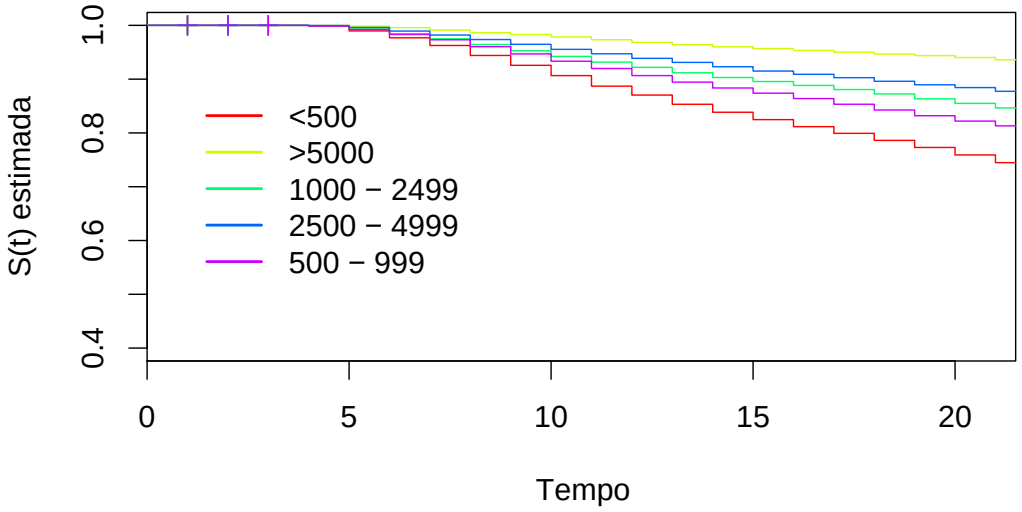


Figura 3.14: Curvas de sobrevivência estimadas a partir do estimador de Kaplan-Meier para a covariável Limite com base na variável resposta tempo até o cancelamento por inadimplência.

Com os gráficos apresentados até o momento podemos observar os níveis, Não-Correntista (covariável Grupo), Auto (covariável Produto) e Telemarketing Ativo (covariável Canal), possuem uma característica em comum, como já foi dito anteriormente, possuiem um caimento brusco em suas curvas de sobrevida a partir do tempo $t = 15$ meses para a variável resposta cancelamento voluntário. A hipótese levantada para esse fenômeno é o fato de que após um ano com a linha de crédito o cliente começa a receber em sua fatura os custos da anuidade para manutenção dessa linha de crédito, pedindo assim o cancelamento da mesma. Outra covariável que possui um comportamento interessante é o Canal Internet sob as duas variáveis respostas, essa covariável apresenta curvas de sobrevida bem distintas em cada variável resposta, fato este que deve ser investigado pela FINANCEIRA.

Com o objetivo de comparar as curvas de sobrevida de cada covariável realizamos o teste de *log-rank*. Este teste apontou diferenças significativas para todas as curvas,

isso acontece devido o tamanho da amostra. Com isso prosseguiremos as análises com todas as covariáveis no estudo.

O próximo passo é encontrarmos um modelo de probabilidades para modelar o tempo até o cancelamento da linha de crédito e o tempo até a inadimplência. Desse modo, na próxima seção iremos encontrar esses modelos para as duas variáveis respostas.

3.3 Ajuste dos Modelos Probabilísticos

Neste momento, o passo mais importante da modelagem é encontrar uma distribuições de probabilidades adequada para os dados em estudo. Somente após encontrar esta distribuição é que será possível estimar e testar as quantidades de interesse.

Inicialmente, serão testadas as distribuição exponencial, Weibull e log-normal, que foram apresentadas nas seções 2.3.1, 2.3.2 e 2.3.3, respectivamente. Essas distribuições foram escolhidas primeiramente por serem as distribuições comumente utilizadas para modelagem de dados de sobrevivência, caso nenhuma dessas distribuições se ajuste bem aos dados, tentaremos usar outras distribuição, como: gama, gama generalizada, dentre outras.

Como vimos na seção 2.3.7, iremos escolher o melhor modelo probabilístico através da análise gráfica e de algumas estatísticas.

O primeiro gráfico que analisaremos é o gráfico das sobrevivências estimadas por Kaplan-Meier *versus* as sobrevivências estimadas pelos modelos exponencial, Weibull e log-normal. Primeiramente, iremos analisar esse gráfico para o evento cancelamento voluntário.

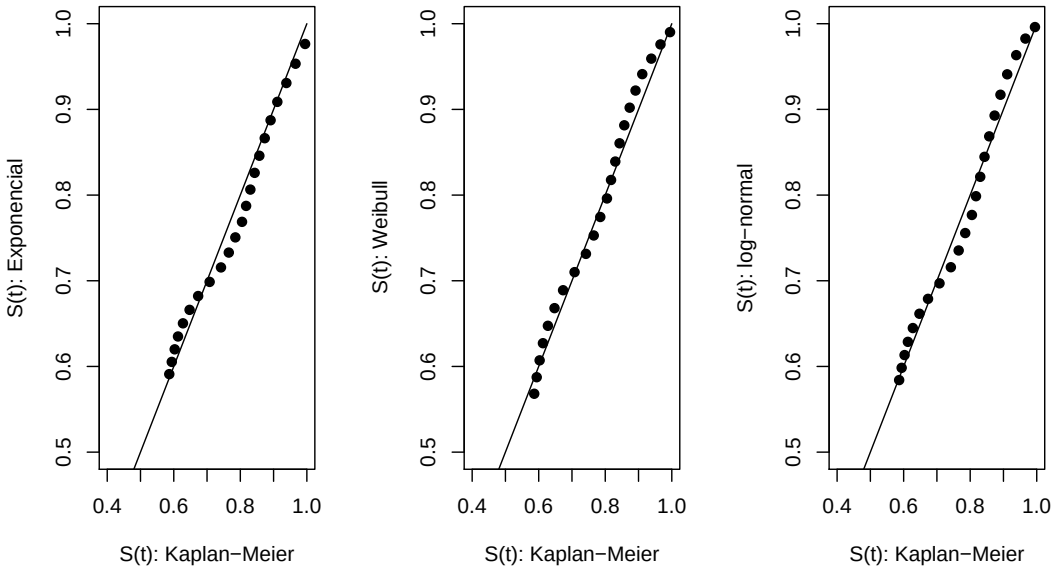


Figura 3.15: Gráficos das sobrevivências estimadas por Kaplan-Meier *versus* as sobrevivências estimadas pelos modelos exponencial, de Weibull e log-normal, com base no evento cancelamento voluntário.

A partir da figura 3.15, é possível observar que os modelos exponencial e log-normal parecem não serem adequados para esses dados, pois a curva se afasta um pouco da reta $y = x$ em alguns pontos. Já para o modelo Weibull a curva se aproxima da reta $y = x$, indicando ser possivelmente o modelo adequado para a variável resposta tempo até o cancelamento voluntário.

Agora iremos analisar o mesmo gráfico para o evento cancelamento por inadimplência.

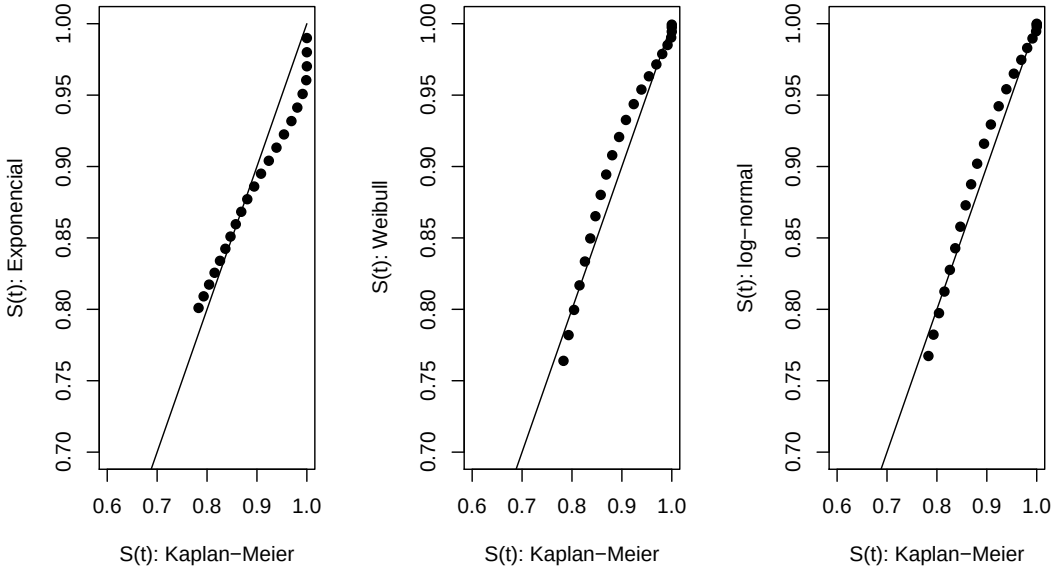


Figura 3.16: Gráficos das sobrevivências estimadas por Kaplan-Meier *versus* as sobrevivências estimadas pelos modelos exponencial, de Weibull e log-normal, com base na variável resposta cancelamento por inadimplência.

Podemos verificar pela figura 3.16 que o modelo exponencial é o modelo que apresenta os pontos mais afastados da reta $y = x$, indicando assim os modelos Weibull e log-normal, possivelmente, adequados para a variável resposta tempo até o cancelamento por inadimplência.

Os próximos gráficos também irão ajudar na escolha do melhor modelo, são os gráficos que utilizam o método de linearização da função sobrevivência.

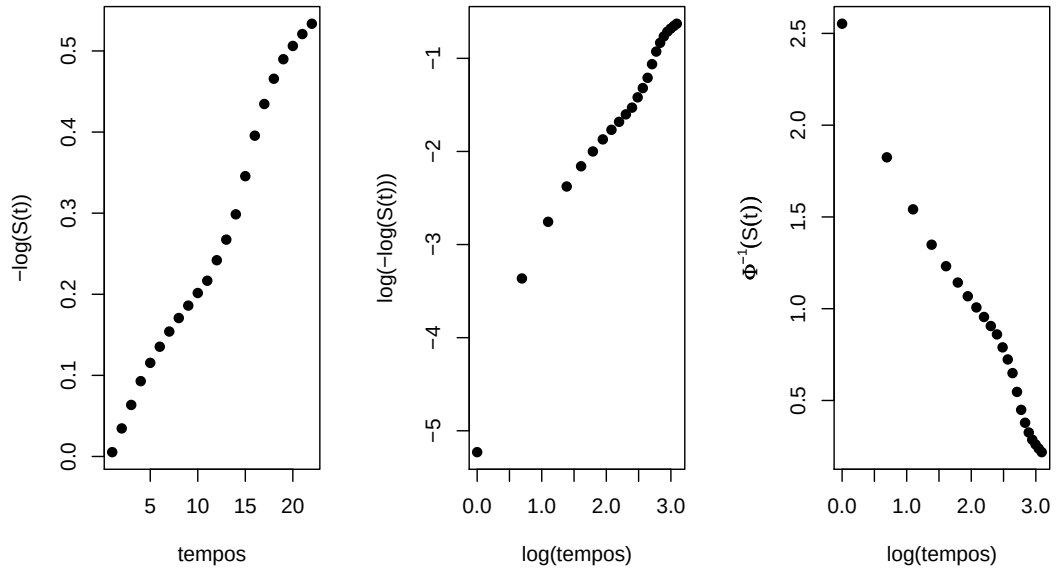


Figura 3.17: Gráficos de t *versus* $-log(\widehat{S}(t))$, $log(t)$ *versus* $log(-log(\widehat{S}(t)))$ e $log(t)$ *versus* $\Phi^{-1}(\widehat{S}(t))$, com base no evento cancelamento voluntário.

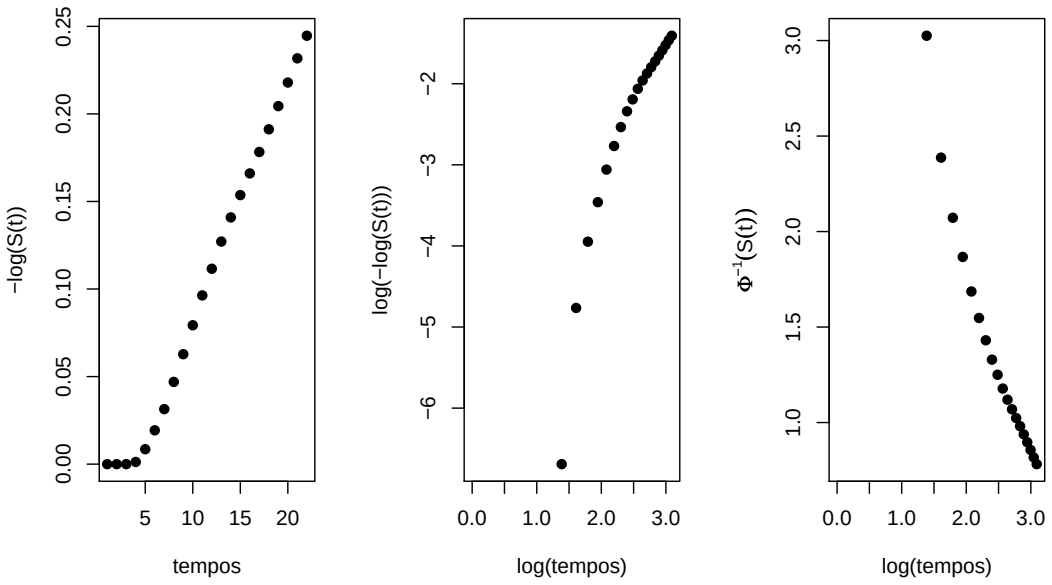


Figura 3.18: Gráficos de t *versus* $-log(\widehat{S}(t))$, $log(t)$ *versus* $log(-log(\widehat{S}(t)))$ e $log(t)$ *versus* $\Phi^{-1}(\widehat{S}(t))$, com base nevento cancelamento por inadimplência.

Os "melhores" modelos são aqueles cujo comportamento aproxima de uma reta. As figuras 3.17 e 3.18, não mostram afastamentos marcantes de uma reta. O último conjunto de gráficos a ser analiso são os gráficos de $S(t)$ estimada por Kaplan-Meier e pelos métodos parámetricos *versus* os tempos.

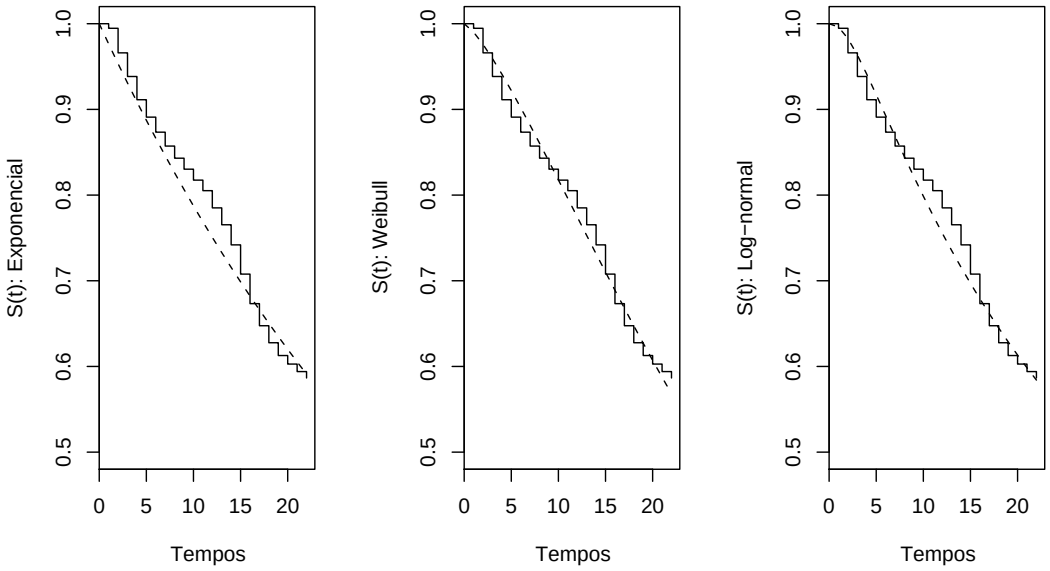


Figura 3.19: Curvas de sobrevivência estimadas pelos modelos exponencial, Weibull e log-normal *versus* a curva de sobrevivência estimada por Kaplan-Meier, com base no evento cancelamento voluntário.

Na figura 3.19 a curva de $S(t)$ estimada pelo método paramétrico que mais se aproxima da curva de Kaplan-Meier é a da distribuição Weibull, indicando assim que a distribuição de probabilidade Weibull é a distribuição mais adequada para a variável resposta tempo até o cancelamento voluntário. Já para a variável resposta tempo até o cancelamento por inadimplência, pela figura 3.20, observamos que a distribuição log-normal é a distribuição mais indicada.

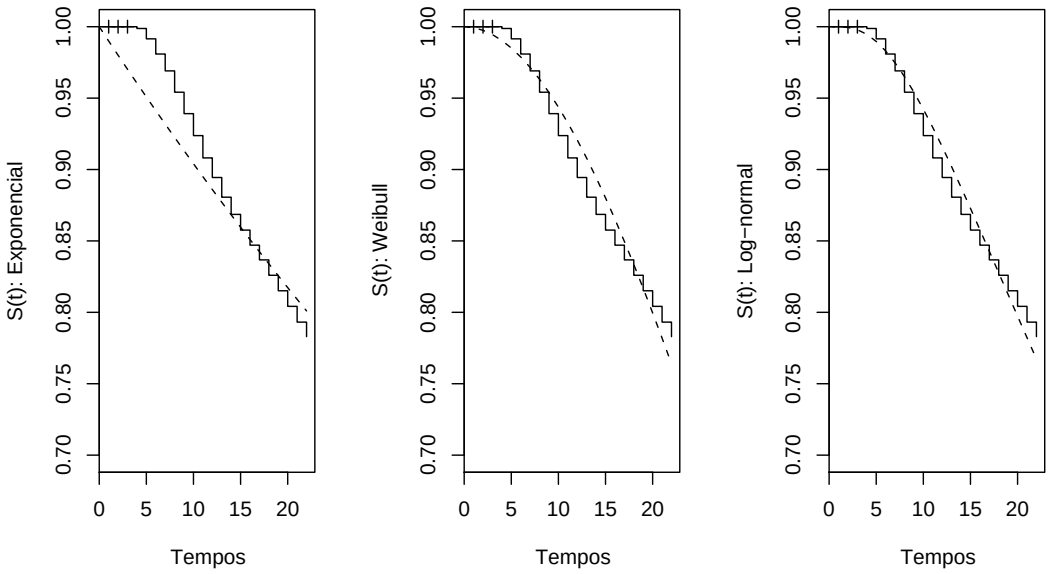


Figura 3.20: Curvas de sobrevivência estimadas pelos modelos exponencial, Weibull e log-normal *versus* a curva de sobrevivência estimada por Kaplan-Meier, com base no evento cancelamento por inadimplência.

Segundo as técnicas gráficas os modelos indicados são: modelo de Weibull para a variável resposta tempo até o cancelamento voluntário e log-normal para o tempo até o cancelamento por inadimplência. Para confirmar a escolha desses modelos foram construídas duas tabelas, uma para cada variável resposta, comparando os valores das estimativas de $S(t)$ para cada tempo, pelo método de Kaplan-Meier, exponencial, Weibull e log-normal. Estas duas tabelas encontram-se nos apêndices deste trabalho, A.1 e A.2. Foram também calculados os valores do $-log(L(\theta))$ e os erros quadráticos, como mostra a tabela abaixo:

Model	$-log(L(\theta))$	EQ
Voluntário ~ exponencial	455240	0.00872
Voluntário ~ Weibull	451514	0.00633
Voluntário ~ lognormal	451146	0.00767
Inadimplência ~ exponencial	227114	0.01076
Inadimplência ~ Weibull	218380	0.00513
Inadimplência ~ lognormal	216273	0.00328

Desse modo, concluímos que os modelos que melhor se ajustam as variáveis respostas, tempo até o cancelamento voluntário e tempo até o cancelamento por inadimplência, são os modelos Weibull e log-normal, respectivamente. Como sabemos que existem fatores que influenciam no tempo de sobrevida, o próximo passo é ajustar esses modelos encontrados nesta seção com as covariáveis apresentadas em 3.1.

3.4 Ajuste dos Modelos de Regressão

Como o interesse desse trabalho é identificar fatores que possam indicar previamente eventuais perdas, iremos nessa seção introduzir as covariáveis apresentadas em 3.1 nos modelos probabilísticos encontrados na seção anterior.

O primeiro passo foi construir todos os modelos possíveis com as seis covariáveis para cada variável resposta, e com algumas interações onde suspeita-se que sejam significativas para a FINANCEIRA. Todos os modelos ajustados podem ser encontrados nos apêndices deste trabalho, B.1 e B.2.

Para encontrar o modelo que melhor se ajusta aos dados foram utilizadas as estatísticas $-log(L(\theta))$ e AIC. Os modelos que tiverem os menores valores dessas estatísticas serão os modelos utilizados. Dentre todos os modelos, os que apresentaram os menores valores de $-log(L(\theta))$ e AIC, foram os seguintes:

$$Voluntário \sim GRUPO + PROD * CANAL + LIMITE + CS + DEBT \quad (3.1)$$

e

$$Inadimplência \tilde{} GRUPO * PROD + CANAL + LIMITE + CS + DEBT \quad (3.2)$$

para as variáveis respostas tempo até o cancelamento voluntário (distribuição Weibull) e tempo até o cancelamento por inadimplência (distribuição log-normal).

As estimativas de ambos os modelos podem ser encontradas nos apêndices, C.1 e C.2. Encontrados os modelos, antes de interpretá-los, devemos verificar a adequação dos mesmos. Para isso foram construídos os gráficos de Cox-Snell, método este discutido na seção 2.4.2. Esses resíduos devem seguir a distribuição exponencial padrão para que os modelos de regressão Weibull e log-normal possam ser considerados adequados. A partir dos gráficos apresentados nas figuras 3.21 e 3.22. Notamos que o modelo Weibull não se ajusta bem aos dados sob a variável resposta tempo até o cancelamento voluntário, vide 3.21. E no gráfico apresentado na figura 3.22, vemos a distribuição log-normal está se ajustando bem aos dados sob a variável resposta tempo até o cancelamento por inadimplência.

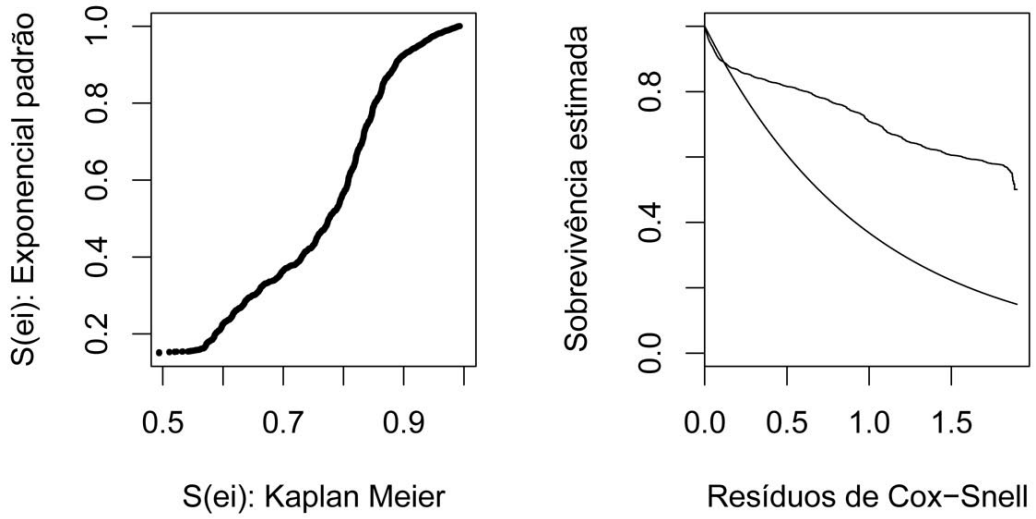


Figura 3.21: Sobrevivência dos resíduos de Cox-Snell estimadas pelo método de Kaplan-Meier e pelo modelo exponencial padrão (gráfico à esquerda) e respectivas curvas de sobrevivência estimadas (gráfico à direita) para a variável resposta tempo até o cancelamento voluntário.

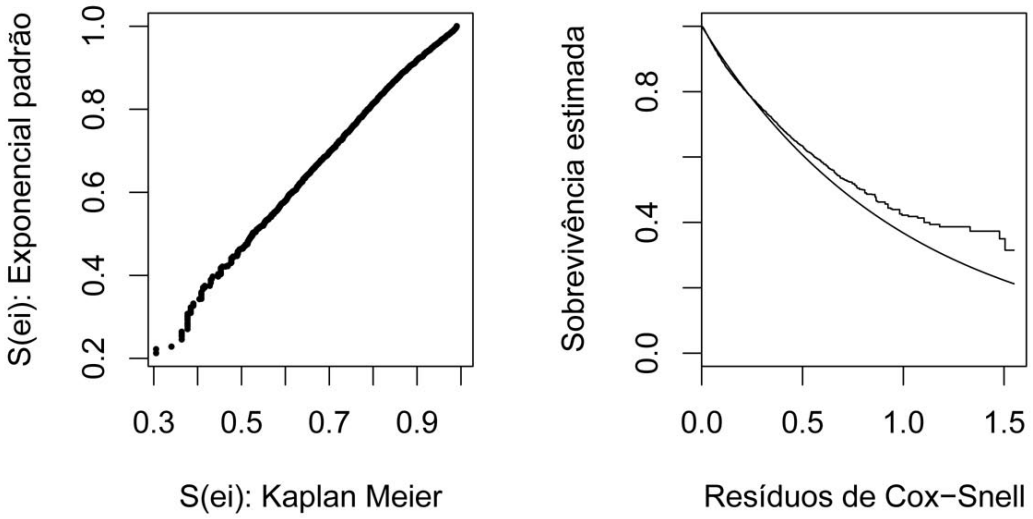


Figura 3.22: Sobrevivência dos resíduos de Cox-Snell estimadas pelo método de Kaplan-Meier e pelo modelo exponencial padrão (gráfico à esquerda) e respectivas curvas de sobrevivência estimadas (gráfico à direita) para a variável resposta tempo até o cancelamento por inadimplência.

Como observamos anteriormente, a análise dos resíduos não foi satisfatória, para testar a adequação dos modelos, utilizaremos outros meios descritos a seguir.

Como foi explicado em 3.1, antes de se iniciar o estudo quantitativo, foram sorteados e reservados 200 clientes para serem utilizados como validação do modelo, com a seguinte combinação de covariáveis: Grupo = CC , Produto = O, Canal = A, Débito = N, Limite = <500 e Credit Score = 3, escolheu-se este grupo, pois combinando as seis covariáveis esse foi o maior grupo encontrado. As figuras 3.23 e 3.24 apresentam as estimativas de $S(t)$ por Kaplan-Meier, com seus respectivos intervalos de confiança de 95%, e pelos modelos encontrados: 3.1 e 3.2.

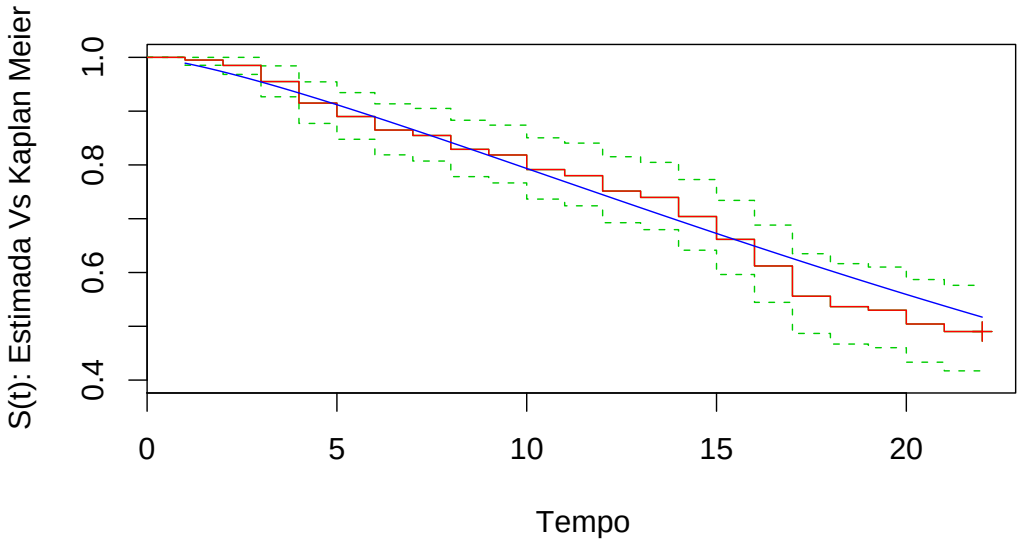


Figura 3.23: Sobrevivência dos 200 clientes reservados estimada por Kaplan-Meier e seus respectivos intervalos de confiança de 95%, e sobrevivência estimada pelo modelo paramétrico encontrado para a variável resposta tempo até o cancelamento voluntário.

As figuras 3.23 e 3.24, mostram que os modelos ajustados 3.1 e 3.2, se adequaram bem aos dados reservados para a validação, pois suas curvas estão dentro dos intervalos de confiança e muito próximas às curvas de Kaplan-Meier.

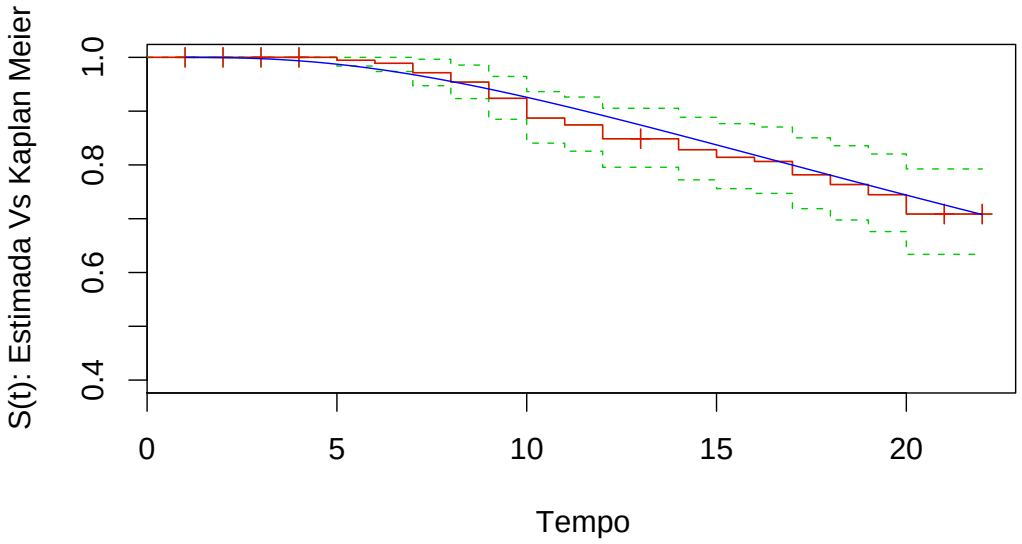


Figura 3.24: Sobrevivência dos 200 clientes reservados estimada por Kaplan-Meier e seus respectivos intervalos de confiança de 95%, e sobrevivência estimada pelo modelo paramétrico encontrado para a variável resposta tempo até o cancelamento por inadimplência.

Foram ainda retiradas quatro amostras do conjunto de dados (com 260.915 observações), sendo duas amostras provenientes da combinação das covariáveis que geram as maiores sobrevidas, uma para cada variável resposta, e duas com as menores sobrevidas, também uma para cada variável resposta. As covariáveis selecionadas para as quatro amostras foram:

- **Variável resposta tempo até o cancelamento voluntário, grupo com MAIOR sobrevida**

  Grupo = CH, Produto = P, Canal = A, Débito = S, Limite = >5000 e Credit Score = média ponderada dos coeficientes.

  Tamanho da amostra = 132 indivíduos

- **Variável resposta tempo até o cancelamento voluntário, grupo com ME-**

NOR sobrevida

Grupo = CC, Produto = S, Canal = A, Débito = N, Limite = <500 e Credit Score = 1.

Tamanho da amostra = 89 indivíduos

- **Variável resposta tempo até o cancelamento por inadimplência, grupo com MAIOR sobrevida**

  Grupo = CH, Produto = P, Canal = A, Débito = S, Limite = >5000 e Credit Score = média ponderada dos coeficientes.

  Tamanho da amostra = 132 indivíduos

  Este grupo é o mesmo grupo da variável resposta tempo até o cancelamento voluntário, com MAIOR sobrevida.

- **Variável resposta tempo até o cancelamento por inadimplência, grupo com MENOR sobrevida**

  Grupo = CN, Produto = S, Canal = I, Débito = N, Limite = média ponderada dos coeficientes e Credit Score = 3.

  Tamanho da amostra = 53 indivíduos

Os gráficos dos modelos com as covariáveis acima, comparados com as curvas de Kaplan-Meier dessas amostras são estes:

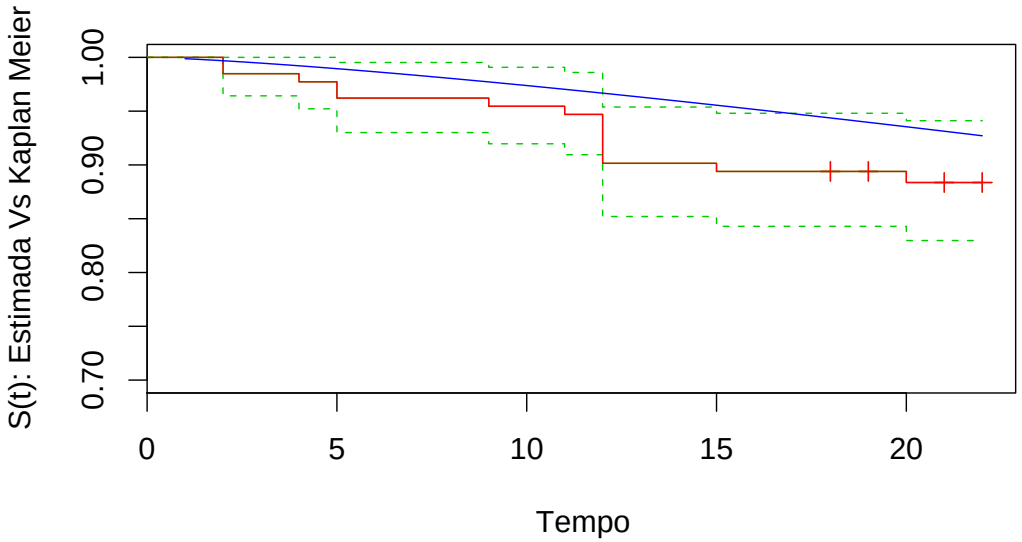


Figura 3.25: Sobrevivência do grupo cuja combinações de covariáveis gera a MAIOR sobrevida, estimada por Kaplan-Meier e seus respectivos intervalos de confiança de 95%, e estimada pelo modelo paramétrico encontrado para a variável resposta tempo até o cancelamento voluntário.

Na figura 3.25 observamos que a curva de $S(t)$ estimada pelo modelo paramétrico superestima a curva estimada por Kaplan-Meier, porém mantén-se dentro do intervalo de confiança de 95%.

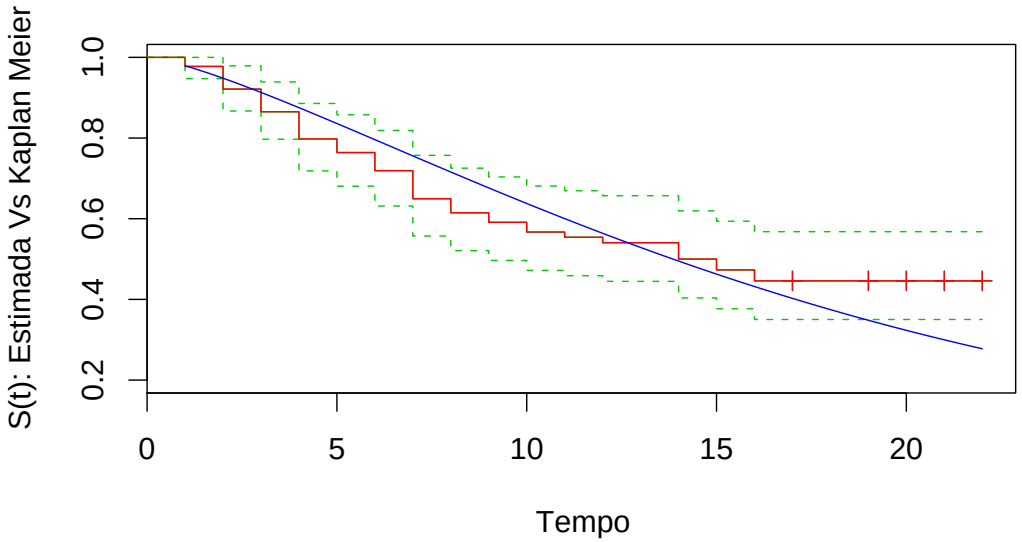


Figura 3.26: Sobrevivência do grupo cuja combinações de covariáveis gera a MENOR sobrevida, estimada por Kaplan-Meier e seus respectivos intervalos de confiança de 95%, e estimada pelo modelo paramétrico encontrado para a variável resposta tempo até o cancelamento voluntário.

Na figura 3.26 a curva de $S(t)$ estimada pelo modelo paramétrico mantém-se aproxima da curva de estimada por Kaplan-Meier.

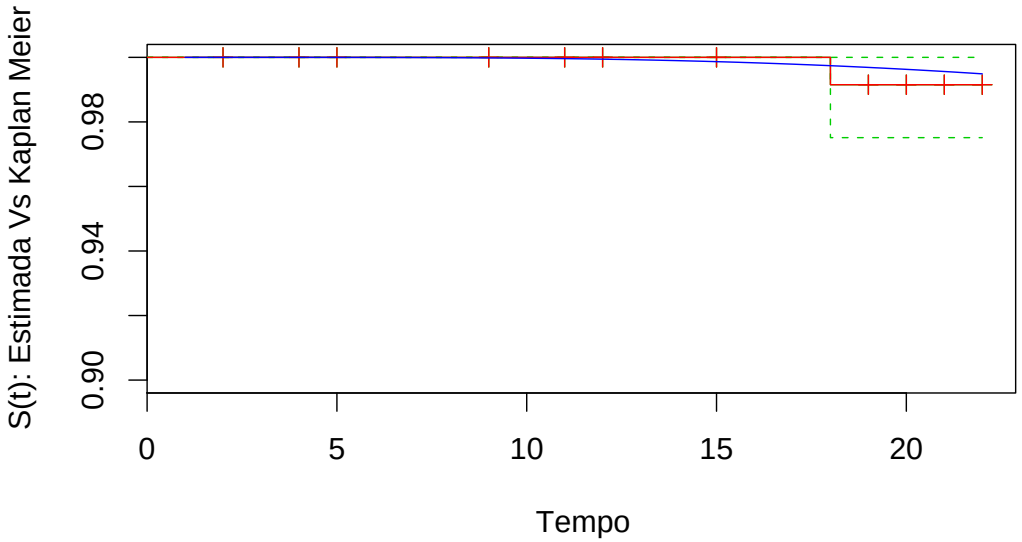


Figura 3.27: Sobrevivência do grupo cuja combinações de covariáveis gera a MAIOR sobrevida, estimada por Kaplan-Meier e seus respectivos intervalos de confiança de 95%, e estimada pelo modelo paramétrico encontrado para a variável resposta tempo até o cancelamento por inadimplência.

Na figura 3.27 podemos observar que a curva estimada pelo modelo paramétrico está muito próxima da curva de Kaplan-Meier.

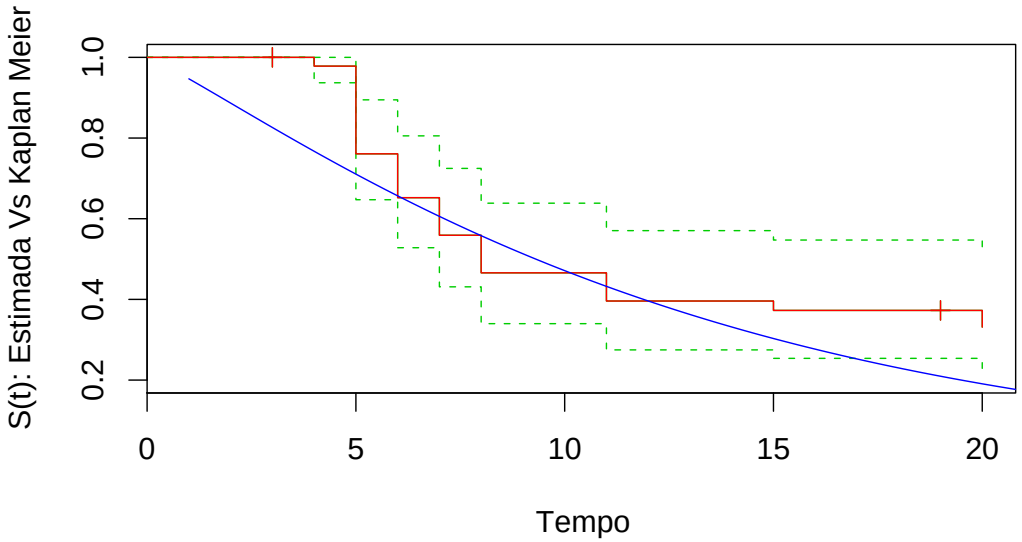


Figura 3.28: Sobrevivência do grupo cuja combinações de covariáveis gera a MENOR sobrevida, estimada por Kaplan-Meier e seus respectivos intervalos de confiança de 95%, e estimada pelo modelo paramétrico encontrado para a variável resposta tempo até o cancelamento por inadimplência.

Nesta última figura 3.28 observamos que a curva estimada pelo modelo paramétrico segue a mesma tendência de curvatura encontra na curva estimada por Kaplan-Meier.

Ressaltamos que os tamanhos dessas quatro amostras são bem pequenos comparados ao tamanho do banco de dados.

Ao fim de toda a análise de adequação dos modelos, podemos concluir que os modelos encontrados, 3.1 e 3.2, adequam-se bem aos dados deste trabalho. Agora buscaremos interpretação dos coeficientes estimados.

3.5 Interpretação dos Coeficientes Estimados

A interpretação dos coeficientes é bastante complexa, principalmente pela natureza das covariáveis, todas são categóricas e cada uma delas possui de dois a seis níveis, temos também a complexidade das combinações entre as covariáveis, que podem não representar grupos reais do estudo. Para facilitar o entendimento, com base nos coeficientes estimados pelos modelos, serão apresentados os principais níveis das covariáveis que mais influenciam positivamente ou negativamente na sobrevida dos clientes segundo cada variável resposta.

Para a variável resposta tempo até o cancelamento voluntário, são observados os seguintes fatores:

- Grupo: o nível Facility (F) apresenta a menor sobrevida. Observamos também o nível CN como sendo a maior sobrevida;
- Canal: destacando-se o nível Internet (I) para uma maior sobrevida e o nível Célula de Financiamento de automóveis (F) para a menor sobrevida;
- Produto: o nível Premium (P) apresentou uma sobrevida superior aos demais fatores, e o nível Social (S) apresentou a menor sobrevida;
- Credit Score: os níveis 5 e H se destacam com uma sobrevida superior, e o fator 1 apresenta a maior queda na sobrevida;
- Limite: o nível contendo limites inferiores a R$500,00 possui a menor sobrevida, e o nível contendo limites acima de R$5.000,00 apresenta a maior sobrevida.

Além disso, como foi apresentado anteriormente, a combinação de Grupo = CC, Produto = S, Canal = A, Débito = N, Limite = $<$500 e Credit Score = 1 representa a menor sobrevida e combinação de Grupo = CH, Produto = P, Canal = A, Débito = S, Limite = $>$5000 apresenta a maior sobrevida.

Para a variável resposta tempo até o cancelamento por inadimplência, os fatores que mais se destacam são os seguintes:

- Grupo: o nível Facility (F) apresenta a maior sobrevida e o nível CN a menor sobrevida;
- Canal: destaca-se o nível Internet (I) com a menor sobrevida e o nível Célula de Financiamento de Automóveis (F) com a maior sobrevida;
- Produto: Premium (P) com a maior sobrevida e Social (S) com a menor sobrevida;
- Credit Score: os níveis 5 e H apresentam as menores sobrevidas;
- Limite: o nível >5000 apresenta uma sobrevida superior aos demais níveis;

Assim como foi apresentado anteriormente, a combinação de Grupo = CH, Produto = P, Canal = A, Débito = S, Limite = >5000 apresenta a maior sobrevida e a combinação Grupo = CN, Produto = S, Canal = I, Débito = N e Credit Score = 3 representa o grupo com menor sobrevida.

4 Conclusões

Com base nos resultados apresentados podemos observar que existem alguns fatores que influenciam de forma significativa nas curvas de sobrevivências da variáveis respostas cancelamento voluntário e por inadimplência. Para a variável cancelamento voluntário vimos que tanto na estimação não-paramétrica como na estimação paramétrica os fatores Facility e Não-correntista são os fatores da covariável Grupo que apresentam uma sobrevida que cai rapidamente, o mesmo acontece com os fatores Auto e Solidário da covariável Produto, e Telemarketing e Célula de Financiamento de Automóveis para a covariável Canal. Sendo esses os fatores que devem ser atacados pela FINANCEIRA para um aumento na curva de sobrevida da variável cancelamento voluntário. Já para a variável cancelamento por inadimplência, o fator Internet da covariável Canal, apresenta uma sobrevida que cai muito em relação aos outros fatores.

Através das técnicas de Análise de Sobrivência chegou-se a dois modelos parámetricos que modelam as variáveis respostas, ou seja, através desses modelos colocamos as covariáveis de um certo grupo de clientes que desejam obter a linha de crédito (estudada) junto a FINANCEIRA e assim podemos avaliar como será o comportamento desse grupo nos próximos meses, se esse grupo de indivíduos com um certo perfil tem um grande probabilidade de cancelar a linha de crédito ou de se tornar um cliente inadimplente. Ressalta-se que os modelos encontrados neste trabalho devem ser utilizados em grupos de indivíduos com um certo perfil, e não somente para um indivíduo.

Uma sugestão para este estudo é realizar uma análise de rentabilidade, para que se possa saber se mesmo os clientes que apresentam uma sobrevida menor estão sendo rentáveis ou não para a FINANCEIRA. Uma outra sugestão é encontrar formas de reter os clientes que apresentam baixa sobrevida para a variável cancelamento vo-

luntário. Um dos objetivos desse estudo era testar os atuais modelos de Credit Score utilizados pela FIANCEIRA, que é o principal fator discriminante para a entrada dos clientes, vimos que o Credit Score é um ótimo preditor para a inadimplência, porém algumas regras nas políticas de concessão de crédito poderão ser revistas visando alterar a entrada de clientes combinando a informação do Credit Score com as covariáveis estudadas Grupo, Canal, Produto, Limite e Débito.

Referências Bibliográficas

[1] R. D. C. T. (2007). *R: A language and environment for statistical computing*. ISBN 3-900051-07-0, URL http://www.R-project.org, Vienna, Austria.

[2] M. T. Bustamante-Teixeira, E. Faerstein, and M. R. Latorre. Técnicas de análise de sobrevivencia. *Cad. Saúde Pública, 18(3):579-594*, mai-jun, 2002.

[3] S. R. Giolo and E. A. Colosimo. *Análise de Sobrevivência Aplicada*. Edgard Blücher, São Paulo, SP, first edition, 2006.

[4] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censured and Truncated Data*. Springer, New York, 1997.

[5] J. F. Lawless. *Statistical Models and Methods for Lifetime Data*. Jonh Wiley Sons, Inc., United States of America, 1982.

[6] S. E. Shimakura, M. S. Carvalho, V. L. Andreozzi, C. T. Codeço, and M. T. S. Barbosa. *Análise de Sobrevida: Teoria e Aplicacões em Saúde*. Editora Fiocruz, Rio de Janeiro, RJ, 2005.

[7] M. H. Tarumoto. Mini curos de análise de sobrevivência. FCT / UNESP Departamento de mat., São Paulo, SP, ago, 2003.

APÊNDICE A – Comparativo $S(t)$

A.1 Cancelamento Voluntário

TEMPOS	KAPLAN-MEIER	EXPONENCIAL	WEIBULL	LOG-NORMAL
1	0.994	0.976	0.990	0.996
2	0.965	0.953	0.975	0.982
3	0.938	0.930	0.959	0.963
4	0.911	0.908	0.941	0.940
5	0.890	0.887	0.922	0.917
6	0.873	0.866	0.902	0.892
7	0.857	0.845	0.881	0.868
8	0.843	0.825	0.860	0.844
9	0.830	0.806	0.839	0.821
10	0.817	0.787	0.817	0.798
11	0.805	0.768	0.796	0.776
12	0.785	0.750	0.774	0.755
13	0.765	0.732	0.752	0.735
14	0.741	0.715	0.731	0.715
15	0.707	0.698	0.710	0.696
16	0.673	0.682	0.688	0.678
17	0.647	0.666	0.668	0.661
18	0.627	0.650	0.647	0.644
19	0.612	0.634	0.627	0.628
20	0.602	0.619	0.607	0.613
21	0.594	0.605	0.587	0.598
22	0.586	0.591	0.568	0.584

A.2 Cancelamento por Inadimplência

TEMPOS	KAPLAN-MEIER	EXPONENCIAL	WEIBULL	LOG-NORMAL
1	1.000	0.989	0.999	0.999
2	1.000	0.980	0.997	0.999
3	1.000	0.970	0.994	0.997
4	0.998	0.960	0.990	0.994
5	0.991	0.950	0.985	0.989
6	0.980	0.941	0.978	0.982
7	0.969	0.931	0.971	0.974
8	0.954	0.922	0.963	0.965
9	0.939	0.913	0.953	0.954
10	0.923	0.904	0.943	0.942
11	0.908	0.894	0.932	0.929
12	0.894	0.886	0.920	0.915
13	0.880	0.877	0.907	0.901
14	0.868	0.868	0.894	0.887
15	0.857	0.859	0.880	0.872
16	0.847	0.850	0.865	0.857
17	0.836	0.842	0.849	0.842
18	0.825	0.833	0.833	0.827
19	0.815	0.825	0.816	0.812
20	0.804	0.817	0.799	0.797
21	0.793	0.809	0.781	0.782
22	0.782	0.800	0.763	0.767

APÊNDICE B – Modelos ajustados

B.1 Cancelamento Voluntário

Model	$-Log(L(\theta))$	AIC	N
SURV1 ~ 1	451514	903029	1
SURV1 ~ GRUPO	451121	902245	2
SURV1 ~ PROD	448802	897606	3
SURV1 ~ CANAL	451418	902838	4
SURV1 ~ LIMITE	449652	899307	5
SURV1 ~ CS	448108	896219	6
SURV1 ~ DEBT	451455	902912	7
SURV1 ~ GRUPO + PROD + CANAL + LIMITE + CS + DEBT	442726	885464	8
SURV1 ~ GRUPO + PROD	448505	897014	9
SURV1 ~ GRUPO + CANAL	451054	902113	10
SURV1 ~ GRUPO + LIMITE	449292	898588	11
SURV1 ~ GRUPO + CS	447442	894888	12
SURV1 ~ GRUPO + DEBT	451084	902173	13
SURV1 ~ PROD + CANAL	448719	897443	14
SURV1 ~ PROD + LIMITE	448524	897053	15
SURV1 ~ PROD + CS	444248	888501	16
SURV1 ~ PROD + DEBT	448778	897560	17
SURV1 ~ CANAL + LIMITE	449457	898919	18
SURV1 ~ CANAL + CS	447957	895919	19
SURV1 ~ CANAL + DEBT	451360	902725	20

Model	$-Log(L(\theta))$	AIC	N
SURV1 ~ LIMITE + CS	444338	888680	21
SURV1 ~ LIMITE + DEBT	449635	899274	22
SURV1 ~ CS + DEBT	447919	895843	23
SURV1 ~ GRUPO + PROD + CANAL	448419	896845	24
SURV1 ~ GRUPO + PROD + LIMITE	448223	896453	25
SURV1 ~ GRUPO + PROD + CS	443901	887808	26
SURV1 ~ GRUPO + PROD + DEBT	448488	896982	27
SURV1 ~ GRUPO + CANAL + LIMITE	449139	898285	28
SURV1 ~ GRUPO + CANAL + CS	447372	894750	29
SURV1 ~ GRUPO + CANAL + DEBT	451016	902039	30
SURV1 ~ GRUPO + LIMITE + DEBT	449281	898569	31
SURV1 ~ GRUPO + LIMITE + DEBT	449281	898569	32
SURV1 ~ GRUPO + CS + DEBT	447288	894583	33
SURV1 ~ PROD + CANAL + LIMITE	448445	896896	34
SURV1 ~ PROD + CANAL + CS	444066	888138	35
SURV1 ~ PROD + CANAL + DEBT	448695	897397	36
SURV1 ~ PROD + LIMITE + CS	443086	886179	37
SURV1 ~ PROD + LIMITE + DEBT	448509	897025	38
SURV1 ~ PROD + CS + DEBT	444114	888234	39
SURV1 ~ CANAL + LIMITE + CS	443921	887848	40
SURV1 ~ CANAL + LIMITE + DEBT	449445	898896	41
SURV1 ~ CANAL + CS + DEBT	447774	895554	42
SURV1 ~ LIMITE + CS + DEBT	444210	888426	43
SURV1 ~ GRUPO + PROD + CANAL + LIMITE	448126	896260	44
SURV1 ~ GRUPO + PROD + CANAL + CS	443801	887610	45
SURV1 ~ GRUPO + PROD + CANAL + DEBT	448402	896812	46
SURV1 ~ GRUPO + PROD + LIMITE + CS	442966	885940	47
SURV1 ~ GRUPO + PROD + CS + DEBT	443773	887554	48
SURV1 ~ GRUPO + PROD + LIMITE + DEBT	448212	896433	49
SURV1 ~ GRUPO + CANAL + CS + DEBT	447217	894443	50

Model	$-Log(L(\theta))$	AIC	N
SURV1 ~ GRUPO + CANAL + LIMITE + CS	443847	887703	51
SURV1 ~ GRUPO + CANAL + LIMITE + DEBT	449129	898267	52
SURV1 ~ GRUPO + LIMITE + CS + DEBT	443961	887931	53
SURV1 ~ CANAL + LIMITE + CS + DEBT	443815	887638	54
SURV1 ~ PROD + CANAL + LIMITE + CS	442925	885858	55
SURV1 ~ PROD + CANAL + LIMITE + DEBT	448429	896866	56
SURV1 ~ PROD + LIMITE + CS + DEBT	442972	885953	57
SURV1 ~ PROD + CANAL + CS + DEBT	443935	887879	58
SURV1 ~ GRUPO + PROD + CANAL + LIMITE + CS	442842	885695	59
SURV1 ~ GRUPO + PROD + CANAL + LIMITE + DEBT	448114	896238	60
SURV1 ~ GRUPO + PROD + CANAL + CS + DEBT	443673	887356	61
SURV1 ~ GRUPO + PROD + LIMITE + CS + DEBT	442850	885710	62
SURV1 ~ GRUPO + CANAL + LIMITE + CS + DEBT	443739	887488	63
SURV1 ~ PROD + CANAL + LIMITE + CS + DEBT	442813	885637	64
SURV1 ~ GRUPO * PROD + CANAL + LIMITE + CS + DEBT	442581	885174	65
SURV1 ~ GRUPO + PROD * CANAL + LIMITE + CS + DEBT	442463	884938	66
SURV1 ~ GRUPO * CANAL + PROD + LIMITE + CS + DEBT	442683	885378	67
SURV1 ~ GRUPO + CANAL + PROD * CS + LIMITE + DEBT	442588	885188	68
SURV1 ~ GRUPO * CS + CANAL + PROD + LIMITE + DEBT	442511	885035	69
SURV1 ~ GRUPO + CANAL * CS + PROD + LIMITE + DEBT	442694	885401	70

B.2 Cancelamento por Inadimplência

Model	$-Log(L(\theta))$	AIC	N
SURV2 ~ 1	216273	432547	1
SURV2 ~ GRUPO	216002	432007	2
SURV2 ~ PROD	214501	429005	3
SURV2 ~ CANAL	215445	430892	4
SURV2 ~ LIMITE	213999	428001	5
SURV2 ~ CS1	211095	422192	6
SURV2 ~ DEBT	215901	431805	7
SURV2 ~ GRUPO + PROD + CANAL + LIMITE + CS1 + DEBT	208082	416176	8
SURV2 ~ GRUPO + PROD	214195	428395	9
SURV2 ~ GRUPO + CANAL	215167	430338	10
SURV2 ~ GRUPO + LIMITE	213573	427150	11
SURV2 ~ GRUPO + CS1	210503	421010	12
SURV2 ~ GRUPO + DEBT	215661	431326	13
SURV2 ~ PROD + CANAL	213687	427379	14
SURV2 ~ PROD + LIMITE	213194	426393	15
SURV2 ~ PROD + CS1	209805	419614	16
SURV2 ~ PROD + DEBT	214088	428180	17
SURV2 ~ CANAL + LIMITE	213132	426268	18
SURV2 ~ CANAL + CS1	210085	420174	19
SURV2 ~ CANAL + DEBT	215091	430187	20

Model	$-Log(L(\theta))$	AIC	N
SURV2 ~ LIMITE + CS1	210435	420875	21
SURV2 ~ LIMITE + DEBT	213738	427481	22
SURV2 ~ CS1 + DEBT	210985	421975	23
SURV2 ~ GRUPO + PROD + CANAL	213505	427017	24
SURV2 ~ GRUPO + PROD + LIMITE	212899	425804	25
SURV2 ~ GRUPO + PROD + CS1	209238	418482	26
SURV2 ~ GRUPO + PROD + DEBT	213818	427642	27
SURV2 ~ GRUPO + CANAL + LIMITE	212981	425968	28
SURV2 ~ GRUPO + CANAL + CS1	209704	419414	29
SURV2 ~ GRUPO + CANAL + DEBT	214848	429702	30
SURV2 ~ GRUPO + LIMITE + DEBT	213321	426648	31
SURV2 ~ GRUPO + LIMITE + DEBT	213321	426648	32
SURV2 ~ GRUPO + CS1 + DEBT	210401	420808	33
SURV2 ~ PROD + CANAL + LIMITE	212390	424787	34
SURV2 ~ PROD + CANAL + CS1	208929	417865	35
SURV2 ~ PROD + CANAL + DEBT	213318	426643	36
SURV2 ~ PROD + LIMITE + CS1	209512	419030	37
SURV2 ~ PROD + LIMITE + DEBT	212852	425711	38
SURV2 ~ PROD + CS1 + DEBT	209638	419283	39
SURV2 ~ CANAL + LIMITE + CS1	209443	418893	40
SURV2 ~ CANAL + LIMITE + DEBT	212902	425811	41
SURV2 ~ CANAL + CS1 + DEBT	209979	419965	42
SURV2 ~ LIMITE + CS1 + DEBT	210336	420678	43
SURV2 ~ GRUPO + PROD + CANAL + LIMITE	212302	424612	44
SURV2 ~ GRUPO + PROD + CANAL + CS1	208603	417215	45
SURV2 ~ GRUPO + PROD + CANAL + DEBT	213154	426317	46
SURV2 ~ GRUPO + PROD + LIMITE + CS1	208802	417613	47
SURV2 ~ GRUPO + PROD + CS1 + DEBT	209099	418206	48
SURV2 ~ GRUPO + PROD + LIMITE + DEBT	212571	425150	49
SURV2 ~ GRUPO + CANAL + CS1 + DEBT	209612	419233	50

Model	$-Log(L(\theta))$	AIC	N
SURV2 ~ GRUPO + CANAL + LIMITE + CS1	208878	417765	51
SURV2 ~ GRUPO + CANAL + LIMITE + DEBT	212744	425497	52
SURV2 ~ GRUPO + LIMITE + CS1 + DEBT	209436	418881	53
SURV2 ~ CANAL + LIMITE + CS1 + DEBT	209354	418717	54
SURV2 ~ PROD + CANAL + LIMITE + CS1	208617	417242	55
SURV2 ~ PROD + CANAL + LIMITE + DEBT	212084	424176	56
SURV2 ~ PROD + LIMITE + CS1 + DEBT	209353	418714	57
SURV2 ~ PROD + CANAL + CS1 + DEBT	208785	417579	58
SURV2 ~ GRUPO + PROD + CANAL + LIMITE + CS1	208203	416416	59
SURV2 ~ GRUPO + PROD + CANAL + LIMITE + DEBT	211994	423999	60
SURV2 ~ GRUPO + PROD + CANAL + CS1 + DEBT	208478	416967	61
SURV2 ~ GRUPO + PROD + LIMITE + CS1 + DEBT	208668	417347	62
SURV2 ~ GRUPO + CANAL + LIMITE + CS1 + DEBT	208798	417607	63
SURV2 ~ PROD + CANAL + LIMITE + CS1 + DEBT	208479	416968	64
SURV2 ~ GRUPO * PROD + CANAL + LIMITE + CS1 + DEBT	207917	415846	65
SURV2 ~ GRUPO + PROD * CANAL + LIMITE + CS1 + DEBT	207924	415861	66
SURV2 ~ GRUPO * CANAL + PROD + LIMITE + CS1 + DEBT	208052	416116	67
SURV2 ~ GRUPO + CANAL + PROD * CS1 + LIMITE + DEBT	207942	415897	68
SURV2 ~ GRUPO * CS1 + CANAL + PROD + LIMITE + DEBT	207948	415909	69
SURV2 ~ GRUPO + CANAL * CS1 + PROD + LIMITE + DEBT	208013	416038	70

APÊNDICE C – Summary

C.1 Cancelamento Voluntário

```
> melhor_1 = survreg(SURV1 ~ GRUPO + PROD * CANAL + LIM_CAT +
                     CS1 + DEBT, dist= "weibull")
> summary(melhor_1)

Call:
survreg(formula = SURV1 ~ GRUPO + PROD * CANAL + LIM_CAT +
        CS1 + DEBT, dist = "weibull")

                     Value Std. Error        z         p
(Intercept)         2.7553    0.05016  54.9260  0.00e+00
GRUPOCH            -0.0348    0.00758  -4.5850  4.54e-06
GRUPOCN             0.0217    0.00682   3.1807  1.47e-03
GRUPOF             -0.2000    0.05974  -3.3485  8.13e-04
GRUPON             -0.0926    0.04167  -2.2235  2.62e-02
GRUPOU             -0.1905    0.01777 -10.7185  8.33e-27
PRODC               0.4321    0.05095   8.4808  2.24e-17
PRODG               0.6416    0.05081  12.6262  1.51e-36
PRODO               0.2508    0.04949   5.0672  4.04e-07
PRODP               1.3757    0.06339  21.7019 1.97e-104
PRODS               0.1490    0.05029   2.9639  3.04e-03
```

CANALF	0.0563	0.06490	0.8670	3.86e-01
CANALI	1.1125	0.27309	4.0738	4.63e-05
CANALO	0.5301	0.09389	5.6462	1.64e-08
CANALT	0.2195	0.06373	3.4436	5.74e-04
LIM_CAT>5000	0.5603	0.01700	32.9525	3.89e-238
LIM_CAT1000 - 2499	0.2543	0.00853	29.7999	3.92e-195
LIM_CAT2500 - 4999	0.4230	0.01373	30.8085	2.01e-208
LIM_CAT500 - 999	0.1123	0.00721	15.5781	1.02e-54
CS12	0.2241	0.00707	31.7189	8.52e-221
CS13	0.3982	0.00723	55.0508	0.00e+00
CS14	0.6124	0.00899	68.0962	0.00e+00
CS15	0.9299	0.01718	54.1416	0.00e+00
CS1H	0.9715	0.01425	68.1624	0.00e+00
DEBTS	0.1044	0.00650	16.0631	4.63e-58
PRODC:CANALF	0.0000	0.00000	NaN	NaN
PRODG:CANALF	0.0000	0.00000	NaN	NaN
PRODO:CANALF	0.0000	0.00000	NaN	NaN
PRODP:CANALF	0.0000	0.00000	NaN	NaN
PRODS:CANALF	0.1267	0.07575	1.6726	9.44e-02
PRODC:CANALI	-0.6077	0.80045	-0.7592	4.48e-01
PRODG:CANALI	-1.3214	0.28683	-4.6071	4.08e-06
PRODO:CANALI	-0.6027	0.27527	-2.1894	2.86e-02
PRODP:CANALI	-1.4436	0.80182	-1.8004	7.18e-02
PRODS:CANALI	-0.3484	0.30202	-1.1536	2.49e-01
PRODC:CANALO	-0.5271	0.09540	-5.5255	3.29e-08
PRODG:CANALO	-0.4267	0.09858	-4.3288	1.50e-05
PRODO:CANALO	-0.5052	0.09401	-5.3744	7.68e-08
PRODP:CANALO	-0.7303	0.18986	-3.8466	1.20e-04
PRODS:CANALO	-0.1743	0.09616	-1.8123	6.99e-02
PRODC:CANALT	6.5493	74.78169	0.0876	9.30e-01
PRODG:CANALT	-0.5976	0.04258	-14.0335	9.72e-45
PRODO:CANALT	0.0000	0.00000	NaN	NaN

```
PRODP:CANALT        0.0000    0.00000      NaN       NaN
PRODS:CANALT        0.0000    0.00000      NaN       NaN
Log(scale)         -0.2839    0.00296 -96.0596  0.00e+00

Scale= 0.753

Weibull distribution
Loglik(model)= -442463.1   Loglik(intercept only)= -451514.8
        Chisq= 18103.24 on 44 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 8
n= 260915
```

C.2 Cancelamento por Inadimplência

```
> melhor_2 = survreg(SURV2 ~ GRUPO * PROD + CANAL + LIM_CAT +
                      CS1 + DEBT, dist= "lognormal")
> summary(melhor_2)

Call:
survreg(formula = SURV2 ~ GRUPO * PROD + CANAL + LIM_CAT +
        CS1 + DEBT, dist = "lognormal")

                    Value Std. Error       z         p
(Intercept)         4.07600    0.06477  62.926  0.00e+00
GRUPOCH             0.09457    0.03890   2.431  1.51e-02
GRUPOCN            -0.22127    0.08086  -2.737  6.21e-03
GRUPOF              0.48810    0.10150   4.809  1.52e-06
GRUPON             -0.00908    0.06557  -0.139  8.90e-01
GRUPOU              0.02859    0.23320   0.123  9.02e-01
```

PRODC	0.21878	0.06940	3.152	1.62e-03
PRODG	0.35056	0.07362	4.762	1.92e-06
PRODO	0.05943	0.06403	0.928	3.53e-01
PRODP	0.60056	0.11428	5.255	1.48e-07
PRODS	-0.27615	0.05878	-4.698	2.63e-06
CANALF	0.13714	0.05369	2.554	1.06e-02
CANALI	-0.83572	0.03101	-26.954	5.12e-160
CANALO	-0.09268	0.00711	-13.033	7.95e-39
CANALT	-0.07789	0.09296	-0.838	4.02e-01
LIM_CAT>5000	0.39075	0.01988	19.659	4.82e-86
LIM_CAT1000 - 2499	0.19786	0.00984	20.112	5.84e-90
LIM_CAT2500 - 4999	0.22130	0.01520	14.558	5.17e-48
LIM_CAT500 - 999	0.13516	0.00824	16.398	1.99e-60
CS12	-0.39075	0.01032	-37.850	0.00e+00
CS13	-0.56428	0.00991	-56.933	0.00e+00
CS14	-0.66791	0.01051	-63.540	0.00e+00
CS15	-0.81895	0.01348	-60.768	0.00e+00
CS1H	-0.86273	0.01258	-68.571	0.00e+00
DEBTS	0.11674	0.00752	15.530	2.18e-54
GRUPOCH:PRODC	0.12501	0.05440	2.298	2.16e-02
GRUPOCN:PRODC	0.05049	0.08602	0.587	5.57e-01
GRUPOF:PRODC	-0.25971	0.28667	-0.906	3.65e-01
GRUPON:PRODC	0.11396	0.22393	0.509	6.11e-01
GRUPOU:PRODC	0.00000	0.00000	NaN	NaN
GRUPOCH:PRODG	0.22519	0.06754	3.334	8.56e-04
GRUPOCN:PRODG	-0.05518	0.08992	-0.614	5.39e-01
GRUPOF:PRODG	-0.57980	0.10008	-5.793	6.91e-09
GRUPON:PRODG	-0.56124	0.10773	-5.210	1.89e-07
GRUPOU:PRODG	0.00000	0.00000	NaN	NaN
GRUPOCH:PRODO	0.05753	0.03986	1.443	1.49e-01
GRUPOCN:PRODO	0.20083	0.08106	2.478	1.32e-02
GRUPOF:PRODO	-0.12853	0.08792	-1.462	1.44e-01

```
GRUPON:PRODO        0.14826    0.07393   2.005  4.49e-02
GRUPOU:PRODO        0.21175    0.23350   0.907  3.64e-01
GRUPOCH:PRODP       0.36302    0.16123   2.252  2.43e-02
GRUPOCN:PRODP      -0.04756    0.12827  -0.371  7.11e-01
GRUPOF:PRODP        0.00000    0.00000     NaN       NaN
GRUPON:PRODP        0.00000    0.00000     NaN       NaN
GRUPOU:PRODP        0.00000    0.00000     NaN       NaN
GRUPOCH:PRODS       0.00000    0.00000     NaN       NaN
GRUPOCN:PRODS       0.24887    0.07745   3.213  1.31e-03
GRUPOF:PRODS       -0.24656    0.22882  -1.078  2.81e-01
GRUPON:PRODS        0.00000    0.00000     NaN       NaN
GRUPOU:PRODS        0.00000    0.00000     NaN       NaN
Log(scale)         -0.13034    0.00401 -32.484 1.79e-231

Scale= 0.878

Log Normal distribution
Loglik(model)= -207917.2   Loglik(intercept only)= -216273.9
        Chisq= 16713.37 on 49 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 5
n= 260915
```