Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

RISCO DE CRÉDITO: CREDIT SCORING E APLICAÇÕES EM ANÁLISE DE SOBREVIVÊNCIA

por

Caio Martins Chagas

Brasília
2013

Caio Martins Chagas

RISCO DE CRÉDITO: CREDIT SCORING E APLICAÇÕES EM ANÁLISE DE SOBREVIVÊNCIA

Relatório apresentado à disciplina Estágio Supervisionado II do curso de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Prof. Doutor Raul Yukihiro Matsushita

Brasília
2013

Dedico este trabalho à minha família e amigos.

Caio Martins Chagas

Agradecimentos

Gostaria de agradecer ao Professor Doutor Raul Yukihiro Matsushita pela sua orientação e críticas. À Professora Doutora Juliana Betini Fachini por ser sempre prestativa e ter me ajudado muito.

Um muito obrigado também aos colegas da Diretoria de Crédito da instituição financeira que me forneceu o banco de dados e me mostrou a visão do mercado sobre o tema.

Ainda um agradecimento aos meus colegas de curso pelos dias e noites de estudo na UnB e pelos diversos momentos de alegria. E um agradecimento muito especial à minha família por terem me dado educação e condições para ter chegado aonde cheguei.

Resumo

A avaliação do risco de crédito é fundamental para qualquer instituição financeira. Para tanto, os modelos de *Credit Scoring* são as principais ferramentas de suporte à concessão de crédito.

Este trabalho apresenta uma análise de clientes de uma grande instituição financeira brasileira utilizando métodos da análise de sobrevivência. E através desse procedimento estatístico pode-se modelar o risco de crédito, entender melhor a relação da inadimplência com potenciais covariáveis e ter maior precisão na hora de decidir qual a classificação de risco de cada cliente.

Palavras-chave: Análise de sobrevivência; Risco de crédito; Modelos de Cox; Modelos discretos.

Abstract

The evaluation of the credit risk is essential for any financial institution. To this end, the models of Credit Scoring are the main tools of support to the concession of credit.

This work presents an analysis of customers of a great Brazilian financial institution using methods of the survival analysis. And through this statistical procedure we can model the credit risk, understand the relationship of default with potential covariates and get greater accuracy in deciding which risk rating of each customer.

Keywords: Survival analysis; Credit risk; Cox models; Discrete models.

SUMÁRIO

1. INTRODUÇÃO

Os modelos de *Credit Scoring* são as principais ferramentas de suporte à concessão de crédito atualmente. Através de características e comportamentos dos clientes cria-se um perfil para o pagador e são definidas quais dessas características estão, efetivamente, relacionadas com o risco de crédito e qual a intensidade e direção desse relacionamento.

A análise de sobrevivência consiste em uma coleção de procedimentos estatísticos para a análise de dados relacionados ao tempo decorrido desde um tempo inicial, pré-estabelecido, até a ocorrência de um evento de interesse. No contexto de *Credit Scoring,* o tempo relevante é medido entre o ingresso do cliente na base de usuários de um produto de crédito até a ocorrência de um evento de interesse, como por exemplo, um problema de inadimplência.

As principais características das técnicas de análise de sobrevivência são sua capacidade de extrair informações de dados censurados, ou seja, daqueles clientes para os quais, no final do acompanhamento no período do desempenho, o problema de crédito não foi observado, além de levar em consideração os tempos para a ocorrência dos eventos. De maneira geral, um tempo censurado corresponde ao tempo decorrido entre o início e o término do estudo ou acompanhamento de um indivíduo sem ser observada a ocorrência do evento de interesse para ele. Na visão de risco de crédito, dados censurados podem indicar a liquidação antecipada da linha de crédito ou a liquidação no prazo certo sem ocorrência da inadimplência.

Similar à regressão logística, é comum em dados de análise de sobrevivência a presença de covariáveis representando também a heterogeneidade da população. Assim, os modelos de regressão em análise de sobrevivência têm como objetivo identificar a relação e a influência dessas variáveis com os tempos de sobrevida, ou com alguma

função dos mesmos. Covariáveis como a utilização de cartão de crédito, estado civil, tempo de conta e idade podem ter bastante influencia no resultado do risco de crédito de cada cliente.

Stepanova e Thomas (2002) e Noh *et al*. (2005) comparam os resultados obtidos através do modelo de sobrevivência com outros métodos mais tradicionais, como a regressão logística e redes neurais, e concluíram que os modelos de sobrevivência apresentavam desempenhos muito semelhantes. Porém, ao contrário destes dois, como o modelo de sobrevivência não possui uma variável dependente binária, mas antes a duração T (tempo de sobrevivência), permite calcular a taxa de risco acumulada para qualquer tempo observado, semelhante à probabilidade de *default* do modelo de redes neurais e da regressão logística. Em 2001, Stepanova e Thomas, num estudo sobre clientes de uma instituição financeira do Reino Unido, realçaram ainda outra vantagem dos modelos de sobrevivência: a possibilidade de calcular a rentabilidade esperada de uma operação, uma vez que permite estimar, através da função sobrevivência, a probabilidade do cliente falhar (inadimplência) em determinado mês.

Este trabalho é composto por 5 capítulos. No primeiro capítulo é introduzido o tema e destacado sua relevância. No Capítulo 2 é formulada uma revisão literária e são definidos os conceitos necessários para o desenvolvimento das técnicas de análise de sobrevivência em risco de crédito. No Capítulo 3 temos a descrição dos dados. No Capítulo 4 são demonstrados os resultados da análise de sobrevivência e a interpretação do modelo de Cox para dados discretos. As conclusões e as propostas para trabalhos futuros são apresentadas no Capítulo 5.

Toda a análise, gráficos, resultados e tabelas do Capítulo 4 foram obtidos através do *software* estatístico SAS 9.3.

2. METODOLOGIA

2.1 Evento de Interesse

Em estudos de análise de sobrevivência, o primeiro passo é definir o evento de interesse e consequentemente o que seria a falha e a censura. Na ótica de risco de crédito, o evento de interesse é a inadimplência, ou seja, quando é dito que o cliente falhou, significa que o cliente se tornou inadimplente.

É informado que os dados foram censurados quando o evento de interesse não ocorre durante todo o tempo de observação dos indivíduos, portanto, neste estudo o indivíduo censurado é aquele que paga suas dívidas em dia ou que pelo menos pagou durante o tempo de observação.

2.2 Censura

A censura pode ser à esquerda, intervalar ou à direita. A primeira ocorre quando o evento de interesse já aconteceu quando o indivíduo foi observado. Intervalar é caracterizada quando não podemos identificar o tempo exato quando ocorreu o evento, mas sim o intervalo de tempo onde ocorreu. Já a última é notada quando o tempo de ocorrência do evento de interesse está à direita do tempo registrado. O nosso foco neste trabalho está na censura à direita, devido ser a mais frequente em dados financeiros e ser a censura observada em nossos dados.

A censura pode ser de três tipos: do tipo I, do tipo II e do tipo aleatório. Censura do tipo I é aquela em que o estudo será terminado após um período pré-estabelecido de tempo. Censura do tipo II é aquela em que o estudo será terminado após ter ocorrido o evento de interesse em um número pré-estabelecido de indivíduos. Um terceiro mecanismo de censura, o do tipo aleatório, acontece quando a observação do indivíduo é

interrompida por algum motivo e quando alguns indivíduos não experimentam o evento até o final do estudo.

A Figura 1.1 representa o cenário que é comumente notado em dados financeiros:

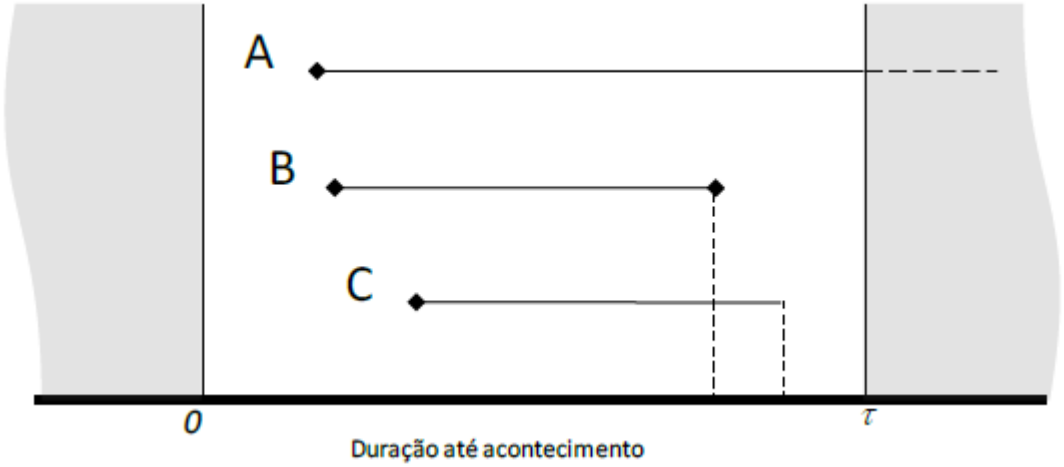


Figura 1.1. Comportamento das falhas e censuras (Alves, 2010, p.9)

Notamos dados com censura à direita nesta ilustração, onde a censura em “A” ocorre pelo fim do tempo de observação e em “C” ocorre por um motivo aleatório, que em dados financeiros pode ser a liquidação antecipada da operação de crédito ou fim das parcelas combinadas em contrato. A falha ou inadimplência é observada em “B”.

2.3 Representação dos Dados de Sobrevivência

Os dados de sobrevivência para o indivíduo $i\ (i = 1, \dots, n)$ sob estudo são representados, em geral, pelo par (t_i, δ_i) sendo t_i o tempo de falha ou de censura e δ_i a variável indicadora de falha ou censura, isto é,

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

Dessa forma, a variável aleatória resposta em análise de sobrevivência é representada por duas colunas no banco de dados.

2.4 Funções

A variável aleatória contínua e não-negativa T, que representa o tempo de falha, é geralmente especificada em análise de sobrevivência pela sua função de sobrevivência ou pela função de taxa de falha (ou risco). Também pode ser expressa pela função densidade de probabilidade e função de taxa de falha acumulada, sendo todas essas funções matematicamente equivalentes, tais que, se uma delas é especificada, as outras podem ser derivadas. Essas funções são utilizadas na prática para descrever diferentes aspectos apresentados pelo conjunto de dados.

2.4.1 Função Densidade de Probabilidade

Para uma distribuição contínua, a função densidade é definida como o limite da probabilidade de observar o evento de interesse em um indivíduo no intervalo de tempo $[t, t + \Delta t]$ por unidade de tempo, podendo ser expressa por

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

A função densidade de probabilidade é definida, também, como a derivada da função densidade de probabilidade acumulada, isto é

$$f(t) = \frac{\partial F(t)}{\partial t}.$$

2.4.2 Função de Sobrevivência

Esta é uma das principais funções probabilísticas usadas para descrever dados de tempo de sobrevivência. Tal função é definida como a probabilidade de não ser observado o evento de interesse para um indivíduo até certo tempo t, ou seja, a probabilidade de um indivíduo sobreviver ao tempo t sem o evento. Em termos probabilísticos, a função é dada por

$$S(t) = P(T \geq t) = 1 - F(t).$$

2.4.3 Função de Risco

A função de risco, ou taxa de falha, é definida como o limite da probabilidade de ser observado o evento de interesse para um indivíduo no intervalo de tempo $[t, t + \Delta t]$ dado que o mesmo tenha sobrevivido até o tempo t, e expressa por

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}.$$

Devido a sua interpretação, a função de risco é muitas vezes utilizada para descrever o comportamento dos tempos de sobrevivência. Essa função descreve como a probabilidade instantânea de falha, ou taxa de falha, se modifica com o passar do tempo, sendo conhecida também como taxa de falha instantânea, força de mortalidade e taxa de mortalidade condicional (Cox & Oakes, 1994).

2.4.4 Função de Taxa de Falha Acumulada

Outra função útil em análise de dados de sobrevivência é a função de taxa de falha acumulada. Esta função, como o próprio nome sugere, fornece a taxa de falha acumulada do indivíduo e é definida por

$$\mathrm{H}(t) = \int_0^t h(u)du = -\log S(t).$$

A função de taxa de falha acumulada não tem uma interpretação direta, mas pode ser útil na avaliação da função de maior interesse que é a taxa de falha, $h(t)$. Isto ocorre essencialmente na estimação não-paramétrica em que $\mathrm{H}(t)$ apresenta um estimador com propriedades ótimas e $h(t)$ é difícil de ser estimada (Colossimo & Giolo, 2006).

2.4.5 Relações entre as Funções

Em termos das funções definidas anteriormente, algumas relações matemáticas importantes são definidas, dentre elas:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}(\log S(t))$$

e

$$S(t) = \exp\big(-H(t)\big) = \exp\left(\int_0^t h(u)du\right).$$

Tais relações mostram que o conhecimento de uma das funções, por exemplo $S(t)$, implica no conhecimento das demais, isto é, de $F(t), f(t), H(t)$ e $h(t)$.

2.5 Modelos de regressão

Em dados de análise de sobrevivência é comum a presença de covariáveis representando a heterogeneidade da população. Assim, os modelos de regressão em análise de sobrevivência têm como objetivo identificar a relação e a influência dessas variáveis com os tempos de sobrevida, ou com alguma função dos mesmos. Desta forma, Cox (1972) propôs o seguinte modelo

$$h(t; x) = \exp(\beta' x) h_0(t),$$

em que $\beta' = (\beta_1, \beta_2, \dots, \beta_p)$ é o vetor dos parâmetros para cada uma das p covariáveis disponíveis e $h_0(t)$ é uma função não-conhecida que reflete, na área financeira, o risco básico de inadimplência inerente a cada cliente.

2.5.1 Modelo de Cox

Um dos principais objetivos ao se modelar a função de risco é determinar potenciais covariáveis que influenciam na sua forma. Outro importante objetivo é mensurar o risco individual de cada cliente. Além do interesse específico na função de risco, é de interesse estimar, para cada cliente, a função de sobrevivência.

Um modelo clássico para dados de sobrevivência, proposto por Cox (1972), é o de riscos proporcionais, também conhecido como modelo de regressão de Cox. Este modelo baseia-se na suposição de proporcionalidade dos riscos, para diferentes perfis de clientes, sem a necessidade de assumir distribuição de probabilidade para os tempos de sobrevida. Por isso, é dito um modelo semi-paramétrico.

O modelo de regressão de Cox é caracterizado pelos coeficientes β's, que medem os efeitos das covariáveis sobre a função de taxa de falha. Estas quantidades devem ser estimadas a partir das observações amostrais para que o modelo fique determinado.

Um método de estimação é necessário para se fazer inferências acerca dos parâmetros do modelo. O método de máxima verossimilhança é bastante conhecido e frequentemente utilizado para este propósito. No entanto, a presença do componente não-paramétrico $h_0(t)$ na função de verossimilhança torna este método inapropriado. Ou seja, sabe-se que:

$$L(\beta) = \prod_{i=1}^{n} [f(t_i|x_i)]^{\delta_i} \, [S(t_i|x_i)]^{1-\delta_i}$$

$$= \prod_{i=1}^{n} [h(t_i|x_i)]^{\delta_i} \, S(t_i|x_i).$$

No modelo de Cox:

$$S(t_i|x_i) = \exp\left\{-\int_0^{t_i} h_0(u) exp\,\{x_i{}'\,\beta\}\, du\right\} = [S_0(t_i)]^{\exp\{x_i'\beta\}}.$$

Assim, aplicando na fórmula da verossimilhança, temos:

$$L(\beta) = \prod_{i=1}^{n} [h_0(t_i)\exp\{x_i'\beta\}]^{\delta_i} [S_0(t_i)]^{\exp\{x_i'\beta\}},$$

que é função do componente não-paramétrico $h_0(t)$.

Uma solução razoável consiste em condicionar a construção de verossimilhança ao conhecimento da história passada de falhas e censuras para eliminar esta função de

perturbação de verossimilhança. Foi exatamente isto que Cox propôs no seu artigo original e formalizou em um artigo subsequente (Cox, 1975), denominando de método de máxima verossimilhança parcial.

2.5.2 Método de Máxima Verossimilhança Parcial

Considere que em uma amostra de n indivíduos, existam $k \leq n$ falhas distintas nos tempos $t_1 < t_2 \ldots < t_k$. Uma forma simples de entender a verossimilhança parcial considera o seguinte argumento condicional: a probabilidade condicional da i-ésima observação vir a falhar no tempo t_i conhecendo quais observações estão sob risco em t_i é:

$$P[indivíduo\ falhar\ em\ t_i \mid uma\ falha\ em\ t_i \text{ e história até } t_i] =$$

$$= \frac{P[indivíduo\ falhar\ em\ t_i \mid sobreviveu\ a\ t_i\ e\ história\ até\ t_i]}{P[uma\ falha\ em\ t_i \mid história\ até\ t_i]} =$$

$$= \frac{h_i(t \mid x_i)}{\sum_{j \in R(t_i)} h_j(t \mid x_j)} = \frac{h_0(t)\exp\{x_i'\beta\}}{\sum_{j \in R(t_i)} h_0(t)\exp\{x_j'\beta\}} =$$

$$= \frac{\exp\{x_i'\beta\}}{\sum_{j \in R(t_i)} \exp\{x_j'\beta\}}.$$

em que $R(t_i)$ é o conjunto dos índices das observações sob risco no tempo t_i. Observe que condicional à história de falhas e censuras até o tempo t_i, o componente não-paramétrico $h_0(t)$ desaparece na última equação.

A função de verossimilhança a ser utilizada para se fazer inferência acerca dos parâmetros do modelo é, então, formada pelo produto de todos os termos representados por essa última equação associados aos tempos distintos de falha, isto é,

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp\{x_i'\beta\}}{\sum_{j \in R(t_i)} \exp\{x_j'\beta\}} = \prod_{i=1}^{n} \left(\frac{\exp\{x_i'\beta\}}{\sum_{j \in R(t_i)} \exp\{x_j'\beta\}} \right)^{\delta_i},$$

em que δ_i é o indicador de falha. Os valores de β que maximizam a função de verossimilhança parcial, $L(\beta)$, são obtidos resolvendo-se o sistema de equações definido

por $U(\beta) = 0$, em que $U(\beta)$ é o vetor escore de derivadas de primeira ordem da função $l(\beta) = \log(L(\beta))$. Isto é,

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left[x_i - \frac{\sum_{j \epsilon R(t_i)} x_j \exp\{x_j' \hat{\beta}\}}{\sum_{j \in R(t_i)} \exp\{x_j' \hat{\beta}\}} \right] = 0.$$

A função de verossimilhança parcial assume que os tempos de sobrevivência são contínuos e, consequentemente, não pressupõe a possibilidade de empates nos valores observados. Na prática, empates podem ocorrer nos tempos de falha ou censura devido à escala de medida. Por exemplo, o tempo não é necessariamente registrado em horas, podendo, em alguns estudos, ser medido em dias, meses ou até mesmo em anos, dependendo da dificuldade em se obter a medida. Da mesma forma, podem ocorrer empates entre falhas e censuras. Quando ocorrer empates entre falhas e censuras, usa-se a convenção de que a censura ocorreu após a falha, o que define as observações a serem incluídas no conjunto de risco em cada tempo de falha.

A função de verossimilhança parcial deve ser modificada para incorporar as observações empatadas quando estas estão presentes. A aproximação proposta por Breslow (1972) e Peto (1972) da função de verossimilhança parcial é simples e frequentemente usada nos pacotes estatísticos comerciais, mas é adequada quando o número de observações empatadas em qualquer tempo não é grande.

Em dados de risco de crédito temos um número muito grande de empates. Isso é consequência das observações serem obtidas mensalmente, ou seja, no fim de cada mês que teremos a resposta se o cliente pagou ou não suas dívidas. Quando o número de empates em qualquer tempo é grande, o modelo de regressão de Cox para dados grupados deve ser usado (Lawless, 1982, Prentice e Gloeckler, 1978).

2.5.3 Tratamento de empates

A função de verossimilhança exata na presença de empates entre os eventos foi proposta por Kalbfleisch e Prentice (1980) e inclui todas as possíveis ordens dos eventos empatados, exigindo, consequentemente, muito esforço computacional, principalmente quando um número grande de empates é verificado em um ou mais dos tempos em que se observa a ocorrência do evento.

Em uma situação com 5 eventos, ocorrendo em um mesmo instante, existem 120 possíveis ordens a serem consideradas; para 10 eventos empatados, esse valor ficaria acima de 3 milhões (Allison, 1995). Algumas aproximações para a função de verossimilhança parcial foram desenvolvidas e trazem vantagens computacionais sobre o método exato.

Seja s_j o vetor que contém a soma de cada uma das p covariáveis para os indivíduos nos quais foram observados o evento no j-ésimo tempo, $t_{(j)}$, $j = 1,2,\dots,r$. O número de eventos no instante $t_{(j)}$ é denotado por d_j. O h-ésimo elemento de s_j é dado por $s_{hj} = \sum_{k=1}^{d_j} x_{hjk}$, em que x_{hjk} é o valor da h-ésima covariável, $h = 1,2,\dots,p$, para o k-ésimo dos d_j indivíduos, $k = 1,2,\dots,d_j$, para os quais foram observados o evento no j-ésimo tempo.

A aproximação proposta por Peto (1972) e Breslow (1972) considera a seguinte função de verossimilhança parcial

$$L_B(\beta) = \prod_{j=1}^{r} \frac{\exp(\beta' s_j)}{[\sum_{l \in R(t_{(j)})} \exp(\beta' x_l)]^{d_j}}.$$

Nesta aproximação, os d_j eventos de interesse, clientes que se tornaram inadimplentes, por exemplo, observados em $t_{(j)}$, são considerados distintos e ocorrem sequencialmente. Esta verossimilhança pode ser diretamente calculada e é adequada quando o número de observações empatadas, em qualquer tempo em que ocorrem os eventos, não é muito

grande. Por isso, esse método está normalmente implementado nos módulos de análise de sobrevivência dos *softwares* estatísticos. Farewell & Prentice (1980) mostram que os resultados dessa aproximação deterioram quando a proporção de empates aumenta em relação ao número de indivíduos sob risco, em alguns dos tempos em que os eventos são observados.

Efron (1977) propõe a seguinte aproximação para a verossimilhança parcial do modelo de riscos proporcionais

$$L_E(\beta) = \prod_{j=1}^{r} \frac{\exp(\beta' s_j)}{\prod_{k=1}^{d_j}[\sum_{l\in R(t_{(j)})} \exp(\beta' s_l) - (k-1)d_k^{-1}\sum_{l\in D(t_{(j)})} \exp(\beta' x_l)]^{d_j}},$$

em que $D(t_{(j)})$ é o conjunto de todos os clientes para os quais foram observados o evento de interesse no instante t_j. Este método fornece resultados mais próximos do exato do que o de Breslow.

Cox (1972) sugeriu a aproximação

$$L_c(\beta) = \prod_{j=1}^{r} \frac{\exp(\beta' s_j)}{\sum_{l\in R(t_{(j)};d_j)} \exp(\beta' s_l)},$$

em que $R(t_{(j)}; d_j)$ denota um conjunto de d_j indivíduos retirados do conjunto de risco no instante $t_{(j)}$. O somatório no denominador corresponde a todos os possíveis conjuntos de d_j indivíduos retirados do conjunto de risco $R(t_{(j)})$. A aproximação de Cox é baseada no modelo para a situação em que a escala de tempo é discreta, permitindo assim a presença de empates. A função de risco para um indivíduo, com vetor de covariáveis $x_i, h_i(t; x)$, é interpretada como a probabilidade de abandono em um intervalo de tempo unitário $(t, t+1)$, dado que esse indivíduo estava sob risco até o instante t, ou seja,

$$h_i(t) = P(t \le T < t+1 \,|T \ge t),$$

sendo T uma variável aleatória que representa o tempo de sobrevivência.

Quando não existem empates em um conjunto de dados de análise de sobrevivência, as aproximações são reduzidas a função de verossimilhança parcial do modelo de Cox. Entretanto, quando existem um número muito elevado de empates, os modelos de regressão discretos são necessários para se realizar uma análise precisa e confiável.

2.5.4 Modelos de Regressão Discretos

A natureza discreta dos tempos de falha deve ser explicitamente reconhecida quando existe um grande número de empates. Métodos para tratar dados discretos ou grupados são apresentados por Lawless (1982, pp. 372-390) e Collet (2003, Cap.9). A estrutura de regressão é especificada em termos da probabilidade de um indivíduo sobreviver a um certo tempo condicional a sua sobrevivência ao tempo anterior observado.

Considere que os tempos de vida são grupados em k intervalos denotados por $I_i = [a_{i-1}, a_i)$, $i = 1, \dots, k$, em que $0 = a_0 < a_1 < \cdots < a_k = \infty$, e assuma que as censuras ocorrem no final do intervalo. Seja δ_{li} uma variável indicadora para o tempo de vida do l-ésimo indivíduo no I_i-ésimo intervalo de tempo, ($\delta_{li} = 0$, se for censurado e $\delta_{li} = 1$, caso contrário). A função de verossimilhança é frequentemente escrita em termos da probabilidade de morte (falha) do l-ésimo indivíduo em I_i, dado que ele estava vivo em a_{i-1} e os valores das covariáveis x_l, ou seja,

$$p_i(x_l) = P[T_l < a_i \mid T_l \geq a_{i-1}, x_l].$$

Então, a função de verossimilhança pode ser obtida considerando as covariáveis x_l tal que: a contribuição de uma observação não-censurada (em I_i) para a função de verossimilhança é:

$$[\{1 - p_1(x_l)\} \dots \{1 - p_{i-1}(x_l)\}]p_i(x_l)$$

e a contribuição de uma observação censurada (em a_i) para a função de verossimilhança é:

$$[\{1 - p_1(x_l)\} \dots \{1 - p_i(x_l)\}].$$

A função de verossimilhança é, então, dada por:

$$L_D(\beta) = \prod_{i=1}^{k} \prod_{l \in R_i} \{p_i(x_l)\}^{\delta_{li}} \{1 - p_i(x_l)\}^{(1-\delta_{li})}.$$

Essa função corresponde à função de verossimilhança de uma variável aleatória com uma distribuição de Bernoulli, cuja variável resposta é δ_{li} e a probabilidade de sucesso é $p_i(x_l)$. A estrutura de regressão representada pela probabilidade $p_i(x_l)$ pode ser modelada por meio de um modela de riscos proporcionais ou de chances proporcionais (Collett, 1991).

Assumindo o modelo de riscos proporcionais de Cox para o tempo de vida T, a função de sobrevivência tem a seguinte forma:

$$S(t \mid x_l) = \exp\left(-\int_0^t h(u \mid x)\, du\right) = [S_0(t)]^{\exp(x_l'\beta)},$$

em que $S_0(t)$ é a função de sobrevivência básica. Então, $p_i(x_l)$ assume a seguinte forma:

$$p_i(x_l) = 1 - \gamma_i^{\exp(x_l'\beta)},$$

em que $\gamma_i = S_0(a_i)/\, S_0(a_{i-1})$, para $i = 1, \dots, k$.

A equação de $p_i(x_l)$ pode ser linearizada utilizando-se uma transformação complemento log-log. Isto é,

$$\log[-log\{1 - p_i(x_l)\}] = \gamma_i^* + x_l'\beta = \eta_{li},$$

em que $\gamma_i^* = \log(-log\gamma_i)$ é o efeito do intervalo e η_{li}, para $i = 1, \dots, k$ e $l = 1, \dots, n$, é o preditor linear.

A função de verossimilhança para este modelo é obtida substituindo-se o resultado de $p_i(x_l)$ na função $L_D(\beta)$ e, então, o logaritmo da função de verossimilhança, $\mathcal{L}(\beta, \gamma)$, pode ser escrito como:

$$\mathcal{L}(\beta,\gamma) = \sum_{i=1}^{k}\sum_{l\,\epsilon\,R_i}\left[\delta_{li}\log\left(1-\gamma_i^{\exp(x_l'\beta)}\right)+(1-\ \delta_{li})\log\left(\gamma_i^{exp(x_l'\beta)}\right)\right].$$

Prentice e Gloeckler (1978) sugeriram o uso da seguinte reparametrização $\gamma_i^* = \log(-log\gamma_i)$, que torna os γ_i^*s irrestritos e a convergência do processo iterativo de estimação dos parâmetros mais rápida.

A expressão reparametrizada é dada por:

$$\mathcal{L} = \sum_{i=1}^{k}\sum_{l\,\epsilon\,R_i}[-(1-\delta_{li})\exp(\gamma_i^* + x_l'\beta) + \delta_{li}\log(1-\exp\{-\exp\{\gamma_i^* + x_l'\beta\}\})].$$

2.5.5 Intervalos de Confiança

Com as estimativas dos parâmetros e os respectivos erros-padrão, $EP(\hat{\beta})$, os intervalos de confiança são construídos dos elementos do vetor de parâmetros β.

Um intervalo de $100(1-\alpha)\%$ de confiança para um determinado parâmetro β_j é obtido fazendo $\widehat{\beta_j} \pm Z_{\alpha/2}EP(\widehat{\beta_j})$, em que $\widehat{\beta_j}$ é o valor da estimativa de máxima verossimilhança do j-ésimo parâmetro e $Z_{\alpha/2}$ o percentil superior $\alpha/2$ de uma distribuição normal padrão.

Se um intervalo $100(1-\alpha)\%$ para β_j não incluir o valor zero, pode ser dito que há evidencias de que o valor real de β_j é estatisticamente diferente de zero. A hipótese nula $H_0: \beta_j = 0$ pode ser testada calculando o valor da estatística $\widehat{\beta_j}/EP(\widehat{\beta_j})$. Esta estatística tem, assintoticamente, distribuição normal padrão.

Geralmente, as estimativas individuais $\widehat{\beta_1}, \widehat{\beta_2}, \widehat{\beta_3}, \ldots, \widehat{\beta_p}$, em um modelo de riscos proporcionais não são todas independentes entre si. Isso significa que testar hipóteses separadamente pode não ser facilmente interpretável.

2.5.6 Estimação da Função de Risco e Sobrevivência

Uma vez ajustado o modelo, a função de risco e a correspondente função de sobrevivência podem, se necessário, ser estimadas.

Suponha que o escore de risco de um modelo de riscos proporcionais contém p covariáveis $x_1, x_2, \dots, x_p$ com as respectivas estimativas para seus coeficientes $\widehat{\beta_1}, \widehat{\beta_2}, \widehat{\beta_3}, \dots, \widehat{\beta_p}$. A função de risco para o i-ésimo indivíduo no estudo é dada por

$$\widehat{h_i}(t) = exp\{\widehat{\beta}' x_i\}\widehat{h_0}(t),$$

em que x_i é o vetor de valores observados das p covariáveis para o i-ésimo indivíduo, $i = 1,2, \dots, n$ e $\widehat{h_0}(t)$ é a estimativa para a função de risco básica. Por meio dessa equação, a função de risco pode ser estimada para um indivíduo, após a função de risco básica ter sido estimada.

Em um problema de *Credit Scoring,* a utilização do escore de risco do modelo de Cox como escore final é uma opção bastante viável de ser utilizada, uma vez que a partir desses valores uma ordenação dos clientes pode ser obtida com relação ao risco de crédito.

Uma estimativa da função de risco básica foi proposta por Kalbfleisch e Prentice (1973) utilizando uma metodologia baseada no método de máxima verossimilhança. Suponha que foram observados r tempos de sobrevida distintos dos clientes que se tornaram inadimplentes, os quais, ordenados, são denotados $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$, existindo d_j eventos e n_j clientes sob risco no instante $t_{(j)}$. A estimativa da função de risco básica no tempo $t_{(j)}$ é dada por

$$\widehat{h_0}(t_j) = 1 - \widehat{\xi_j},$$

sendo $\widehat{\xi_j}$ a solução da equação

$$\sum_{l \epsilon D(t_{(j)})} \frac{\exp(\widehat{\beta}' x_l)}{1 - \widehat{\xi_j}^{\exp(\widehat{\beta}' x_l)}} = \sum_{l \epsilon R(t_{(j)})} \exp(\widehat{\beta}' x_l),$$

para $j = 1,2, \dots, r$, sendo $D\left(t_{(j)}\right)$ o conjunto de todos os d_j indivíduos que em um problema de *Credit Scoring,* por exemplo, se tornaram inadimplentes, no j-ésimo tempo, $t_{(j)}$, e $R\left(t_{(j)}\right)$ representando os n_j indivíduos sob risco no mesmo instante $t_{(j)}$.

Quando o evento é observado para mais de um cliente em um mesmo instante de tempo, ou seja, $d_j > 1$ para algum j, o somatório do lado esquerdo da equação acima compreende a soma de uma série de frações na qual $\widehat{\xi_j}$ esta no denominador elevado a diferentes potências. Assim, a equação não pode ser solucionada explicitamente, e métodos iterativos são necessários.

A suposição de que o risco de ocorrência de eventos entre dois tempos consecutivos é constante, permite considerar $\widehat{\xi_j}$ como uma estimativa da probabilidade de que não seja observado o evento de interesse no intervalo $t_{(j)}$ e $t_{(j+1)}$. A função de sobrevivência básica pode ser estimada por

$$\widehat{S_0}(t) = \prod_{j=1}^{k} \widehat{\xi_j},$$

para $t_{(k)} \leq t < t_{(k+1)}$, $k = 1, 2, \dots, r-1$. Assim, a função de sobrevivência para o i-ésimo indivíduo é dada por

$$\widehat{S_i}(t) = \left[\widehat{S_0}(t)\right]^{\exp\left(\widehat{\beta}' x_i\right)},$$

para $t_{(k)} \leq t < t_{(k+1)}$, $k = 1, 2, \dots, r-1$.

3. DADOS

O banco de dados foi obtido de um grande banco brasileiro e possui informações de 10 mil clientes, sendo 5 mil considerados *bons* pagadores e 5 mil *maus* no primeiro ano de observação. Lewis (1994) sugere que, em geral, amostras com tamanhos menores de 1500 clientes *bons* e 1500 *maus*, podem inviabilizar a construção de modelos com capacidade preditiva aceitável para um modelo de *Credit Scoring,* além de não permitir sua divisão. Normalmente, em grande parte das aplicações de modelagem com variável resposta binária, um desbalanceamento significativo, muitas vezes de ordem de 20 *bons* para 1 *mau*, é observado entre o número de *bons* e *maus* pagadores nas bases de clientes das instituições. Essa situação pode prejudicar o desenvolvimento do modelo, uma vez que o número de *maus* pode ser muito pequeno e insuficiente para estabelecer perfis com relação às variáveis explanatórias e também para observar possíveis diferenças em relação aos *bons* clientes. Thomas *et al.* (2002) sugere que as amostras em um modelo de *Credit Scoring* tendem a estar em uma proporção 1:1, de *bons* e *maus* clientes, ou algo em torno desse valor.

Os clientes foram observados de Junho de 2008 a Dezembro de 2010 (30 meses) e foi verificado se entraram em inadimplência ou não. Nas instituições financeiras, em geral, é considerado inadimplente aquele cliente que tem no final do mês alguma operação em atraso superior a 90 dias, que é o critério adotado para o banco de dados. Todos os produtos destes clientes foram considerados (cheque especial, cartão de crédito e CDC em geral), independentemente da data que tenham sido contratados. São todos clientes não novos e, para controle, clientes que não tinham nenhum produto a serem acompanhados na data-base (t0) e após 6 meses (t0+6), foram excluídos da modelagem. A data-base ou data da análise é 31/05/2008. Logo, todos os clientes foram observados na mesma data de início.

Um fator importante a ser considerado na construção do modelo é o horizonte de previsão. Este será o intervalo para o qual o modelo permitirá fazer as previsões de quais indivíduos serão mais ou menos prováveis de se tornarem inadimplentes ou de serem menos rentáveis. A regra é de 12 a 18 meses, porém na prática observamos que um intervalo de 12 meses é o mais utilizado.

Thomas *et al.* (2002) também propõem um período de 12 meses para modelos de *Credit Scoring*, sugerindo que a taxa de inadimplência dos clientes das empresas financeiras em função do tempo aumenta no início, estabilizando somente após 12 meses. Assim, qualquer horizonte mais breve do que esse pode não refletir de uma forma real o percentual de *maus* clientes prejudicando uma possível associação entre as características dos indivíduos e o evento de interesse modelado, no caso, a inadimplência. Por outro lado, a escolha de um intervalo de tempo muito longo para o horizonte de previsão também pode não trazer benefícios, fazendo com que a eficácia do modelo diminua, uma vez que, pela distância temporal, os eventos se tornam pouco correlacionados com potenciais variáveis cadastrais.

São especificadas 17 covariáveis para cada cliente que podem ser potenciais influenciadoras à inadimplência. São elas: idade, estado civil, tempo de conta, nível de instrução, renda líquida, dentre outras (tabela completa no Apêndice).

4. RESULTADOS

4.1 Tempos de sobrevivência

A Tabela 4.1 dá um grande indicativo de qual a modelagem adequada para a análise da amostra. Nota-se grande número de empates e empates em todos os tempos.

Tabela 4.1. Distribuição dos tempos de sobrevivência

Tempos	Frequência	Porcentagem	Frequência acumulada	Porcentagem acumulada
1	457	4.57	457	4.57
2	493	4.93	950	9.50
3	413	4.13	1363	13.63
4	495	4.95	1858	18.58
5	473	4.73	2331	23.31
6	413	4.13	2744	27.44
7	457	4.57	3201	32.01
8	319	3.19	3520	35.20
9	331	3.31	3851	38.51
10	318	3.18	4169	41.69
11	408	4.08	4577	45.77
12	423	4.23	5000	50.00
13	49	0.49	5049	50.49
14	34	0.34	5083	50.83
15	34	0.34	5117	51.17
16	36	0.36	5153	51.53
17	30	0.30	5183	51.83
18	44	0.44	5227	52.27
19	34	0.34	5261	52.61
20	27	0.27	5288	52.88
21	25	0.25	5313	53.13
22	35	0.35	5348	53.48
23	33	0.33	5381	53.81
24	37	0.37	5418	54.18
25	24	0.24	5442	54.42
26	43	0.43	5485	54.85
27	20	0.20	5505	55.05
28	24	0.24	5529	55.29
29	18	0.18	5547	55.47
30	26	0.26	5573	55.73
31	4427	44.27	10000	100.00

4.2 Distribuição dos tempos de inadimplência

A Figura 4.1 mostra o comportamento da percentagem de clientes que se tornam inadimplentes durante os meses de observação. Nota-se um aumento no inicio do acompanhamento até o prazo de 11 a 13 meses e depois a inadimplência se estabiliza.

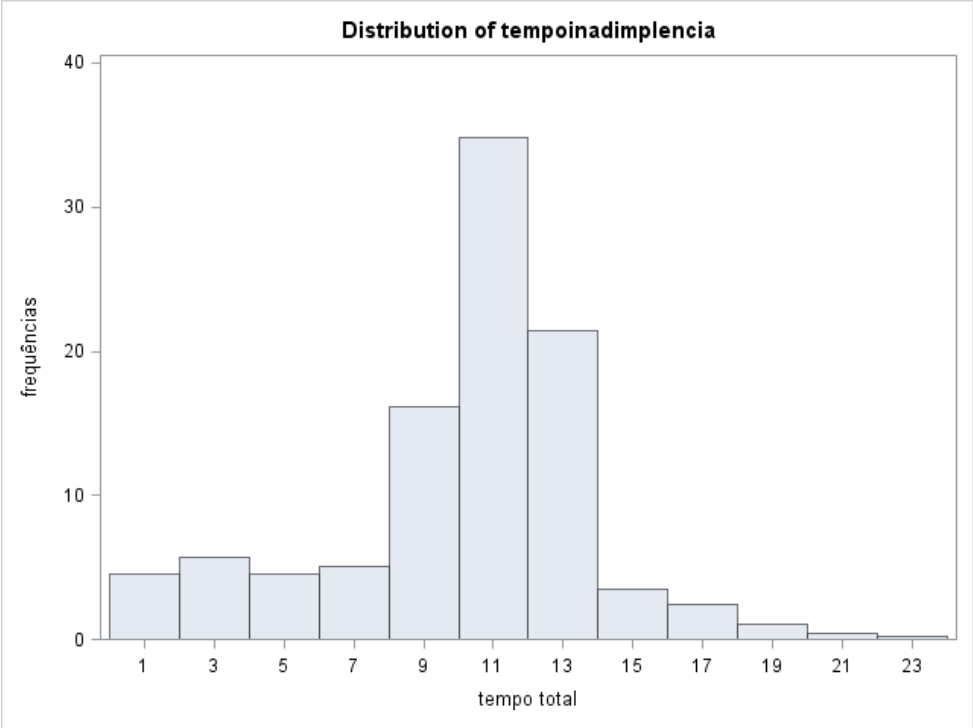


Figura 4.1. Distribuição dos tempos de inadimplência

4.3 Curvas de risco e sobrevivência

Em estudos de análise de sobrevivência sobre alguma amostra, o ponto de partida é a observação exploratória das curvas de sobrevivência e de risco. Para estimar as funções de sobrevivência será utilizado o método de Kaplan-Meier (KM) que é o mais utilizados em *softwares* estatísticos e é o *default* do SAS 9.3. O estimador de KM pode ser definido como

$$\hat{S}(t) = \prod_{j:\, t_j < t} \left(\frac{n_j - d_j}{n_j} \right),$$

em que d_j é o número de inadimplentes no tempo t_j e n_j é o número de indivíduos sob risco em t_j, ou seja, neste caso, os indivíduos que não se tornaram inadimplentes e não foram censurados até o instante imediatamente anterior a t_j. O estimador de KM é estimador de máxima verossimilhança de $S(t)$ e não-viciado para amostras grandes.

Já as funções de risco são obtidas pelo estimador de Nelso-Aalen (NA). O estimador de risco acumulado de NA, definido até o maior tempo observado no estudo, é

$$\tilde{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i},$$

e sua variância estimada é

$$\hat{V}\left(\tilde{H}(t)\right) = \sum_{t_i \leq t} \frac{d_i}{n_i^2},$$

onde d_i é o número de inadimplentes no tempo t_i e n_i é o número de indivíduos sob risco em t_i. O *default* do SAS 9.3 utiliza a função de risco suavizada de Kernel (*Kernel-Smoothed Hazard)*, baseada no estimador de NA para estimar a função de risco. Considere os pulos de $\tilde{H}(t)$ e $\hat{V}\left(\tilde{H}(t)\right)$ nos eventos com tempo $t_1 < t_2 < \cdots < t_D$ do seguinte modo:

$$\Delta\tilde{H}(t_i) = \tilde{H}(t_i) - \tilde{H}(t_{i-1}),$$

$$\Delta\hat{V}\left(\tilde{H}(t_i)\right) = \hat{V}\left(\tilde{H}(t_i)\right) - \hat{V}\left(\tilde{H}(t_{i-1})\right),$$

onde $t_0 = 0$.

O estimador suavizado de Kernel (ESK) de $h(t)$ é uma média ponderada dos $\Delta\tilde{H}(t)$ sobre os tempos que estão dentro de uma janela ou banda (*bandwidth)* de largura *b.* As ponderações ou pesos são controlados pela função de Kernel, $K(\,)$, escolhida. Foi utilizada a função *Epanechnikov kernel* (K_E), definida no intervalo [-1,1]:

$$K_E(x) = \frac{3}{4}(1 - x^2), \qquad -1 \leq x \leq 1.$$

O ESK da função de risco é definido para todos os tempos de $(0, t_D)$. A função de risco estimada é dada por

$$\hat{h}(t) = \frac{1}{b} \sum_{i=1}^{D} K\left(\frac{t - t_i}{b}\right) \Delta\tilde{H}(t_i),$$

e a variância de $\hat{h}(t)$ é estimada por

$$\widehat{\sigma^2}\left(\hat{h}(t)\right) = \frac{1}{b^2} \sum_{i=1}^{D} K\left(\frac{t - t_i}{b}\right)^2 \Delta\hat{V}\left(\tilde{H}(t_i)\right).$$

Os cálculos em detalhe são demonstrados em Gasser e Müller (1979).

Quando há mais de uma curva de sobrevivência no gráfico $\hat{S}(t)$ se faz necessário saber se elas são significativamente iguais ou não para identificar se aquela estratificação ou agrupamento é interessante. Uma forma de avaliar se as curvas de sobrevivência diferem significativamente entre cada grupo é analisando os respectivos intervalos de confiança e verificar se eles se sobrepõem ou não. Também, com recurso a testes estatísticos, é possível avaliar diferenças entre as curvas de sobrevivência. O SAS informa duas estatísticas alternativas para testar a hipótese nula de que a curva de sobrevivência é a mesma para todos os *g* grupos: o teste de log-rank e o teste de Wilcoxon. Um terceiro teste, a estatística da razão de verossimilhança (-2Log(LR)), é calculada sob a suposição adicional de que os tempos dos eventos de interesse têm uma distribuição exponencial. O teste de log-rank é o mais indicado para o estudo por assumir riscos proporcionais, que é a suposição do modelo de regressão de Cox que será utilizado para a modelagem. A estatística do teste segue uma distribuição χ^2 com *g-1* graus de liberdade (*g = grupos ou estratos)* e todos os resultados assumiram um nível de significância de 5%.

4.3.1 Curva Geral

A representa gráfica de $\hat{S}(t)$ e $\hat{h}(t)$, respectivamente, estão representadas na Figura 4.2.

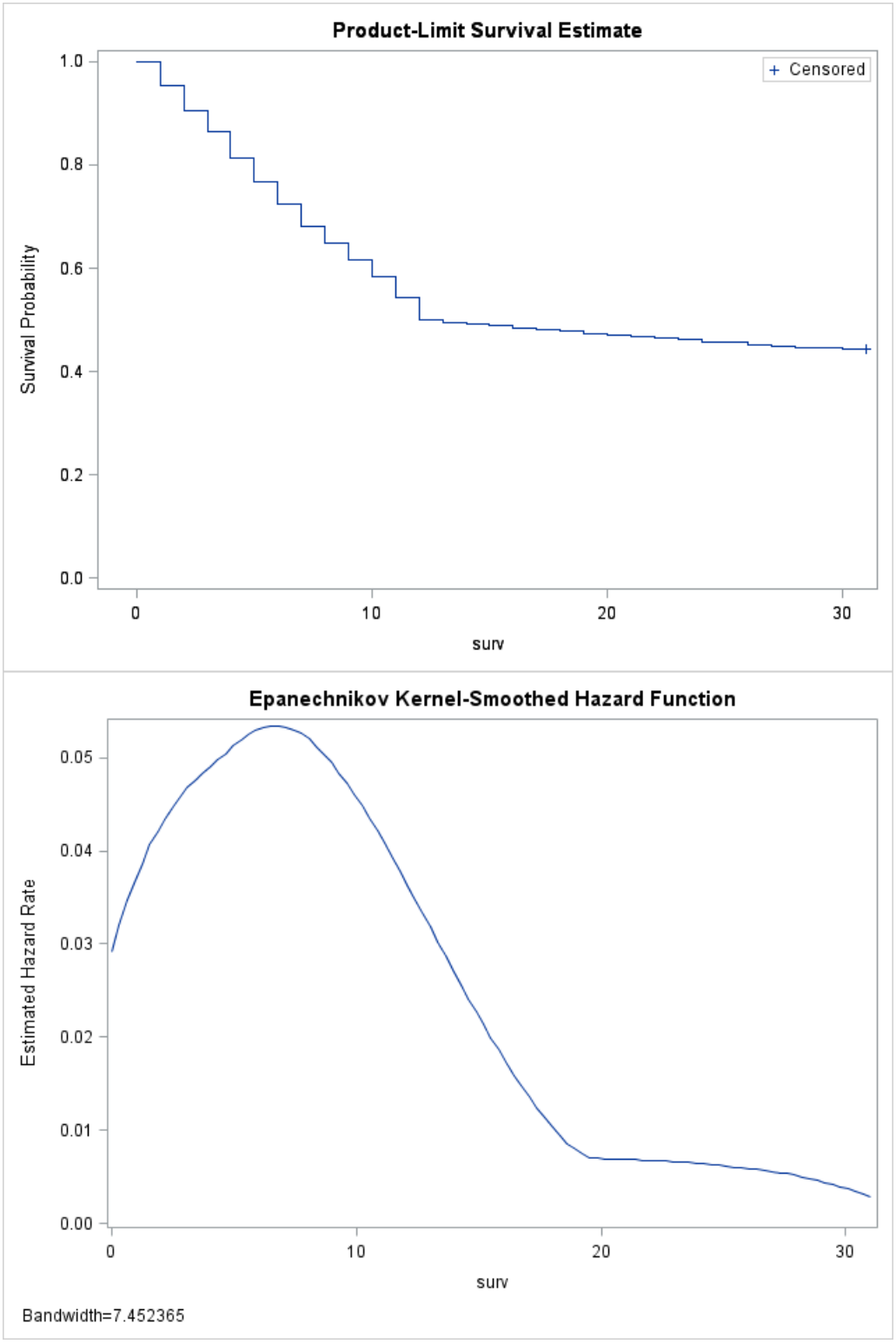


Figura 4.2. Curvas de sobrevivência e de risco gerais

Tabela 4.2. Função de sobrevivência estimada

Tempos	Indicador Censura	Sobrevivência Estimada	Intervalo de Confiança	
			LI	LS
0	.	1,00000	1,00000	1,00000
1	0	0,95430	0,95002	0,95822
2	0	0,90500	0,89909	0,91059
3	0	0,86370	0,85682	0,87027
4	0	0,81420	0,80644	0,82169
...				
30	0	0,44270	0,43294	0,45241
31	1	.	.	.

Pelo gráfico de sobrevivência nota-se o decaimento da curva até o mês 12 e depois a estabilização dos inadimplentes, que é confirmado pelo gráfico de risco que sofre uma queda drástica após 12 meses. Por ser uma amostra controlada, com clientes que possuem todos os produtos de interesse e que permaneceram no banco durante todo o tempo de observação, a censura só é observada no tempo 31 que é a data fim da análise, demonstrada pelo traço no final da curva de sobrevivência. 44,27% dos indivíduos foram censurados, ou seja, foram adimplentes com suas dívidas durante o período de 30 meses. O *output* do SAS nos apresenta uma mediana de 12,5 com intervalo de confiança de 12 a 15 meses, ou seja, 50% dos clientes ficaram inadimplentes depois de 12 meses e meio, mas esta probabilidade não é real devido a amostra ser balanceada entre bons e maus clientes. A Tabela 4.2 é gerada e informa a função de sobrevivência estimada para cada tempo e seu respectivo intervalo de confiança. Por exemplo, a probabilidade do cliente ser bom pagador no quarto mês de observação é de 81,42% com um intervalo de confiança de 80,64% a 82,16%. Essa probabilidade também não é muito informativa devido o tempo inicial (t0) não ser de entrada do cliente em uma linha de crédito específica, ou seja, são analisados vários produtos ao mesmo tempo com diferentes datas de contratação.

4.3.2 Curva por Idade

Na Figura 4.3 temos os gráficos $\hat{S}(t)$ e $\hat{h}(t)$ por cada faixa etária: menores de 20 anos, 20 a 25, 25 a 35, 35 a 45 e maiores de 45.

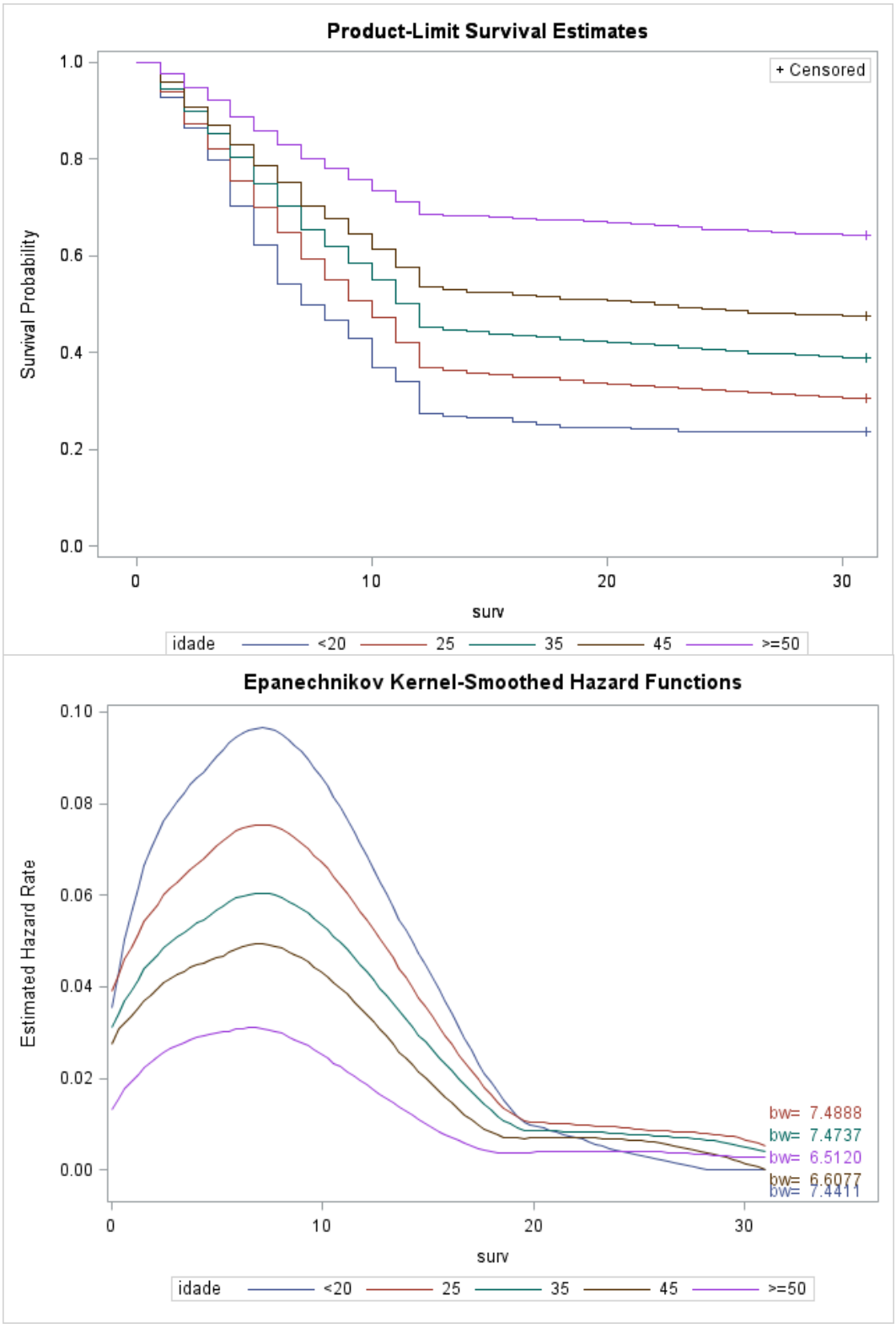


Figura 4.3. Curvas de sobrevivência e de risco por idade

Tabela 4.3. Distribuição de clientes por idade

Somatório de Falhas e Censuras				
IDADE	Total	Falhas	Censuras	Censura%
<20	223	170	53	23.77
25	2895	2010	885	30.57
35	2462	1505	957	38.87
45	1845	966	879	47.64
>=50	2575	922	1653	64.19
TOTAL	10000	5573	4427	44.27

Nitidamente a probabilidade de sobrevivência de pessoas com mais idade é maior que as demais, indicando que quanto maior a idade, menor a probabilidade de inadimplência. Os jovens menores de 20 anos possuem uma curva de risco mais elevada que os de 25, que têm curva mais elevada que os de 35 e assim por diante. Essas curvas aumentam até o mês 10 e depois descem drasticamente até próximo do mês 20, onde se estabilizam e tomam valores próximos. A Tabela 4.3 mostra como se distribuíram as falhas e censuras por cada faixa etária e novamente é exposto que com o aumento da idade, aumenta-se a porcentagem de censura, ou seja, a porcentagem de bons pagadores.

A tabela 4.4 mostra que todos os testes de igualdade rejeitaram a hipótese nula devido o p-valor ser menor que 0.05, ou seja, há diferença significativa entre as curvas de sobrevivência por faixa etária.

Tabela 4.4 Teste de igualdade por idade

Teste de Igualdade			
Teste	Chi-quadrado	GL	p-valor
Log-Rank	698.1726	4	<.0001
Wilcoxon	632.4062	4	<.0001
-2Log(LR)	978.7487	4	<.0001

4.3.3 Curva por Nível de Instrução

A Figura 4.4 demonstra os gráficos $\hat{S}(t)$ e $\hat{h}(t)$ por nível de instrução do cliente.

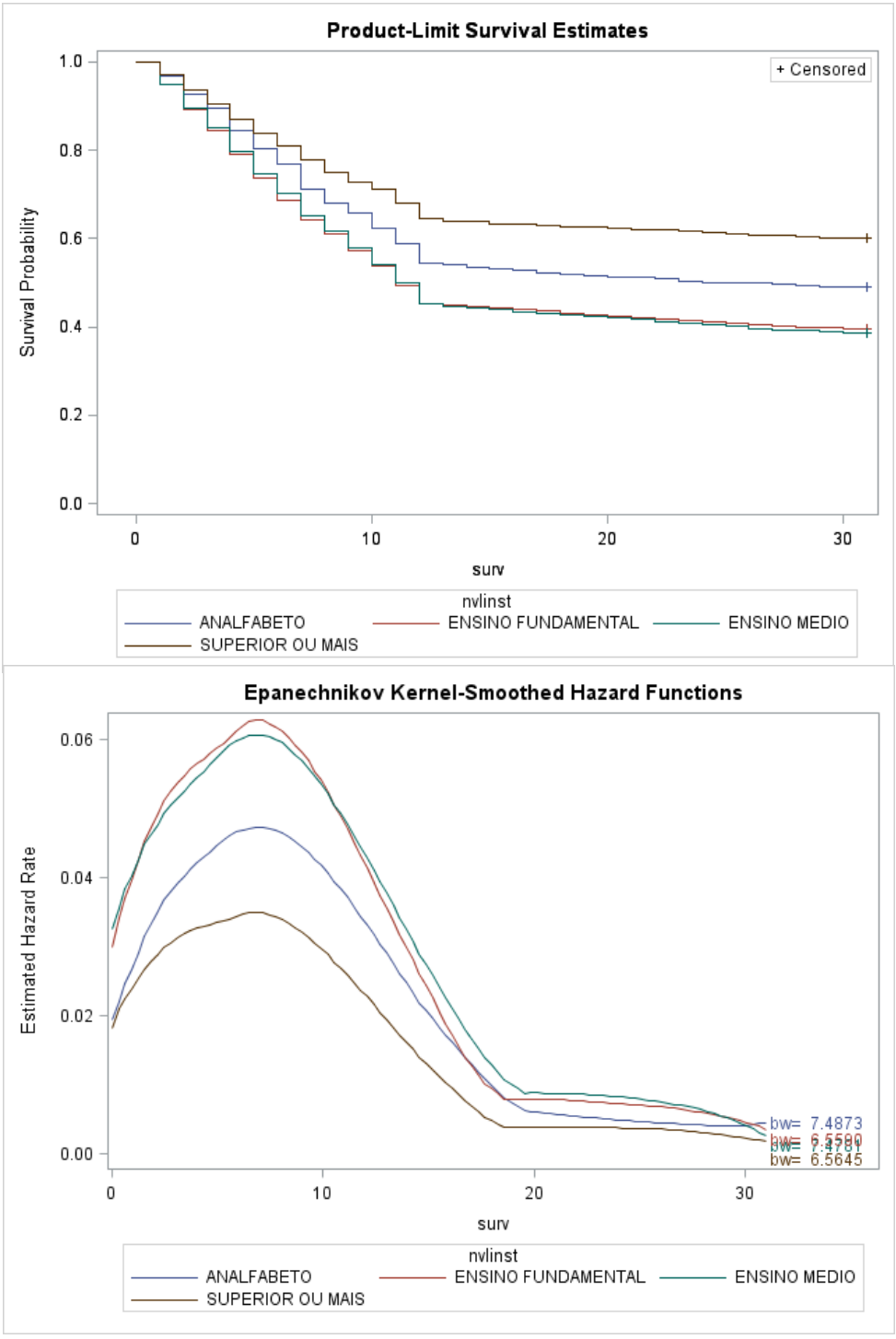


Figura 4.4. Curvas de sobrevivência e de risco por nível de instrução

Tabela 4.5. Distribuição de clientes por nível de instrução

Somatório de Falhas e Censuras				
NÍVEL DE INTRUÇÃO	Total	Falhas	Censuras	Censura%
Analfabeto	152	76	76	50.00
Ensino Fundamental	3156	1907	1249	39.58
Ensino Médio	3986	2440	1546	38.79
Superior ou mais	2055	820	1235	60.10
TOTAL	9349	5243	4106	43.92

Quando se trata de nível de instrução, se espera que quanto maior for a instrução da pessoa, maior educação financeira ela terá e, consequentemente, menos propícia a inadimplência ela será. Mas de acordo com a amostra, a probabilidade de sobrevivência de pessoas analfabetas é maior que de pessoas no ensino fundamental e médio. Os gráficos demonstram que pessoas com ensino superior têm menos risco de serem inadimplentes. Logo depois temos as analfabetas. E as pessoas com maior risco de se tornarem inadimplentes são aquelas com ensino fundamental e médio, tendo curvas de sobrevivência e risco bem parecidas. A tabela 4.5 informa a distribuição de falhas e censuras por nível de instrução. 651 observações com tempos de falha, censura ou estrato inválidos foram deletadas.

Esse comportamento dos clientes analfabetos não é esperado mas pode ser explicado pelo perfil do indivíduo. Pessoas analfabetas têm idade mais elevada e renda muito baixa (Na amostra, 71,7% maiores de 60 anos e 76,9% com renda menor que 500 reais mensais) . Têm medo de tomar empréstimos, principalmente por serem de uma geração que não tinha contato com bancos na juventude e terem orgulho de pagar suas contas em dia, ter o “nome limpo”.

Novamente a hipótese de igualdade entre as funções é rejeitada pelos três testes, indicando que há diferença significativa entre as curvas, como podemos notar na Tabela 4.6.

Tabela 4.6. Teste de igualdade por nível de instrução

Teste de Igualdade			
Teste	Chi-quadrado	GL	p-valor
Log-Rank	260.2528	3	<.0001
Wilcoxon	235.8740	3	<.0001
-2Log(LR)	377.8722	3	<.0001

4.3.4 Curva por Estado Civil

Os gráficos $\hat{S}(t)$ e $\hat{h}(t)$ por estado civil são apresentados na Figura 4.5.

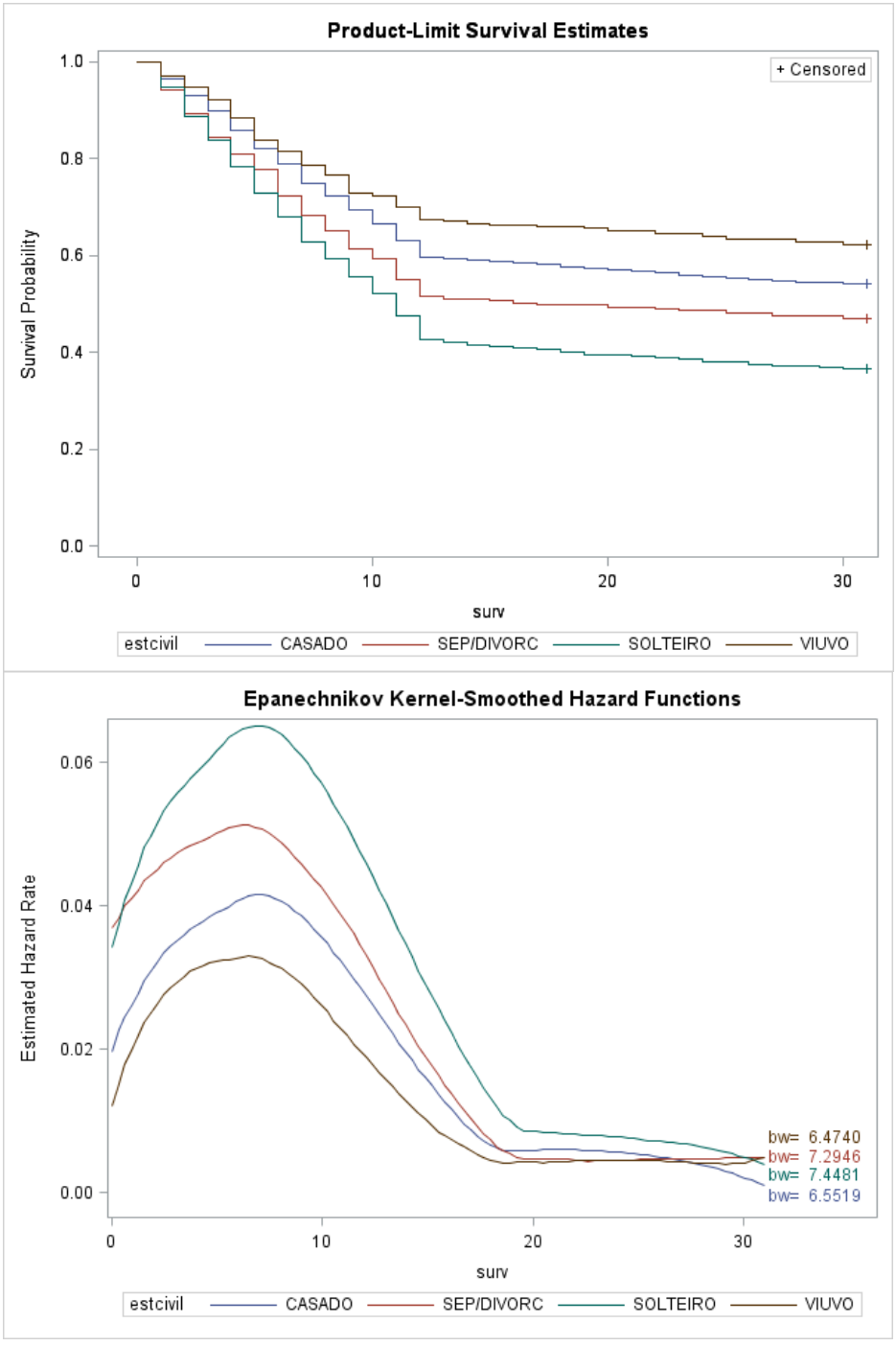


Figura 4.5. Curvas de sobrevivência e de risco por estado civil

Tabela 4.7. Distribuição de clientes por estado civil

Somatório de Falhas e Censuras				
ESTADO CIVIL	Total	Falhas	Censuras	Censura%
Casado	3347	1533	1814	54.20
Separado/Divorciado	476	252	224	47.06
Solteiro	5673	3598	2075	36.58
Viúvo	504	190	314	62.30
TOTAL	10000	5573	4427	44.27

De acordo com os gráficos $\hat{S}(t)$ e $\hat{h}(t)$ por estado civil, clientes viúvos têm maior probabilidade de sobrevivência que os demais, provavelmente por serem de idades mais elevadas e com maior experiência com o crédito. Os que têm segunda maior probabilidade são as pessoas casadas e logo após as separadas ou divorciadas. Os solteiros são os indivíduos com maior risco de inadimplência. A Tabela 4.7 demonstra a distribuição de falhas e censuradas de cada estado civil.

O teste de igualdade rejeita a hipótese nula e confirma que as funções são diferentes significativamente, assim como mostra a Tabela 4.8.

Tabela 4.8. Teste de igualdade por estado civil

Teste de Igualdade			
Teste	Chi-quadrado	GL	p-valor
Log-Rank	333.5410	3	<.0001
Wilcoxon	308.5318	3	<.0001
-2Log(LR)	465.1342	3	<.0001

4.3.5 Curva por Tempo de Conta

A Figura 4.6 traz os gráficos $\hat{S}(t)$ e $\hat{h}(t)$ por tempo de conta do cliente.

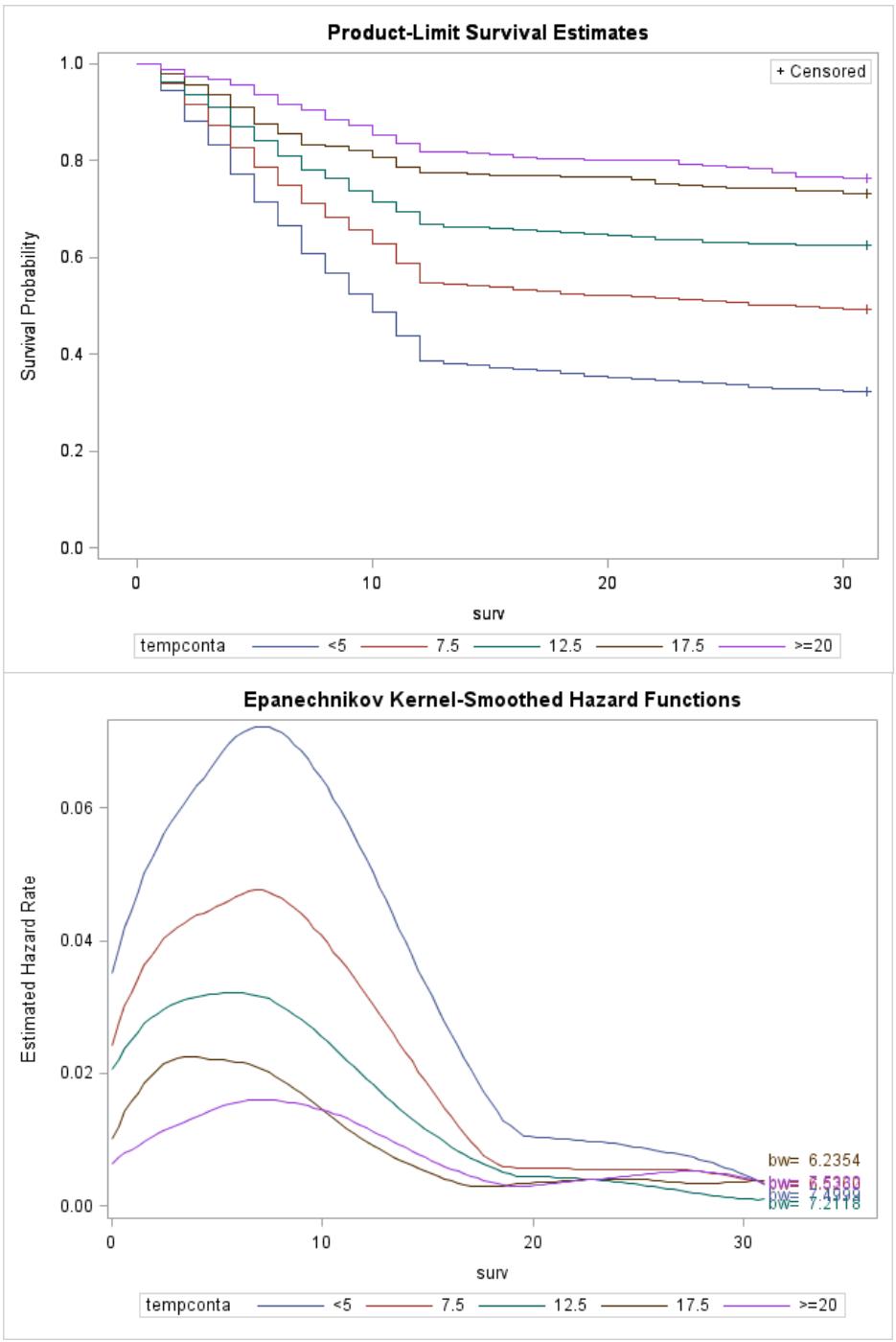


Figura 4.6. Curvas de sobrevivência e de risco por tempo de conta

Tabela 4.9. Distribuição de clientes por tempo de conta

Somatório de Falhas e Censuras				
TEMPO DE CONTA (anos)	Total	Falhas	Censuras	Censura%
<5	5117	3462	1655	32.34
7.5	2812	1427	1385	49.25
12.5	1203	451	752	62.51
17.5	392	105	287	73.21
>=20	405	96	309	76.30
TOTAL	9929	5541	4388	44.19

Quando a amostra é dividida por tempo de conta, aqueles com maior tempo têm também as maiores probabilidades de sobrevivência. Portanto, quanto maior o tempo de conta do cliente, menos propício a inadimplência ele será. A Tabela 4.9 mostra a distribuição de inadimplentes. No início, clientes com mais de 17 anos e meio de conta têm um risco maior que os com mais de 20, mas depois de 10 meses de observação, as curvas se cruzam e se estabilizam com valores aproximados. 71 observações com tempos de falha, censura ou estrato inválidos foram deletadas.

Através das informações contidas na Tabela 4.10, rejeita-se a hipótese nula de igualdade entre as curvas.

Tabela 4.10. Teste de igualdade por tempo de conta

Teste de Igualdade			
Teste	Chi-quadrado	GL	p-valor
Log-Rank	724.4236	4	<.0001
Wilcoxon	648.1285	4	<.0001
-2Log(LR)	1065.9343	4	<.0001

4.3.6 Curva por Tempo de Ocupação Principal

Os gráficos $\hat{S}(t)$ e $\hat{h}(t)$ por tempo de ocupação principal são apresentados na Figura 4.7.

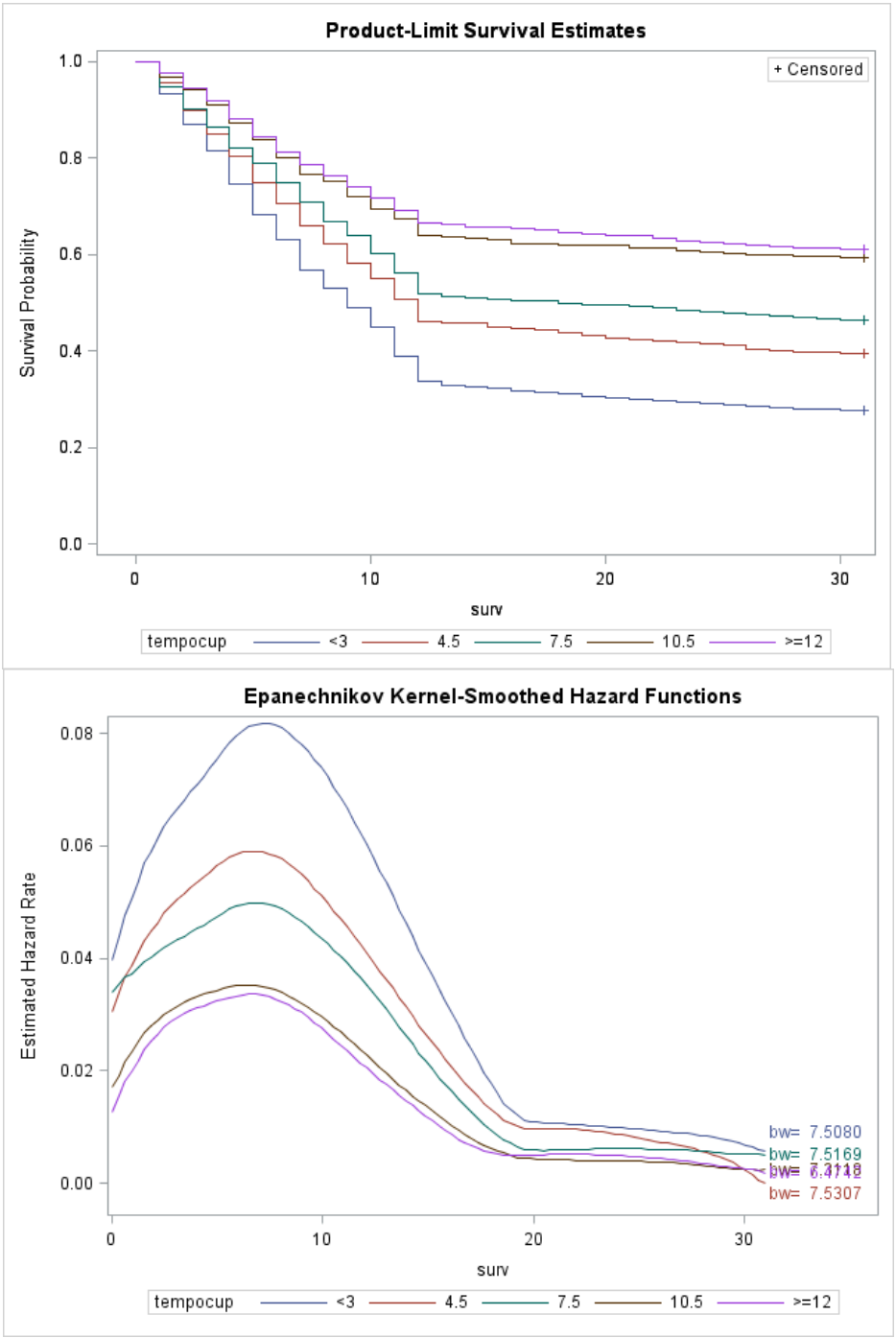


Figura 4.7. Curvas de sobrevivência e de risco por tempo de ocupação principal

Tabela 4.11. Distribuição de clientes por tempo de ocupação principal

Somatório de Falhas e Censuras				
TEMPO DE OCUPAÇÃO (anos)	Total	Falhas	Censuras	Censura%
<3	3155	2285	870	27.58
4.5	1801	1087	714	39.64
7.5	1381	741	640	46.34
10.5	1034	421	613	59.28
>=12	2171	844	1327	61.12
TOTAL	9542	5378	4164	43.64

O agrupamento por tempo de ocupação principal segue a mesma analogia do tempo de conta. Quanto maior o tempo de ocupação, maior a probabilidade de sobrevivência. As curvas das pessoas com 10 anos e meio de ocupação e mais de 12 anos são bem próximas. A Tabela 4.11 mostra o comportamento dos clientes por tempo de ocupação principal de acordo com os casos de inadimplência. 458 observações com tempos de falha, censura ou estrato inválidos foram deletadas.

O teste de igualdade rejeitou a hipótese nula para todos os testes. Logo, há diferença significativa entre as curvas, como informa a Tabela 4.12.

Tabela 4.12. Teste de igualdade por tempo de ocupação principal

Teste de Igualdade			
Teste	Chi-quadrado	GL	p-valor
Log-Rank	736.2435	4	<.0001
Wilcoxon	660.7323	4	<.0001
-2Log(LR)	1009.5152	4	<.0001

4.3.7 Curva por Somatório de Rendas Líquidas

A Figura 4.8 detalha os gráficos $\hat{S}(t)$ e $\hat{h}(t)$ por somatório de rendas líquidas cadastrais dos clientes.

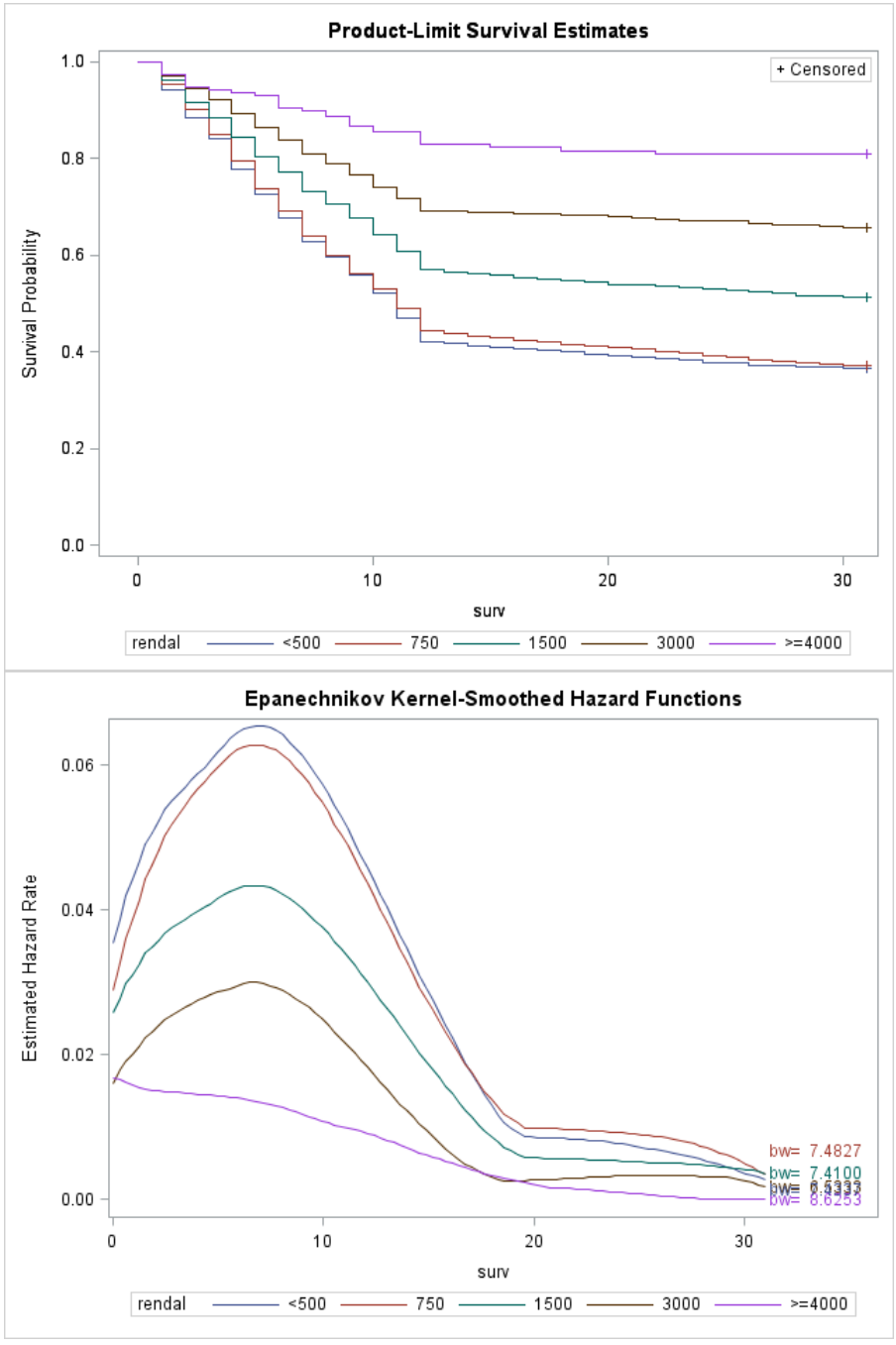


Figura 4.8. Curvas de sobrevivência e de risco por somatório de rendas líquidas

Tabela 4.13. Distribuição de clientes por somatório de rendas líquidas

Somatório de Falhas e Censuras				
RENDA LÍQUIDA (R$)	Total	Falhas	Censuras	Censura%
<500	3271	2076	1195	36.53
750	3085	1934	1151	37.31
1500	2153	1048	1105	51.32
3000	931	319	612	65.74
>=4000	158	30	128	81.01
TOTAL	9598	5407	4191	43.67

Os gráficos $\hat{S}(t)$ e $\hat{h}(t)$ por somatório de rendas líquidas cadastrais nos informam que quanto maior a renda do cliente, menor será seu risco de inadimplência. Pessoas com mais de 4 mil reais de renda líquida têm probabilidade de sobrevivência bem maior que as demais. Já os clientes que possuem os maiores riscos de falha, com renda menores de 500 e até 750 reais, apresentam curvas bem próximas. A distribuição de clientes por renda é apresentada na Tabela 4.13. 402 observações com tempos de falha, censura ou estrato inválidos foram deletadas.

Novamente, nota-se diferença significativa entre as curvas de sobrevivência como demonstra a Tabela 4.14.

Tabela 4.14. Teste de igualdade por somatório de rendas líquidas

Teste de Igualdade			
Teste	Chi-quadrado	GL	p-valor
Log-Rank	392.8221	4	<.0001
Wilcoxon	356.1051	4	<.0001
-2Log(LR)	587.6217	4	<.0001

4.4 Regressão

4.4.1 Escolha do modelo

Como foi dito anteriormente, a aproximação de Breslow e a de Efron, ambas aproximações da função de verossimilhança parcial, trazem resultados satisfatórios quando o número de empates é relativamente pequeno. Quando o número de empates é grande, como podemos notar na Tabela 4.1, se faz necessário utilizar modelos mais precisos como o modelo Exato e o modelo Discreto. A diferença entre os dois é que o modelo Exato assume que existe uma verdadeira ordenação para os eventos empatados mas que não é conhecida, tendo um tempo contínuo. Enquanto o modelo Discreto assume que os eventos realmente ocorreram exatamente no mesmo tempo, tendo um tempo discreto.

Para auxiliar na escolha do modelo mais indicado Chalita *et al.* (2002) propõem uma regra empírica baseada na seguinte definição para a proporção de empates: $pe = (d - k)/n$, em que d é o número total de falhas, k é o número de falhas distintas e n é o número total de eventos. Observe que, se não houver empates, $d = k$ e $pe = 0$. Por outro lado, se todas as observações forem empatadas, $d = n$ e $k = 1$ e, então, $pe = (n - 1)/n$. Uma sugestão empírica foi proposta pelos autores para decidir entre os modelos discretos e as aproximações para a função de verossimilhança parcial utilizando o valor de pe. A proposta está na Tabela 4.15.

Tabela 4.15. Proposta empírica para a decisão entre os modelos discretos e as aproximações para a função de verossimilhança parcial.

pe (%)	Modelos
< 20	**Deve ser** usado o modelo contínuo com aproximações para a função de verossimilhança parcial
20 a 25	**Pode ser** usado o modelo contínuo com aproximações para a função de verossimilhança parcial
> 25	**Deve ser** usado um modelo discreto

Como há empates em todas as observações, $k = 1$. Portanto:

$$pe = \frac{5573 - 1}{10000} = 0.5572 = 55.72\%,$$

de acordo com a proposta empírica, os modelos discretos são mais indicados para esta amostra.

Normalmente, quando se trata de modelos discretos, o modelos de riscos proporcionais de Cox e o modelo logístico são os mais conhecidos e utilizados. A partir de simulações de Monte Carlos, Chalita *et al.* (2002) chegaram a conclusão de que o modelo de Cox se ajusta melhor do que o logístico.

4.4.2 Análise da Tabela Comparativa

A Tabela 4.16 traz um comparativo entre os modelos propostos para amostras com empates. Os testes de chi-quadrado são testes de Wald para a hipótese nula de que cada coeficiente (β) é igual a 0. Esta estatística é baseada na distribuição assintótica de $\hat{\beta}$ e é uma generalização do teste t de Student (Wald, 1943). É calculada simplesmente pelo quadrado do parâmetro estimado dividido pela variância estimada. Todos os cálculos assumem 5% de nível de significância.

Todos os modelos têm a mesma conclusão para as hipóteses testadas exceto pela variável *media* que é rejeitada pelo modelo Discreto e não rejeitada pelos demais. Para o modelo de Breslow o p-valor é de 0.2974, para o de Efron é de 0.3468 e para o Exato é de 0.3279. Já no Discreto, o p-valor = 0.0038 está dentro do limite de rejeição. O grau de liberdade das variáveis categóricas é calculado pela quantidade de níveis menos 1. Um exemplo disso é a variável “estcivil” que se refere ao estado civil do cliente. Devido ser uma variável categórica com quatro níveis, o grau de liberdade é 3, assim como ocorre com a variável “nvlinst” (nível de instrução).

Tabela 4.16. Análise comparativa entre os modelos propostos para amostras com empates

ANÁLISE DE MÁXIMA VEROSSIMILHANÇA									
MODELOS		BRESLOW		EFRON		EXATO		DISCRETO	
Variáveis	GL	Chi-quadrado	P-valor	Chi-quadrado	P-valor	Chi-quadrado	P-valor	Chi-quadrado	P-valor
idade	1	71.7459	<.0001	78.0973	<.0001	78.0969	<.0001	77.7171	<.0001
estcivil	3	13.7237	0.0033	15.3689	0.0015	15.4907	0.0014	11.1837	0.0108
nvlinst	3	79.4465	<.0001	87.4181	<.0001	87.7746	<.0001	84.9091	<.0001
tempconta	1	144.0581	<.0001	155.2520	<.0001	155.4929	<.0001	152.8227	<.0001
tempocup	1	4.4584	0.0347	4.7767	0.0288	4.8031	0.0284	5.3190	0.0211
rendal	1	60.0922	<.0001	60.4938	<.0001	60.2227	<.0001	60.8538	<.0001
vrestri	1	107.97604	<.0001	1188.3181	<.0001	1186.4108	<.0001	1031.6761	<.0001
cheque	1	84.4978	<.0001	100.2829	<.0001	101.1087	<.0001	74.1306	<.0001
chequeesp	1	326.6246	<.0001	356.0415	<.0001	356.5827	<.0001	349.1436	<.0001
smcc	1	26.7186	<.0001	27.2505	<.0001	27.1413	<.0001	24.2846	<.0001
ftcartao	1	0.1318	0.7166	0.1381	0.7102	0.1365	0.7118	0.1051	0.7458
nparc1	1	6.8488	0.0089	7.8304	0.0051	7.9355	0.0048	8.8142	0.0030
vpago	1	28.2632	<.0001	30.3511	<.0001	30.2601	<.0001	26.6845	<.0001
nparc2	1	7.9040	0.0049	8.5699	0.0034	8.6057	0.0034	9.4563	0.0021
nparc3	1	0.3434	0.5579	0.1447	0.7037	0.1155	0.7340	0.0301	0.8623
media	1	1.0858	0.2974	0.8850	0.3468	0.9573	0.3279	8.3986	0.0038
comp	1	0.2499	0.6171	0.1644	0.6851	0.1535	0.6952	0.1076	0.7428

Esta diferença de resultado do modelo de Cox com os demais pode ser explicado pelo grande número de empates que acabam deteriorando as aproximações da verossimilhança parcial e o modelo Exato.

4.4.3 Análise da Máxima Verossimilhança Estimada pelo Modelo Discreto

Quando o modelo de riscos proporcionais é utilizado, os coeficientes das covariáveis podem ser interpretados como o logaritmo da razão do risco do evento de dois indivíduos com características diferentes para uma covariável específica. Dessa forma, o coeficiente de uma covariável específica é interpretado como o logaritmo da razão do risco do evento de um indivíduo, que assume determinado valor para esta covariável, em relação a outro indivíduo para o qual foi observado um outro valor que é assumido como referência.

As estimativas da razão de risco e seus respectivos intervalos de confiança são normalmente obtidos a partir do modelo múltiplo final ajustado. A interpretação dos parâmetros depende do tipo de covariável considerada, podendo ser contínua ou categórica.

Suponha um modelo de riscos proporcionais com apenas uma variável contínua x. A função de risco para o i-ésimo indivíduo para o qual $x = x_i$ é

$$h_i(t) = \exp\left(\hat{\beta}' x_i\right) h_0(t).$$

Considere a razão de risco entre dois indivíduos i e j, os quais assumem os valores $x = x + 1$ e $x = x$ respectivamente, ou seja,

$$\frac{h_i(t)}{h_j(t)} = \frac{\exp\left[\hat{\beta}(x+1)\right] h_0(t)}{\exp\left[\hat{\beta}(x)\right] h_0(t)} = \frac{\exp\left[\hat{\beta}(x+1)\right]}{\exp\left[\hat{\beta}(x)\right]} = \exp\left(\hat{\beta}\right).$$

Assim, $\exp\left(\hat{\beta}\right)$ estima a razão de risco de clientes que assumem o valor $x = x + 1$ em relação aos que tem $x = x$, para qualquer valor de x. Pode-se dizer que o risco de se observar o evento de interesse para os clientes que assumem $x = x + 1$ é $\exp\left(\hat{\beta}\right)$ vezes o risco para os clientes com $x = x$. Dessa forma, a razão de risco quando o valor de x é acrescido em r, é $\exp\left(r\hat{\beta}\right)$. O parâmetro β pode ser interpretado como o logaritmo da razão de risco dos dois indivíduos considerados.

Quando a covariável classifica os clientes em um entre m grupos, estes grupos podem ser considerados como níveis de um fator. No modelo de riscos proporcionais, a função de risco para um indivíduo no j-ésimo grupo, $j = 1, 2, \ldots, m$, é dado por

$$h_j(t) = \exp\left(\gamma_j\right) h_0(t),$$

em que γ_j é o efeito referente ao j-ésimo nível do fator e $h_0(t)$ a função de risco básica. Adotando essa parametrização do modelo, temos que um dos parâmetros assume valor igual a zero para uma determinada categoria ou grupo, denominada referência. As razões de riscos das demais categorias são obtidas em relação a essa categoria adotada como

referência. O risco para esse grupo de referência é dado pela função de risco básica. Assim, a razão de risco, em um determinado t, de um cliente pertencente a um grupo diferente ao de referência em relação ao de referência é $\exp(\gamma_j)$. Similar ao caso de uma variável contínua, podemos dizer que o risco dos indivíduos pertencentes a algum grupo $j, j \geq 2,$ é $\exp(\widehat{\gamma_j})$ vezes o risco do grupo adotado como referência. Consequentemente, o parâmetro γ_j é o logaritmo da razão do risco do evento de interesse de um cliente do grupo j para outro pertencente ao grupo um adotado como referência, ou seja,

$$\gamma_j = \log\left\{\frac{h_j(t)}{h_0(t)}\right\}.$$

Das 10 mil observações, 8745 foram usadas. Isso ocorreu devido 1255 observações terem sido deletadas por terem tempos de falha ou censura inválidos ou variáveis explicativas inválidas. Das 8745 observações analisadas, 43.20% foram censuradas.

Para testar a hipótese nula de que todos o coeficientes são nulos, o SAS disponibiliza três teste: Wald, Score e Razão de Verossimilhança. A estatística de Wald é calculada usando certas funções (formas quadráticas) dos parâmetros estimados e suas variâncias e covariâncias estimadas. A estatística Score é baseada em funções similares a primeira e segunda derivada da função de verossimilhança. E finalmente, a estatística da Razão de Verossimilhança é calculada maximizando a verossimilhança duas vezes: sob a hipótese nula. A estatística é então duas vezes a diferença positiva no log da razão da verossimilhança vezes dois.

Como informa a Tabela 4.17, as três estatísticas rejeitam a hipótese nula para a amostra em estudos, ou seja, há evidências de pelo menos um dos coeficientes é diferente de 0, com 21 graus de liberdade e nível de significância de 5%. Este grau de liberdade é obtido pelas 15 variáveis contínuas mais os três níveis de cada uma das duas variáveis categóricas.

Tabela 4.17. Estatísticas para testar se todos os coeficientes são nulos

TESTE GLOBAL PARA A HIPÓTESE NULA: BETA = 0			
Teste	Chi-quadrado	GL	P-valor
Razão de Verossimilhança	3484.2795	21	<.0001
Score	3887.3384	21	<.0001
Wald	3085.5346	21	<.0001

As variáveis explicativas selecionadas para o modelo Discreto, como informa a Tabela 4.16, são: *idade, estcivil, nvlinst, tempconta, tempocup, rendal, vrestri, cheque, chequeesp, smcc, nparc1, vpago, nparc2* e *media.* A variável *idade* se refere à idade do cliente, *estcivil* ao estado civil, *nvlinst* ao nível de instrução*, tempconta* ao tempo de conta, *tempocup* ao tempo de ocupação principal, *rendal* ao somatório das rendas líquidas cadastrais, *vrestri* ao valor de restrição, *cheque* à quantidade de cheques devolvidos nos últimos 6 meses, *chequeesp* à utilização média do cheque especial no último semestre, *smcc* ao saldo médio de conta corrente por renda líquida, *nparc1* ao número de parcelas pagas por número de parcelas do contrato do crédito parcelado, *vpago* ao somatório de valores pagos pelo total do contrato do crédito parcelado, *nparc2* ao número de parcelas pagas por número de parcelas do contrato do crédito consignado e *media* à média nos últimos 6 meses. Para informações mais detalhadas vide Apêndice.

Após selecionar as variáveis, se faz necessário interpretá-las. A interpretação do quando cada variável explicativa influencia na variável resposta pode ser obtida através dos valores da razão de risco da Tabela 4.18.

Tabela 4.18. Resultados do ajuste do modelo Discreto de Cox e correspondentes razões de risco

ANÁLISE DAS ESTIMATIVAS DE MÁXIMA VEROSSIMILHANÇA							
Parâmetros	GL	Parâmetros Estimados	Erro Padrão	Chi-quadrado	P-valor	Razão de Risco	Categoria
idade	1	-0.01357	0.00154	77.7171	<.0001	0.987	-
estcivil	1	0.08279	0.08815	0.8821	0.3476	1.086	CASADO
estcivil	1	0.30473	0.10766	8.0123	0.0046	1.356	SEP/ DIVORC
estcivil	1	0.13079	0.08988	2.1175	0.1456	1.140	SOLTEIRO
nvlinst	1	0.60233	0.13639	19.5016	<.0001	1.826	ANALFABETO
nvlinst	1	0.46976	0.05150	83.2102	<.0001	1.600	ENSINO FUNDAMENTAL
nvlinst	1	0.34370	0.04789	51.4991	<.0001	1.410	ENSINO MEDIO
tempconta	1	-0.05222	0.00422	52.8227	<.0001	0.949	-
tempocup	1	-0.00586	0.00254	5.3190	0.0211	0.994	-
rendal	1	-0.0001938	0.0000248	60.8538	<.0001	1.000	-
vrestri	1	0.03244	0.00101	1031.6761	<.0001	1.033	-
cheque	1	0.02272	0.00264	74.1306	<.0001	1.023	-
chequeesp	1	0.79213	0.04239	349.1436	<.0001	2.208	-
smcc	1	-0.01948	0.00395	24.2846	<.0001	0.981	-
ftcartao	1	-0.0007307	0.00225	0.1051	0.7458	0.999	-
nparc1	1	-0.31825	0.10719	8.8142	0.0030	0.727	-
vpago	1	0.52547	0.10172	26.6845	<.0001	1.691	-
nparc2	1	-0.28562	0.09288	9.4563	0.0021	0.752	-
nparc3	1	0.03080	0.17752	0.0301	0.8623	1.031	-
media	1	0.00207	0.0007139	8.3986	0.0038	1.002	-
comp	1	-0.00108	0.00328	0.1076	0.7428	0.999	-

Para as variáveis categóricas, *estcivil* e *nvlinst,* as categorias que possuíam maior probabilidade de sobrevivência foram utilizadas como referência, ou seja, a categoria “viúvo” para estado civil e “ensino superior ou mais” para nível de instrução. Portanto, analisando os clientes por estado civil, o risco de um solteiro ser inadimplente é 1,14 vezes o risco dos viúvos. Já o risco dos separados ou divorciados é 1,35 vezes. E o risco dos casados é apenas 1,08 vezes o risco dos viúvos. Analisando por nível de instrução, o risco de um analfabeto se tornar inadimplente é 1,82 vezes o risco de quem tem nível superior ou mais. Já o risco de quem tem apenas ensino fundamental é 1,6 vezes. E o risco de quem tem até o ensino médio é 1,41 vezes o risco de quem é graduado ou possui um ensino de nível mais elevado.

Para variáveis quantitativas como idade, tempo de conta, tempo de ocupação principal, renda líquida e quantidade de cheques devolvidos, a interpretação é direta e mais objetiva. Analisando os clientes por idade, a razão de risco é 0.987 . Logo, a cada ano a mais de idade, o risco de inadimplência diminui 1.3% $(100[0.987 - 1] = -1.3)$. Por tempo de conta, a cada ano a mais do cliente no banco, o risco de se tornar inadimplente diminui 5.1%. A cada ano a mais com a mesma profissão, diminui 0.6%. Para o somatório de rendas líquidas, o valor da razão de risco é 1 devido o SAS informar um valor aproximado. O valor mais preciso é: $\exp$ (-0.0001938) = 0.9998062 . Para facilitar na interpretação, devido a renda ser informada com valores em escala maior, é interessante multiplicar o parâmetro por uma constante: $\exp([1000][-0.0001938]) = 0.8238$. Logo, a cada mil reais a mais na renda líquida do cliente, o risco de inadimplência diminui 17.62%. Em relação a quantidade de cheques devolvidos, a cada cheque a mais devolvido nos últimos seis meses de conta, o risco aumenta 2.3%.

As variáveis *vrestri, chequeesp, smcc, nparc1, vpago, nparc2* e *media* são interpretadas de uma forma mais subjetiva por seus valores terem sido obtidos de proporções, médias ou fórmulas específicas. Para o valor de restrição, a cada unidade a mais de valor, o risco aumenta 3.3%. A cada unidade a mais no resultado da utilização média do cheque especial no último semestre, aumenta 120,8% do risco. E a cada unidade a mais nos valores de *smcc, nparc1* e *nparc2,* o risco diminui 1.9%, 27.3% e 24.8%, respectivamente. Já para as variáveis *vpago e media*, a cada unidade a mais, o risco aumentou 69.1% e 0.2%, respectivamente.

4.4.4 Proporcionalidade dos Riscos

A suposição básica para realização de qualquer estudo com o modelo de regressão de Cox é a de riscos proporcionais. A violação desta suposição pode acarretar sérios vícios na estimação dos coeficientes do modelo (Struthers e Kalbfleisch, 1986).

Um forma de avaliar se os riscos são proporcionais é analisando as curvas de sobrevivência e risco estimadas. As Figuras 4.3 a 4.8 indicam que há proporcionalidade de riscos, ou seja, as curvas não se cruzam, são paralelas.

Outro método bastante utilizado é a análise dos resíduos de Schoenfeld (1982). Para definir tais resíduos no modelo de Cox, considere que se o i-ésimo indivíduo com vetor de covariáveis $x_i = (x_{1i}, x_{2i}, \dots, x_{pi})$, é observado falhar, tem-se para este indivíduo um vetor de resíduos de Schoenfeld $r_i = (r_{i1}, r_{i2}, \dots, r_{ip})$ em que cada componente r_{iq}, para $q = 1, \dots, p$, é definido por

$$r_{iq} = x_{iq} - \frac{\sum_{j \in R(t_i)} x_{jq} \exp\left(x_j' \hat{\beta}\right)}{\sum_{j \in R(t_i)} \exp\left(x_j' \hat{\beta}\right)}.$$

Os resíduos são definidos para cada falha e não são definidos para censuras. Como usual para resíduos, $\sum_i r_i = 0$.

Esta técnica gráfica envolve, como qualquer outra, conclusões subjetivas, pois depende da interpretação dos gráficos.

A dispersão aleatória dos dados no gráfico indica proporcionalidade dos riscos. Segue nas Figuras 4.9 a 4.12 os resíduos de Schoenfeld para a maioria das covariáveis selecionadas pelo modelo. A partição notada nos gráficos, formando colunas, é explicada pela natureza discreta dos tempos. Pode-se concluir que os gráficos não apresentam nenhuma tendência acentuada durante o tempo e que os dados têm dispersão bastante aleatória. Resíduos de variáveis categóricas são pouco informativos, por isso não foram mostrados.

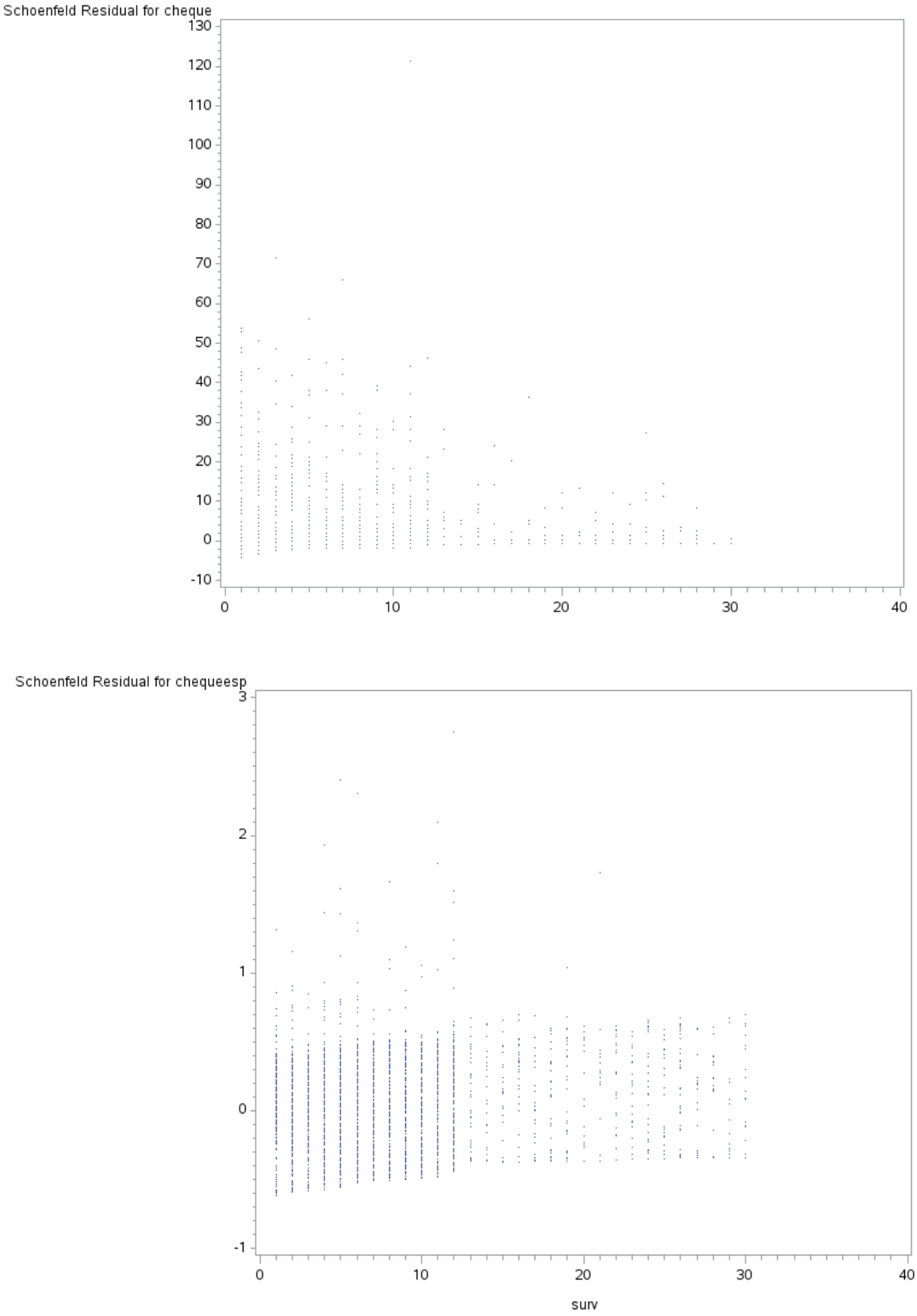
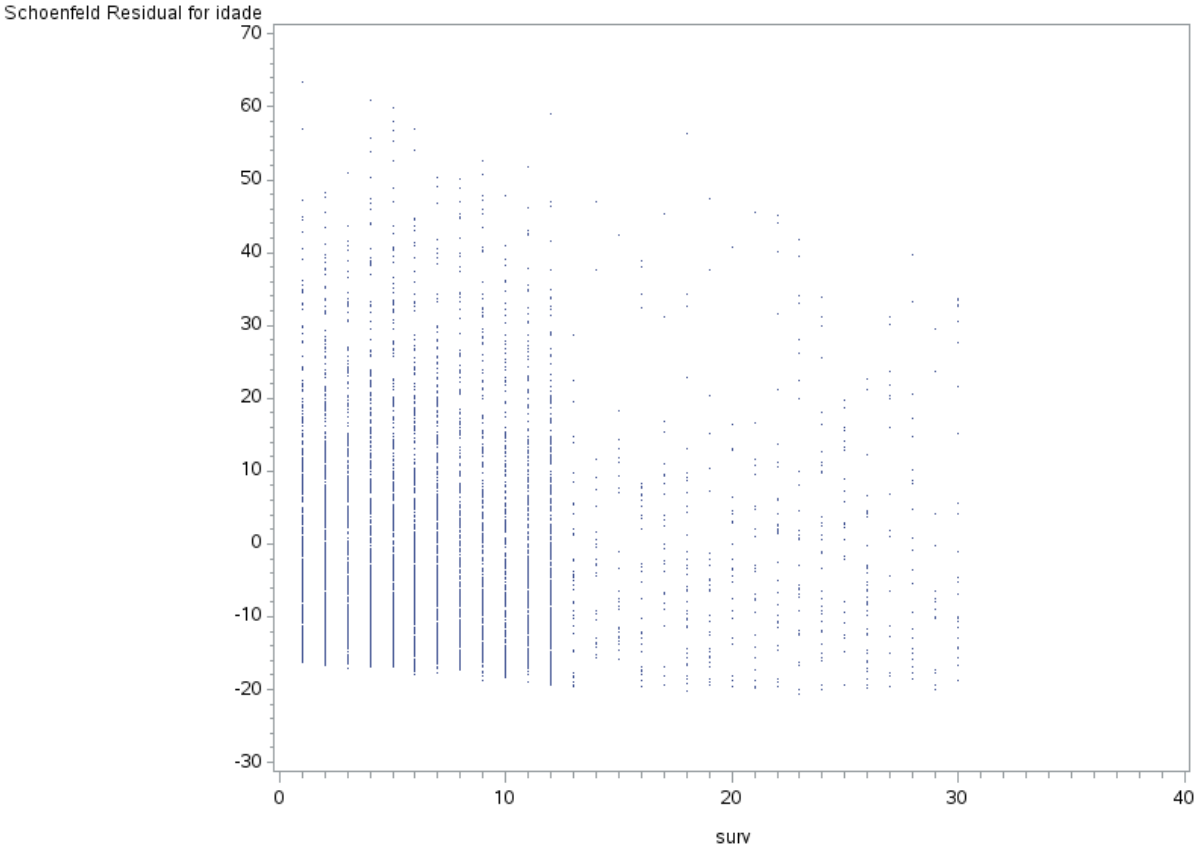


Figura 4.9. Resíduos de Schoenfeld das variáveis *cheque* e *chequeesp*



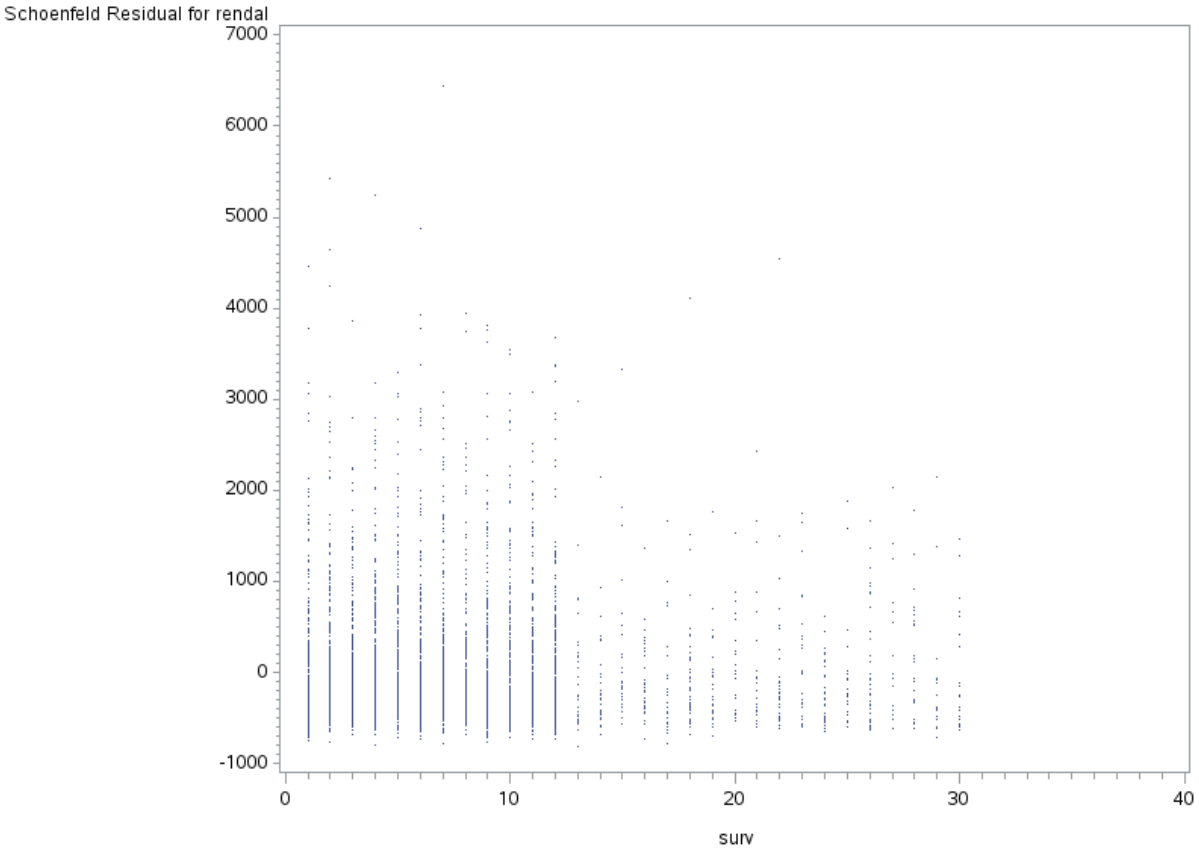


Figura 4.10. Resíduos de Schoenfeld das variáveis *idade* e *rendal*

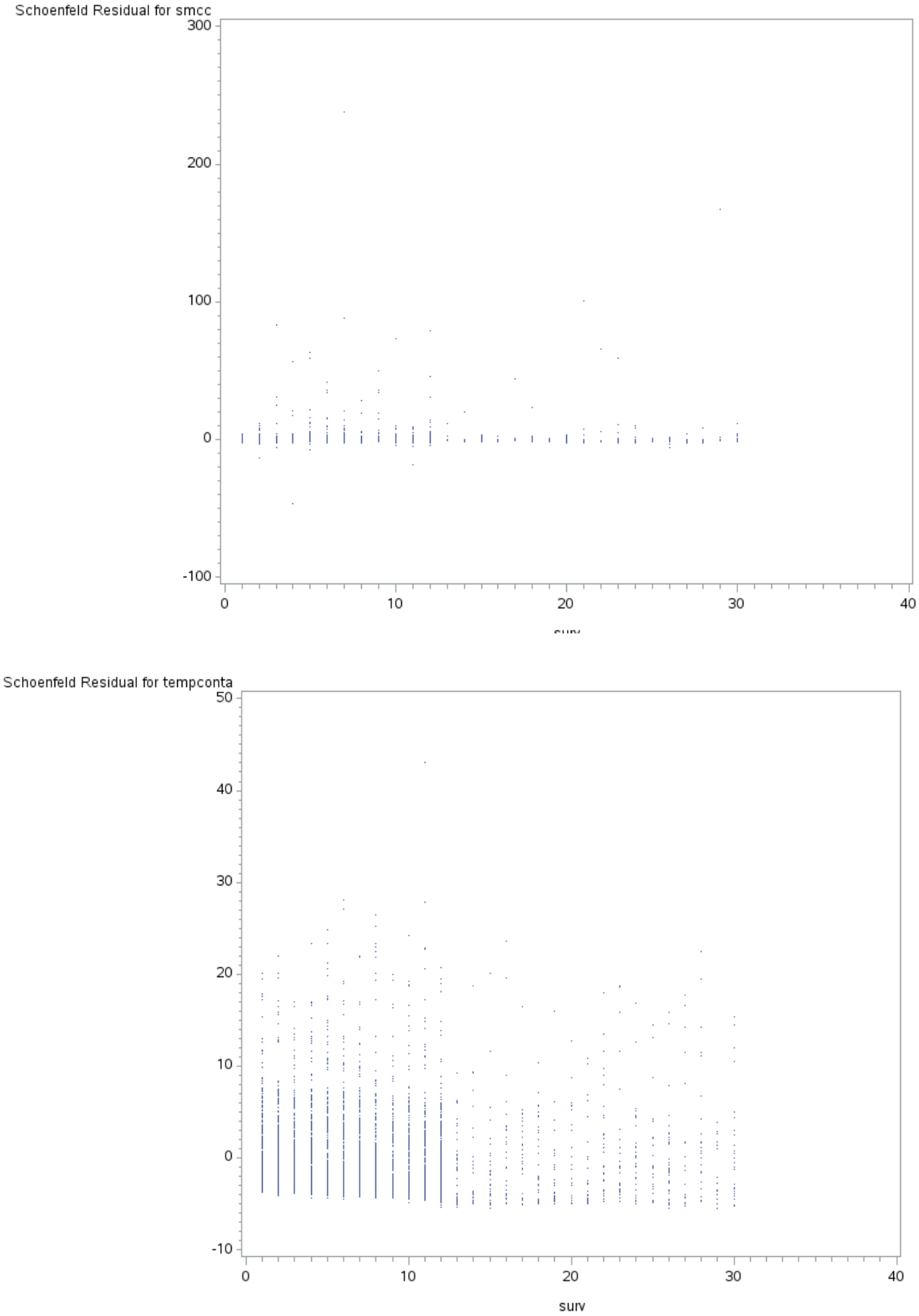
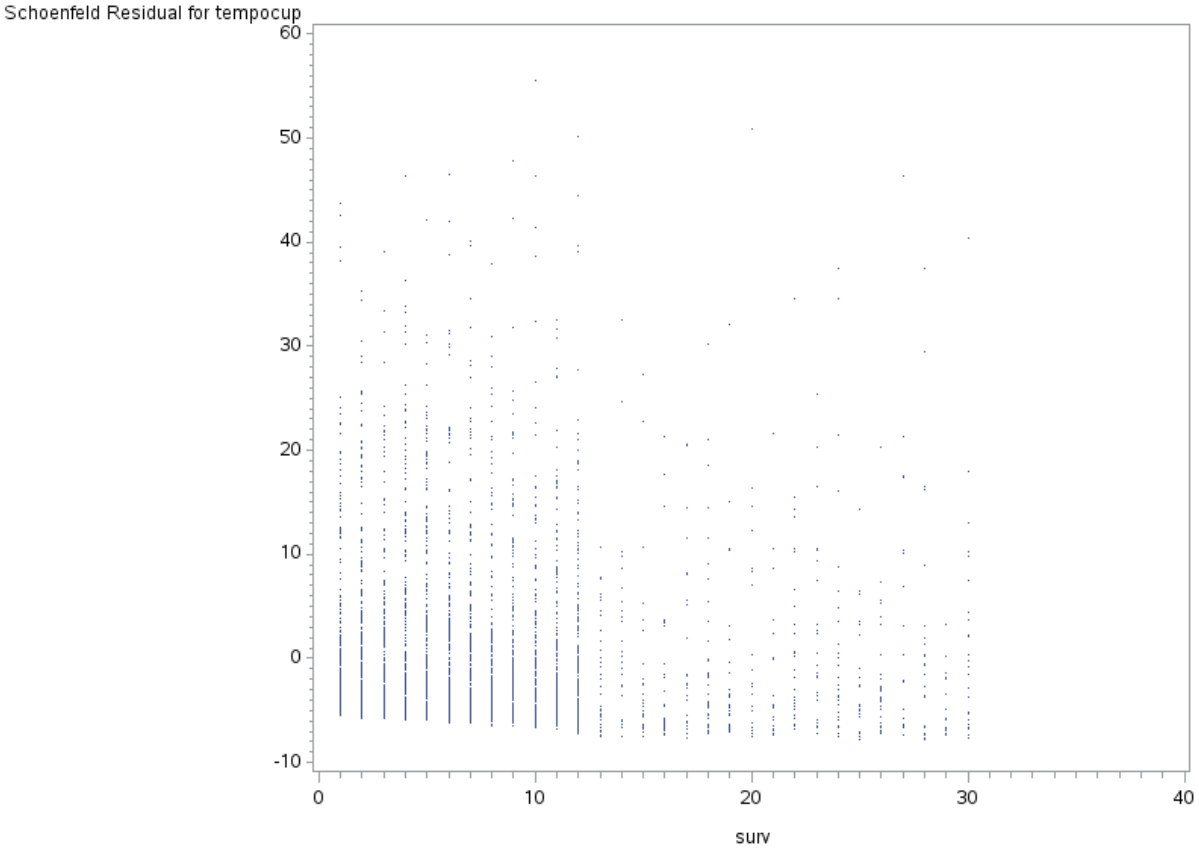


Figura 4.11. Resíduos de Schoenfeld das variáveis *smcc* e *tempconta*



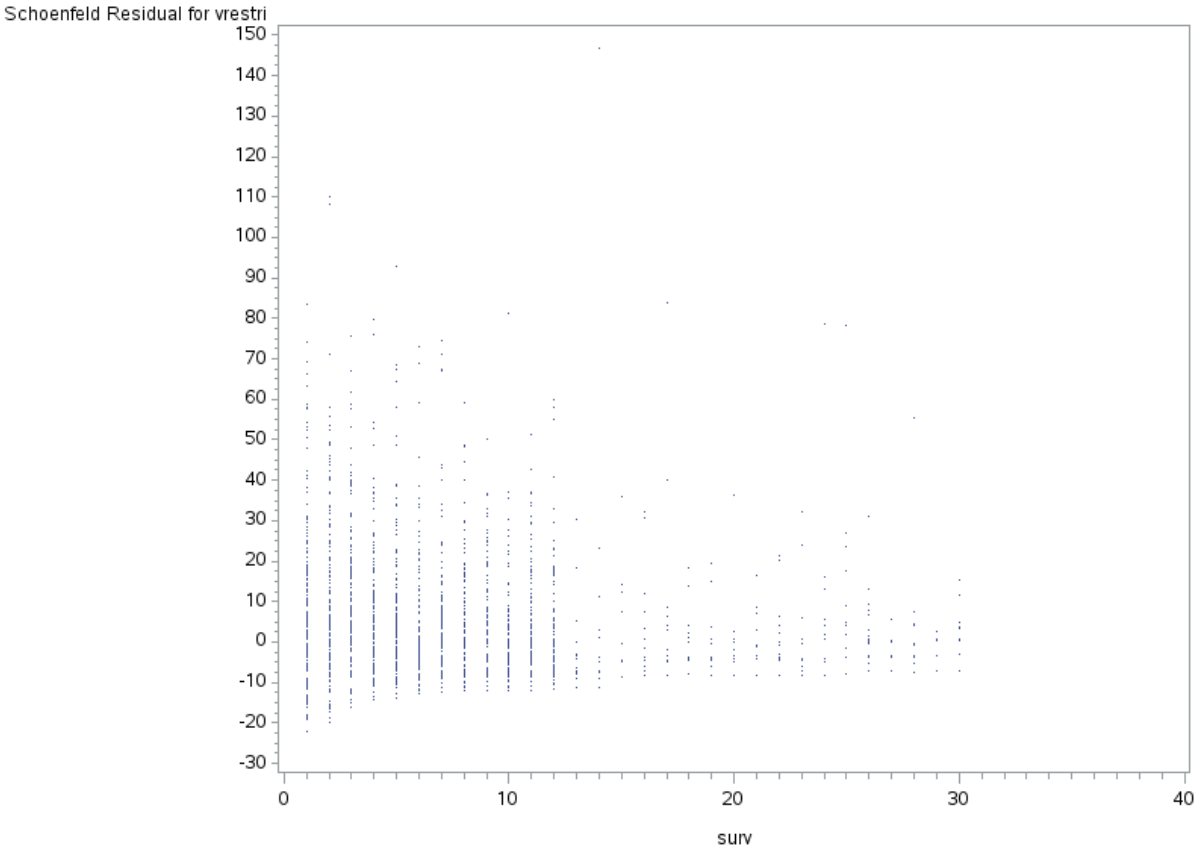


Figura 4.12. Resíduos de Schoenfeld das variáveis *tempocup* e *vrestri*

5. CONCLUSÃO

A metodologia de análise de sobrevivência tem como vantagem a utilização do tempo até a ocorrência do evento de interesse que se pretende estimar, apresentando assim uma visão contínua do comportamento do cliente, e dessa forma sendo possível, se necessário, a avaliação do risco de crédito dos clientes em qualquer dos tempos dentro do intervalo observado, o que, de certa forma, provoca uma mudança no paradigma da análise de dados de crédito. Para tanto, os modelos de regressão de Cox são amplamente utilizados na análise financeira, justamente pela sua flexibilidade devido o componente não-paramétrico.

Este trabalho demonstrou o comportamento de covariáveis com a inadimplência dos clientes. Quanto maior a idade, o tempo de conta, o tempo de ocupação principal e a renda, menor será a probabilidade do cliente não pagar suas contas. Normalmente, espera-se que quanto maior for o nível de instrução maior será a educação financeira do indivíduo e, consequentemente, menos propício a inadimplência ele será. Mas os dados mostraram uma informação interessante, que é comum no mercado financeiro, onde os analfabetos têm menos risco de se tornar inadimplente do que as pessoas com ensino fundamental e médio.

Através da análise de máxima verossimilhança estimada pelo modelo Discreto, a influência das potenciais covariáveis na inadimplência foi medida e interpretada. Como exemplo, a cada mil reais a mais na renda líquida de um cliente, o risco de inadimplência diminui 17.62%. Essa intensidade pode auxiliar diretamente nas tomadas de decisão da política de risco de crédito de uma instituição financeira.

Como o cenário econômico é muito dinâmico, a obtenção de informações mais atualizadas para serem utilizadas na validação dos modelos poderia trazer ganhos para a

metodologia como um todo, fazendo com que os resultados das medidas de avaliação fossem mais próximas e fiéis à realidade atual.

Nas instituições financeiras em geral existem grandes bancos de dados com uma enorme quantidade de covariáveis para cada cliente e métodos de seleção de variáveis como o *stepwise,* que utiliza um algoritmo para avaliar a importância de cada covariável, são de grande utilidade e agregaria mais precisão ao modelo.

Existem vários métodos para avaliação da qualidade do ajuste do modelo. Dentre eles, o resíduo de Cox-Snell é bastante frequente nas literaturas. Este método avalia a qualidade geral de ajuste do modelo, mas é mais indicado para avaliar o ajuste de modelos paramétricos de análise de sobrevivência, sendo pouco informativo para os modelos de Cox estimados pela verossimilhança parcial (Allison, 1995). Os métodos de resíduos *martingal e deviance* são modificações dos resíduos de Cox-Snell e são utilizados para verificar a presença de pontos atípicos (*outliers*). Para estudos futuros, a avaliação do modelo utilizando esses métodos também seria interessante. Como os modelos Discretos foram mais indicados para a análise, também seria interessante comparar o resultados com outros tipos de modelos Discretos, como o modelo de chances proporcionais logístico.

APÊNDICE

Nome Variável	Construção
Estado civil	Código do estado civil do cliente: 1 SOLTEIRO(A) 2 CASADO(A)-COMUNHAO UNIVERSAL 3 CASADO(A)-COMUNHAO PARCIAL 4 CASADO(A)-SEPARACAO DE BENS 5 VIUVO(A) 6 SEPARADO/A EXTRA/JUDICIALMENTE 7 DIVORCIADO(A) 8 CASADO(A) - REGIME DOTAL 9 CAS REGIME MISTO OU ESPECIAL 11CASADO(A)-PART.FINAL AQUESTO
Idade (anos)	Diferença entre a data da análise e a data de nascimento em anos
Nível de instrução	Código do nível de instrução: 01 ANALFABETO 02 ENSINO FUNDAMENTAL 03 ENSINO MEDIO 04 SUPERIOR INCOMPLETO 05 SUPERIOR COMPLETO 06 POS GRADUADO 07 MESTRADO 08 DOUTORADO 09 SUPERIOR EM ANDAMENTO
Somatório de rendas líquidas cadastrais	Somatório das rendas líquidas mensais atualizadas cadastradas
Tempo de conta (anos).	Diferença em anos da data da análise e da data de abertura da primeira conta como correntista
Tempo de ocupação principal (anos)	Diferença entre a data da análise e a data de início da ocupação principal
Valor de restrição	Valor de restrição dos pesos das anotações baixadas e em ser, ocorridas nos últimos 5 anos. Valor das restrições = $\sum_{i=1}^{n} peso_i * \exp(0{,}05 * \ delta\_t_i)$. Onde: *delta_ti* (anotação baixada)= total de dias entre a data de análise e a data de baixa da anotação dividido por (–365) (deve ser um número negativo) ou *delta_ti* (anotação em ser)= total de dias entre a data da análise e data da ocorrência da anotação dividido por 365. Se a data da ocorrência estiver em branco, considerar 31 maio de 2003. *n* = número de anotações no cadastro do cliente; *i* = *i* –ésima anotação do cliente (varia de 1 a *n*).

	$peso_i$ = peso referente a *i* –ésima anotação do cliente.
Quantidade de cheques devolvidos nos últimos 6 meses	Quantidade de cheques devolvidos pelas alíneas 11, 12, 13 ou 14 nos 6 meses anteriores a data de análise.
Saldo médio em conta corrente e investimento (incluindo cheque especial e adiantamento a depositante) nos últimos 6 meses/ Renda líquida da ocupação principal nos últimos 6 meses	Saldo médio em conta corrente e investimento (incluindo cheque especial e adiantamento a depositante) nos últimos 6 meses dividida pela renda líquida da ocupação principal nos últimos 6 meses.
Utilização média do cheque especial no último semestre	- Captura-se o valor médio de utilização no mês (soma de todos os saldos devedores diários, dividido pelo número de dias úteis) e o limite contratado existente ao final do mês. - Caso o cliente possua mais de um cheque especial, somam-se tanto os valores médios como os limites contratados. - Calcula-se a utilização média para cada um dos últimos 6 meses, dividindo-se o valor médio de utilização pelo limite contratado. - Efetua-se a contagem relativa à quantidade de meses dentre os últimos 6, em que o cliente possuiu o cheque especial. - Soma-se a utilização média nesses meses e divide-se pela quantidade de meses apurada no item anterior. Obs: verifica-se se o cliente possuiu cheque especial durante o mês através da existência de limite.
Valor médio faturado cartão nos últimos 6 meses/ Exposição no semestre	Para cada um dos 6 meses anteriores a data da análise, capturou-se, o somatório do valor total faturado no mês. Caso o cliente possuísse mais de uma conta cartão, somaram-se todos os valores faturados do cliente. Fez-se a contagem da quantidade de meses em que o cliente teve limite de cartão de crédito. Dividiu-se o somatório pela quantidade de meses apurada anteriormente. Calculou-se a média semestral das operações de crédito em aberto utilizando-se o saldo do ultimo dia do mês das operações em aberto. Dividiu-se o valor médio faturado pela exposição média do semestre.
Número de parcelas pagas / Número parcelas do contrato (Crédito Parcelado)	Para cada um dos 6 meses anteriores ao mês da data da análise, contou-se a quantidade de parcelas pagas e a quantidade total de parcelas das operações pertencentes ao grupamento Crédito Parcelado do porftólio de limites. Dividiu-se a somatória da quantidade de parcelas pagas pela quantidade total de parcelas do contrato para cada mês do semestre. Contou-se a quantidade de meses em que o cliente possuía operações do grupamento crédito parcelado em aberto. Dividiu-se o somatório da razão quantidade parcelas pagas/quantidade total de parcelas dos contratos de crédito parcelado do semestre pela quantidade de meses em que o cliente possuía operações do grupamento parcelado.

Somatório dos valores pagos/ Total do contrato (Crédito Parcelado)	Para as operações pertencentes ao grupamento parcelado em ser no mês anterior ao da análise ou anterior ao mês anterior da análise quando tiver sido liquidada no mês anterior ao da análise, dividiu-se o somatório do valor principal pago pelo somatório do valor principal total dos contratos de produtos pertencentes ao grupamento crédito parcelado.
Número de parcelas pagas / Número parcelas do contrato (Crédito Consignado)	Para cada um dos 6 meses anteriores ao mês da data da análise, contou-se a quantidade de parcelas pagas e a quantidade total de parcelas das operações pertencentes ao grupamento Crédito Consignado do porftólio de limites. Dividiu-se a somatória da quantidade de parcelas pagas pela quantidade total de parcelas do contrato para cada mês do semestre. Contou-se a quantidade de meses em que o cliente possuía operações do grupamento crédito consignado. Dividiu-se o somatório da razão quantidade parcelas pagas/quantidade total de parcelas dos contratos de crédito consignado do semestre pela quantidade de meses em que o cliente possuía operações do grupamento consignado.
Número de parcelas pagas / Número parcelas do contrato (Crédito Reescalonamento)	Para cada um dos 6 meses anteriores ao mês da data da análise, contou-se a quantidade de parcelas pagas e a quantidade total de parcelas das operações pertencentes ao grupamento Crédito Reescalonamento do porftólio de limites. Dividiu-se a somatória da quantidade de parcelas pagas pela quantidade total de parcelas do contrato para cada mês do semestre. Contou-se a quantidade de meses em que o cliente possuía operações do grupamento crédito reescalonamento em aberto. Dividiu-se o somatório da razão quantidade parcelas pagas/quantidade total de parcelas dos contratos de crédito reescalonamento do semestre pela quantidade de meses em que o cliente possuía operações do grupamento reescalonamento.
Média (Quantidade de dias atraso / quantidade de operações em ser) nos últimos 6 meses	Para cada um dos 6 meses anteriores ao mês da data da análise, contou-se o número de dias de atraso e a quantidade total de operações. Dividiu-se o somatório da quantidade de dias de atraso pela quantidade total de operações para cada mês do semestre. Somaram-se os resultados das razões acima e dividiu-se por 6.
Comprometimento mensal da somatória de rendas líquidas cadastrais do mês	Com base nos dados do mês anterior ao da análise, apurou-se: - 10% do valor refinanciado de cartão de crédito no último mês, - 10% do saldo devedor de cheque especial no último mês, - valor das prestações das demais operações de crédito e, - os valores das prestações em atraso. Somaram-se os valores obtidos e dividiu-se pelo somatório das rendas líquidas registradas no módulo “Dados Profissionais” do cadastro.

REFERÊNCIAS BIBLIOGRÁFICAS

Allison, P. D. (1995) *Survival Analysis Using the SAS® System: A Practical Guide,* Cary, NC: SAS Institute Inc., 292 pp.

Alves, B.C. (2010) Modelos heterogénos de sobrevivência: uma aplicação ao risco de crédito. Tese de Mestrado em Prospecção e Análise de Dados. *ISCTE Business School. Instituto Universitário de Lisboa*.

Cao, R., Vilar, J.M. & Devia, A. (2009). Modelling consumer credit risk via survival analysis. *Universidade da Coruña.*

Chalita, L.V.A.S., Colosimo, E.A., Demétrio, C.B.G. (2002) Likelihood Approximations and Discrete Models for Tied Survival Data, *Communications in Statistics – Theory and Methods,* 31, 1215-1229.

Collett, D.(1991). *Modelling Binary Data,* Chapman and Hall, Boca Raton.

Collett, D.(2003). *Modelling Survival Data in Medical Research,* 2ed., Chapman and Hall, London.

Colosimo, E.A. & Giolo, S.R. (2006). Análise de Sobrevivência Aplicada. São Paulo: Edgard Blücher.

Cox, D.R. (1972), Regression models and life tables, *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 34(2) 187-220.

Dai, Z. J. (2010). About Time - Building credit score cards with survival analysis. *School of Mathematics and Statistics, The University of New South Wales.*

Diniz, C. & Louzada, F. (2012). *Modelagem Estatística para Risco de Crédito.* 20º SINAPE. João Pessoa. ABE.

Gasser, T. & Müller, H.G. (1979). *Kernel Estimation of Regression Functions,* in Smoothing Techniques for Curve Estimation, Lecture Notes in Mathematics. Berlin: Springer-Verlag. 757, 23-68.

Kalbfleisch, J.D. e Prentice, R.L. (1973). Marginal likelihoods based on cox's regression and life model. *Biometrika,* 60(2), 267-278.

Kalbfleisch, J.D. e Prentice, R.L. (1980) *The statistical analysis of failure time data.* John Wiley, New York.

Lawless, G.A.(1982). *Statistical Models and Methods for Lifetime Data.* John Wiley and Sons, New York.

Noh, H.J., T.H. Roh e I. Han (2005), Prognostic personal credit risk model considering censored information, *Expert Systems with Applications*, 28(4), 753-762.

Prentice, R.L., Gloeckler, L.A. (1978) Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data. *Biometrics,* 34, 57-67.

Schoenfeld, D. (1982). Partial Residuals for the Proporcional Hazard Regression Model. *Biometrika,* 69, 239-241.

Stepanova, M. & Thomas, L. (2001), PHAB scores: Proportional hazard analysis behavioural scores, *Journal of the Operational Research Society*, 52(9), 1007-1016.

Stepanova, M. & Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Researc,* Vol. 50, 277–289.

Struthers, C.A., Kalbfleisch, J.D. (1986). Misspecified Proportional Hazards Models. *Biometrika,* 73, 363-369.

Thomas, L.C., D.B. Eldman e J.N. Crook (2002), *Credit Scoring and Its Applications*, Society for Industrial and Applied Mathematics, Philadelphia.

Zhang, A. (2009). Statistical methods in credit risk modeling. *A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Statistics) in The University of Michigan.*

Wald, A. (1943). Tests of Statistical Hypotheses concerning Several Parameters when the Number of Observations is Large. *Trans. Amer. Math. Soc.,* 54, 426-482.