U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

RENDIBILIDADE DE TRANSACÇÕES DE CRÉDITO PESSOAL COM RECURSO A ANÁLISE DE SOBREVIVÊNCIA

Alberto Filipe Neves Correia

MESTRADO EM ENGENHARIA MATEMÁTICA

ABRIL 2007

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

RENDIBILIDADE DE TRANSACÇÕES DE CRÉDITO PESSOAL COM RECURSO A ANÁLISE DE SOBREVIVÊNCIA

Alberto Filipe Neves Correia

MESTRADO EM ENGENHARIA MATEMÁTICA

ABRIL 2007

falta versão

Tese submetida à Faculdade de Ciências da
Universidade do Porto para obtenção do grau de
Mestre em Engenharia Matemática

Dissertação realizada sob a supervisão da
Professora Doutora Maria do Carmo Guedes
Departamento de Matemática Aplicada
Faculdade de Ciências da Universidade do Porto
Abril de 2007

dedicado à minha mãe, Lucília

Agradecimentos

Gostaria de agradecer à Professora Doutora Maria do Carmo Guedes pela sua orientação e sugestões que muito ajudaram a realizar esta dissertação.

Também gostaria de agradecer ao Dr. Manuel Gonçalves que, pela sua inspiração e experiência no negócio bancário, sempre me forneceu valiosos conselhos e apoio neste trabalho. Os agradecimentos estendem-se naturalmente à instituição de crédito de que faz parte, nomeadamente no que respeita à cedência dos dados usados experimentalmente.

Por último, e não menos importante, agradeço à minha família, amigos e colegas de trabalho pela sua paciência, disponibilidade e recomendações que me foram dando ao longo deste trabalho. Muito obrigado a todos.

Resumo

Actualmente as instituições financeiras classificam os seus clientes baseando-se em sistemas de *scoring* de crédito, avaliando o perfil de risco dos clientes que, por sua vez, é utilizado na decisão de crédito. Este trabalho pretende mostrar que a rendibilidade das transacções de crédito ou, mais propriamente, a rendibilidade esperada, pode ser outro indicador de apoio à decisão de crédito e pode ainda ser utilizado para calcular o *pricing* das transacções.

O pagamento antecipado e o incumprimento são acontecimentos que afectam negativamente a rendibilidade das transacções de crédito. O primeiro acontecimento implica a perda de juros e o segundo implica não só os juros, mas também o capital. Mais do que saber se estes acontecimentos vão ocorrer, para o cálculo da rendibilidade é mais importante saber quando é que vão ocorrer. A análise de sobrevivência e, em particular, os modelos de vida acelerada ou de *hazards* proporcionais permitem estimar funções de sobrevivência que, neste caso, se traduzem como probabilidades das transacções 'sobreviverem' aos acontecimentos referidos em função do tempo e com base num conjunto de variáveis explicativas. Essas funções são depois utilizadas de forma determinante no cálculo da rendibilidade esperada como probabilidades de receber as componentes de capital e juro das prestações de cada transacção de crédito.

Palavras-chave: Rendibilidade esperada; análise de sobrevivência; modelos de vida acelerada; modelos de *hazards* proporcionais; *pricing*.

Abstract

Financial institutions currently establish customer's classification using credit scoring systems that evaluate their risk profile which in turn will be used in credit decisions. This study intends to show that profitability of loans or, more specifically, the expected profit, can be another indicator supporting credit decisions and can also be used to calculate loan pricing. Early repayment and default effect the profitability of loans negatively. The former implies the loss of interest and the latter implies not only the loss of interest, but of capital also. For profitability computation, better than knowing if these events will occur, is knowing when they will occur. Survival analysis and, in particular, accelerated life models or proportional hazard models permit the estimation of survival functions which, in this case, are the loan 'survival' probabilities to the referred events as function of time and based on a given set of variables. These functions are then used in the computation of expected profit as probabilities of receiving the capital and interest of the instalments of each loan.

Keywords: Expected profit; survival analysis; accelerated life models; proportional hazard models; pricing

Conteúdo

Introdução

Actualmente as instituições de crédito baseiam as decisões de concessão de crédito no perfil de risco dos seus clientes. Os pareceres dos decisores de crédito com base em regras como os cinco C's referidos por Thomas [11] (carácter do cliente, capital pedido, colateral, capacidade financeira e condições do mercado) manifestam-se claramente insuficientes quando falamos de mercados muito grandes e heterogéneos como o retalho, em que a consistência e o tempo das decisões são aspectos decisivos.

A solução passa pelos chamados sistemas de *scoring* de crédito. Trata-se de processos de modelação estatística que permitem prever o comportamento futuro dos clientes com base no seu desempenho no passado. Através da selecção de um conjunto de variáveis explicativas e de métodos estatísticos é possível atribuir um *score* que classifica o cliente em termos de risco e que está relacionado com a sua probabilidade de incumprimento.

Num ambiente competitivo como o das instituições de crédito, os desafios actuais são cada vez mais exigentes. Espera-se conseguir aumentar o volume de vendas e rendimento, logo o lucro, reduzindo despesas operacionais e perdas resultantes de inadimplemento. Tudo isto melhorando os níveis de serviço e promovendo a boa relação com o cliente.

Os bancos começam a perceber que há vantagens em considerar como principal objectivo a maximização do lucro em vez da minimização do risco. Quer isto dizer que, mantendo as perdas sob controlo, é possível ainda abranger um grupo de clientes que possuem factores de risco mais elevados, mas que ainda assim poderão dar lucro ao banco, permitindo à instituição expandir o seu portefólio de crédito [1]. Pretende-se deste modo olhar para as transacções de crédito não só sob o ponto de vista de risco, mas também sob o ponto de vista da rendibilidade que poderão proporcionar ao banco.

A rendibilidade das transacções de crédito está dependente de diversos factores como

o montante pedido, prazo, taxas de juro, probabilidades de pagamento antecipado e de incumprimento, etc. Estes dois últimos factores são acontecimentos que podem ocorrer em determinada altura do empréstimo causando perdas para o banco. Essas perdas serão tanto maiores quanto mais cedo esses acontecimentos ocorrerem, daí ser necessário modelar, não se estes acontecimentos vão acontecer, mas quando.

Este trabalho invoca por isso um tipo de análise estatística conhecido como análise de sobrevivência, cujos modelos permitem estimar funções de sobrevivência, eventualmente sob influência de variáveis explicativas, em função do tempo. Essas funções são probabilidades de 'sobreviver' a determinados acontecimentos, que neste caso serão o pagamento antecipado e o incumprimento.

Cada transacção de crédito deverá ter associada uma função de sobrevivência para o pagamento antecipado, que indica a probabilidade de receber a componente de juros de cada prestação, e uma função de sobrevivência para o incumprimento, que indica a probabilidade de receber cada prestação (capital e juro). Com base no montante, prazo, taxas de juro e as probabilidades de receber as componentes de capital e juro em cada prestação, é possível obter uma forma de calcular a rendibilidade esperada das transacções de crédito.

O trabalho encontra-se estruturado em três capítulos, sendo o primeiro uma introdução à análise de sobrevivência, nomeadamente aos seus termos e conceitos, estimação não-paramétrica de funções de sobrevivência e aos modelos de vida acelerada e *hazards* proporcionais. O segundo capítulo faz uma aplicação da análise de sobrevivência a dados referentes a crédito pessoal. Mais especificamente, são estimadas funções de sobrevivência para o pagamento antecipado e para o incumprimento, com base num conjunto de variáveis explicativas. São depois analisados vários modelos paramétricos e o modelo de *hazards* proporcionais de Cox (semi-paramétrico), sendo este último bastante competitivo, em termos de poder discriminante, com a tradicional regressão logística.

No terceiro capítulo explicita-se o papel das funções de sobrevivência no cálculo da rendibilidade esperada das transacções de crédito. Pode ver-se ainda a evolução das rendibilidades médias e acumuladas em função das probabilidades de pagamento antecipado e de incumprimento, podendo identificar-se *cut-offs* capazes de diferenciar transacções de crédito com maior ou menor rendibilidade.

O cálculo da rendibilidade esperada pode ainda ser utilizada para o cálculo de *pricing* associado às transacções de crédito. Em vez de se querer saber a rendibilidade esperada de uma dada transacção, a questão é agora saber que taxa juro deve ser cobrada de modo a proporcionar determinado nível de rendibilidade ao banco. Dado que é difícil saber à partida a rendibilidade esperada de uma determinada transacção, a ideia é torná-la relativa, por exemplo, à rendibilidade máxima esperada, isto é, a rendibilidade obtida no caso de não haver pagamento antecipado nem incumprimento. Esta questão é deixada em aberto, mas é seguramente uma aplicação do cálculo da rendibilidade esperada.

Capítulo 1

Análise de sobrevivência

1.1 O que é a análise de sobrevivência?

A análise de sobrevivência é um conjunto de métodos e técnicas estatísticas que analisam uma variável aleatória positiva. Tipicamente essa variável é o tempo até ocorrência de um determinado acontecimento de interesse, também chamado *tempo de sobrevivência.*

As suas origens estão associadas aos campos da Biologia e Medicina e ao uso de tabelas de mortalidade (o acontecimento usual era a 'morte' do indivíduo), mas estas técnicas não se restringem apenas a estes campos e encontram aplicações em áreas tão diversas como Ciências Sociais e Económicas ou Engenharias.

Além da variável tempo, os dados de sobrevivência podem ainda conter um conjunto de variáveis independentes relacionadas com a variável aleatória. Nesse caso, o objectivo passa por modelar a distribuição associada ao tempo até o acontecimento e estabelecer uma dependência desse tempo com as variáveis independentes através de uma regressão.

Uma característica deste tipo de análise é a possibilidade de ter dados *censurados*, que fornecem apenas informação parcial da variável aleatória de interesse. É o que se tem no caso do acontecimento não ser observado durante o período de estudo, podendo eventualmente acontecer no futuro. São dados que têm grande importância na análise e não podem ser simplesmente ignorados, uma vez que, além de muitas outras considerações, os indivíduos com maior tempo de sobrevivência têm também maior probabilidade de ter os dados censurados.

De acordo com Miller [4], a censura pode ser classificada em vários tipos : *tipo I*, quando existe um tempo t_c a partir do qual todos os dados estão censurados; *tipo II*, quando a partir de determinada ordem n_c (indivíduos ordenados por tempo de sobrevivência) todos os dados estão censurados; *tipo III* ou *censura aleatória*, quando a censura não está relacionada com o tempo de sobrevivência e ocorre de forma aleatória. As razões para considerar este último tipo de censura prendem-se com a chegada ao fim do estudo ou por outro qualquer motivo, não relacionado com o tempo de sobrevivência, que impossibilite a recolha de mais informação.

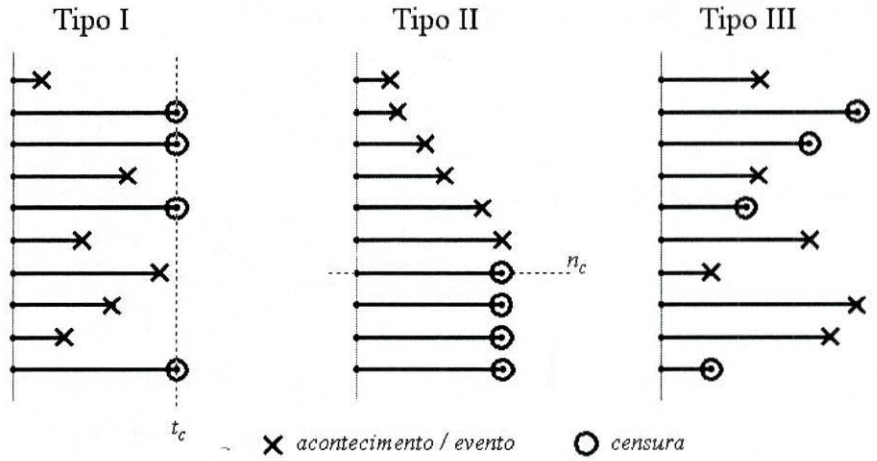

Figura 1.1: Tipos de censura. Os esquemas mostram que os dados podem estar censurados a partir de um determinado tempo t_c (Tipo I), a partir de determinada ordem, n_c, de tempos de sobrevivência (Tipo II), ou aleatoriamente (Tipo III).

Existem ainda outros tipos de censura. Estes que já foram referidos inserem-se na classe da *censura à direita*. No entanto também pode acontecer *censura à esquerda*, por exemplo, quando a variável de interesse é muito grande e não é possível observar o seu início.

Para aplicação da maior parte dos resultados é condição essencial que a censura seja não informativa ou independente (censura aleatória). Segundo Allison [2], um indivíduo com censura num tempo t_c deve ser representativo de todos os outros indivíduos com o mesmo conjunto de variáveis explicativas que sobreviveu até t_c.

1.2 Funções de sobrevivência e de *hazard*

Seja T uma variável que representa o tempo de um acontecimento de interesse, isto é, o tempo medido desde um instante inicial até à ocorrência desse acontecimento. Podemos então descrever a variável T de três formas que acabam por ser equivalentes:

- A função de distribuição e a função de sobrevivência dadas por

$$F(t) = P(T \leq t) \qquad \text{e} \qquad S(t) = 1 - F(t) \tag{1.1}$$

- A função densidade de probabilidade (fdp) definida por

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \tag{1.2}$$

- A função *hazard* definida por

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{1.3}$$

Esta última forma é bastante popular na análise de sobrevivência e define-se como o limite da probabilidade de um acontecimento ocorrer num intervalo muito pequeno de tempo, $[\, t, t + \Delta t\,[$, por unidade de tempo, assumindo que esse acontecimento não ocorreu até ao início desse intervalo. A função $h(t)$ é referida por Lee [3] como uma forma de quantificar o *risco* instantâneo de um acontecimento ocorrer por unidade de tempo, desempenhando um papel importante na análise de sobrevivência.

Allison [2] exibe o seguinte exemplo para clarificar a noção de *hazard* (chamemos-lhe 'risco' para facilitar o texto): suponhamos que o risco de um determinado acontecimento ocorrer é $0,1$ com o tempo medido em meses. Isto significa que num mês espera-se obter o acontecimento $0,1$ vezes. Se o risco for agora de $1,3$ com o tempo medido em anos, isto quer dizer que num ano o acontecimento deverá ocorrer $1,3$ vezes (assumindo o risco constante ao longo do período de referência).

Alternativamente, a função *hazard* também se pode expressar em função de $f(t)$ e da função de sobrevivência

$$h(t) = \frac{f(t)}{S(t)} \tag{1.4}$$

onde também se tem que

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt} \tag{1.5}$$

o que implica

$$h(t) = -\frac{d}{dt} \log S(t) \tag{1.6}$$

A função cumulativa de *hazard* pode então escrever-se para $t > 0$ como

$$H(t) = \int_0^t h(u)du = -\log S(t) \tag{1.7}$$

É possível ainda relacionar as funções de sobrevivência de probabilidade através das expressões

$$S(t) = \exp\left(-\int_0^t h(u)\ du\right) \tag{1.8}$$

donde se obtém naturalmente que

$$f(t) = h(t) \exp\left(-\int_0^t h(u)\ du\right) \tag{1.9}$$

Convém notar que, ao contrário do que acontece com as funções de distribuição e sobrevivência, $F(t)$ e $S(t)$, respectivamente, a função cumulativa de *hazard*, $H(t)$, não é uma probabilidade. De facto, apesar de não tomar valores negativos, pode tomar valores superiores a 1, uma vez que

$$\lim_{t\to+\infty} F(t) = 1 \Leftrightarrow \lim_{t\to+\infty} S(t) = 0 \Leftrightarrow \lim_{t\to+\infty} H(t) = +\infty \tag{1.10}$$

1.3 Estimação não paramétrica de funções de sobrevivência

O método de Kaplan-Meier é o mais utilizado para estimar a função de sobrevivência, e não é mais do que um método de máxima verosimilhança não-paramétrico.

O método é simples e intuitivo. Se não houver dados censurados, $\hat{S}(t)$ é apenas a proporção de indivíduos com tempo de sobrevivência maior que t.

A situação é ligeiramente diferente no caso de haver dados censurados. Suponhamos que existem $t_1, \ldots, t_m$ tempos de sobrevivência distintos. Em cada tempo pode dizer-se que há n_j indivíduos em *risco* (que não tiveram o acontecimento ou censura) antes de t_j. Seja d_j o número de indivíduos que tiveram o acontecimento t_j. Nestes termos o estimador de Kaplan-Meier é dado por

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right), \quad \text{para} \quad t_1 \leq t \leq t_m \tag{1.11}$$

Resumindo, para um dado tempo t, basta tomar todos os tempos de sobrevivência menores que t, calcular para cada um deles a quantidade entre parênteses da fórmula, que se interpreta como probabilidade condicionada de sobreviver até t_{j+1}, dado que sobreviveu até t_j, e multiplicá-las entre si. Note-se que para $t < t_1$ (o menor tempo de sobrevivência), $\hat{S}(t) = 1$. Para $t > t_m$ (o maior tempo de sobrevivência) $\hat{S}(t) = 0$, se não houver dados censurados maiores que t_m.

Considere-se um caso meramente ilustrativo deste estimador. Os valores seguintes representam tempos de sobrevivência de 10 indivíduos, dos quais 4 estão censurados (assinalados com o sinal +):

$$\frac{1}{2} \quad \frac{3}{4}+ \quad \frac{3}{4}+ \quad \frac{3}{4} \quad \frac{1}{2}+ \quad \frac{3}{4} \quad \frac{1}{4} \quad \frac{3}{4} \quad 1 \quad \frac{3}{8}+$$

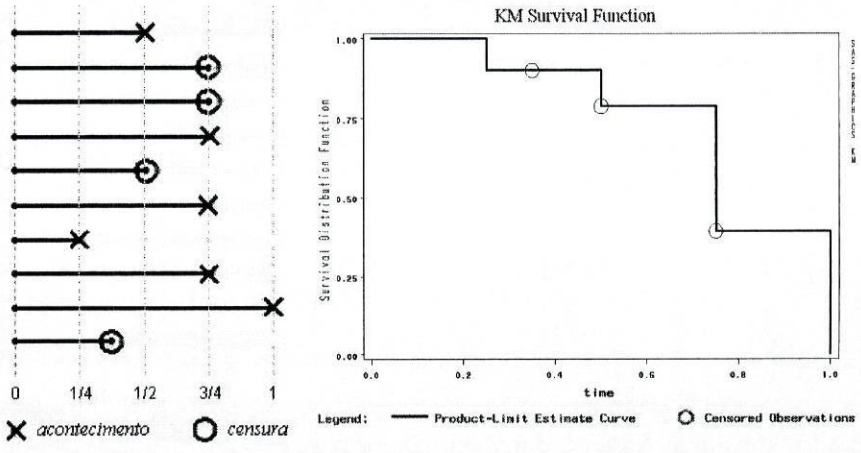

Figura 1.2: Na figura da esquerda esquematizaram-se 10 tempos de sobrevivência dos quais 4 se encontram censurados. Na figura da direita está representada a respectiva função de sobrevivência estimada pelo método Kaplan-Meier.

Neste caso, obtém-se os seguintes valores da função de sobrevivência:

$S(0) = 1$

$S(\frac{1}{4}) = \left(1 - \frac{1}{10}\right) = 0,9$

$S(\frac{1}{2}) = \left(1 - \frac{1}{10}\right)\left(1 - \frac{1}{8}\right) = 0,7875$

$S(\frac{3}{4}) = \left(1 - \frac{1}{10}\right)\left(1 - \frac{1}{8}\right)\left(1 - \frac{3}{6}\right) = 0,39375$

$S(1) = 0$

1.4 Modelos paramétricos de funções de sobrevivência

Seguidamente são apresentados alguns exemplos de modelos paramétricos bastante utilizados na análise de sobrevivência. Para cada modelo são exibidas as funções de sobrevivência, S, e de densidade de probabilidade, f. A parametrização sugerida relativamente às distribuições Exponencial, Weibull, Gama e Loglogística não corresponde à dos parâmetros estimados nas secções posteriores[1].

1. Exponencial

   Este modelo assume o risco constante, isto é, a função *hazard* é dada por $h(t) = \lambda$, com $\lambda > 0$, o que equivale a dizer que a função *hazard* cumulativa é dada por $H(t) = \lambda t$. Deste modo, as funções de sobrevivência e densidade de densidade de probabilidade são dadas por

   $$S(t) = e^{-\lambda t} \tag{1.12}$$

   $$f(t) = \lambda e^{-\lambda t} \tag{1.13}$$

   Tem-se ainda que média e variância são dados por $\frac{1}{\lambda}$ e $\frac{1}{\lambda^2}$, respectivamente.

2. Weibull

   O modelo de Weibull é uma generalização do modelo exponencial. A função de *hazard*

[1] No entanto, para os modelos Exponencial, Weibull e LogLogístico é possível fazer uma reparametrização fazendo $\lambda = e^{-\mu}$ e $\gamma = \frac{1}{\sigma}$, sendo μ e σ parâmetros estimados nas secções posteriores

cumulativa é dada por $H(t) = (\lambda t)^{\gamma}$ para $\lambda, \gamma > 0$. Pode-se então expressar as funções de sobrevivência e probabilidade por

$$S(t) = e^{-(\lambda t)^{\gamma}} \tag{1.14}$$

$$f(t) = \gamma\lambda(\lambda t)^{\gamma-1}e^{-(\lambda t)^{\gamma}} \tag{1.15}$$

A média e a variância são dadas respectivamente por

$$\frac{1}{\lambda}\Gamma\left(1+\frac{1}{\gamma}\right) \quad \text{e} \quad \frac{1}{\lambda^2}\left(\Gamma\left(1+\frac{2}{\gamma}\right) - \Gamma^2\left(1+\frac{1}{\gamma}\right)\right)$$

em que $\Gamma(\gamma)$ é a função Gama[2] definida por

$$\Gamma(\gamma) = \int_0^{\infty} u^{\gamma-1}e^{-u}du \tag{1.16}$$

3. Gama

O modelo Gama é outra generalização do modelo exponencial. A sua função de densidade de probabilidade é dada por

$$f(t) = \frac{\lambda(\lambda t)^{\gamma-1}e^{-\lambda t}}{\Gamma(\gamma)} \tag{1.17}$$

e sua função de sobrevivência por

$$\begin{aligned} S(t) &= 1 - \int_0^t \frac{\lambda}{\Gamma(\gamma)}(\lambda x)^{\gamma-1}e^{-\lambda x}dx \\ &= 1 - \frac{1}{\Gamma(\gamma)}\int_0^{\lambda t} u^{\gamma-1}e^{-u}du \\ &= 1 - I(\lambda t, \gamma) \end{aligned} \tag{1.18}$$

onde $\Gamma(\gamma)$ está definido em (1.16) e

$$I(s,\gamma) = \frac{1}{\Gamma(\gamma)}\int_0^s u^{\gamma-1}e^{-u}du \tag{1.19}$$

é a chamada *função Gama incompleta.*

A média e variância são dadas por $\frac{\gamma}{\lambda}$ e $\frac{\gamma}{\lambda^2}$ respectivamente.

[2]Para γ inteiro positivo tem-se que $\Gamma(\gamma+1) = \gamma!$.

4. Lognormal

   Como sugere o próprio nome da distribuição, assumindo que $\log T \sim N(\mu, \sigma^2)$, as funções de sobrevivência e de densidade de probabilidade podem exprimir-se do seguinte modo:

$$S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right) \tag{1.20}$$

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log t - \mu}{\sigma}\right)^2\right) \tag{1.21}$$

   onde Φ é a função de distribuição cumulativa da distribuição normal.

5. Log-logístico

   Quando $X = \log T$ segue uma distribuição logística dada por $F(x) = \frac{e^x}{1+e^x}$ com média μ e variância σ^2, diz-se que a distribuição de T tem uma distribuição log-logística dada pela expressão $F(z) = \frac{z}{1+z}$ em que $z = \exp\left(\frac{\log t - \mu}{\sigma}\right)$. Fazendo $\gamma = \frac{1}{\sigma}$ e $\lambda = e^{-\mu}$ pode obter-se as seguintes funções de sobrevivência e de densidade de probabilidade

$$S(t) = \frac{1}{1 + (\lambda t)^\gamma} \tag{1.22}$$

$$f(t) = \frac{\lambda\gamma(\lambda t)^{\gamma-1}}{(1 + (\lambda t)^\gamma)^2} \tag{1.23}$$

Relativamente a estes modelos paramétricos, a Figura 1.3 exibe algumas formas típicas de funções *hazard*.

1.5 Modelos de vida acelerada e *hazards* proporcionais

Acontece muitas vezes que o tempo de sobrevivência é influenciado por variáveis que traduzem determinadas características dos indivíduos. Elas devem ser consideradas de alguma forma no modelo de modo a potenciar o seu poder preditivo.

Os modelos considerados no capítulo anterior podem facilmente ser adaptados de forma a permitir a influência destas variáveis explicativas através de um vector de *covariáveis*, $\mathbf{x} = (x_1, x_2, \ldots, x_k)'$, e de parâmetros, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$. O vector de covariáveis pode

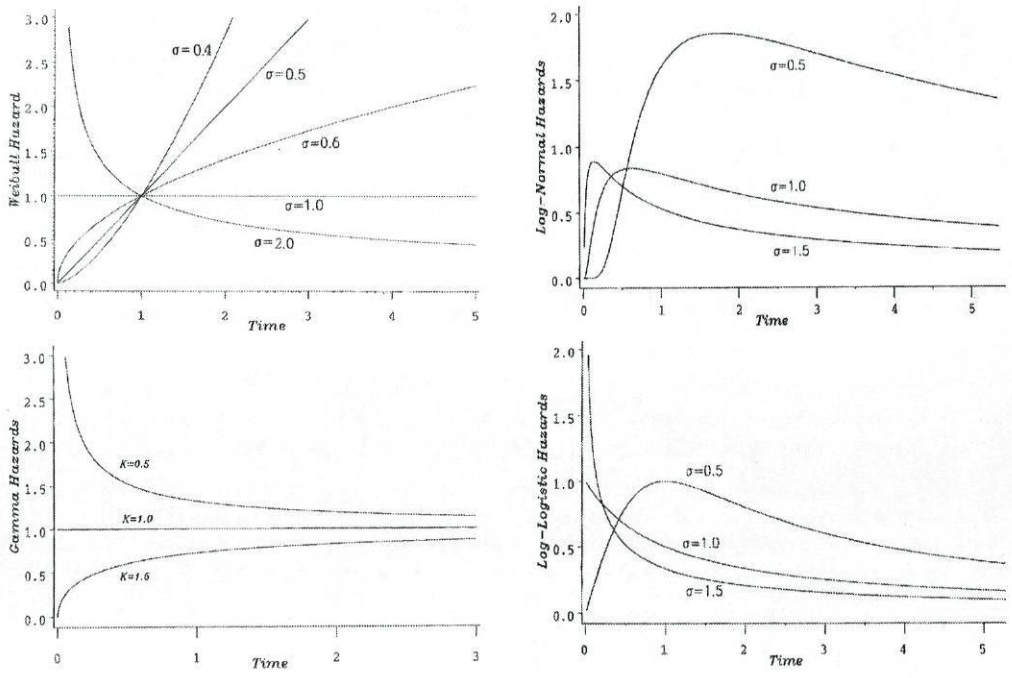

Figura 1.3: Diversos gráficos de funções *hazard* dos modelos paramétricos Weibull, Gama, Log-logístico e Log-normal. [Fonte: Allison, Survival Analysis using SAS]

ser obtido fazendo $\mathbf{x} = \mathbf{y} - \bar{\mathbf{y}}$, onde $\bar{\mathbf{y}}$ é um vector de valores de referência (eventualmente a média) das variáveis dadas por $\mathbf{y} = (y_1, y_2, \ldots, y_k)'$. Stepanova e Thomas [5] sugerem alternativamente que cada variável seja dividida em subgrupos, sendo depois substituídos por variáveis binárias 0/1.

Na análise de sobrevivência salientam-se duas classes de modelos no relacionamento das covariáveis com os tempos de sobrevivência: modelos de *hazards* proporcionais e modelos de vida acelerada (conhecidos como *Accelerated Life models* ou *Accelerated Failure Time Models - AFT Models*).

Nos modelos de vida acelerada a função *hazard* é dada por

$$h(t) = e^{\boldsymbol{\beta}'\mathbf{x}} h_0(t.e^{\boldsymbol{\beta}'\mathbf{x}}) \tag{1.24}$$

em que h_0 é uma função base de *hazard* que se obtém no caso das covariáveis serem todas zero. Em termos da função de sobrevivência fica estabelecida a relação $S(t) = S_0(\alpha t)$, onde $\alpha = e^{\boldsymbol{\beta}'\mathbf{x}}$ e S_0 é a função de sobrevivência associada à função de *hazard* h_0 da expressão (1.24). Banasik [7] refere que neste caso as covariáveis têm o papel de acelerar ou tornar

mais lento o processo de vida do sistema.

Os parâmetros deste tipo de modelos podem ser estimados de forma semelhante ao que acontece com a regressão linear usual. Seja T_i a variável aleatória denotando o tempo de sobrevivência do i-ésimo indivíduo e $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})'$ o vector de covariáveis associado a esse mesmo indivíduo. O modelo é dado por

$$\log T_i = \mu + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \sigma\epsilon_i \tag{1.25}$$

onde $\mu, \beta_1, \ldots, \beta_k$ e σ são parâmetros estimados e ϵ_i é o termo de perturbação aleatória.

Nos modelos de *hazards* proporcionais tem-se que a função *hazard* é dada por

$$h(t) = e^{\boldsymbol{\beta}'\mathbf{x}} h_0(t) \tag{1.26}$$

As covariáveis têm neste caso um efeito 'multiplicador' na função base de *hazard*. Relativamente à função de sobrevivência tem-se que $S(t) = [S_0(t)]^\alpha$, onde $\alpha = e^{\boldsymbol{\beta}'\mathbf{x}}$ e S_0 é a função de sobrevivência associada à função de *hazard* h_0 na expressão (1.26).

A razão de se chamar *hazards proporcionais* vem do facto de nestes modelos o quociente das funções de *hazard* de dois indivíduos i e j ser constante.

$$\frac{h_i(t)}{h_j(t)} = \exp\left(\beta_1(x_{i1} - x_{j1}) + \ldots + \beta_k(x_{ik} - x_{jk})\right) \tag{1.27}$$

Como consequência, os gráficos dos logaritmos das funções *hazard* em função do tempo deverão ser paralelos (Figura 1.4).

$$\log h_i(t) = \log h_0(t) + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \tag{1.28}$$

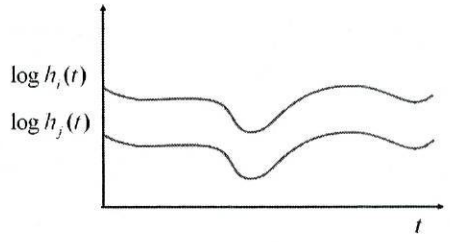

Figura 1.4: Gráficos do logaritmo de *hazards* proporcionais. [Fonte: Allison, Survival Analysis using SAS]

A diferença entre as duas classes de modelos é que nos modelos de *hazards* proporcionais os que estão mais em risco permanecem sempre mais em risco ao longo do tempo,

enquanto que nos modelos de vida acelerada o risco pode ser variável com o tempo. Em 1980, Kalbfleisch e Prentice mostraram que as únicas distribuições que são simultaneamente *hazards* proporcionais e vida acelerada são as Exponencial e Weibull [7].

Ciampi e Etezadi-Amoli [18] propuseram ainda um outro modelo mais geral que acaba por ser uma mistura destes já referidos, sendo a função de *hazard* é dada por

$$h(t) = e^{\beta' \mathbf{x}} h_0(t.e^{\varphi' \mathbf{x}}) \tag{1.29}$$

onde β e φ são vectores de parâmetros e h_0 é a função base de *hazard*. Obtém-se o modelo de vida acelerada quando $\varphi = \beta$ e o modelo *hazards* proporcionais quando $\varphi = 0$.

1.6 Máxima verosimilhança e verosimilhança parcial

Os modelos paramétricos utilizados na análise de sobrevivência são estimados por máxima verosimilhança. Este método é já bastante conhecido, mas requer algum cuidado na sua aplicação em dados de sobrevivência, no que respeita às observações censuradas.

Suponhamos por um momento a inexistência de observações censuradas. Admitindo a independência das n observações tem-se que a função de verosimilhança seria dada por

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i|\mathbf{x}_i) \tag{1.30}$$

em que $\boldsymbol{\theta}$ é um vector composto por todos os parâmetros a estimar (inclui os parâmetros das covariáveis e da distribuição da variável aleatória) e $\mathbf{x}_i$ é o vector de covariáveis do i-ésimo indivíduo.

Considerando a presença de dados censurados (censura aleatória), a função de verosimilhança seria dada por

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[f(t_i|\mathbf{x}_i) S_C(t_i|\mathbf{x}_i)\right]^{\delta_i} \left[f_C(t_i|\mathbf{x}_i) S(t_i|\mathbf{x}_i)\right]^{1-\delta_i} \tag{1.31}$$

onde

$$\delta_i = \begin{cases} 1 & \text{se dados não estão censurados} \\ 0 & \text{se os dados estão censurados} \end{cases}$$

em que $f(t_i|\mathbf{x}_i)$ e $S(t_i|\mathbf{x}_i)$ são as funções de densidade e sobrevivência do tempo até ocorrência do acontecimento de interesse, respectivamente, e $f_C(t_i|\mathbf{x}_i)$ e $S_C(t_i|\mathbf{x}_i)$ são as funções

de densidade e sobrevivência do tempo até censura, respectivamente. Como é referido por Miller [4], não estando a censura relacionada com o tempo até ocorrência do evento de interesse, os produtos $\prod_{i=1}^{n} [S_C(t_i|\mathbf{x}_i)]^{\delta_i}$ e $\prod_{i=1}^{n} [f_C(t_i|\mathbf{x}_i)]^{1-\delta_i}$ não envolvem os parâmetros a estimar, de modo que podem ser tratados como constantes na maximização da função de verosimilhança, podendo simplificar-se a expressão anterior

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [f(t_i|\mathbf{x}_i)]^{\delta_i} [S(t_i|\mathbf{x}_i)]^{1-\delta_i} = \prod_{i=1}^{n} [h(t_i|\mathbf{x}_i)]^{\delta_i} S(t_i|\mathbf{x}_i) \tag{1.32}$$

Note-se que a segunda forma para a função de verosimilhança dada em (1.32) permite exprimí-la unicamente em termos da função de *hazard* (a função de sobrevivência está relacionada com a função de *hazard* conforme (1.8)). O método de máxima verosimilhança consiste depois em encontrar estimativas de $\boldsymbol{\theta}$, que maximizam o logaritmo de $L(\theta)$ (designada por *Loglikelihood*)

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \log h(t_i|\mathbf{x}_i) + \log S(t_i|\mathbf{x}_i) \tag{1.33}$$

através de um sistema de equações

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial(\boldsymbol{\theta})} = 0 \tag{1.34}$$

cuja resolução requer normalmente a utilização de processos iterativos, como o método de Newton-Raphson.

Um outro método, designado de *partial likelihood* ou verosimilhança parcial, foi proposto em 1972 por Sir David Cox [9] para estimar os parâmetros do modelo semi-paramétrico de *hazards* proporcionais (que, no entanto, permitia uma fácil generalização para modelos de não-proporcionalidade). Essa forma de estimação permitia obter os parâmetros dados pelo vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$ sem ter de especificar a função base de *hazard* $h_0(t)$, dada em (1.26). O seu método baseava-se na ordenação, ou *ranking*, dos tempos de sobrevivência. Deste modo, alterações ao tempo como somar ou multiplicar por uma constante, ou mesmo tomar o seu logaritmo, não produzia efeito sobre o valor dos parâmetros.

Uma função de verosimilhança típica é o produto das verosimilhanças de todos os indivíduos da amostra. No caso da verosimilhança parcial e para n indivíduos da amostra tem-se

que

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{h_i(t)}{\sum_{j=1}^{n} Y_{ij}\, h_j(t)} \right)^{\delta_i} = \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{\beta}' \mathbf{x}_i}}{\sum_{j=1}^{n} Y_{ij}\, e^{\boldsymbol{\beta}' \mathbf{x}_j}} \right)^{\delta_i} \tag{1.35}$$

onde

$$Y_{ij} = \begin{cases} 1 & \text{se } t_j \geq t_i \\ 0 & \text{caso contrário} \end{cases} \quad \text{e} \quad \delta_i = \begin{cases} 1 & \text{se dados não estão censurados} \\ 0 & \text{se os dados estão censurados} \end{cases}$$

A introdução de Y_{ij} na expressão permite, de forma conveniente, excluir do denominador os indivíduos que já tiveram o acontecimento de interesse, e δ_i serve para excluir as observações com censura.

Cox sugere que se trate a função dada em (1.35) como uma função de verosimilhança ordinária, podendo ser maximizada relativamente a $\boldsymbol{\beta}$. Como normalmente acontece, é mais conveniente maximizar o logaritmo da verosimilhança, ou seja,

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left(\boldsymbol{\beta}' \mathbf{x}_i - \log \left(\sum_{j=1}^{n} Y_{ij}\, e^{\boldsymbol{\beta}' \mathbf{x}_j} \right) \right) \tag{1.36}$$

A função de verosimilhança anterior assume que não há 'empates' (*tied data*) nos tempos de sobrevivência, ou seja, que é possível estabelecer uma ordem única desses tempos. Mas em muitos casos a variável aleatória é discreta ou está agrupada (por exemplo, quando o tempo é dado em meses ou anos) e a ordenação dos tempos não é clara, tendo a função de verosimilhança que incluir todas as ordens possíveis. Seguidamente é usada a notação dada por Stepanova e Thomas [5] para simplificar a expressão de $L(\boldsymbol{\beta})$: considere-se a ordenação $t_{(1)} < t_{(2)} < \ldots < t_{(m)}$ dos tempos de sobrevivência e $R(t_{(i)})$ o conjunto de observações em risco em $t_{(i)}$; seja d_i o número de acontecimentos no tempo t_i e seja $R(t_{(i)}; d_i)$ o conjunto de todos os subconjuntos constituídos pelas d_i observações que poderiam ter tido o acontecimento de interesse em $t_{(i)}$; seja $R \in R(t_{(i)}; d_i)$ o conjunto das observações que poderiam ter tido o acontecimento em $t_{(i)}$ e seja $\mathbf{s}_R = \sum_{l \in R} \mathbf{x}_l$ a soma dos vectores das covariáveis das observações em R; denote-se por D_i o conjunto dos indivíduos d_i que tiveram o acontecimento em t_i, e seja $\mathbf{s}_{D_i} = \sum_{l \in D_i} \mathbf{x}_l$ a soma dos vectores de covariáveis destes indivíduos. Deste modo, a função de verosimilhança é dada por:

$$L_{Cox}(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp(\boldsymbol{\beta}' \mathbf{s}_{D_i})}{\left(\sum_{R \in R(t_{(i)};d_i)} \exp(\boldsymbol{\beta}' \mathbf{s}_R)\right)} \tag{1.37}$$

Segundo Miller [4], o denominador da expressão (1.37) pode gerar um número excessivo de combinações possíveis e que pode ser muito pouco eficiente a nível computacional. Por essa razão são consideradas aproximações propostas por Breslow (1974) e Efron (1977), sendo esta última, segundo Allison [2], mais rigorosa à custa de um pouco mais de tempo computacional.

$$L_{Breslow}(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp(\boldsymbol{\beta}' \mathbf{s}_{D_i})}{\left(\sum_{l \in R(t_{(i)})} \exp(\boldsymbol{\beta}' \mathbf{x}_l)\right)^{d_i}} \tag{1.38}$$

$$L_{Efron}(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp(\boldsymbol{\beta}' \mathbf{s}_{D_i})}{\prod_{j=1}^{d_i} \left(\sum_{l \in R(t_{(i)})} \exp(\boldsymbol{\beta}' \mathbf{x}_l) - \frac{j-1}{d_i} \sum_{l \in D_i} \exp(\boldsymbol{\beta}' \mathbf{x}_l)\right)} \tag{1.39}$$

Outra forma de lidar com os empates é considerar o tempo de sobrevivência como uma variável discreta. Cox sugeriu mesmo a substituição de $h(t) = e^{\boldsymbol{\beta}' \mathbf{x}} h_0(t)$ pela expressão de um modelo logístico discreto dado por

$$\frac{h(t)}{1 - h(t)} = e^{\boldsymbol{\beta}' \mathbf{x}} \frac{h_0(t)}{1 - h_0(t)} \tag{1.40}$$

1.7 Diagnóstico dos modelos

Ao utilizar um modelo matemático devemos questionar se este está correctamente ajustado ao problema. No caso dos modelos de *hazards* proporcionais quer-se ver se é verificado o pressuposto de proporcionalidade, se algumas covariáveis requerem algum tipo de transformação ou se há *outliers* (observações com um tempo de sobrevivência muito diferente do esperado) que podem ter impacto indesejado nos resultados.

O resíduo de Cox-Snell é definido por

$$r_{C_i} = \exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_i) \hat{H}_0(t_i) = \hat{H}_i(t_i) = -\log \hat{S}_i(t_i) \tag{1.41}$$

em que, para o i-ésimo indivíduo, se tem que $\hat{\boldsymbol{\beta}}$ e $\mathbf{x}_i$ são os vectores de parâmetros estimados e covariáveis, respectivamente. Para o tempo de sobrevivência observado t_i, tem-se que as

funções $\hat{H}_0$, $\hat{H}_i$ e $\hat{S}_i$ são as funções estimadas de base de *hazard* cumulativa, de *hazard* cumulativa e de sobrevivência, respectivamente. Pode mostrar-se que $-\log S(t_i)$ tem uma distribuição exponencial de média unitária[3], independentemente da forma da função S [17]. Se o modelo estiver correctamente ajustado, a função de sobrevivência estimada será semelhante e terá as mesmas propriedades de $S(t)$. Assim, será de esperar que $-\log \hat{S}(t_i) = r_{C_i}$ tenha também uma distribuição exponencial de média unitária. Para verificar esta propriedade calculam-se estimativas para $\hat{S}(r_{C_i})$, por exemplo através do método de Kaplan-Meier. Conforme referido por Stepanova [5], o ajuste do modelo será tanto maior quanto maior a proximidade do gráfico de $\log(-\log \hat{S}(r_C))$ em função de $\log(r_C)$ a uma recta com declive unitário que passa pela origem.

O resíduo de Schoenfeld é calculado para cada covariável e é especialmente importante no que respeita à investigação da proporcionalidade de *hazard*, de eventuais covariáveis dependentes do tempo ou transformações de covariáveis. Podem ser definidos segundo o vector $\mathbf{r}_{S_i}$ considerando as k covariáveis do modelo para o i-ésimo indivíduo.

$$\mathbf{r}_{S_i} = (r_{S_{i1}}, \ldots, r_{S_{ik}}) \tag{1.42}$$

onde

$$r_{S_{ip}} = x_{ip} - E(x_{ip}|R_i) \quad , \quad p = 1, \ldots, k \tag{1.43}$$

Para um dado indivíduo i, o resíduo de cada covariável é calculado fazendo a diferença entre o valor da covariável x_{ik} e o seu valor esperado, condicionado ao 'conjunto de risco' R_i, isto é, o conjunto de indivíduos que não tiveram o acontecimento até t_i. Farrington [14] refere que os resíduos não devem mostrar nenhum tipo de padrão sistemático se for válido o pressuposto dos *hazards* proporcionais. Caso contrário, se ao longo do tempo esse

[3] Assumindo a existência da função de sobrevivência inversa e considerando a variável aleatória positiva T com função de sobrevivência S (contínua) e $Y = -\log S(T)$ tem-se que

$$\begin{aligned} P(Y > y) &= P(-\log S(T) > y) = P(S(T) < \exp(-y)) \\ &= P(T > S^{-1}(\exp(-y))) = S(S^{-1}(\exp(-y))) \\ &= \exp(-y) \end{aligned}$$

pressuposto não se verificar, o(s) gráfico(s) dos resíduos deverão reflectir uma tendência positiva (negativa) conforme aumente (diminua) a razão de *hazard* .

1.8 Medidas de discriminação: curvas ROC

Quando a variável de saída é dicotómica (por exemplo: 0/1, sim/não, bom/mau, etc) e as previsões são probabilidades de ocorrência de um acontecimento, os modelos podem ser avaliados segundo dois conceitos gerais: discriminação e calibração.

A discriminação refere-se à capacidade do modelo distinguir correctamente as classes de saída enquanto que a calibração avalia a proximidade numérica entre as probabilidades previstas e as reais. Apesar de, num modo geral, um modelo com boa discriminação possuir boa calibração e vice-versa, D'Agostino [13] refere que é sempre preferível obter um modelo com bom poder discriminante, uma vez que este pode sempre ser recalibrado.

Uma das medidas de discriminação mais usadas para um modelo é a área debaixo da curva ROC (*Receiver Operating Characteristic*). Vejamos como construir uma curva deste tipo para um modelo de regressão.

		Estado classificado	
		$\oplus$	$\ominus$
Estado	$\oplus$	a	b
real	$\ominus$	c	d

Tabela 1.1: Matriz entre os estados reais e classificados.

Suponhamos que temos n indivíduos. Através da regressão é possível estimar e ordenar as probabilidades de ocorrência de um determinado acontecimento $(Q_1, Q_2, \ldots, Q_n)$ de modo que $Q_i \leq Q_{i+1}$, para todo $i = 1, \ldots, n-1$. Sob a regra de se classificar como positivos ($\oplus$) todos os que verificarem $Q_i > Q^*$, para um valor Q^* (*cut-off*), e como negativos ($\ominus$) os que não verificarem a regra, pode apresentar-se o resultado segundo uma matriz 2×2 como a da Tabela 1.1.

Daqui pode calcular-se a *sensibilidade* dada por $\frac{a}{a+b}$ e a *especificidade* dada por $\frac{d}{c+d}$, que não é mais do que a razão dos que se prevêm positivos relativamente ao total de positivos e

a razão dos que se esperam negativos relativamente ao total de negativos, respectivamente. Se seleccionarmos todos os valores possíveis de *cut-off* e desenharmos o gráfico da sensibilidade em função de $1-$ especificidade obtêm-se curvas como as da Figura 1.5.

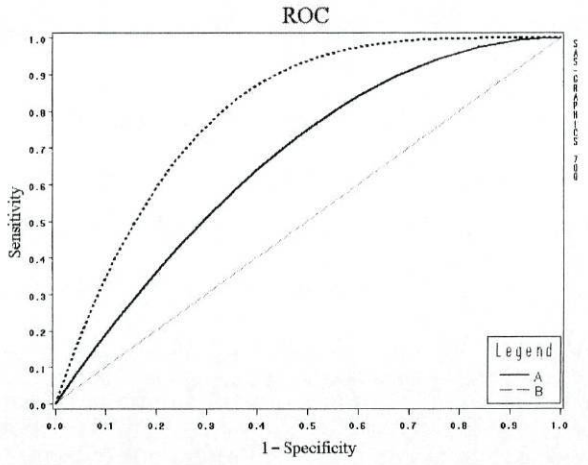

Figura 1.5: Exemplos de curvas ROC *Receiver Operating Characteristic* utilizadas na comparação do poder discriminante de modelos. No exemplo da figura, um modelo que estivesse representado pela curva B teria maior poder discriminante que um outro representado pela curva A.

A área debaixo desta curva, conhecida por AUROC (*Area Under ROC*) ou estatística C é uma medida de discriminação e pode ser interpretada como a probabilidade estimada da classificação positiva ser maior que a classificação negativa, isto é,

$$\text{estatística } C = \hat{P}(Q_{\oplus} > Q_{\ominus}) \tag{1.44}$$

onde $Q_{\oplus}$ são as probabilidades estimadas dos que tiveram os acontecimentos e $Q_{\ominus}$ são as probabilidades dos que não tiveram os acontecimentos.

O valor da estatística C pode variar entre $0,5$, no caso de não haver discriminação, e 1, no caso de discriminação perfeita, e está relacionado unicamente com a ordenação das probabilidades previstas.

Esta forma de avaliar o poder discriminante de um modelo também pode ser aplicado aos modelos de *hazards* proporcionais de Cox. A maiores probabilidades de ocorrência de determinado acontecimento correponderão certamente menores tempos de sobrevivência, daí ser possível estabelecer uma ordem dos valores previstos. Para se obter algum tipo de conclusão será conveniente comparar com outros métodos (p.e. regressão logística) com a

condição de não considerar nessa análise comparativa os dados censurados (característica dos modelos de análise de sobrevivência).

Capítulo 2

Aplicação dos modelos de análise de sobrevivência

2.1 Dados da análise

Os dados seleccionados para esta análise consistem em cerca de 58.000 transacções de crédito pessoal de uma instituição bancária de referência, subdivididas em dois conjuntos de treino e teste na proporção 50/50. A informação recolhida até Ago'06 (transacções iniciadas no período Ago'03 - Dez'05) contempla características relativas a cada transacção e ao cliente. Parte dessa informação será utilizada como variáveis nos modelos deste trabalho, a saber, o prazo original (entre 12 e 36 meses), o montante pedido, a idade do cliente, tempo de permanência no banco (esta variável binária apenas serve para diferenciar se se trata de um cliente recente ou não), indicador de telefone e, finalmente, um *score* interno do banco (traduz o perfil de risco do cliente ou a sua probabilidade de incumprir baseado num modelo comportamental).

Outro tipo de informação que é importante para qualquer modelo de análise de sobrevivência é o já referido tempo de sobrevivência que vai desde o início da transacção de crédito até ao primeiro incumprimento (conceito a definir), ou até ao pagamento antecipado (se houver), ou até se deixar de ter mais informação (censura), ou, simplesmente, até ao final do prazo.

Neste trabalho, uma transacção de crédito será considerada com incumprimento quando

Descritivo	Nome	Designação
Prazo original	prazo	x_1
Montante original	montante	x_2
Idade do cliente	idade	x_3
Antiguidade na instituição (variável binária)	antiguidade	x_4
Indicador de telefone (variável binária)	telefone	x_5
Score comportamental interno do banco	score	x_6

Tabela 2.1: Descrição dos dados utilizados.

tiver três ou mais prestações em atraso consecutivas, ainda que entretanto essa situação seja regularizada. Este é o critério elegido por Thomas [11], também mencionado em bastante bibliografia sobre o assunto e adoptado por inúmeras instituições financeiras. Quanto ao pagamento antecipado, este será considerado quando a transacção de crédito for totalmente liquidada (e não parcialmente).

Como já foi referido, a análise de sobrevivência permite a introdução de dados censurados. Neste caso tratam-se de transacções de crédito de clientes que até à data do fim do estudo ainda não tinham acabado de pagar o empréstimo sem que tivessem tido algum dos acontecimentos: pagamento antecipado ou incumprimento.

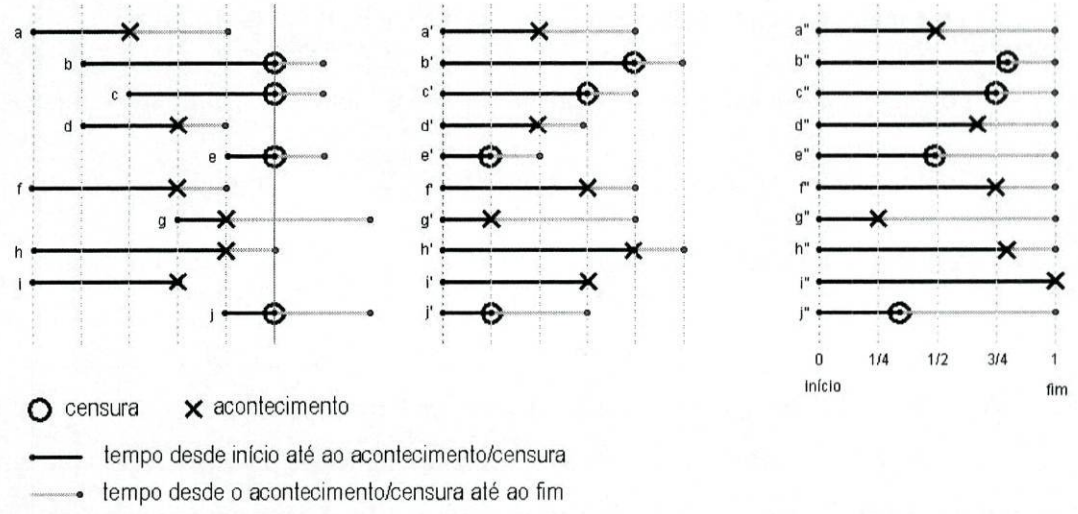

Figura 2.1: Representação de tempos de sobrevivência das transacções de crédito. O primeiro esquema mostra que podem ter inícios em diferentes alturas do tempo, mas que podem ser comparadas assumindo o mesmo instante inicial, conforme se mostra no segundo esquema. O terceiro esquema representa uma alteração na escala do tempo, visando homogeneizar os diferentes prazos, de modo a terem o mesmo instante inicial e final (teórico).

O facto das transacções terem inícios diferentes (primeiro esquema da Figura 2.1) não

constitui problema porque pode sempre assumir-se que começam ao mesmo tempo (segundo esquema). No entanto, há uma particularidade diferente dos dados tradicionais de análise de sobrevivência. As transacções de crédito pessoal têm um prazo associado, isto é, teoricamente sabe-se quando é que eles vão terminar. Uma forma de as poder comparar, apesar de haver diferentes prazos, foi de proceder à sua homogeneização numa mesma escala de tempo, de modo a que todas tivessem o mesmo início e o mesmo fim (terceiro esquema). Basta para isso dividir o tempo de sobrevivência de cada empréstimo pelo seu prazo original, cujo resultado deverá ser um número entre 0 e 1 (por conveniência pode ser multiplicado por 100). Nos resultados apresentados no decorrer deste trabalho a escala de tempo considerada será entre 0 e 100 para indicar o início e fim (teórico) do empréstimo, respectivamente. Por exemplo, se num empréstimo com prazo de 2 anos se registar incumprimento ao fim de 1 ano, então o incumprimento ficará registado em $t = 50$; se, em vez disso, tiver antecipado o pagamento do empréstimo ao fim de 6 meses, então o pagamento antecipado registar-se-á em $t = 25$.

No fim do estudo pode assim observar-se um dos seguintes casos em cada transacção:

- Teve incumprimento pela primeira vez numa dada altura do tempo;
- Foi pago antecipadamente (sem incumprimento) numa dada altura do tempo;
- Foi pago (sem incumprimento nem pagamento antecipado) no final do prazo;
- Ainda não foi totalmente pago (sem ter tido incumprimento nem pagamento antecipado).

A razão de se distinguir os acontecimentos pagamento antecipado e incumprimento é óbvia: ambos têm impacto negativo sobre a rendibilidade das operações, mas, de um modo geral, o segundo acarreta muito mais prejuízo para o banco do que o primeiro.

Na análise de sobrevivência uma forma de lidar com dois (ou mais) acontecimentos de interesse no mesmo conjunto de dados é considerar separadamente a análise de cada um, tomando o(s) outro(s) como censura, conforme sugestão de Stepanova e Thomas [5]. Assim, no caso do incumprimento, consideram-se censurados os tempos de sobrevivência relativos a pagamento antecipado e os que já estariam censurados à partida. O procedimento é análogo

no caso do pagamento antecipado. Sem grandes modificações a nível técnico é relativamente simples fazer esta dupla análise aos dados, incluindo um indicador (*flag*) em cada observação que identifica se o tempo de sobrevivência está censurado ($flag = 0$), se refere a incumprimento ($flag = 1$), ou pagamento antecipado ($flag = 2$).

2.2 Estimativas não paramétricas das funções de sobrevivência

Fazendo uma análise aos dados, nomeadamente estimando a função de sobrevivência através do método Kaplan-Meier é possível obter representações de $S(t)$ para os acontecimentos pagamento antecipado e incumprimento.

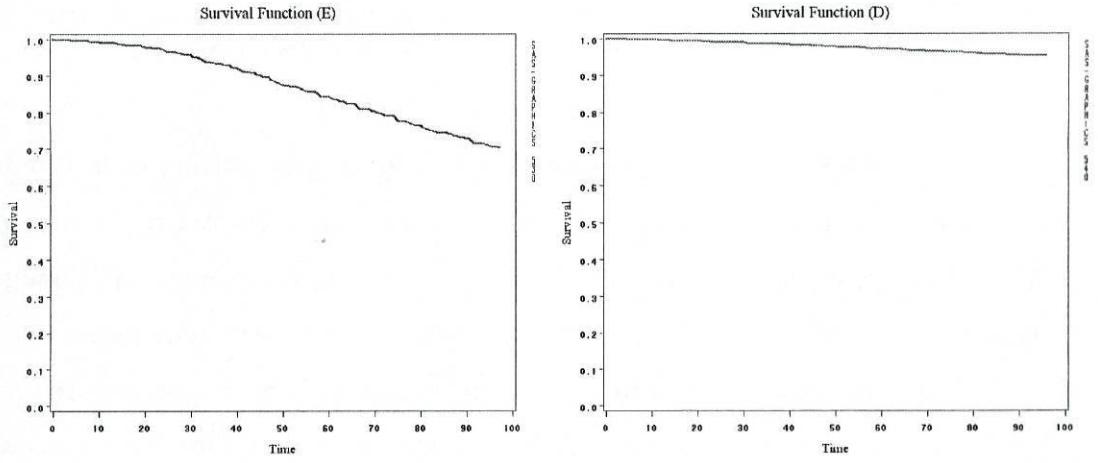

Figura 2.2: Funções de sobrevivência estimadas pelo método Kaplan-Meier dos acontecimentos pagamento antecipado (esquerda) e incumprimento (direita).

Analisando os gráficos pode verificar-se que há mais clientes que antecipam o pagamento do que aqueles que incumprem. Nota-se ainda que o pagamento antecipado acontece com mais frequência nos últimos $\frac{3}{4}$ do tempo de vida da transacção, ou seja, quando $t > 25$. No caso do incumprimento, a ocorrência deste acontecimento aparenta ser constante ao longo do tempo. Apesar dos gráficos baseados nas funções de sobrevivência serem bastante úteis, também não é de desprezar os gráficos das funções de *hazard*, pois permitem identificar propriedades interessantes dos acontecimentos em questão.

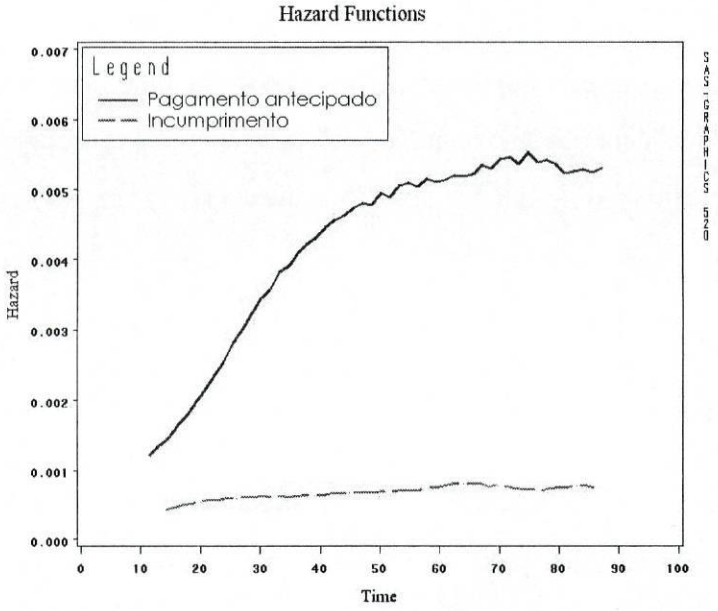

Figura 2.3: Funções *hazard* alisadas por um processo de médias móveis conhecido por *kernel smoothing* (descrito por Ramlau-Hansen em 1983) para os dois tipos de acontecimentos considerados.

O gráfico das funções de *hazard* permitem investigar o risco instantâneo de determinado acontecimento ocorrer. Na Figura 2.3 constata-se que o risco do pagamento antecipado é, de um modo geral, crescente ao longo do tempo, com uma ligeira quebra na parte final da transacção de crédito. No que respeita ao risco do incumprimento, este parece ser apenas ligeiramente crescente com o decorrer do tempo, contrariando neste caso a noção preconcebida de que "se uma transacção de crédito vai correr mal, então ela vai correr mal cedo"[7]. Segundo este gráfico o risco de ocorrência de incumprimento não é mais elevado na fase inicial do empréstimo, o que sugere haver boas decisões de crédito nas transacções de crédito consideradas para este trabalho. Se o risco de incumprimento fosse mais elevado na fase inicial, isso poderia ser indicador de que o banco estaria a decidir mal o seu crédito e que os seus modelos de classificação poderiam não estar a discriminar bem.

Um primeiro passo na análise dos dados é encontrar a distribuição do tempo de sobrevivência. A relação entre as covariáveis explicativas e esse tempo pode ser investigada preliminarmente através de subgrupos de covariáveis. Consideremos, por exemplo, as transacções de crédito divididas em dois subgrupos relativamente ao prazo original. O primeiro sendo constituído por transacções com prazo inferior a 30 meses, e as restantes no segundo sub-

grupo (poder-se-ia ter escolhido outros subgrupos quaisquer).

O efeito em termos de funções de sobrevivência e a análise da proporcionalidade de *hazards* pode ser observada graficamente.

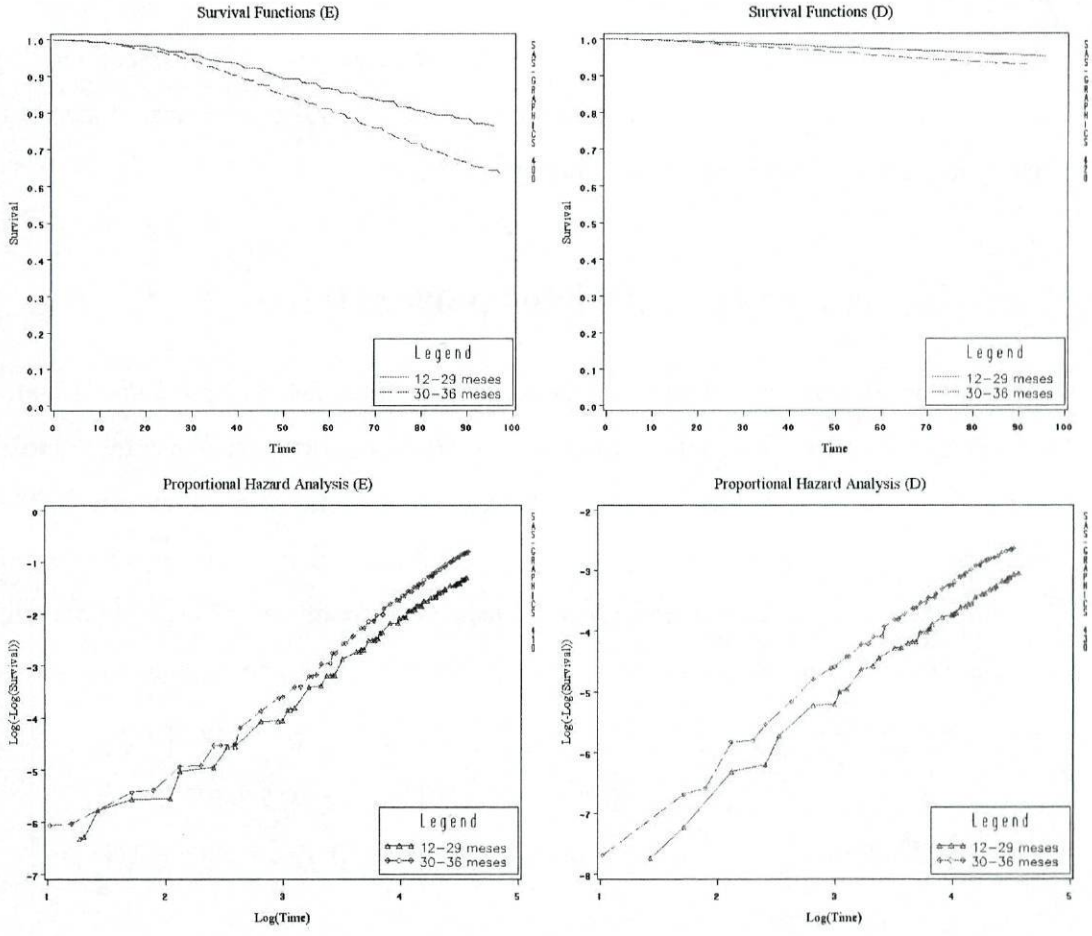

Figura 2.4: Em cima: Funções de sobrevivência estimadas pelo método Kaplan-Meier dos acontecimentos pagamento antecipado (esquerda) e incumprimento (direita) para os dois subgrupos. Em baixo: Gráficos de $\log(-\log(S))$ em função de $\log(t)$ para pagamento antecipado (esquerda) e incumprimento (direita) para os dois subgrupos.

Os gráficos da Figura 2.4 referentes às funções de sobrevivência (em cima) evidenciam que as transacções com prazo inferior a 30 meses tendem a pagar antecipadamente e a incumprir menos do que as restantes. Por outro lado, se se quiser verificar a proporcionalidade de *hazards*, os gráficos de $\log(-\log S(t))$ em função de $\log(t)$ (em baixo) desses subgru-

pos deverão ser linhas paralelas[1]. O paralelismo dos gráficos é mais evidente no caso do incumprimento do que no caso do pagamento antecipado.

Se o modelo de Weibull for apropriado (cuja função de sobrevivência foi já dada na expressão (1.14)), tem-se que $\log(-\log S(t)) = \gamma \log \lambda + \gamma \log t$, ou seja, os gráficos de $\log(-\log S(t))$ em função de $\log(t)$ deverão ser linhas rectas. No pagamento antecipado e no incumprimento observa-se um comportamento aproximadamente linear, donde se conclui que o modelo de Weibull poderá ser uma opção válida.

2.3 Regressão com modelos paramétricos

A análise de sobrevivência permite também efectuar regressão com base num conjunto de covariáveis explicativas. Além das diferenças em termos de sobrevivência entre grupos, o programa SAS® permite estudar o efeito das covariáveis no tempo de sobrevivência. Mais propriamente, são calculadas estatísticas de χ^2 de Wald[2] que testam a hipótese nula do parâmetro ser zero. Ou, por outras palavras, que a variável correspondente não tem grande efeito sobre o tempo de sobrevivência, dado que as outras variáveis estão no modelo. A Tabela 2.2 apresenta esses resultados para o conjunto de covariáveis $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6)'$.

	Pagamento antecipado		**Incumprimento**	
Effect	Wald Chi-Square	Pr > ChiSq	Wald Chi-Square	Pr > ChiSq
x_1 (prazo)	496,0894	<,0001	137,7454	<,0001
x_2 (montante)	25,4640	<,0001	15,8935	<,0001
x_3 (idade)	184,8070	<,0001	40,1365	<,0001
x_4 (antiguidade)	8,9992	0,0027	2,9499	0,0859
x_5 (telefone)	2,6863	0,1012	0,0085	0,9265
x_6 (score)	106,5811	<,0001	485,5781	<,0001

Tabela 2.2: Teste χ^2 de Wald para verificar o efeito das covariáveis. Esta estatística testa a hipótese nula de cada coeficiente ser zero, calculando o quadrado do quociente entre o parâmetro estimado e o seu erro estimado.

[1]Basta pensar que para dois subgrupos A e B em que se verifique o pressuposto de *hazards* proporcionais tem-se que $S_A(t) = [S_B(t)]^\alpha$, donde se obtém naturalmente que $\log(-\log S_A(t)) = \log(\alpha) + \log(-\log S_B(t))$.

[2]O teste χ^2 de Wald também permite testar a hipótese nula global $H_0 : \boldsymbol{\beta} = \mathbf{0}$. Sob condições gerais, esta estatística tem uma distribuição assimptótica de χ^2 com k graus de liberdade (sendo k a dimensão de $\boldsymbol{\beta}$), dada a hipótese nula [22, SAS® PHREG Procedure]: $\chi^2_{Wald} = \hat{\boldsymbol{\beta}}'[V(\hat{\boldsymbol{\beta}})]^{-1}\hat{\boldsymbol{\beta}}$.

A hipótese nula deve ser rejeitada quando o nível de significância é inferior a $0,05$ (valor usualmente utilizado). Analisando os valores pode concluir-se que no pagamento antecipado a covariável x_5 não vai de encontro a este critério, enquanto que no incumprimento são as variáveis x_4 e x_5. Deste modo, estas covariáveis não farão parte da estimação dos parâmetros nos modelos apresentados.

Além de identificar as variáveis mais significativas, o programa SAS® permite obter estimativas para os parâmetros das covariáveis e dos modelos através de máxima verosimilhança. A Tabela 2.3 mostra os resultados obtidos com modelo Weibull para o pagamento antecipado e o incumprimento (no *output* do SAS® têm-se as designações $\mu = Intercept$ e $\sigma = Scale$).

Pagamento antecipado

Parameter	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
μ (*Intercept*)	5,0908	0,0120	5,0672	5,1144	178967	<,0001
β_1 (prazo)	-0,0208	0,0009	-0,0226	-0,0190	496,21	<,0001
β_2 (montante)	-0,0097	0,0019	-0,0134	-0,0059	25,76	<,0001
β_3 (idade)	0,0088	0,0006	0,0075	0,0100	186,93	<,0001
β_4 (antiguidade)	-0,1638	0,0547	-0,2709	-0,0566	8,98	0,0027
β_6 (score)	0,0025	0,0002	0,0020	0,0030	108,12	<,0001
σ (*Scale*)	0,5722	0,0066	0,5595	0,5852		
d (*Shape*)	1,7475	0,0200	1,7087	1,7873		

Incumprimento

Parameter	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
μ (*Intercept*)	6,5206	0,0643	6,3945	6,6467	10267,9	<,0001
β_1 (prazo)	-0,0312	0,0027	-0,0364	-0,0260	136,36	<,0001
β_2 (montante)	0,0274	0,0069	0,0139	0,0409	15,80	<,0001
β_3 (idade)	0,0114	0,0018	0,0077	0,0150	38,05	<,0001
β_6 (score)	0,0157	0,0007	0,0143	0,0171	493,70	<,0001
σ (*Scale*)	0,7039	0,0186	0,6684	0,7413		
d (*Shape*)	1,4207	0,0375	1,3490	1,4962		

Tabela 2.3: Estimativas com SAS® dos parâmetros para o modelo de Weibull relativamente ao pagamento antecipado e incumprimento.

Os valores apresentados incluem estimativas dos parâmetros e respectivos erros associados, limites para intervalos de confiança dos parâmetros e estatísticas χ^2 que testam a hipótese nula do valor do parâmetro ser zero. O valor da coluna *Chi-Square* representa assim a significância das covariáveis e é obtido calculando o quadrado do quociente entre o parâmetro estimado e o respectivo erro.

Pode confirmar-se que no pagamento antecipado as covariáveis x_1, x_3, x_6, x_2 e x_4 (por ordem decrescente de significância) têm níveis de significância inferiores a $0,05$. Isto é também verificado no caso do incumprimento, mas a ordem de significância das variáveis é diferente, a saber, x_6, x_1, x_3 e x_2 (por ordem decrescente de significância).

	Pagamento antecipado				**Incumprimento**			
Parameter	Weibull	Gama	Log-logistic	Log-normal	Weibull	Gama	Log-logistic	Log-normal
μ (*Intercept*)	5,0908	5,1004	4,9482	5,0909	6,5206	7,1222	6,4370	7,0030
β_1 (prazo)	-0,0208	-0,0209	-0,0209	-0,0205	-0,0312	-0,0335	-0,0316	-0,0330
β_2 (montante)	-0,0097	-0,0093	-0,0095	-0,0085	0,0274	0,0288	0,0276	0,0285
β_3 (idade)	0,0088	0,0093	0,0091	0,0098	0,0114	0,0119	0,0115	0,0118
β_4 (antiguidade)	-0,1638	-0,1660	-0,1645	-0,1639				
β_6 (score)	0,0025	0,0027	0,0027	0,0029	0,0157	0,0185	0.0161	0,0178
σ (*Scale*)	0,5722	0,7988	0,5272	1,0500	0,7039	2,0284	0,6845	1,5847
d (*Shape*)	1,7475	0,4958			1,4207	-0,4520		
Log Likelihood	-16636	-16600	-16605	-16645	-5079	-5038	-5072	-5041

Tabela 2.4: Estimativas com SAS® dos parâmetros para os modelos de Weibull, Gama, Loglogístico e Lognormal dos acontecimentos pagamento antecipado e incumprimento.

Os resultados obtidos com outros modelos foram bastante semelhantes em termos do valor dos parâmetros, conforme a Tabela 2.4. Os modelos que melhor se ajustam aos dados têm menores valores absolutos de *Loglikelihood* dada na expressão (1.33). No entanto, quando se verifica uma proximidade muito grande entre os modelos, a escolha normalmente recai sobre aquele que é matematicamente mais simples. Segundo o critério de simplicidade é preferível o modelo Weibull até por ser um modelo de *hazards* proporcionais.

Outra forma de avaliar o ajuste do modelo tem a ver com o resíduo de Cox-Snell dado em (1.41). Conforme referido anteriormente, os modelos que melhor se ajustam aos dados são aqueles cujo gráfico de $\log \hat{H}(r_C) = \log(-\log \hat{S}(r_C))$ em função de $\log(r_C)$ se aproxima de uma recta que passa pela origem com declive unitário. Note-se que as figuras referentes aos resíduos de Cox-Snell da Figura 2.5 revelam um desvio relativamente a essa recta de referência. Este desvio pode ser justificado pela diminuição do risco de pagamento antecipado e incumprimento quando a transacção de crédito se encontra próxima de chegar ao fim do prazo (visível na Figura 2.3). Não é pois de estranhar que os modelos paramétricos acabem por não se ajustar muito bem a esta 'quebra' final de tendência de monotonia . No entanto, cerca de 99% das observações têm valores de $\log(r_C)$ superiores a -5 (apenas um número

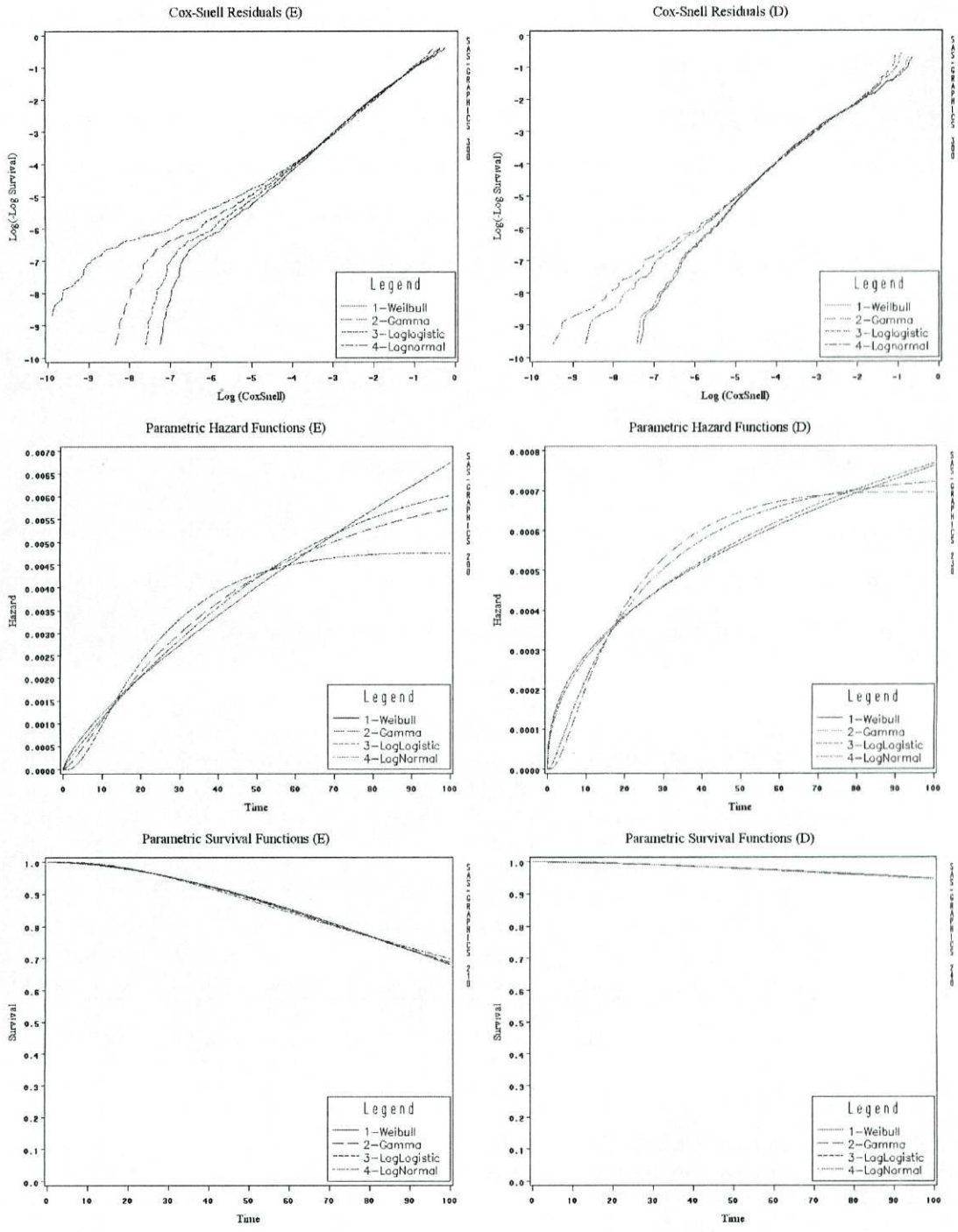

Figura 2.5: Gráficos relativos aos modelos paramétricos Weibull, Gama, Log-logístico e Lognormal relativamente aos acontecimentos pagamento antecipado (esquerda) e incumprimento (direita). Em cima: Resíduos de Cox-Snell; No centro: Funções de *hazard* (note-se que o eixo das ordenadas são diferentes nos dois acontecimentos); Em baixo: Funções de sobrevivência.

reduzido de observações é responsável por esse desvio), onde o ajuste dos modelos à recta referida é bastante melhor.

Com base nos parâmetros estimados destes quatro modelos é possível ainda obter gráficos para as funções base de *hazard* e de sobrevivência também apresentadas na Figura 2.5. As diferenças entre os modelos são bastante mais visíveis nos gráficos das funções de *hazard* do que nas funções de sobrevivência, em que quase não é possível distinguí-las.

2.4 Regressão com o modelo de *hazards* proporcionais de Cox

O modelo de *hazards* proporcionais de Cox é um modelo semi-paramétrico e os parâmetros estimados por verosimilhança parcial referem-se apenas às covariáveis do modelo (como foi visto anteriormente, este método não depende da função base de *hazard*).

Pagamento antecipado

Parameter	Estimate	Standard Error	Chi-Square	Pr>ChiSq	Hazard Ratio	95% Haz. Ratio Conf. Limits	
β_1 (prazo)	0,0314	0,00165	361,51	<,0001	1,032	1,029	1,035
β_2 (montante)	0,0169	0,00332	26,10	<,0001	1,017	1,010	1,024
β_3 (idade)	-0,0151	0,00110	187,12	<,0001	0,985	0,983	0,987
β_4 (antiguidade)	0,2542	0,09443	7,24	0,0071	1,289	1,072	1,552
β_6 (score)	-0,0041	0,00042	95,20	<,0001	0,996	0,995	0,997

Incumprimento

Parameter	Estimate	Standard Error	Chi-Square	Pr>ChiSq	Hazard Ratio	95% Haz. Ratio Conf. Limits	
β_1 (prazo)	0,0400	0,00383	109,06	<,0001	1,041	1,033	1,049
β_2 (montante)	-0,0392	0,00974	16,17	<,0001	0,962	0,943	0,980
β_3 (idade)	-0,0161	0,00258	38,87	<,0001	0,984	0,979	0,989
β_6 (score)	-0,0221	0,00085	677,00	<,0001	0,978	0,977	0,980

Tabela 2.5: Estimativas com SAS® dos parâmetros do modelo *hazards* proporcionais para pagamento antecipado (tabela superior) e incumprimento (tabela inferior). As três colunas da direita dizem respeito à chamada *hazard ratio* que é dada por e^{β_i}, onde β_i é o valor do parâmetro estimado, e respectivos limites inferior e superior com 95% de confiança.

À semelhança dos resultados obtidos com os modelos paramétricos, na Tabela 2.5 obtida pelo SAS®[3] são apresentadas estimativas dos parâmetros e respectivos erros associados

[3] Apenas está apresentado parte do *output* exibido pelo *software*. A restante informação respeita a estatísticas dos dados e convergência dos algoritmos, que têm interesse meramente técnico.

e estatísticas χ^2 sob a hipótese nula de cada parâmetro ser zero, onde se constata que a ordem de significância das covariáveis se mantém. As magnitudes dos parâmetros estimados são normalmente pouco informativas, mas uma transformação simples permite dar uma interpretação bastante útil e intuitiva. Sendo β_i o parâmetro associado à i-ésima covariável, o valor e^{β_i} está representado na coluna *Hazard ratio* e pretende indicar o aumento/diminuição de risco por acréscimo de uma unidade na covariável. Por exemplo, o aumento de 1 unidade na covariável x_1 (aumento de 1 mês no prazo) está associado a um aumento de $(e^{0,0314} - 1) \approx 3,2\%$ do risco (ou *hazard*) da ocorrência de pagamento antecipado, mantendo as outras covariáveis constantes.

O sinal dos parâmetros das covariáveis também dá informação acerca do tempo de sobrevivência. Um sinal positivo (negativo) indica que um aumento da covariável conduz a menores (maiores) tempos de sobrevivência.

Comparativamente aos parâmetros estimados com os modelos paramétricos, os respectivos sinais são recíprocos. Isso não é surpreendente atendendo à própria formulação dos modelos (conforme as expressões (1.25) e (1.28)). Nos modelos paramétricos ou de vida acelerada os parâmetros estão num formato de log-tempo, enquanto nos *hazards* proporcionais o formato é log-*hazard*.

	Pagamento antecipado			**Incumprimento**		
Parâmetro	Weibull paramétrico (Log-Tempo)	Weibull paramétrico (Log-Hazard)	Cox semi-paramétrico (Log-Hazard)	Weibull paramétrico (Log-Tempo)	Weibull paramétrico (Log-Hazard)	Cox semi-paramétrico (Log-Hazard)
β_1 (prazo)	-0,0208	0,0364	0,0314	-0,0312	0,0443	0,0400
β_2 (montante)	-0,0097	0,0170	0,0169	0,0274	-0,0389	-0,0392
β_3 (idade)	0,0088	-0,0154	-0,0151	0,0114	-0,0162	-0,0161
β_4 (antiguidade)	-0,1638	0,2863	0,2542			
β_6 (score)	0,0025	-0,0044	-0,0041	0,0157	-0,0223	-0,0221

Tabela 2.6: Comparação entre parâmetros estimados pelo modelo Weibull (paramétrico) e pelo modelo de Cox (semi-paramétrico), ambos *hazards* proporcionais, para pagamento antecipado e incumprimento.

Dos modelos paramétricos experimentados, o de Weibull é o único que também é um modelo de *hazards* proporcionais. A Tabela 2.6 permite avaliar a semelhança entre os valores dos parâmetros estimados através do modelo de Weibull e do modelo semi-paramétrico de *hazards* proporcionais de Cox. No entanto, as estimativas obtidas pelo modelo Weibull necessitam de uma transformação para serem directamente comparáveis com as estimativas

do modelo semi-paramétrico. Essa transformação consiste em fazer $\beta_i^* = -\frac{\beta_i}{\sigma}$, sendo β_i e σ parâmetros estimados pelo modelo de Weibull.

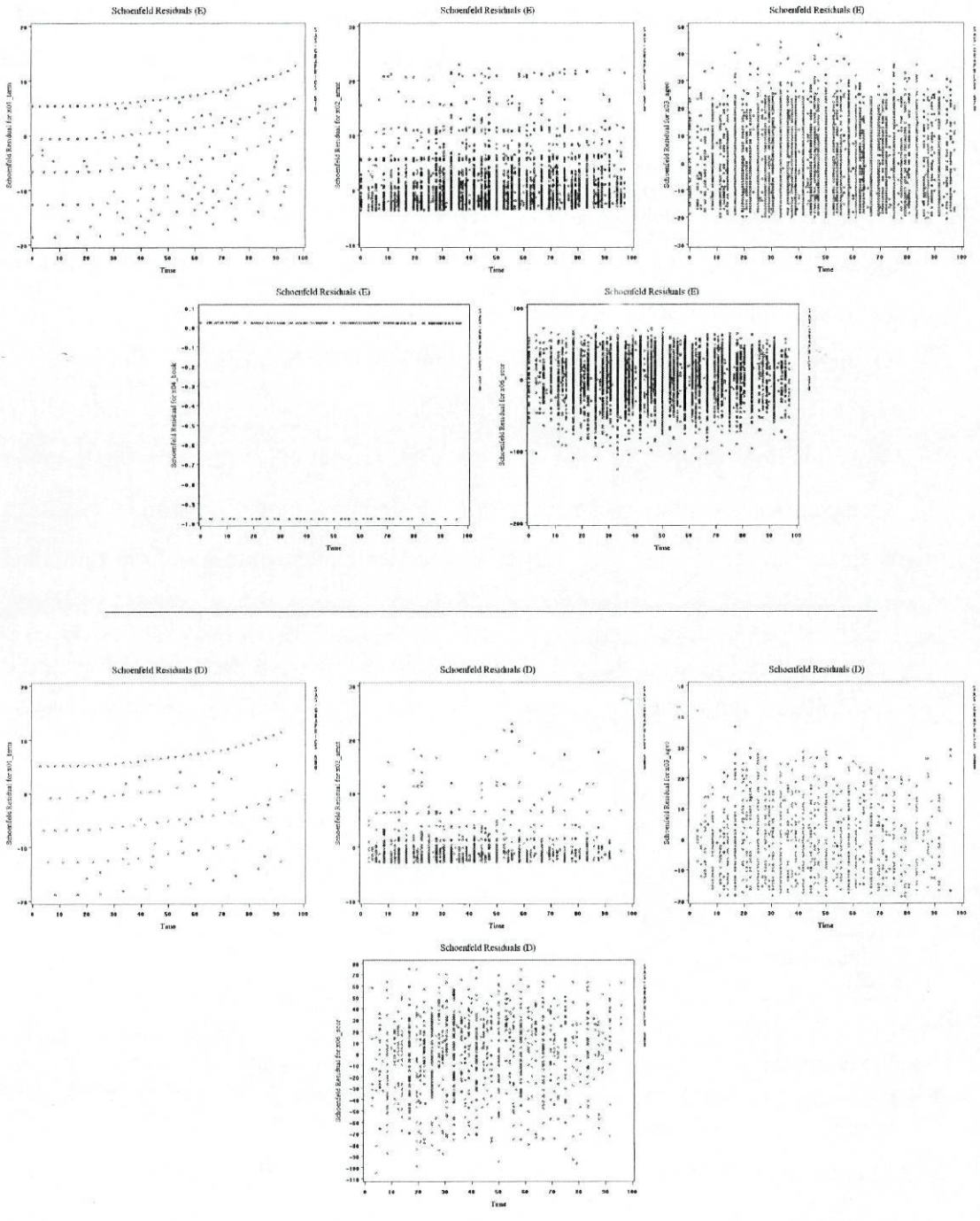

Figura 2.6: Resíduos de Schoenfeld das covariáveis x_1, x_2, x_3, x_4 e x_6 relativamente ao pagamento antecipado (grupo superior) e das variáveis x_1, x_2, x_3 e x_6 relativamente ao incumprimento (grupo inferior).

Os resíduos de Schoenfeld permitem detectar possíveis desvios do pressuposto de *hazards* proporcionais e são calculados apenas para as observações não censuradas. Examinando os gráficos de r_S estes não devem apresentar nenhum tipo de padrão sistemático. Se não se verificar esse pressuposto e se a razão de *hazard* se alterar ao longo do tempo, esse efeito será reflectido no gráfico destes resíduos.

Observando os residúdos de Schoenfeld da Figura 2.6 pode constatar-se que a variável x_4 apresenta um aspecto distintamente diferente da usual nuvem de pontos (característica das variáveis contínuas), mas isso decorre unicamente por se tratar de uma variável binária.

Na mesma figura pode contatar-se que a covariável x_1 apresenta uma tendência sistemática positiva nos dois tipos de acontecimentos, sugerindo que esta covariável esteja de alguma forma relacionada com o tempo.

Já anteriormente foi referido que o modelo de Cox pode utilizado em casos de não proporcionalidade de *hazards*. Acontece que um modelo de *hazards* proporcionais assume que o efeito de cada covariável é o mesmo em todas as alturas do tempo. Allison [2] refere que esse pressuposto dificilmente é verificado e uma ou mais covariáveis têm de facto algum tipo de interacção com o tempo. Quando se utiliza um modelo de *hazards* proporcionais em que o pressuposto seja violado numa dada covariável (suprimindo a sua interacção com o tempo), o parâmetro estimado acaba por reflectir um efeito 'médio' dessa interacção com o tempo. Segundo o mesmo autor, isso não constitui um caso assim tão problemático, uma vez que é prática comum ao efectuar-se uma regressão não considerar as dependências do tempo. No entanto, em casos em que a interacção com o tempo é muito forte, capaz de produzir resultados substancialmente diferentes, ou em casos em que o investigador esteja explicitamente interessado nessa interacção, torna-se necessário levar em consideração essa dependência.

Num exemplo dado por Stepanova, caso o modelo só tivesse uma covariável, z_1, poder-se-ia averiguar se esta variável estaria relacionada com o tempo incluindo no modelo uma outra variável, $z_2 = z_1 t$, representando a interacção da variável z_1 com o tempo. Neste caso particular, o modelo dado por $h(t) = e^{\beta_1 z_1} h_0(t)$ seria alterado para $h(t) = e^{\beta_1 z_1 + \beta_2 z_2} h_0(t) = e^{(\beta_1 + \beta_2 t) z_1} h_0(t)$. Se o parâmetro estimado referente à variável dependente do tempo, β_2, fosse significante, então o pressuposto dos *hazards* proporcionais não se verificaria.

2.5 *Hazards* proporcionais e regressão logística

Tradicionalmente a regressão logística é utilizada para estimar a probabilidade de ocorrência de determinado acontecimento. No que diz respeito à decisão de crédito, o acontecimento que tem merecido maior atenção é o incumprimento, sendo a probabilidade de incumprimento o principal indicador de risco dos clientes.

O modelo logístico relaciona a probabilidade de incumprir, p, com uma combinação linear de variáveis $\mathbf{x} = (x_1, x_2, \ldots, x_k)'$ e parâmetros $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$, através da expressão

$$\log\left(\frac{p}{1-p}\right) = \mu + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{2.1}$$

ou, equivalentemente,

$$p = \frac{\exp(\mu + \boldsymbol{\beta}'\mathbf{x})}{1 + \exp(\mu + \boldsymbol{\beta}'\mathbf{x})} \tag{2.2}$$

A regressão logística tenta prever a probabilidade de um acontecimento ocorrer, enquanto que o modelo de *hazards* proporcionais estima o tempo de sobrevivência, isto é, o tempo de ocorrência do acontecimento. Em todo o caso é sempre possível estabelecer uma ordenação das observações, quer seja através probabilidade de ocorrência de um acontecimento, quer seja do tempo estimado de sobrevivência. Espera-se, por exemplo, que a maiores probabilidades de incumprimento correspondam menores tempos de sobrevivência (analogamente para pagamento antecipado). Deste modo é possível comparar estes modelos relativamente ao seu poder discriminante entre as classes 'bom' e 'mau'. Para cada um dos acontecimentos (pagamento antecipado e incumprimento) estas classes servem apenas para classificar de 'bons' o facto de não terem o respectivo acontecimento e de 'maus' caso contrário.

Note-se que os dados censurados não são tratáveis pela regressão logística usual, ao contrário do que acontece nos modelos de análise de sobrevivência, pelo que se optou por excluí-los desta análise comparativa. Seguidamente fez-se uma partição do conjunto de dados na proporção $60/40$ para os chamados conjuntos de treino e de teste. O primeiro serviu para estimar os parâmetros constituintes dos modelos logístico e de *hazards* proporcionais e o segundo serviu para testar o desempenho desses modelos com base nos parâmetros entretanto estimados.

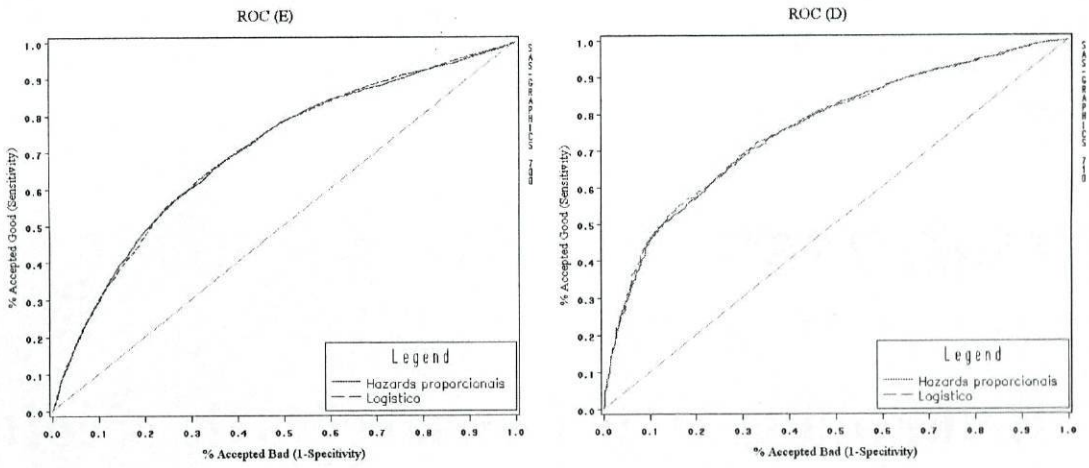

Figura 2.7: Curvas ROC relativamente ao pagamento antecipado (esquerda) e incumprimento (direita). A maior proximidade da linha diagonal é indicador de reduzido poder discriminante dos modelos.

Uma forma simples de avaliar e comparar o desempenho destes modelos é através das curvas ROC, cuja descrição foi dada no capítulo anterior e se apresentam na Figura 2.7.

Neste caso, o poder discriminante dos modelos é idêntico, quer no caso do pagamento antecipado, quer no caso do incumprimento, uma vez que as curvas de cada gráfico são muito semelhantes. Isto reforça a ideia que os modelos de *hazards* proporcionais são competitivos com a regressão logística na classificação das classes 'bom' e 'mau', isto é, têm um poder discriminante bastante semelhante, com a vantagem do primeiro poder ser utilizado para estimar o tempo de ocorrência dos acontecimentos, importante para o cálculo da rendibilidade das transacções de crédito.

Capítulo 3

Rendibilidade de transacções de crédito

As instituições de crédito sempre tiveram preocupações relativamente à decisão e acompanhamento de crédito em função do perfil de risco dos clientes ou, mais propriamente, na sua probabilidade de incumprir. No entanto, vários autores salientam inúmeras vantagens na passagem da 'minimização do risco' para a 'maximização do lucro', contrapondo o tradicional *Credit Scoring* ao chamado *Profit Scoring*. Apesar do advento das ferramentas de *data mining* e de sistemas mais completos de *data warehouse*, essa transição não é tão fácil como se poderia julgar. Na realidade, enquanto que a taxa de incumprimento depende principalmente das decisões de concessão, limites e recuperação de crédito, Thomas [11] aponta ainda outras decisões que afectam o lucro e que incluem os níveis de serviço, *marketing* e *pricing*.

Neste capítulo apresenta-se o cálculo da rendibilidade esperada de transacções de crédito pessoal recorrendo a métodos de análise de sobrevivência e mostrar que este pode ser um indicador útil na decisão de crédito, podendo ainda ser utilizado para o cálculo de *pricing* deste tipo de transacções.

A utilização da análise de sobrevivência apresenta grandes vantagens. Do ponto de vista matemático, proporciona facilidades em lidar com dados censurados e variáveis dependentes do tempo, relativamente aos métodos tradicionais. Além disso, e no contexto deste trabalho, Banasik [7] salienta outras vantagens deste tipo de análise:

- uma melhor estimativa da rendibidade da operação;

- previsão dos níveis de incumprimento e pagamento antecipado como função do tempo;

- as decisões podem levar em conta o tempo previsto da transacção de crédito;

- maior facilidade na incorporação de estimativas de factores económicos.

Espera-se assim poder prever, não se vai ocorrer o incumprimento, mas quando. Além do incumprimento, o pagamento antecipado é outra causa da perda de rendibilidade das operações que interessa considerar e que deverá estar incluída no modelo.

3.1 Cálculo da rendibilidade esperada com funções de sobrevivência

O cálculo da rendibilidade das transacções de crédito pessoal necessitará de alguns pressupostos para simplificar a formulação do problema. Deste modo, todas as transacções serão consideradas com pagamentos mensais de prestações fixas, postcipadas e sem período de carência (como acontece na maior parte dos casos). As taxas de juros serão consideradas fixas ao longo do empréstimo e as despesas operacionais, de manutenção e comissões não serão incorporadas no modelo.

Dadas as condições referidas e no caso ideal das transacções de crédito seguirem sempre o plano de pagamentos, a rendibilidade das operações seria fácil de calcular. Isto é, no caso de não haver nem pagamento antecipado nem incumprimento, um empréstimo de montante L e prazo de T meses com uma taxa de juro nominal mensal r e taxa de *funding* ou de custo de capital r', deveria proporcionar ao banco um lucro de

$$\begin{aligned} P_{max} &= \sum_{k=1}^{T} \frac{V_k}{(1+r')^k} - L \\ &= \sum_{k=1}^{T} \frac{c_k + j_k}{(1+r')^k} - L \end{aligned} \tag{3.1}$$

em que V_k é a prestação do empréstimo que pode ser subdividida nas componentes de capital e juro[1], c_k e j_k, respectivamente.

[1] No caso das prestações fixas (ver apêndice para mais detalhe) tem-se que

$$V_k = V = rL\frac{(1+r)^T}{(1+r)^T - 1} \quad , \quad c_k = rL\frac{(1+r)^{k-1}}{(1+r)^T - 1} \quad , \quad j_k = rL\frac{(1+r)^T - (1+r)^{k-1}}{(1+r)^T - 1}$$

Na realidade o que acontece é que nem todos os empréstimos seguem esse plano, porque alguns clientes pagam antecipadamente e, pior do que isso, incumprem. Estes factores representam perdas para o banco. Mais precisamente, o pagamento antecipado implica a perda do pagamento dos juros referentes às prestações remanescentes do empréstimo, enquanto o incumprimento implica não só a perda dos juros mas também do capital. Note-se que as perdas nunca são totais, uma vez que o pagamento antecipado pode ter penalizações e, em caso de incumprimento, muitas vezes é possível recuperar o montante em dívida. No entanto isso também não vai ser considerado no modelo.

Incorporando agora os factores de pagamento antecipado e incumprimento nas prestações tem-se que o lucro esperado (expressão adaptada de Stepanova e Thomas [6]) seria dado por

$$P_{exp} = \sum_{k=1}^{T} \frac{c_k\, S_k^D + j_k \left(S_k^E + S_k^D - 1\right)}{(1+r')^k} - L \tag{3.2}$$

onde, relativamente à k-ésima prestação, se tem que c_k e j_k são as componentes de capital e juro da prestação, S_k^E e S_k^D são as funções de sobrevivência para pagamento antecipado e incumprimento[2], respectivamente. Note-se que pela forma como se definiu os acontecimentos 'pagamento antecipado' (E) e 'incumprimento' (D) não é possível que ambos ocorram em simultâneo porque são acontecimentos incompatíveis. Por essa razão a componente de juros da expressão (3.2) é multiplicada pela probabilidade de não ter pagamento antecipado nem incumprimento.

$$P(E \cap D) = 0 \quad \Rightarrow \quad P(\bar{E} \cap \bar{D}) = P(\bar{E}) + P(\bar{D}) - 1$$

Por outras palavras, o lucro é a soma das parcelas de capital e juro de cada prestação multiplicadas pelas probabilidades de as receber (através das funções de sobrevivência), actualizadas para o instante inicial, menos o montante original.

Assumindo o modelo de *hazards* proporcionais, tem-se que a expressão (3.2) pode ser alterada para

$$P_{exp} = \sum_{k=1}^{T} \frac{c_k\, (S_{0k}^D)^{\alpha_D} + j_k \left((S_{0k}^E)^{\alpha_E} + (S_{0k}^D)^{\alpha_D} - 1\right)}{(1+r')^k} - L \tag{3.3}$$

[2]A notação E e D vem do inglês *Early repayment* e *Default*.

onde, relativamente à k-ésima prestação, S_{0k}^{E} e S_{0k}^{D} são as funções de sobrevivência base para pagamento antecipado e incumprimento (isto é, quando as covariáveis são todas zero), $\alpha_E = \exp(\boldsymbol{\beta}'_E \mathbf{x})$ e $\alpha_D = \exp(\boldsymbol{\beta}'_D \mathbf{x})$.

Esta forma de calcular a rendibilidade esperada enfrenta um problema que é comum às técnicas já conhecidas e utilizadas do *scoring* de crédito. Uma vez que os parâmetros do modelo são estimados com base numa amostra de transacções aceites pela instituição de crédito, os resultados podem sofrer um enviesamento quando aplicados a todas as transacções (inclusivamente as rejeitadas). O problema da *inferência de rejeição* foi investigado por Crook e Banasik [16] que utilizaram uma amostra rara concedida por uma instituição de crédito que ocasionalmente atribuía o crédito a virtualmente toda a procura de crédito. Uma das técnicas testadas, *re-weighting*, baseava-se no facto de haver uma desproporção de clientes com determinadas características, dada a selecção ser manifestamente não-aleatória (p.e. clientes desempregados dificilmente seriam aceites). Este efeito poderia ser compensado pela atribuição de 'pesos' mais elevados às transacções com características mais frequentemente rejeitadas. Outra técnica referida pelos autores, extrapolação, consistia em estimar um modelo inicial com as transacções aceites pela instituição. Aplicando esse modelo às transacções rejeitadas seria possível classificá-las nas categorias 'bom-mau' e, a partir daí, estimar um modelo final envolvendo todas as transacções (aceites e rejeitadas). Os resultados dos autores, porém, mostraram não haver uma grande melhoria nos modelos ao incluir-se uma técnica de inferência de rejeição, em especial se a taxa de rejeição ou recusa for reduzida.

Um outro método de inferência de rejeição proposto por Sohn e Shin [15] recorre aos mesmos métodos já utilizados neste trabalho, a análise de sobrevivência. A partir da amostra de transacções aceites constrói um modelo de sobrevivência para o incumprimento aplicando-o depois às transacções rejeitadas. Se o limite inferior de um intervalo de confiança de 90% do tempo de sobrevivência mediano (poderia ser outro percentil) da transacção rejeitada for maior que o tempo de sobrevivência mediano das transacções aceites, considerar-se-ia que a transacção seria aceite.

3.2 Resultados do cálculo da rendibilidade esperada

A rendibilidade esperada dada pela expressão (3.2) será agora calculada com base nos resultados obtidos em secções anteriores. O modelo de *hazards* proporcionais de Cox (semi-paramétrico) necessita de uma função base de sobrevivência. Optou-se pela função base dada pelo modelo de Weibull por este ser um modelo de *hazards* proporcionais e por já se ter mostrado que o ajuste aos dados era bastante aceitável.

Mais exactamente, utilizando os parâmetros estimados do modelo de Weibull (Tabela 2.3) para as funções base de sobrevivência referente ao pagamento antecipado e incumprimento tem-se que

$$S_{0k}^{E} = \exp(-(e^{-5,0908} \times \tfrac{k}{T} \times 100)^{\frac{1}{0,5722}})$$
$$S_{0k}^{D} = \exp(-(e^{-6,5206} \times \tfrac{k}{T} \times 100)^{\frac{1}{0,7039}})$$

Utilizando também os parâmetros estimados pelo modelo de *hazards* proporcionais de Cox (Tabela 2.5) relativamente às covariáveis para os mesmos acontecimentos tem-se que

$$\alpha_E = e^{\beta'_E \mathbf{x}} = e^{0,0314 \times x_1 + 0,0169 \times x_2 - 0,0151 \times x_3 + 0,2542 \times x_4 - 0,0041 \times x_6}$$
$$\alpha_D = e^{\beta'_D \mathbf{x}} = e^{0,0400 \times x_1 - 0,0392 \times x_2 - 0,0161 \times x_3 - 0,0221 \times x_6}$$

Os valores $\beta'_E\mathbf{x}$ e $\beta'_D\mathbf{x}$ representam assim uma espécie de preditor linear, ou *score*, para cada empréstimo relativamente aos acontecimentos em análise. Quanto maior for o valor destes *scores* maior será a probabilidade de pagamento antecipado e incumprimento, respectivamente.

A Figura 3.1 exibe vários gráficos, relativamente a pagamento antecipado (esquerda) e incumprimento (direita), incluindo a distribuição das transacções de crédito em função de intervalos dos *scores* (em cima) e referentes à rendibilidade média em função desses *scores*, considerando o custo de capital igual a zero (ao centro) e igual a 4% (em baixo). Claro que os valores de rendibilidade são mais elevados quando $r' = 0$, por isso o segundo valor deverá ser uma aproximação mais realista. Comparam-se a média da rendibilidade esperada calculadas através da expressão (3.3), com a média da rendibilidade máxima calculada através da expressão (3.1) e com a média da rendibilidade actual. Esta última corresponde à que se verificou de facto, assumindo a perda dos juros a partir do momento em que se registasse o

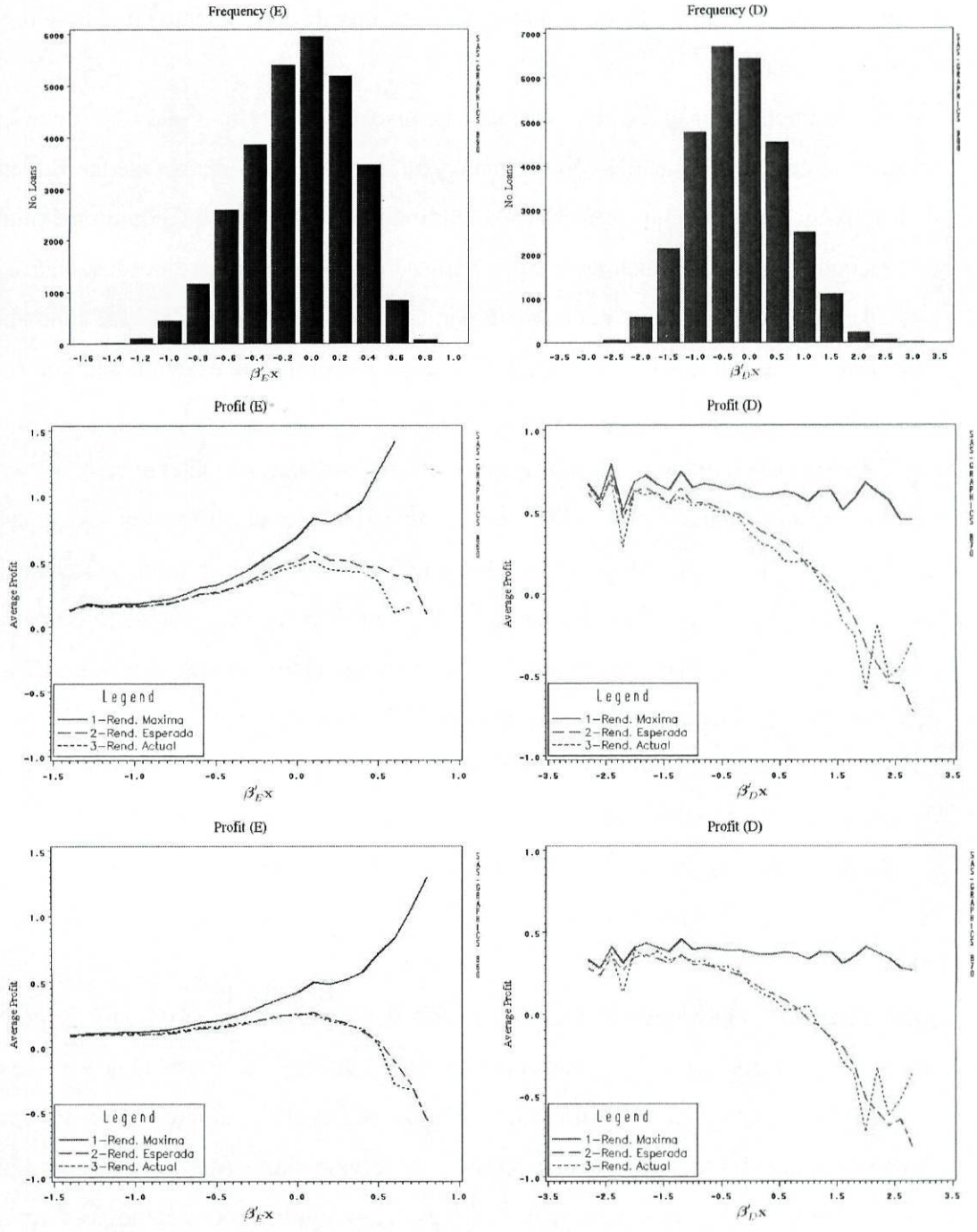

Figura 3.1: Em cima: Número de empréstimos por intervalos de *score* $\beta'_E \mathbf{x}$ e $\beta'_D \mathbf{x}$ para pagamento antecipado (esquerda) e incumprimento (direita), respectivamente. Ao centro: Rendibilidade média em função dos *scores* com o custo de capital igual a zero. Em baixo: Rendibilidade média em função dos *scores* com o custo de capital igual a 4%. A rendibilidade máxima é obtida quando não há incumprimento nem pagamento antecipado; a rendibilidade esperada é o valor estimado pelo modelo; a rendibilidade actual ou real foi a que se verificou de facto.

pagamento antecipado e a perda do capital e juros a partir do momento em que se registasse o incumprimento[3].

Relativamente ao pagamento antecipado nota-se que *scores* mais elevados, onde a probabilidade deste acontecimento ocorrer é maior, estão associadas maiores médias de rendibilidade máxima, enquanto que as médias da rendibilidade esperada atingem um máximo e vão decrescendo. No caso do incumprimento, para valores de *score* mais elevados, onde a probabilidade deste acontecimento ocorrer é maior, as médias da rendibilidade máxima tendem a permanecer constantes, enquanto que as médias da rendibilidade esperada são notoriamente decrescentes com o aumento do *score*.

Uma explicação para as diferenças entre os gráficos destes acontecimentos poderá ter a ver com as variáveis explicativas. O prazo é a variável mais explicativa no caso do pagamento antecipado (empréstimos mais longos têm maior probabilidade de liquidação antecipada), enquanto que o incumprimento tem como variável mais explicativa o *score* interno do banco para classificação do risco do cliente (independente do prazo do empréstimo). Além disso, convém notar que os valores mais extremados dos *scores* $\beta'_E\mathbf{x}$ e $\beta'_D\mathbf{x}$ possuem muito menor número de empréstimos (como se vê nos gráficos de barras) e são por isso mais susceptíveis de sofrer variações nos valores médios das rendibilidades.

Consideremos agora, não a rendibilidade média, mas a rendibilidade acumulada em função dos *scores* subdividida em dois grupos em termos de montante (mais ou menos que 10.000€) e prazo (mais ou menos que 30 meses).

A Figura 3.2 mostra que os empréstimos de montante inferior a 10.000€ proporcionam maior rendibilidade que os restantes, mas isso está relacionado com o facto de estes serem em número superior aos outros. Em termos de prazo evidencia-se uma característica peculiar. Pode ver-se que, ao contrário do que acontece com o montante, os gráficos da rendibilidade para diferentes prazos se intersectam. Este resultado, também constatado por Stepanova e Thomas [6], é justificado pelo facto da rendibilidade das transacções de crédito de montantes semelhantes ter de levar em conta, quer o *score*, quer o prazo.

Em todo o caso os gráficos permitem identificar *cut-offs* a partir dos quais já não se regista

[3]No caso dos dados censurados considerou-se a rendibilidade esperada a partir do momento em que se registasse o tempo censurado.

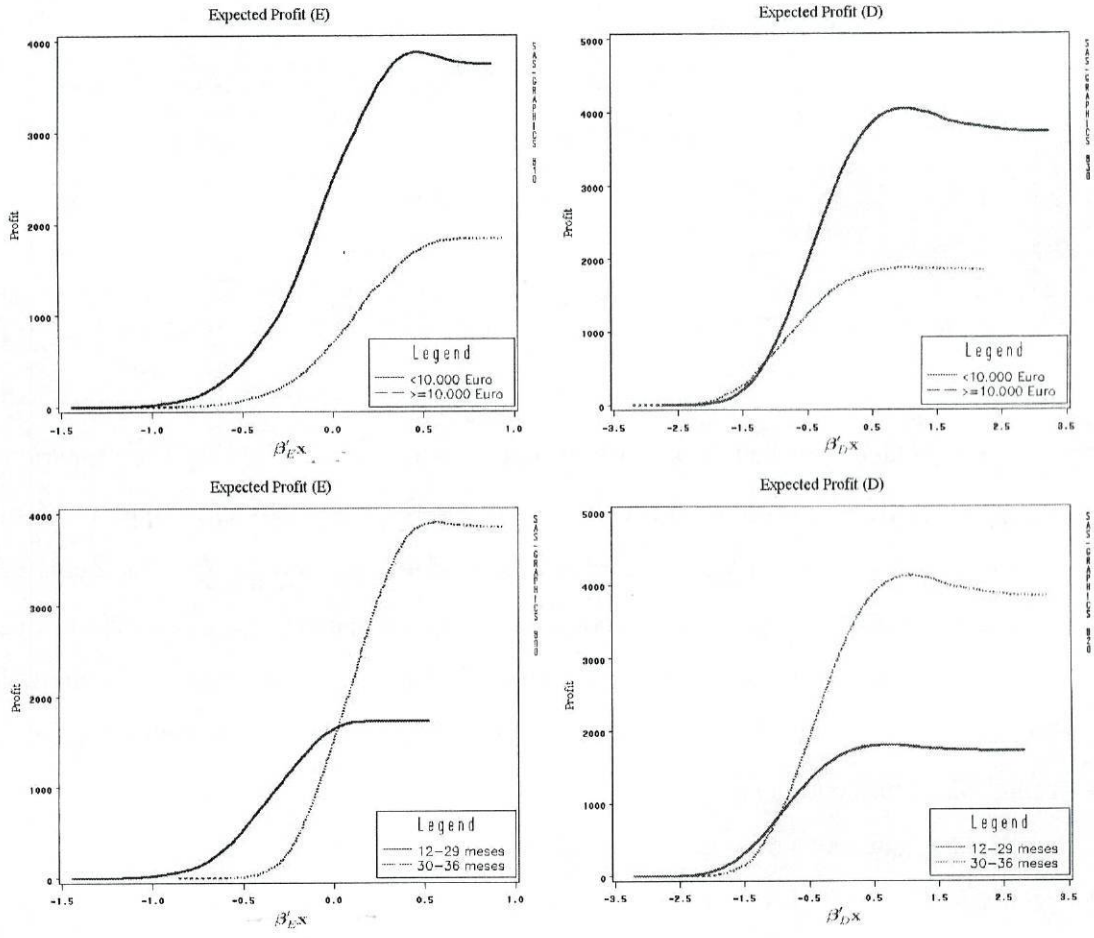

Figura 3.2: Rendibilidade esperada acumulada em função dos *scores* de pagamento antecipado $\beta'_E\mathbf{x}$ (esquerda) e de incumprimento $\beta'_D\mathbf{x}$ (direita) para dois subgrupos em termos de montante (em cima) e prazo (em baixo), coniderando o custo de capital igual a 4%.

aumentos de rendibilidade.

A Tabela 3.1 mostra uma comparação do que poderia ser a decisão com base na rendibilidade esperada das transacções de crédito. A partir da amostra de teste, constituída por 29.041 transacções de crédito pessoal, fez-se uma simulação do que poderiam ser as decisões automáticas de aprovação com base no risco do cliente, isto é, com base num *score* para a probabilidade de incumprimento. Segundo este método correntemente utilizado, assume-se que uma transacção deverá ser rentável ($\oplus$) no caso do cliente ser considerado 'bom' por esse *score* e não rentável ($\ominus$) no caso contrário. Utilizando a rendibilidade esperada para o mesmo efeito, considera-se rentável uma transacção com rendibilidade esperada positiva e

não rentável no caso contrário.

	Valores actuais		Risco do cliente		Rend. esperada	
Rendibilidade	Número	Montante	Número	Montante	Número	Montante
⊕ - prevista ⊕	**26.644**	**100,0**	**23.518**	**89,7**	**24.422**	**96,1**
⊕ - prevista ⊖	0	0,0	3.126	10,3	2.222	3,9
⊖ - prevista ⊕	**0**	**0,0**	**1.045**	**-20,0**	**934**	**-24,4**
⊖ - prevista ⊖	2.397	-32,5	1.352	-12,5	1.463	-8,1

Tabela 3.1: Comparação entre uma classificação baseada no risco do cliente e outra baseada na rendibilidade esperada das transacções de crédito em termos de número e montante (unidades monetárias).

Os resultados apresentados mostram que considerando o risco do cliente para decisão de crédito seriam aprovadas 24.563 transacções, das quais 1.045 não seriam rentáveis para o banco. Através da rendibilidade esperada seriam aprovadas 25.356, das quais 934 não seriam rentáveis para o banco. Nota-se uma ligeira vantagem último método que pode ainda ser comprovado pelas respectivas rendibilidades. No primeiro caso a rendibilidade das transacções seria de 69,7 unidades monetárias[4] (u.m.), enquanto que no segundo caso a rendibilidade ascenderia a 71,7 u.m.

Note-se que os valores apresentados não devem ser considerados valores absolutos, mas relativos. Os resultados foram efectuados segundo pressupostos muito específicos que simplificaram os cálculos, ignorando ainda a possibilidade de *overrides*, isto é, a possibilidade de existirem factores impeditivos da aprovação automática, apesar do modelo assim o indicar.

3.3 *Pricing* com base na rendibilidade esperada

O *pricing* baseado no risco, conhecido por *Risk-based pricing*, é uma modalidade de atribuição de preços diferenciados que tende a reflectir o risco ou rendibilidade potencial das transacções de crédito. Logicamente, isto leva a atribuir taxas de juros mais baixas aos melhores clientes e mais elevadas aos piores. Este processo está longe de ser simples e tem subjacente vários desafios. Por exemplo, o fenómeno da *selecção adversa* pode estar presente quando a taxa de juro oferecida a um cliente é mais elevada que a taxa *standard*. Aqueles que a aceitam não são de uma amostra aleatória da população, mas sim de um grupo de risco elevado, pois

[4]Para estes efeitos comparativos optou-se por relativizar o montante envolvido, considerando-se unidades monetárias (u.m.) em vez de, por exemplo, milhares ou milhões de euros.

não conseguiram uma taxa mais baixa noutro sítio. Significa que dever-se-ia usar uma taxa ainda maior para compensar esse risco, o que agravaria mais a situação, e que levaria a uma escalada ascendente da taxa de juro.

Uma outra situação pode ocorrer com os bons clientes, porque ao atribuir um preço mais reduzido que o cliente consideraria aceitável, na realidade está a perder-se alguma da rendibilidade possível de obter. Estes exemplos servem para referir que, mais do que atribuir um preço com base no risco, igualmente importante é conhecer e compreender as condições do mercado e da concorrência e também saber interagir e explicar ao cliente as razões da taxa de juro serem diferentes do *standard*. Edelman [1] sugere alternativamente que, em vez de ajustar o preço ao risco do cliente, pode ajustar-se o risco ao preço. Por exemplo, para os melhores (piores) clientes poder-se-ia ser menos (mais) exigente nas requisições de garantias, colateral ou provas documentais.

O objectivo deste trabalho não está propriamente no cálculo do *pricing* das transacções de crédito, mas esta é uma possível aplicação. Dado que a rendibilidade é afectada pelo montante, prazo, taxas de juro, bem como outros factores económicos que muitas vezes são difíceis de obter, dever-se-ia considerar uma medida que fosse relativa à rendibilidade máxima.

Hoadley [12] propõe uma medida chamada lucro holístico (*holistic profit*), e que é o quociente entre o lucro e o lucro obtido sob discriminação perfeita. Aqui o contexto é um pouco diferente mas a medida sugerida tem um paralelo com o quociente entre a rendibilidade esperada e a rendibilidade máxima. A razão entre as expressões 3.2 e 3.1,

$$P_{racio} = \frac{P_{exp}}{P_{max}} \tag{3.4}$$

permite responder à questão de encontrar a taxa de juro de uma transacção de crédito, admitindo que o banco está disposto a ganhar uma determinada percentagem da sua rendibilidade máxima, P_{racio}. Isso corresponde a resolver a equação 3.4 em ordem à taxa de juro r, implicitamente envolvida no cálculo das componentes de capital, c_k, e juro, j_k, de cada prestação V_k das expressões (3.1) e (3.2).

Conclusão

As análises efectuadas nas secções anteriores permitiram reconhecer potencialidades nos métodos de análise de sobrevivência quando aplicados ao *scoring* de crédito, através do cálculo da rendibilidade esperada das transacções de crédito pessoal.

A rendibilidade das transacções de crédito é determinantemente condicionada pela ocorrência de pagamento antecipado e incumprimento, entre outros factores. Neste trabalho foi possível a modelação destes dois acontecimentos utilizando modelos de análise de sobrevivência.

O método de Kaplan-Meier (não paramétrico) permitiu estimar funções de sobrevivência para estes dois acontecimentos em função do tempo. Vários modelos de vida acelerada (paramétricos) foram experimentados, incorporando já variáveis explicativas, obtendo-se estimativas para os parâmetros dos modelos. Análises comparativas com base nos resíduos de Cox-Snell mostram que o desempenho dos modelos é relativamente semelhante, sendo o modelo de Weibull aquele que é matematicamente mais simples. Com o modelo de *hazards* proporcionais de Cox (semi-paramétrico) foi possível obter estimativas para os parâmetros das covariáveis, uma vez que o modelo não depende da função base de *hazard*. Esses valores estimados através de verosimilhança parcial revelaram-se bastante próximos dos obtidos com o modelo paramétrico de Weibull, também um modelo de *hazards* proporcionais. Além disso, o poder discriminante do modelo de Cox mostrou-se bastante competitivo com o modelo de regressão logística tradicional, quando avaliado através das curvas ROC.

O cálculo da rendibilidade esperada das transacções de crédito utilizou as funções de sobrevivência estimadas para o pagamento antecipado e incumprimento sob influência das variáveis explicativas. O valor encontrado foi depois comparado com a rendibilidade máxima (quando não ocorre nem pagamento antecipado nem incumprimento) e a rendibilidade

real. A comparação destas rendibilidades foi exibida graficamente, calculando os seus respectivos valores médios em função de *scores* para pagamento antecipado e incumprimento, obtidos pelo modelo de *hazards* proporcionais. Analisando também os valores acumulados da rendibilidade esperada foi ainda possível identificar *cut-offs* a partir dos quais não se deverá esperar um aumento da rendibilidade das transacções. Por isso este poderá também ser um factor a considerar na decisão de crédito.

Uma análise comparativa de decisão de crédito com base no risco do cliente (correntemente utilizado) e com base na rendibilidade esperada, evidenciou uma ligeira vantagem deste último método em número (aumento de cerca 3% na taxa de aprovação) e montante.

O cálculo da rendibilidade esperada pode ter outra aplicação bastante útil. Trata-se de calcular o *pricing* das transacções de crédito assumindo que se espera obter determinado nível de rendibilidade, isto é, uma determinada percentagem da rendibilidade máxima das transacções.

Bibliografia

[1] THOMAS, L. C., EDELMAN, D.B., CROOK, J. N., Credit Scoring and its Applications, *SIAM* (2002)

[2] ALLISON, P., Survival Analysis using SAS®: a practical guide, *Wiley & Sons* (1995)

[3] LEE, E., Statistical Methods for Survival Data Analysis, *Wiley & Sons* (1992)

[4] MILLER, R. G., Survival Analysis, *Wiley & Sons* (1981)

[5] STEPANOVA, M. e THOMAS, L. C., Survival analysis methods for personal loan data, *Proc. Credit Scoring and Credit Control VI*, Credit Research Centre, University of Edimburgh, *Operations Research* (1999) **52**, 277-289

[6] STEPANOVA, M. e THOMAS, L. C., PHAB scores: proportional hazards analysis behavioural scores, *Journal of the Operational Research Society* (2001) **52**, 1007-1016

[7] BANASIK, J., CROOK, J. N. e THOMAS, L. C., Not if but when will borrowers default, *Journal of the Operational Research Society* (1999) **50**, 1185-1190

[8] COX, D. R., Regression models and life-tables (with discussion), *J R Stat Soc Ser B* (1972) **74**, 187-220

[9] COX, D. R., Partial likelihood, *Biometrika* (1975) **62**, 187-202

[10] THOMAS, L. C., HO, J. e SCHERER, W. T., Time will tell: behavioural scoring and the dynamics of consumer credit assessment, *IMA Journal of Management Mathematics* (2001) **12**, 89-103

[11] THOMAS, L. C., A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers, *International Journal of Forecasting* (2000) **16**, 149-172

[12] HOADLEY, B. e OLIVER,R.M., Business measures of scorecards benefit, *IMA Journal of Mathematics Applied in Business & Industry* (1998) **9**, 55-64

[13] D'AGOSTINO, R. B. e NAM, B. H., Evaluation of the performance of survival analysis models: discrimination and calibration measures, *Handbook of Statistics - Elsevier* (2004) Vol. **23**

[14] FARRINGTON, C. P., Residuals for Proportional Hazards Models with interval-censored survival data, *Biometrics* (2002) **56**, 473-482

[15] SOHN, S. Y. e SHIN, H. W., Reject inference in credit operations based on survival analysis, *Expert Systems with Applications* (2006) **31**, 26-29

[16] CROOK, J. e BANASIK, J., Does reject inference really improve the performance of application scoring models?, *Journal of Banking & Finance* (2004) **28**, 857-874

[17] COLLET, D., Modelling Survival Data in Medical Research, *Chapman & Hall*, London, U.K. (1994)

[18] ROCHA, C., Análise de Sobrevivência, *Apresentação do IV Congresso Anual da Sociedade Portuguesa de Estatística* (1996)

[19] OLIVER, R. M. e WELLS, E., Efficient frontier cutoff policies in credit portfolios, *Journal of the Operational Research Society* (2001) **52**, 1025-1033

[20] KALAPODAS, T., Credit risk assessment: a challenge for financial institutions, *IMA Journal of Management Mathematics* (2006) **17**, 25-46

[21] MCNAB, H. e WYNN, A., Principles and practice of consumer credit risk management, *Institute os financial services UMIST*

[22] SAS®, Help , *SAS Institute Inc., Cary, NC, USA.* (2002)

Apêndice A

Capital e juros em empréstimos de prestação fixa

A amortização de um empréstimo segue, de uma forma geral, um regime de juro composto. O modelo mais frequente que se encontra (existem muitas variantes) é o de uma prestação fixa ao longo do tempo de vida do empréstimo dividida em duas componentes: capital e juros.

Se num dado instante for concedido um empréstimo de valor L com pagamento em n termos ou prestações e uma taxa de juro r (geralmente uma taxa nominal) tem-se a seguinte relação para o cálculo das prestações:

$$L = \sum_{k=1}^{n} V_k (1+r)^{-k} \tag{A.1}$$

e, se a prestação for fixa,

$$L = \sum_{k=1}^{n} V (1+r)^{-k} \tag{A.2}$$

ou ainda[1]

$$L = V \frac{1-(1+r)^{-n}}{r} = V a_{n|r} \tag{A.3}$$

A prestação fixa de um determinado empréstimo pode assim ser calculada dado o montante pedido, prazo e taxa de juro.

[1] Da soma dos n primeiros termos de uma progressão geométrica:
$\sum_{k=1}^{n}(1+r)^{-k} = (1+r)^{-1} \times \frac{1-(1+r)^{-n}}{1-(1+r)^{-1}} = \frac{1-(1+r)^{-n}}{r} = a_{n|r}$

A k-ésima prestação de um empréstimo pode sempre subdividir-se nas componentes de capital e juro, $c_k + j_k$. Estas componentes podem ser calculadas dados o montante do empréstimo L, o prazo p e a taxa referente ao período de pagamento r, sendo $V = \frac{L}{a_{\overline{p}|r}} = rL\frac{(1+r)^p}{(1+r)^p-1}$ a prestação (fixa). A seguir estão calculadas as componentes de capital e juro das primeiras quatro prestações. O raciocínio é análogo para o cálculo das restantes.

$$
\begin{aligned}
j_1 &= rL & c_1 &= V - j_1 = rL\tfrac{1}{(1+r)^p-1} \\
j_2 &= r(L - c_1) = rL\tfrac{(1+r)^p-(1+r)}{(1+r)^p-1} & c_2 &= V - j_2 = rL\tfrac{(1+r)}{(1+r)^p-1} \\
j_3 &= r(L - c_1 - c_2) = rL\tfrac{(1+r)^p-(1+r)^2}{(1+r)^p-1} & c_3 &= V - j_3 = rL\tfrac{(1+r)^2}{(1+r)^p-1} \\
j_4 &= r(L - c_1 - c_2 - c_3) = rL\tfrac{(1+r)^p-(1+r)^3}{(1+r)^p-1} & c_4 &= V - j_4 = rL\tfrac{(1+r)^3}{(1+r)^p-1}
\end{aligned}
$$

Mostremos por indução que

$$j_k = rL\frac{(1+r)^p - (1+r)^{k-1}}{(1+r)^p - 1} \tag{A.4}$$

ou equivalentemente, que $c_k = V - j_k = rL\frac{(1+r)^{k-1}}{(1+r)^p-1}$.

Para $k = 1$ a hipótese é trivialmente verificada. Mostremos que a hipótese é hereditária:

$$
\begin{aligned}
j_{k+1} &= r\left(L - \sum_{i=1}^{k}(c_i)\right) \\
&= r\left(L - \sum_{i=1}^{k} rL\frac{(1+r)^{k-1}}{(1+r)^p - 1}\right), \text{ por hipótese} \\
&= r\left(L - rL\frac{1}{(1+r)^p - 1}\sum_{i=1}^{k}(1+r)^{i-1}\right) \\
&= r\left(L - rL\frac{1}{(1+r)^p - 1}\frac{(1+r)^k - 1}{r}\right) \\
&= rL\left(1 - \frac{(1+r)^k - 1}{(1+r)^p - 1}\right) \\
&= rL\frac{(1+r)^p - (1+r)^k}{(1+r)^p - 1}
\end{aligned}
$$

Em conclusão, a k-ésima prestação (fixa) tem componentes de capital e juro dadas por

$$V = rL\frac{(1+r)^p}{(1+r)^p - 1} = \underbrace{rL\frac{(1+r)^{k-1}}{(1+r)^p - 1}}_{Capital} + \underbrace{rL\frac{(1+r)^p - (1+r)^{k-1}}{(1+r)^p - 1}}_{Juro} \tag{A.5}$$

Plano de amortizações

Termo	CapAmort	Prest	Cap	Juro	CapAc	JurAc	%CapAc	%JurAc
1	1000,00	88,85	78,85	10,00	78,85	10,00	8%	15%
2	921,15	88,85	79,64	9,21	158,49	19,21	16%	29%
3	841,51	88,85	80,43	8,42	238,92	27,63	24%	42%
4	761,08	88,85	81,24	7,61	320,16	35,24	32%	53%
5	679,84	88,85	82,05	6,80	402,21	42,04	40%	64%
6	597,79	88,85	82,87	5,98	485,08	48,01	49%	73%
7	514,92	88,85	83,70	5,15	568,78	53,16	57%	80%
8	431,22	88,85	84,54	4,31	653,32	57,48	65%	87%
9	346,68	88,85	85,38	3,47	738,70	60,94	74%	92%
10	261,30	88,85	86,24	2,61	824,93	63,56	82%	96%
11	175,07	88,85	87,10	1,75	912,03	65,31	91%	99%
12	87,97	88,85	87,97	0,88	1000,00	66,19	100%	100%

Tabela A.1: Exemplo de um plano de amortizações; Legenda: CapAmort - capital a amortizar; Prest - valor da prestação; Cap - componente de capital da prestação; Juro - componente de juro da prestação; CapAc - capital amortizado acumulado; JurAc - juros acumulados pagos; %CapAc - percentagem de capital amortizado; %JurAc - percentagem de juros pagos.

A Tabela A.1 mostra um exemplo de um plano de amortizações subdividindo cada prestação nas componentes de capital e juro (ignorando comissões e outras despesas adicionais do empréstimo). Para o efeito considerou-se um empréstimo de montante $L = 1000$ € em $p = 12$ meses com uma taxa de juro anual nominal de $r = 12\%$ (1% ao mês) com prestações fixas de valor igual a $V = 1000/a_{\overline{12}|1}$.