ISCTE Business School
Instituto Universitário de Lisboa

MODELOS HETEROGÉNEOS DE SOBREVIVÊNCIA:

Uma Aplicação ao Risco de Crédito

Bruno Cardoso Alves

Tese de Mestrado

em Prospecção e Análise de Dados

Orientador:

Prof. Doutor José Gonçalves Dias, Prof. Auxiliar, ISCTE Business School,
Departamento de Métodos Quantitativos

Abril 2010

Dedicado à Mónica, ao Salvador e à Beatriz.

Agradecimentos

Gostaria de agradecer ao Professor Doutor José Gonçalves Dias pela sua orientação, críticas, ensinamentos e por me manter no rumo certo durante estes meses de tese.

Um muito obrigado também ao meu Director, Adriano Assunção, que para além do contributo para a minha evolução profissional, sempre demonstrou o seu apoio sem o qual esta tese teria sido muito mais difícil de concluir. Este agradecimento estende-se também à administração da instituição financeira que me ofereceu esta oportunidade, apoiando-me financeiramente e cedendo os dados utilizados nesta tese.

Ainda um agradecimento aos meus colegas de mestrado, Leonor Pereira da Costa, Marcos Henriques e Tiago Salgado Pereira pela sua cumplicidade nos dias e noites de estudo no ISCTE, durante a parte académica deste mestrado.

Um agradecimento muito especial à Mónica, pelo carinho e apoio durante este mestrado em que estive menos presente, ao Salvador e à Beatriz, por serem sempre um factor de inspiração.

Resumo

Para criar modelos de apoio à gestão de cobranças de clientes numa instituição financeira de crédito, foram estimados modelos de sobrevivência heterogéneos, para prever a duração até dois acontecimentos: (i) registo do primeiro atraso no pagamento das mensalidades do contrato de crédito; e (ii) registo de atrasos superiores a 90 dias – *default*. Seguiu-se uma abordagem condicional tipo II, utilizando todos os clientes da amostra para estimar a duração até ao primeiro atraso e uma sub-amostra, com os clientes que registaram esse primeiro atraso, para estimar a duração até *default*. Para cada acontecimento foram testadas as distribuições exponencial, Weibull, log-normal e log-logística, em modelos agregados e de mistura. A duração até ao primeiro incidente (i) foi estimada através de um modelo de sobrevivência com proporção de imunes. Esta proporção resulta de um modelo logístico utilizando o *scoring* interno como variável concomitante. Para os não imunes considerou-se que a duração t segue uma distribuição log-normal, com variáveis explicativas para os parâmetros μ e σ. A duração entre o primeiro incidente e uma situação de *default* (ii) estimou-se através de um modelo de sobrevivência de mistura com 3 segmentos, com uma função de ligação logit multinomial e assumindo também que t segue uma distribuição log-normal. Neste segundo modelo apenas foram modelados os pesos do modelo logit, considerando μ e σ constantes. Os modelos de sobrevivência apresentados incluem maioritariamente informação recolhida na altura da originação, aplicáveis igualmente como modelos de *profit scoring*, estimando o envolvimento na data de *default*, dado um *cash-flow* futuro.

Palavras chave: Análise de sobrevivência; Risco de crédito; Modelos de mistura; *Scoring* comportamental.

Abstract

To create models that support the receivables management in a financial institution, heterogeneous survival models were estimated to predict time until two events: (i) having at least one payment overdue; and (ii) 90 days overdue - default. A conditional 2 approach was followed, using all customers of the sample to estimate time until a first payment overdue. A second model was developed, considering only the sub-sample of clients who experienced the first overdue. The exponential, Weibull, log-normal and log-logistic distributions were tested in estimating the time to each event, in aggregate and mixture models. Time to the first overdue (i) was predicted through a survival analysis with immunes, with a logistic model to estimate probability of immunity, using internal credit scoring as covariate. For the non-immunes, a log-normal function, with covariates for both parameters, μ and σ, was estimated to predict time to first overdue. The time between the first overdue and default (ii) was estimated by survival mixture model with 3 segments, with a multinomial logit link function and assuming that time to default also follows a log-normal distribution. Covariates on the second model were considered on the proportions of the mixture model, setting the parameters μ and σ as constants in each group. The survival models presented in this thesis are estimated with data collected at the beginning of the loan, allowing its application in a profit scoring model, by predicting the exposure at the time the customer enters into a situation of default, given an expected cash-flow.

Keywords: Survival analysis; Credit risk; Mixture models; Behavioural scoring.

Sumário executivo

O risco de crédito ocorre quando um cliente falha em honrar os seus compromissos. Um dos pilares para uma instituição financeira ser bem sucedida assenta no modo como é avaliado o risco dos seus clientes, pelo que possuir modelos que minimizem a incerteza na tomada de decisões, é um factor crítico de sucesso neste tipo de negócio. No âmbito da gestão do risco de crédito, uma instituição necessita de tomar dois tipos de decisões. Numa fase inicial do processo terá de decidir se aprova o crédito e numa segunda fase, depois de eliminar crédito com potencial de risco mais elevado, é necessário definir que tipo de relacionamento adoptar com os clientes em carteira. Do mesmo modo, podemos classificar os modelos utilizados na avaliação do risco de crédito em 2 grupos: (*i*) modelos estáticos – *scoring* aplicacional – utilizados para avaliar o risco de crédito no momento da concessão; (*ii*) modelos dinâmicos – *scoring* comportamental – para avaliar o comportamento do cliente ao longo do contrato.

Num modelo de *scoring* aplicacional é feita a ligação entre duas fotografias, uma com informação relativa ao momento da concessão de crédito e outra com a situação do cliente, por exemplo, 12 meses após o início do contrato. Por seu lado, num modelo de *scoring* comportamental, o comportamento futuro de um cliente é o resultado, não só da fotografia inicial em que o crédito é concedido, mas também de informação recente sobre o seu comportamento. Através da análise de sobrevivência é possível substituir a segunda fotografia relativa à situação do cliente, por um filme do comportamento futuro do cliente, permitindo estimar, não só a probabilidade de um cliente se encontrar em situação de *default* 12 meses após, mas também estimar a probabilidade associada para um determinado momento futuro, o que resulta num *timing* mais eficiente em tomadas de decisões por parte de instituição de crédito.

Nesta tese de mestrado são apresentados modelos de sobrevivência com mistura, que estimam a duração até ao momento do cliente entrar em incumprimento, com o objectivo de apoiar a gestão de cobranças de clientes numa instituição financeira de crédito. Foram considerados dois acontecimentos, associados a uma situação de incumprimento ligeiro e incumprimento avançado, respectivamente: (i) registo do primeiro atraso no pagamento das mensalidades do contrato de crédito, que resulta na transição do cliente para gestão de cobranças; (ii) registo de atrasos superiores a 90 dias – *default*. Seguiu-se uma abordagem condicional tipo II, utilizando todos os clientes da amostra para estimar a duração entre o início do contrato e o registo do primeiro atraso e

criado um segundo modelo, considerando apenas uma sub-amostra com os clientes que registaram o primeiro atraso. Neste segundo modelo, a variável *t* representa a duração entre o registo do primeiro atraso e a transição para uma situação de *default*.

Foram testadas as distribuições exponencial, Weibull, log-normal e log-logística, para duração até cada acontecimento, em modelos agregados e de mistura. Os modelos eleitos foram seleccionados com base nos critérios de informação, AIC e BIC, concluindo-se que a distribuição log-normal é aquela que melhor se ajusta à duração para ambos os acontecimentos. Concluiu-se também que existe heterogeneidade tanto na análise ao primeiro acontecimento, com uma proporção de clientes imunes ao risco de crédito, como na análise à transição para uma situação de *default*, onde foram identificados três grupos de clientes com padrões comportamentais distintos. Assim, a duração até ao primeiro incidente (i) foi estimada através de um modelo de sobrevivência com proporção de imunes, em que esta proporção resulta de um modelo logístico utilizando o nível de *scoring* interno como covariável. Para o grupo dos não imunes foram consideradas covariáveis para modelar os parâmetros da log-normal, μ e σ. Como seria de esperar, a níveis mais elevados de *scoring* está associada uma maior probabilidade de pertencer ao grupo dos imunes. Para o grupo dos não imunes, a duração é influenciada por informação relacionada com a estabilidade do cliente, com as garantias apresentadas e com as características da operação.

A duração entre o primeiro incidente e uma situação de *default* (ii) foi estimada através de um modelo de sobrevivência de mistura com 3 segmentos, utilizando uma função de ligação logit multinomial e assumindo também que a duração *t* segue uma distribuição log-normal. Assim, foram identificados 3 padrões de comportamento na transição para uma situação de *default*: (1) clientes com uma duração mais longa; (2) clientes com uma duração intermédia; e (3) clientes com duração mais curta, ou seja, com uma transição muito rápida, não pagando consecutivamente as mensalidades após registarem um primeiro atraso, transitando para uma situação de *default*, 3 meses após o primeiro acontecimento. Neste segundo modelo apenas foram modeladas as proporções da mistura, considerando os parâmetros, μ e σ, da log-normal como constantes. Assim, a probabilidade de um cliente pertencer ao grupo com uma duração mais longa aumenta nos casos em que o cliente apresente sinais de estabilidade e melhores garantias. A este grupo de clientes estão ainda associadas condições de crédito mais competitivas. Porém, ao contrário do que seria de esperar, clientes pertencentes a um escalão etário mais

elevado têm uma menor probabilidade de pertencerem ao grupo com duração mais longa. Para o grupo com duração mais curta, 3 meses após o primeiro incidente, a probabilidade associada aumenta em contratos de crédito com taxas mais elevadas e maior prazo remanescente no momento em que o cliente regista o primeiro incidente.

A análise de sobrevivência como técnica aplicada em modelos de *scoring* comportamental, ao projectar um filme do desempenho futuro de um cliente de crédito, permite a sua utilização em outros tipos de modelos, como em *profit scoring*. Os modelos de sobrevivência apresentados neste estudo, incluem na sua grande maioria informação recolhida no momento da originação, possibilitando também a sua aplicação nesse tipo de modelos, estimando o envolvimento no momento em que o cliente entra em situação de *default*, calculando desta forma uma perda esperada que deve ser considerada na avaliação da rendibilidade de uma operação.

Lista de Figuras

Lista de Tabelas

Índice

1. Introdução

Em qualquer área de negócio estará sempre presente um factor de incerteza, seja ele de ordem política ou económica. Nos últimos 500 anos, por exemplo, o plano económico e político tem sido assumido por diferentes protagonistas como Portugal, Espanha, Holanda, França, Inglaterra e EUA e, segundo alguns analistas, será assumido nos próximos tempos pela China e pela Índia (Keidel, 2008; Khanna, 2007). O risco de crédito é apenas uma componente de um leque mais vasto de risco, mas que assume uma enorme importância no negócio do crédito, e ocorre quando um cliente falha em honrar os seus compromissos. Um dos pilares para uma instituição financeira ser bem sucedida assenta no modo como avalia o risco dos seus clientes, sendo muitas vezes referido que o risco é o principal negócio de uma instituição financeira. Assim, possuir modelos que apoiem na tomada de decisões, de forma a minimizar a incerteza, é um factor crítico de sucesso neste tipo de negócio.

No âmbito da gestão do risco de crédito, uma instituição necessita de tomar dois tipos de decisões. Numa fase inicial do processo terá de decidir se concede o crédito e numa segunda fase, depois de eliminar crédito com potencial de risco mais elevado, é necessário definir que tipo de relacionamento adoptar com os clientes em carteira. Do mesmo modo, podemos classificar os modelos utilizados na avaliação do risco de crédito em 2 grupos: (*i*) modelos estáticos – *scoring* aplicacional – utilizados para avaliar o risco de crédito no momento da concessão; (*ii*) modelos dinâmicos – *scoring* comportamental – para avaliar o comportamento do cliente ao longo do contrato, identificando, por exemplo, padrões que caracterizam clientes com maior potencial de incumprir (Thomas, 2009*)*.

Consoante o objectivo seja explicar ou prever, a probabilidade de um cliente incumprir, diferentes abordagens podem ser adoptadas na criação de um modelo de *Credit Scoring*, entre as mais utilizadas estão: Análise Discriminante, Regressão Logística, Redes Neuronais e Árvores de Decisão (Thomas, 2009). Nesta primeira fase, os modelos estatísticos utilizados tornam-se úteis não só na aprovação, ou recusa de uma operação, mas também na definição do preço, na medida em que a taxa mínima a que uma instituição financeira está disposta a conceder o crédito é maior em operações com perfil de risco mais elevado. A informação utilizada neste tipo de modelos contém variáveis que espelham características da operação e do cliente no momento da proposta

de crédito e a sua situação em termos de cumprimento numa data posterior, ou seja, faz a ligação entre o início da operação e uma data futura.

Porém, essa é apenas a fase inicial na relação da instituição financeira com o cliente. O modo como a operação é gerida, o tipo de alterações efectuadas na operação, que medidas são tomadas pela instituição quando o cliente falha um pagamento, se o cliente amortiza o empréstimo antecipadamente, todos estes factores também eles afectam a rendibilidade de uma operação. Num modelo estático assume-se que a probabilidade de um cliente incumprir se mantêm inalterada ao longo do tempo, ao passo que num dinâmico essa probabilidade resulta da sua própria dinâmica pelo que pode variar ao longo do tempo.

Num modelo de *scoring* comportamental, o comportamento futuro de um cliente é o resultado, não só da fotografia inicial em que o crédito é concedido, mas também de informação recente sobre o seu comportamento. Deste modo, ao contrário do que sucede nos modelos estáticos, a probabilidade do cliente incumprir assume um comportamento dinâmico fruto da evolução do seu próprio comportamento. Duas das possíveis abordagens neste tipo de modelos são a Análise de Sobrevivência e as Cadeias de Markov (Thomas, 2009). Outra possibilidade seria a utilização de modelos como, por exemplo, uma mistura de regressões de distribuições Poisson ou gama em que o objectivo passa por modelar a contagem de meses até ao momento de um cliente entrar em incumprimento. Porém, estes modelos, ao contrário do que sucede com os modelos de duração, além de ignorarem o factor tempo, não consideram a censura de dados.

Ao mesmo tempo que contribui para uma maior eficácia na concessão de crédito, com uma melhoria da qualidade da carteira de crédito, a utilização de modelos estatísticos na gestão de risco de crédito contribui também para uma maior eficiência, permitindo uma maior automatização de processos, traduzindo-se numa vantagem competitiva, tornando as tomadas de decisão mais rápidas e libertando os analistas de crédito da análise de propostas com perfis de risco mais definidos, ou seja, propostas com muito bom ou muito mau perfil de risco serão, respectivamente, aprovadas ou recusadas automaticamente[1], ficando apenas uma parcela das propostas de crédito sujeita à decisão do analista.

[1] Ficando pendentes apenas questões relacionadas com a validação dos dados dos clientes.

O objectivo deste estudo é criar um modelo preditivo que auxilie a tomada de decisões no seio da gestão de cobranças de uma instituição de crédito, tentando antecipar o comportamento de clientes numa carteira de crédito. Ao assumir que o comportamento de um cliente é dinâmico torna-se necessário incluir o factor tempo na análise e em vez de estimar a probabilidade de um cliente entrar em situação de incumprimento, o objectivo passa por estimar o momento, por exemplo, em que o cliente transita do estado de bom para mau cliente. Um cliente entra em incumprimento no momento em que falha o pagamento de uma prestação, altura em que fica sob vigilância da gestão de cobranças. Porém, considerar apenas este acontecimento, pode ser bastante escasso, na medida em que se ignora uma fase posterior em que o cliente transita, ou não, para uma situação de incumprimento mais avançado, com atrasos superiores a 90 dias, mais próximo de uma situação de incumprimento definitivo, denominado *default*. Assim, com o objectivo de avaliar a dinâmica do cliente ao logo do prazo do contrato de crédito, optou-se por considerar estes dois acontecimentos, registo do primeiro incidente e *default*, possibilitando assim antecipar a entrada de um cliente numa situação de gestão de cobranças e ao mesmo tempo avaliar o seu comportamento futuro, a partir do momento em que está sob gestão de cobranças.

Nesta tese de mestrado são apresentados modelos de sobrevivência de mistura, que estimam a duração até ao momento do cliente transitar para um estado de *default*, dado um conjunto informação sobre ao momento do cliente e a operação de crédito. Assim, este documento está organizado em 4 Capítulos para além desta Introdução e da Conclusão. No Capítulo 2 é feita uma descrição do *scoring* comportamental, explicando as vantagens que advêm da aplicação da análise de sobrevivência a este tipo de modelos de *scoring*. O Capítulo 3 começa com uma breve introdução à análise de sobrevivência e apresenta os conceitos e os tipos de modelos de análise de sobrevivência, desde os mais simples (agregados) aos mais complexos (heterogéneos). A estrutura dos dados e as variáveis utilizadas são apresentados no Capítulo 4, bem como uma pré-selecção das variáveis consideradas como covariáveis nos modelos de sobrevivência. No Capítulo 5 são estimados modelos relativos a cada acontecimento, sendo apresentado um modelo de imunidade para a duração até ao primeiro incidente e um modelo de sobrevivência de mistura, com 3 segmentos, para a duração entre o primeiro incidente e a transição para uma situação de *default*.

2. *Scoring* comportamental

Inicialmente, o principal objectivo dos modelos de *scoring* era apenas prever a probabilidade de um cliente entrar numa situação de *default* (PD), dado um conjunto de características do cliente e da operação, funcionando como modelo de classificação de possíveis clientes em bons ou maus clientes, auxiliando deste modo a decisão de conceder, ou não, crédito a um novo proponente.

Num modelo de *scoring* comportamental, a definição da situação do cliente é normalmente compatível com a definição utilizada num modelo de *scoring* aplicacional, sendo uma das mais comuns para classificar os clientes como maus clientes o facto de não ter pago três prestações mensais durante, por exemplo, os 12 a 18 meses seguintes. Porém, nos modelos comportamentais a instituição já conhece o cliente, sendo possível modelar o seu comportamento futuro com base, não só na informação utilizada para construir um modelo de *scoring* aplicacional, mas também em informação adicional sobre o seu comportamento recente, auxiliando na tomada de decisões relacionadas com os clientes que compõem a sua carteira de crédito. Por exemplo, num produto de cartões de crédito, ou conta corrente, os limites de crédito podem ser fixados em função de um modelo comportamental. Numa óptica de gestão de cobranças/crédito, este modelo pode ser útil ao identificar grupos de clientes com diferentes PD associadas, para os quais se podem definir diferentes estratégias de cobranças. Numa empresa de serviços (ex: comunicações), este tipo de modelos pode ajudar a minimizar perdas provenientes da transferência de clientes para empresas concorrentes, antecipando uma eventual fuga através de campanhas promocionais.

Thomas (2009) define o *scoring* aplicacional como um modo de ligar duas fotografias do cliente, a primeira relacionada com as suas características no momento de avaliação da proposta de crédito (ponto de observação) e a segunda do seu estado de incumprimento, por exemplo, 12 meses após a primeira fotografia, como demonstrado na Figura 2.1.

Figura 2.1. *Scoring* aplicacional

Num modelo comportamental, com a introdução de informação relativa ao desempenho recente do cliente, essa primeira fotografia é substituída por um filme do comportamento do cliente, durante um período de 6 a 12 meses até ao ponto de observação. Assim, o padrão de comportamento vai sofrendo alterações à medida que o contrato vai decorrendo. Essa informação pode ser fornecida por variáveis como, por exemplo, o envolvimento médio, máximo ou mínimo, o valor do crédito vencido, o número de rendas não pagas, entre outras, permitindo deste modo que a PD de um cliente sofra alterações ao longo do tempo (Thomas, 2009). Porém, como apresentado na Figura 2.2, a segunda fotografia mantém-se num modelo de *scoring* comportamental tradicional.

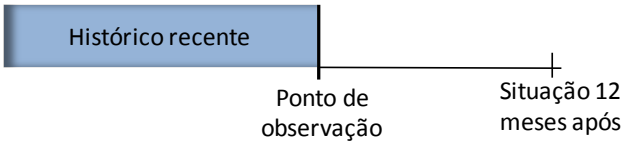

Figura 2.2. *Scoring* comportamental tradicional

A principal diferença entre os modelos representados nas Figuras 2.1 e 2.2 reside apenas na informação utilizada, na medida em que um modelo comportamental tradicional segue a mesma abordagem que um modelo aplicacional ao transformar informação sobre o comportamento recente do cliente num conjunto de medidas do seu comportamento.

Através de um modelo adequado é possível substituir a segunda fotografia por um filme do comportamento futuro do cliente. Este pode ser modelado tanto por modelos de sobrevivência como por modelos de cadeias de Markov (Thomas *et al.*, 2001). Assim, é possível estimar, não só a probabilidade de um cliente se encontrar em situação de *default* 12 meses após, mas também estimar a PD associada para um determinado momento futuro (3, 6 ou 9 meses após o ponto de observação, ver Figura 2.3), permitindo a tomada de decisões por parte da instituição com um *timing* mais eficiente.

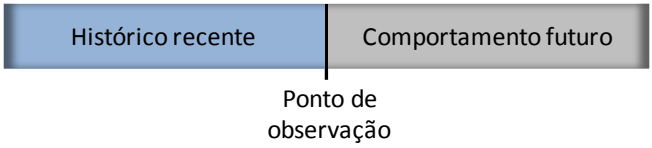

Figura 2.3. Desempenho histórico e comportamento futuro

A inclusão do factor tempo, em vez de uma variável binária que classifica o cliente em bom ou mau, através de um modelo de sobrevivência traduz-se também numa vantagem ao analisar a rendibilidade de uma operação de crédito. Por exemplo, considerando dois clientes com a mesma PD associada segundo um modelo aplicacional, em que o primeiro entra em *default* passados apenas 6 meses e o outro após 3 anos, apesar de um modelo de *scoring* aplicacional lhes atribuir a mesma PD, estas duas operações podem reflectir perdas e, consequentemente, rendibilidades muito diferentes. Bancos e outras instituições financeiras estão cada vez mais interessados um analisar a rendibilidade esperada de uma operação, em vez de estimarem apenas uma PD, pelo que o factor temporal não deve ser descurado na atribuição de novos limites de crédito e na definição do *pricing*. Assim, surge um outro tipo de modelos, *profit scoring*, que em vez de classificarem os clientes de crédito, em bons ou maus clientes, com base na sua PD, classificam os clientes/operações de crédito em função da sua rendibilidade (Thomas *et al.*, 2002).

Tratar a informação do modo apresentado nas Figuras 2.2 e 2.3, implica ignorar clientes que ainda não possuam suficiente informação histórica para modelar a PD, por exemplo nos 12 meses seguintes. Thomas *et al.* (2001) referem que uma das desvantagens dos modelos de *scoring* comportamental é necessitarem de informação histórica de cerca de 2 anos, o que implica que o tipo de clientes utilizados para estimar o modelo seja diferente daqueles a que o modelo será aplicado.

Um modelo de análise de sobrevivência, além de permitir modelar a duração até ao momento em que o cliente entra numa situação de incumprimento, com base apenas em informação utilizada em modelos aplicacionais, permite ainda considerar a censura de informação devido a numa determinada janela de observação o contrato de crédito não ter transitado ainda para uma situação de *default*, ou seja, sabe-se apenas que esse contrato não entrou em incumprimento até um determinado ponto temporal t, podendo ser considerado na estimativa da taxa de risco associada a um qualquer $T \leq t$. Deste modo, é possível incluir clientes em que o ponto de observação, apresentado na Figura 2.1, seja mais recente.

Stepanova e Thomas (2000) e Noh *et al*. (2005) comparam os resultados obtidos através do modelo de sobrevivência com outros métodos mais tradicionais, como a regressão logística e redes neuronais, e concluíram que os modelos de sobrevivência

apresentavam desempenhos muito semelhantes. Porém, ao contrário destes dois, como o modelo de sobrevivência não possui uma variável dependente binária, mas antes a duração *t*, permite calcular a taxa de risco acumulada para qualquer valor de *T*, semelhante à probabilidade de *default* do modelo de redes neuronais e da regressão logística. Neste estudo, os autores utilizaram variáveis explicativas demográficas e variáveis relacionadas com o comportamento passado do cliente (6 meses). Em 2001 Stepanova e Thomas, num estudo sobre clientes de uma instituição financeira do Reino Unido, realçaram ainda outra vantagem dos modelos de sobrevivência: a possibilidade de calcular a rendibilidade esperada de uma operação, uma vez que permite estimar, através da função sobrevivência, a probabilidade do cliente incumprir em determinado mês.

Thomas (2000) refere a importância de acrescentar nos modelos de *scoring* informação relacionada com a envolvente económica. Malik e Thomas (2010) e Bellotti e Crook (2007) utilizaram um modelo de análise de sobrevivência (regressão de Cox com proporcionalidade), com introdução de variáveis macroeconómicas, como a taxa de juro e o nível de desemprego, para estimar a duração até ao momento em que um cliente entra numa situação de *default*. Em particular, Malik e Thomas acrescentaram também no modelo de proporcionalidade um *score* comportamental interno como variável explicativa. Cao *et al.* (2009), seguindo as indicações do Basel II, utilizaram uma regressão de Cox com risco proporcional para modelar a probabilidade de *default* para um horizonte temporal de 12 meses, introduzindo o *scoring* como variável explicativa.

Andreeva *et al.* (2006) modelaram a duração entre a primeira compra por cartão de crédito e a segunda compra ou a entrada em situação de *default*, através de um modelo de análise de sobrevivência paramétrico com base numa distribuição exponencial. Os autores utilizaram variáveis explicativas relacionadas com as características do cliente (telefone fixo, tipo de habitação, estado civil, número de dependentes, profissão, idade do cliente, idade do cônjuge, antiguidade na morada e antiguidade no emprego) e da primeira compra (valor, tipo de produto, entre outros). Nesse estudo, os autores demonstraram que existe uma relação entre a rendibilidade e a duração até uma segunda compra/*default* e concluíram que é possível optimizar a rendibilidade de uma operação através da função de sobrevivência, em vez de considerar uma PD estática estimada por meio de uma regressão logística.

Porém, uma das desvantagens de modelos de análise de sobrevivência simples, em que a duração segue a mesma distribuição para todos os clientes, como os apresentados por Bellotti e Crook (2007) e Andreeva *et al.* (2006), aplicados à gestão de risco, está relacionada com o facto de considerarem que todos os clientes, mais tarde ou mais cedo, transitam para uma situação de *default*. Porém, uma carteira de crédito contém tipicamente uma proporção muito reduzida de clientes em situação de *default*, pelo que o pressuposto de modelos simples de sobrevivência é no mínimo questionável. Assim, Yildirim (2008) propôs um modelo de mistura multinível para incorporar o efeito de sobrevivência a longo prazo, em que cada cliente é agrupado, com base no tipo de propriedade e zona geográfica. Assim, o grupo de imunes ao risco (Z=1) e o grupo de clientes expostos ao risco (Z=0) é definido pela variável latente Z. Esta proporção de clientes imunes ao risco foi considerada por Beran e Djaïdja (2006), que estimou um modelo de sobrevivência com proporção de imunes, aplicado a um conjunto de clientes de um banco suíço. Neste modelo, os autores utilizaram uma função logística para definir a proporção de imunes e uma função exponencial para modelar a duração até ao momento em que o cliente entra em *default*, sendo o parâmetro da função exponencial modelado pelo montante do crédito e pelo *scoring* interno do banco, atribuído no passado a cada cliente.

A existência de um grupo de clientes que não estão expostos ao risco de crédito, é um pressuposto que, apesar de permitir incorporar no modelo de sobrevivência um padrão de comportamento que se traduz no cumprimento do crédito, pode mesmo assim ser escasso dada a heterogeneidade entre os clientes de crédito, podendo ser necessário estimar modelos de mistura com um número mais elevado de segmentos.

3. Análise de sobrevivência

3.1. Introdução

A análise de sobrevivência, ou modelos de duração, foi concebida inicialmente como uma forma de tratar informação relativa à mortalidade e tem como objectivo modelar a duração até à ocorrência de um determinado acontecimento, que implica a transição de um estado inicial para outro estado final[2].

Uma vantagem deste tipo de modelos é permitir a inclusão de observações em que a duração até ao acontecimento se encontra censurada. Esta censura ocorre quando temos apenas alguma informação sobre a duração, desconhecendo-se o momento exacto em que se regista o acontecimento. Ou seja, sabe-se apenas que, durante o período analisado, [0, τ], não tinha transitado de estado, o que significa que o período de sobrevivência é superior ao período de observação - dados censurados à direita. A Figura 3.1 mostra exemplos de episódios censurados resultantes do intervalo de tempo analisado.

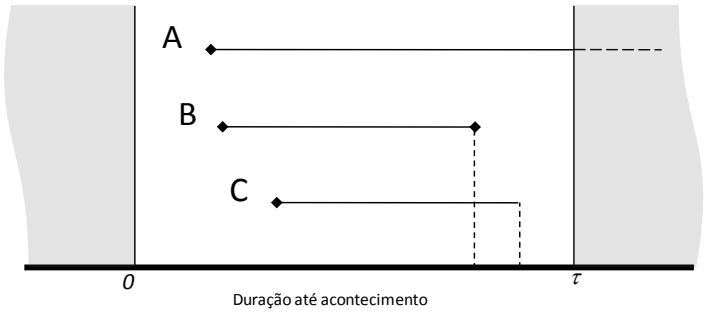

Figura 3.1. Dados censurados à direita

O episódio A é censurado à direita porque a janela temporal analisada limitou a sua observação para além de τ, sabendo-se apenas que até ao momento τ não ocorreu o acontecimento em estudo. O episódio B está completo, não existindo qualquer tipo de censura. O episódio C é censurado à direita porque terminou sem que tenha ocorrido o acontecimento. Se o acontecimento em estudo for a entrada em incumprimento de um cliente de um contrato de crédito, esta situação pode dever-se à rescisão antecipada do

[2] Para uma introdução à análise de sobrevivência, *vide* Kleinbaum e Klein (2005).

contrato de crédito ou então ter-se atingido o fim do prazo inicialmente acordado para o contrato de crédito, sem que o cliente transitasse de estado.

Existem ainda outros dois tipos de censura: à esquerda e intervalar. Dados censurados à esquerda ocorrem quando o período de sobrevivência é inferior ao período de observação. Por exemplo, num estudo sobre a duração até ao momento em que um indivíduo é infectado por um vírus, o período de observação decorre até uma resultado positivo de uma análise a esse vírus. Porém, não se sabe ao certo quando o indivíduo foi infectado, sabe-se apenas que ele foi infectado antes do resultado da análise. O outro tipo de censura, intervalar, está normalmente associada à periodicidade de recolha dos dados (Hosmer e Lemeshow, 1999), sabe-se apenas que o indivíduo transitou de estado algures entre o momento[3] $t_{j-1}t$ e t_j .

Num modelo de análise de sobrevivência, a distribuição do período de tempo decorrido até esse acontecimento, *T*, pode ser representada pela função de sobrevivência, $S(t) = P(T > t)$, que indica a probabilidade do acontecimento ocorrer após o tempo *t.*

A informação recolhida para modelar a função de sobrevivência condicional, $S(t|\boldsymbol{x})$, é representada pelo vectores $\boldsymbol{T}$, $\boldsymbol{X}$ *e* $\boldsymbol{\delta}$, em que T_i é a duração (tempo) observada para o indivíduo *i*, $\boldsymbol{X_i}$ é o vector das variáveis explicativas para o indivíduo *i* e δ_i indica se o indivíduo transitou de estado no momento definido por T_i. Deste modo, a função de sobrevivência condicional apresenta a seguinte fórmula:

$$S(t|\boldsymbol{x}) = P(T > t|\boldsymbol{X} = \boldsymbol{x}) = \int_t^{\infty} f(u|\boldsymbol{x})du, \qquad (3.1)$$

em que $f(t|\boldsymbol{x})$ representa a função densidade de probabilidade.

Além da função de sobrevivência, a distribuição de *T* pode também ser caracterizada pela taxa de risco (*hazard rate*) $h(t|\boldsymbol{x})$ e pela função de risco acumulada (*cumulative hazard function*) $\Lambda(t|\boldsymbol{x})$ do seguinte modo:

$$h(t|\boldsymbol{x}) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t | T \geq t, \boldsymbol{X}=\boldsymbol{x})}{\Delta t} = \frac{f(t|\boldsymbol{x})}{S(t|\boldsymbol{x})} \qquad (3.2)$$

[3] No âmbito do risco de crédito, os dados apresentam-se normalmente numa base mensal pelo que *T* (variável discreta) representa o número de meses decorridos até ao momento em que o cliente transita para um estado de incumprimento, ignorando-se em que dia exactamente um cliente entrou em incumprimento.

$$\Lambda(t|\boldsymbol{x}) = \int_0^t h(u|\boldsymbol{x})du = \int_0^t \frac{f(u|\boldsymbol{x})}{S(u|\boldsymbol{x})}du. \quad (3.3)$$

Assim, a função de sobrevivência e probabilidade acumulada podem ser representadas da seguinte forma:

$$S(t|\boldsymbol{x}) = e^{-\Lambda(t|\boldsymbol{x})} \quad (3.4)$$

$$F(t|\boldsymbol{x}) = 1 - S(t|\boldsymbol{x}). \quad (3.5)$$

Neste estudo foram definidos dois tipos de modelos de sobrevivência: agregados e com heterogeneidade. Os primeiros podem ainda ser classificados como não paramétricos, semi-paramétricos e paramétricos. Os modelos com heterogeneidade caracterizam-se por considerarem uma mistura de distribuições, admitindo a presença de heterogeneidade, permitindo assim a criação de grupos. Dependendo do acontecimento em estudo, estes últimos modelos tornam-se úteis também por permitirem a inclusão de uma proporção de indivíduos imunes ao risco, ou seja, embora esta proporção possa não fazer muito sentido se o objecto de estudo for a mortalidade, no âmbito do risco de crédito é perfeitamente natural considerar uma proporção de imunes.

3.2. Modelos agregados

- **Modelos não paramétricos**

Com o objectivo de evitar perda de informação proveniente de dados censurados à direita, Kaplan e Meier (1958) introduziram um estimador que permite estimar a função de sobrevivência, $S(t)$, considerando simultaneamente dados censurados e não censurados. A fórmula de Kaplan-Meier (KM) para a probabilidade de sobrevivência até ao período $t_j\,t$ é a seguinte:

$$S(t_j) = S(t_{j-1}) \times P(T > t_j | T \geq t_j), \quad (3.6)$$

em que a probabilidade de sobrevivência, $S(t_{j-1})$, pode ser definida como o produto das parcelas que estimam a probabilidade para $t_{j-1}t$ e períodos anteriores:

$$S(t_{j-1}) = \prod_{i=1}^{j-1} P(T > t_i | T \geq t_i) \quad (3.7)$$

$$P(T > t_i | T \geq t_i) = 1 - \frac{Individuos\ \ que\ \ incumpriram\ \ \ em\ t_i}{Individuos\ \ que\ \ sobreviveram\ \ \ até\ t_i}. \quad (3.8)$$

Este estimador (KM) enquadra-se num tipo de análise de sobrevivência não paramétrica. Este tipo de modelos não assumem nenhum pressuposto sobre a distribuição da taxa de risco, nem sobre o modo como as covariáveis a afectam. É uma abordagem que normalmente se enquadra numa fase de análise descritiva da função sobrevivência, em que o efeito das covariáveis é analisado estratificando os dados em grupos, de modo a representar diferentes curvas de sobrevivência e de risco para cada grupo.

Uma forma de avaliar se as curvas de sobrevivência diferem significativamente entre cada grupo é analisando os respectivos intervalos de confiança e verificar se eles se sobrepõem ou não. Também com recurso a testes estatísticos é possível avaliar diferenças entre as curvas de sobrevivência, nomeadamente o teste *log-rank*, que testa a hipótese nula da curva de sobrevivência ser a mesma para todos os *g* grupos. A estatística do teste segue uma distribuição χ^2 com *g-1* graus de liberdade. Tal como outras estatísticas, noutros tipos de teste do χ^2 recorrem-se às frequências observadas e frequências esperadas para cada categoria da variável analisada (Kleinbaum e Klein, 2005). Outros testes alternativos para testar a igualdade da curva de sobrevivência podem ser utilizados como os testes de *Wilcoxon*, *Tarone-Ware* e *Peto* (Lee e Wang, 2003)

Outro modelo não paramétrico de análise de sobrevivência também utilizada como técnica descritiva são as Tábuas de Mortalidade (*Life Tables)*. Uma limitação deste tipo de abordagem é não permitir comparar um número muito elevado de grupos, que tornaria a análise muito confusa. Além disso, para analisar o efeito de variáveis contínuas seria necessário proceder à sua transformação numa variável categorial.

- **Modelos semi-paramétricos**

À semelhança do que sucede com os modelos de sobrevivência não paramétricos, os modelos semi-paramétricos também não assumem qualquer pressuposto sobre a distribuição da taxa de risco, porém, assumem alguns pressupostos sobre o efeito das covariáveis sobre a taxa de risco, nomeadamente, que as taxas de risco são proporcionais entre os grupos para todo o espaço temporal. Este conceito de proporcionalidade na função de risco foi introduzido por Cox (1972) - *proportional hazard* (PH) *function* - e permitiu relacionar características individuais com a taxa de

risco de um acontecimento para determinado período através de um modelo de regressão de Cox à semelhança do que sucede num modelo de *credit scoring*. Assim, sendo $\boldsymbol{x}$ o vector das variáveis explicativas, pode definir-se a taxa de risco (*hazard rate*) do seguinte modo:

$$h(t|\boldsymbol{x}) = g(\boldsymbol{x})\, h_0(t). \tag{3.9}$$

Isto significa que as características de um cliente/operação têm um efeito multiplicador sobre uma taxa de risco base, $h_0(t)$, calculada para o período *t*. Uma vez que a taxa de risco não pode assumir valores negativos, $g(\boldsymbol{x})$ é definida como uma função exponencial das variáveis explicativas:

$$h(t|\boldsymbol{x}) = e^{\boldsymbol{\beta x}}\, h_0(t), \tag{3.10}$$

e a probabilidade de sobrevivência além do período *t* resulta da seguinte fórmula:

$$S(t|\boldsymbol{x}) = [S_0(t)]^{e^{\boldsymbol{\beta x}}}, \tag{3.11}$$

em que $S_0(t)$ representa a função de sobrevivência base para o período *t.*

Enquanto que $h_0(t)$ depende do período de tempo, os coeficientes associados às variáveis explicativas, $\boldsymbol{\beta}$, mantêm-se constantes ao longo do tempo, definindo apenas a magnitude da função ao longo do tempo, ou seja, assume-se que as variáveis explicativas são independentes do factor tempo.

Para aferir sobre o pressuposto de proporcionalidade acima referido é possível recorrer tanto a abordagens gráficas como a testes estatísticos relacionados com o ajustamento. Uma das abordagens gráficas mais comuns é através das curvas de sobrevivência -ln-ln, que consiste em comparar *-ln(-ln)* das curvas de sobrevivência para cada grupo definido por uma variável categorial e verificar se essas curvas são paralelas. Aplicando esta transformação em (3.11) temos:

$$-ln[-ln\, S(t|\boldsymbol{x})] = -ln[-ln\, S_0(t)] - \boldsymbol{\beta x}. \tag{3.12}$$

Outra forma de avaliar graficamente este pressuposto é comparando a curva de sobrevivência observada (*KM*) com a estimada através de uma regressão de *Cox* com proporcionalidade de risco. Uma aproximação dessas duas curvas de sobrevivência,

para uma categoria da variável analisada, sugere que o pressuposto de proporcionalidade é razoável.

Porém, a abordagem gráfica para avaliar do pressuposto de proporcionalidade é um pouco subjectiva, pelo que é conveniente recorrer a testes estatísticos. Assim, uma forma mais objectiva de avaliar o pressuposto da proporcionalidade é através da análise dos resíduos, ou seja, se os resíduos não estiverem correlacionados com a duração então o pressuposto de proporcionalidade não é violado. Através dos resíduos *Schoenfeld*, definidos para cada variável explicativa no modelo estimado, é possível concluir sobre a proporcionalidade de cada uma das variáveis (Hosmer e Lemeshow, 1999).

Caso não se verifique o pressuposto de proporcionalidade relativamente a uma variável categorial com *k* categorias, esta pode ser utilizada como uma variável de estratificação, dando origem a *k* taxas de risco base (*baseline hazards*), mantendo-se porém os pesos $\boldsymbol{\beta}$ constantes em todos os *k* níveis de estratificação.

- **Modelos paramétricos**

Num modelo de sobrevivência paramétrico pressupõe-se que a duração segue uma determinada distribuição com função densidade de probabilidade $f(t)$, aplicando-se a fórmula em (3.1) para calcular a função de sobrevivência e a fórmula definida em (3.2) para calcular a taxa de risco. Relativamente ao modo como as covariáveis afectam a taxa de risco pode assumir-se dois tipos de efeitos: (i) que a covariáveis têm um efeito multiplicador sobre uma taxa de risco (PH); (ii) ou que as covariáveis têm um efeito multiplicador sobre o factor tempo (*accelerated failure time models* - AFT).

Um modelo de risco proporcional (PH) assume uma função da taxa de risco base comum a todos os indivíduos, que se desloca para cima, ou para baixo, como resultado do efeito das covariáveis, semelhante à regressão de Cox, com a diferença da taxa de risco seguir uma distribuição conhecida. A função da taxa de risco e função de sobrevivência são semelhantes às definidas para a regressão de Cox em (3.10) e (3.11), respectivamente. Por seu turno, num modelo AFT as variáveis têm um feito multiplicador sobre o factor tempo, aumentando ou reduzindo a duração até à ocorrência do acontecimento. A taxa de risco resulta da seguinte fórmula:

$$h(t|\boldsymbol{x}) = e^{\boldsymbol{\beta x}}\, h_0\left(t\, e^{\boldsymbol{\beta x}}\right), \qquad (3.13)$$

e a probabilidade de sobrevivência além do período *t* resulta da seguinte fórmula:

$$S(t|\boldsymbol{x}) = S_0\left(t\, e^{\boldsymbol{\beta x}}\right). \tag{3.14}$$

Algumas das distribuições normalmente consideradas num modelo paramétrico de análise de sobrevivência são: a exponencial, a Weibull, a log-logística e a log-normal.

Distribuição exponencial

A distribuição mais simples utilizada em análise de sobrevivência é a distribuição exponencial, conhecida pela "ausência de memória". Se a duração *t* segue uma distribuição exponencial com parâmetro λ, então a função densidade de probabilidade é definida do seguinte modo:

$$f(t) = \lambda e^{-\lambda t}, \tag{3.15}$$

com a função de sobrevivência e função de risco definidas respectivamente por:

$$S(t) = e^{-\lambda t} \tag{3.16}$$

$$h(t) = \lambda. \tag{3.17}$$

A taxa de risco quando *t* segue uma distribuição exponencial é constante e o acontecimento em causa é aleatório e independente de *t*. O valor esperado e a variância são: $E[T] = 1/\lambda$ e $VAR[T] = 1/\lambda^2$.

Distribuição Weibull

A distribuição Weibull é uma generalização da distribuição exponencial. Se a duração *t* segue uma distribuição Weibull, com parâmetros λ e *p*, então a função densidade de probabilidade, função de sobrevivência e função de risco são definidas, respectivamente, do seguinte modo:

$$f(t) = p\lambda^p t^{p-1} e^{-(\lambda t)^p} \tag{3.18}$$

$$S(t) = e^{-(\lambda t)^p} \tag{3.19}$$

$$h(t) = p\lambda^p t^{p-1}. \tag{3.20}$$

e valor esperado e a variância são dados, respectivamente por: $E[T] = \lambda^{-1}\Gamma(1 + \mathrm{p}^{-1})$ e $VAR[T] = \lambda^{-2}\left[\Gamma(1 + 2p^{-1}) - \Gamma^{2}(1 + p^{-1})\right]$.

A Figura 3.2 apresenta taxas de risco associadas a uma distribuição Weibull com λ=0.5 e diferentes valores para *p*.

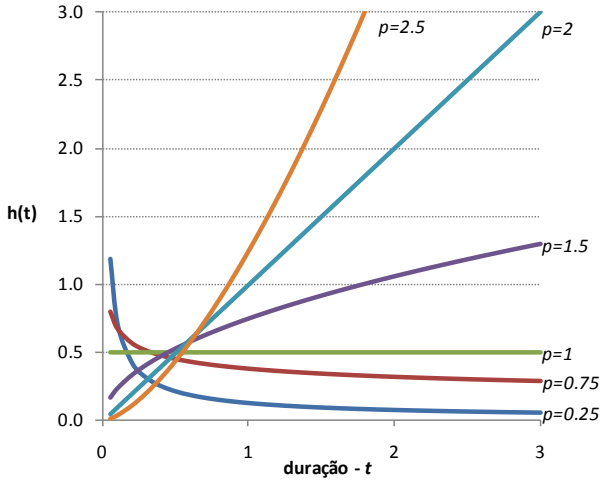

Figura 3.2. Taxas de risco para uma distribuição Weibull (λ=0.5)

Ao contrário do modelo exponencial, o modelo Weibull não assume uma taxa de risco constante, tornando-se deste modo numa distribuição mais flexível. No caso específico em que $p = 1$, os modelos Weibull e exponencial são equivalentes. Se $0 < p < 1$ então a taxa de risco é representada por uma função monótona decrescente e quando $p > 1$ a taxa de risco é representada por uma função monótona crescente.

Distribuição log-normal

Um modelo log-normal assume que o logaritmo da duração (*t*) segue uma distribuição normal com parâmetros μ e σ. Deste modo, a função densidade de probabilidade, função de sobrevivência e função de risco são definidas, respectivamente, da seguinte forma:

$$f(t) = \frac{1}{t\sigma}\phi\left[\frac{\ln(t)-\mu}{\sigma}\right] \tag{3.21}$$

$$S(t) = 1 - \Phi\left[\frac{\ln(t)-\mu}{\sigma}\right] \tag{3.22}$$

$$h(t) = \frac{1}{t\sigma}\frac{\phi\left[\frac{\ln(t)-\mu}{\sigma}\right]}{1-\Phi\left[\frac{\ln(t)-\mu}{\sigma}\right]}, \tag{3.23}$$

onde ϕ e Φ representam, respectivamente, a função densidade de probabilidade e a função distribuição de uma normal-padrão:

$$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{\left(-\frac{t^2}{2}\right)} \tag{3.24}$$

$$\Phi(t) = \int_{-\infty}^{t} \phi(\tau) d\tau . \tag{3.25}$$

O valor esperado e a variância são dados, respectivamente por: $E[T] = \exp(\mu + \sigma^2/2)$ e $VAR[T] = (\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)$. A Figura 3.3 apresenta taxas de risco associadas a uma distribuição log-normal com diferentes valores para os parâmetros μ e σ.

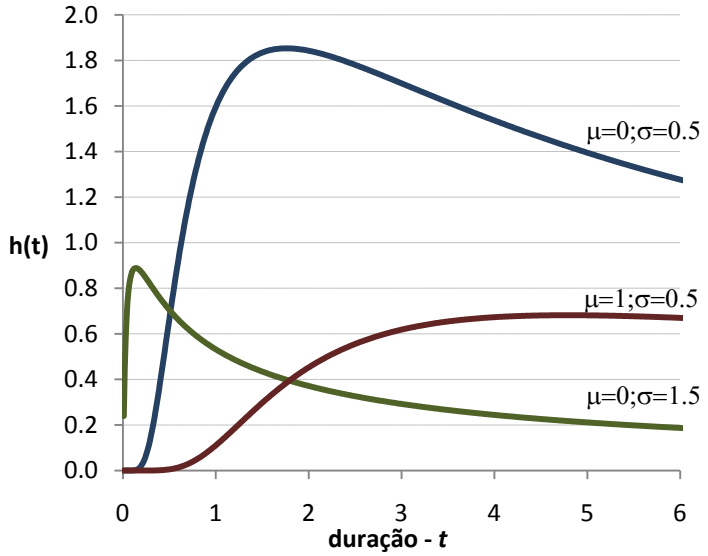

Figura 3.3. Taxas de risco para uma distribuição log-normal

Um modelo log-normal caracteriza-se por uma relação não monótona entre a taxa de risco e a duração, em que a taxa de risco cresce inicialmente até um máximo e decresce posteriormente.

Distribuição log-logística

Um modelo baseado numa distribuição log-logística assume que a duração (*t*) segue uma distribuição log-logística com parâmetros λ e α. A função densidade de probabilidade, função de sobrevivência e função de risco associadas a este tipo de modelos são, respectivamente[4]:

[4] No manual do Stata 10 é referido 1/γ em vez de α.

$$f(t) = \frac{\alpha \lambda^{\alpha} t^{\alpha-1}}{[1+(\lambda t)^{\alpha}]^2} \quad (3.26)$$

$$S(t) = \frac{1}{1+(\lambda t)^{\alpha}} \quad (3.27)$$

$$h(t) = \frac{\alpha \lambda^{\alpha} t^{\alpha-1}}{1+(\lambda t)^{\alpha}}. \quad (3.28)$$

O valor esperado e a variância são dados, respectivamente por: $E[T] = \frac{\pi\alpha}{\lambda \sin(\pi/\alpha)}$ e $VAR[T] = \lambda^{-2}\left[\frac{2\pi/\alpha}{\sin(2\pi/\alpha)} - \frac{(\pi/\alpha)^2}{\sin^2(\pi/\alpha)}\right]$. A Figura 3.4 apresenta as taxas de risco para uma distribuição log-logística para $\lambda = 1$ e diferentes valores para o parâmetro α.

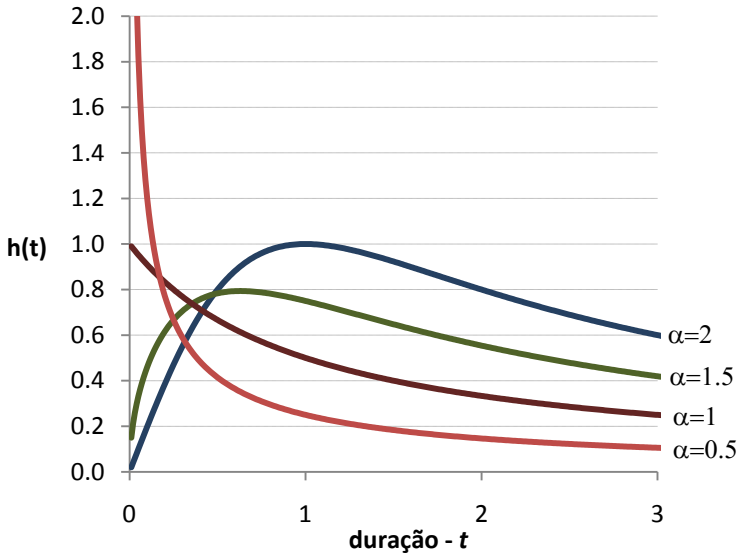

Figura 3.4. Taxas de risco para uma distribuição log-logística (λ=1)

Os modelos baseados nesta distribuição são bastante flexíveis. Como se mostra na Figura 3.4, para $\alpha \leq 1$ obtém-se uma relação monótona decrescente entre a taxa de risco e a duração e para $\alpha > 1$ uma relação não monótona, com um comportamento semelhante ao do modelo log-normal.

3.3. Modelos com heterogeneidade

Os modelos de sobrevivência agregados assumem que a função de risco se encontra completamente especificada dada uma função de risco base e um conjunto de variáveis explicativas e que na sua essência os indivíduos formam uma população homogénea. Lee e Timmermans (2007) referem num estudo sobre deslocações que introduzir uma classe latente é benéfico, por representar as diferenças de comportamentos nos

indivíduos. Os modelos de análise de sobrevivência apresentados nesta secção assumem que não existe apenas uma função de risco base comum para todos os indivíduos, mas antes, consideram a presença de grupos heterogéneos com diferentes funções de risco associadas. Essa heterogeneidade pode ser contínua ou discreta, consoante o número de segmentos associados seja infinito ou finito, respectivamente. O mesmo é dizer que a latente Z segue uma distribuição contínua ou discreta.

- **Modelos de mistura**

Um modelo de sobrevivência de mistura permite considerar vários grupos de indivíduos, em que a duração *t* até ao acontecimento segue uma distribuição com diferentes parâmetros. Vakratsas e Bass (2002) estimaram um modelo de sobrevivência de mistura com o objectivo de identificar e caracterizar diferentes segmentos de agregados familiares em termos da sua taxa de compra (risco), e a sua propensão para reduzir a duração entre compras periódicas em função de campanhas de marketing promocionais. Se considerarmos um modelo de mistura com *S* componentes (grupos), então a função densidade de probabilidade e função de sobrevivência são definidas do seguinte modo:

$$f(t|\boldsymbol{\varphi}) = \sum_{s=1}^{S} \pi_s \, f(t|\boldsymbol{\theta}_s) \tag{3.29}$$

$$S(t|\boldsymbol{\varphi}) = \sum_{s=1}^{S} \pi_s \, S(t|\boldsymbol{\theta}_s), \tag{3.30}$$

em que $\boldsymbol{\varphi}$ representa todos os parâmetros do modelo $(\boldsymbol{\pi}, \boldsymbol{\theta})$, $\boldsymbol{\theta}_s$ representa os parâmetros da função densidade de probabilidade associado ao grupo *s* e π_s a probabilidade *a priori* de um indivíduo pertencer ao grupo *s*, ou a sua proporção, e satisfazem as seguintes condições:

i) $\pi_s > 0$;

ii) $\sum_{s=1}^{S} \pi_s = 1$.

No modelo apresentado em (3.29) assume-se que π_s é constante para cada grupo, o que significa que todos os indivíduos têm a mesma probabilidade de pertencer a cada grupo. O modelo concomitante permite que as proporções de mistura $\boldsymbol{\pi}$ sejam uma função de covariáveis (Dias, 2004), ou seja, $\pi_{is} = g(\boldsymbol{w}_i, \boldsymbol{\gamma}_s)$, em que $\boldsymbol{\gamma}_s$ representa o conjunto de

parâmetros para a componente *s*. Assim, as proporções de mistura podem ser estimadas considerando uma função de ligação logit multinomial:

$$\pi_{is} = \frac{e^{\boldsymbol{\gamma_s w_i}}}{\sum_{s=1}^{S} e^{\boldsymbol{\gamma_s w_i}}} \tag{3.31}$$

em que $\boldsymbol{\gamma_s}$ representa os coeficientes associados às variáveis explicativas $\boldsymbol{w_i}$, para a probabilidade do indivíduo pertencer ao grupo *s*, com $\boldsymbol{\gamma}_S = \mathbf{0}$ para que o modelo esteja identificado.

Considerar a existência de vários grupos, com diferentes parâmetros para uma determinada distribuição, permite por exemplo criar uma função de risco global muito mais flexível, mesmo quando é considerada uma função monótona para cada grupo. Jiang e Murthy (1998) mostram como modelar, por exemplo, a taxa de risco através de uma mistura de duas Weibull e identificaram oito tipos de curvas para a função de risco: duas monótonas e seis do tipo "montanha russa". Uma função de risco com um comportamento não monótono sugere a existência de heterogeneidade dos indivíduos (Vakratsas, 1998).

- **Modelos com proporção de imunes**

Os tipos de modelos de análise de sobrevivência anteriormente apresentados não permitem considerar a existência de um grupo de indivíduos imunes ao risco, ou seja, para os quais não se regista o acontecimento em estudo. Esses modelos assumem que se um indivíduo for observado por um período suficientemente longo ele irá registar um incidente:

$$\lim_{t\to\infty} S(t|\boldsymbol{x}) = 0. \tag{3.32}$$

Um modelo de imunidade, ou cura, permite considerar duas proporções de indivíduos: (i) π - proporção dos imunes (*cured*); e (ii) $(1-\pi)$ proporção dos indivíduos expostos ao risco (*uncured*). Abreu e Rocha (2006) referem que a existência de um elevado número de observações censurada, pode sustentar uma proporção de indivíduos imunes ao risco. As autoras criaram um modelo de cura, considerando a distribuição de Chen, aplicado a pacientes com leucemia mielógena aguda avançada, aos quais foi feito um transplante de medula óssea. De Angelis *et al.* (1999) estimaram um modelo de

sobrevivência relativa com mistura, utilizando dados de pacientes com cancro do cólon. Este tipo de modelos tem também sido aplicado a outras áreas, nomeadamente na área do risco de crédito, como referido no Capítulo 2 o estudo de Beran e Djaïdja (2006).

Assim, a função de sobrevivência neste tipo de modelos é composta por duas parcelas, a dos imunes $S_c(t|\boldsymbol{x})$ e a dos não imunes $S_u(t|\boldsymbol{x})$:

$$\begin{aligned} S(t|\boldsymbol{x}) &= \pi S_c(t|\boldsymbol{x}) + (1-\pi)\, S_u(t|\boldsymbol{x}) \\ &= \pi + (1-\pi)\, S_u(t|\boldsymbol{x}). \end{aligned} \tag{3.33}$$

Uma vez que os indivíduos pertencentes ao grupo dos imunes nunca irão registar o acontecimento, então $\lim_{t\to\infty} S_c(t|\boldsymbol{x}) = 1$. A função de sobrevivência relativa ao grupo dos não imunes pode ser modelada considerando uma das distribuições dos modelos agregados. Sendo Z uma variável latente binária, em que $Z = 1$ representa um indivíduo imune e $Z = 0$ um indivíduo em risco (não imune), é possível modelar $\pi = P(Z = 1)$ através de uma regressão logística:

$$\pi(\boldsymbol{w}_i) = \frac{1}{1+e^{-\gamma \boldsymbol{w}_i}}, \tag{3.34}$$

em que $\boldsymbol{\gamma}$ representa os coeficientes associados às variáveis explicativas $\boldsymbol{w}_i$ da probabilidade do indivíduo pertencer ao grupo dos imunes. Kuk e Chen (1992) utilizaram um modelo de imunidade, combinando uma função logística para estimar a proporção de imunes com um modelo de risco proporcional semi-paramétrico. Ortega *et al.* (2009) estimaram um modelo de imunidade, analisando o efeito de covariáveis simultaneamente na duração e na proporção de imunes, num estudo com pacientes com melanoma de pele, sujeitos a um tratamento pós-operatório. Introduzindo o modelo logístico (3.34) em (3.33), a função sobrevivência total surge da seguinte forma:

$$S(t|\boldsymbol{x}, \boldsymbol{w}) = \pi(\boldsymbol{w}) + [\,1 - \pi(\boldsymbol{w})]\, S_u(t|\boldsymbol{x}). \tag{3.35}$$

Considerando $f_u(t|\boldsymbol{x})$ como a função densidade de probabilidade associada ao grupo dos não imunes, tem-se as seguintes funções densidade de probabilidade e risco global:

$$f(t|\boldsymbol{x}, \boldsymbol{w}) = [1 - \pi(\boldsymbol{w})]\, f_u(t|\boldsymbol{x}) \tag{3.36}$$

$$h(t|\boldsymbol{x}, \boldsymbol{w}) = \frac{[1 - \pi(\boldsymbol{w})]\, f_u(t|\boldsymbol{x})}{\pi(\boldsymbol{w}) + [1 - \pi(\boldsymbol{w})]\, S_u(t|\boldsymbol{x})}. \tag{3.37}$$

- **Modelos de fragilidade**

Nos modelos de fragilidade é introduzido um efeito aleatório não observado (α) na função de risco de modo a considerar a heterogeneidade dos indivíduos. Para este tipo de modelos, a taxa de risco e a função de sobrevivência, considerando o pressuposto de proporcionalidade, são apresentadas da seguinte forma, para o indivíduo *i*:

$$h(t_i|\alpha_i) = \alpha_i h(t_i) \tag{3.38}$$

$$S(t_i|\alpha_i) = S(t_i)^{\alpha_i}. \tag{3.39}$$

Especificar o modelo deste modo permite distinguir na análise entre um nível individual - $h(t_i|\alpha_i)$ e $S(t|\alpha_i)$ - e um nível populacional - $h(t_i)$ e $S(t_i)$. Se $\alpha_i > 1$, então o indivíduo está mais exposto ao risco por motivos que não estão directamente relacionados com as covariáveis incluídas no modelo. Se a componente de fragilidade for comum dentro de um grupo de indivíduos (*shared frailty*) então a taxa de risco, para o modelo multinível, para o indivíduo *i* pertencente ao grupo *j* é apresentada da seguinte forma:

$$h(t_{ij}|\alpha_j) = \alpha_j h(t_{ij}). \tag{3.40}$$

Por questões de identificação, assume-se que $\alpha > 0$ segue uma determinada distribuição $g(\alpha)$, com $\mu = E[\alpha] = 1$ e $\sigma^2 = Var(\alpha) = \theta$ (Gutierrez, 2002), sendo este último o parâmetro usualmente estimado. A função não condicional de sobrevivência para a população é dada por:

$$S_\theta(t) = \int_0^\infty [S(t)]^\alpha \, g(\alpha) d\alpha \tag{3.41}$$

em que θ realça que o modelo depende da variância do coeficiente de fragilidade.

3.4. Estimação

O método de máxima verosimilhança (MLE - *maximum likelihood estimation*) permite estimar um modelo para um tipo de distribuição, que resulta numa função de verosimilhança específica. De um modo geral, a função de verosimilhança é simplesmente o produto das funções densidade de cada indivíduo. No caso concreto da análise de sobrevivência, normalmente caracterizada pela censura de dados, a função de verosimilhança inclui também informação relativa a essa censura. Para uma sub-

amostra de indivíduos com censura, sabe-se apenas que não registaram o acontecimento durante o período de observação *t*, em que probabilidade do acontecimento ocorrer após o tempo *t* é dada pela função de sobrevivência $S(t|x)$. Considerando d_i uma variável binária em que $d_i = 1$ indica que não houve censura e $d_i = 0$ indica que a observação foi censurada, a função de verosimilhança surge da seguinte forma:

$$L(\boldsymbol{\varphi}|t, \boldsymbol{x}) = \prod_{i=1}^{n} f(t_i|\boldsymbol{x}_i)^{d_i}\, S(t_i|\boldsymbol{x}_i)^{1-d_i}. \tag{3.42}$$

O logaritmo da função de verosimilhança é dado por:

$$\ln L(\boldsymbol{\varphi}|t, \boldsymbol{x}) = \sum_{i=1}^{n}\{d_i \ln f(t_i|\boldsymbol{x}_i) + (1-d_i)\ln S(t_i|\boldsymbol{x}_i)\} \tag{3.43}$$

ou de forma equivalente por:

$$\ln L\,(\boldsymbol{\varphi}|t, \boldsymbol{x}) = \sum_{i=1}^{n}\{d_i \ln h(t_i|\boldsymbol{x}_i) + \ln S(t_i|\boldsymbol{x}_i)\}. \tag{3.44}$$

As funções apresentadas em (3.42) e (3.43) consideram apenas a censura de dados à direita. Por exemplo, considerando que *t* segue uma distribuição log-normal com parâmetros $\boldsymbol{\varphi} = (\mu, \sigma)$, a contribuição do indivíduo *i* para o valor do logaritmo da função de verosimilhança, assumindo apenas a censura de dados à direita, resulta na seguinte fórmula[5]:

$$\ln L_{\,i} = -d_i\left\{\ln t_i + \lambda + \ln\sqrt{2\pi} + \frac{(\ln t_i - \mu)^2}{2e^{2\lambda}}\right\} + \\ + (1-d_i)ln\left\{1 - \Phi\left(\frac{\ln t_i - \mu}{e^{\lambda}}\right)\right\}, \tag{3.45}$$

em que $\sigma = exp(\lambda)$ por forma a garantir $\sigma > 0$.

Considerando o logaritmo da função de verosimilhança definida em (3.43), um modelo de análise de sobrevivência com mistura de *S* componentes é estimado considerando a seguinte função log-verosimilhança:

$$\ln L = \sum_{i=1}^{n}\{d_i \ln f(t_i|\boldsymbol{x}_i) + (1-d_i)\ln S(t_i|\boldsymbol{x}_i)\}. \tag{3.46}$$

[5] Substituindo em (3.42), a função densidade de probabilidade e função de sobrevivência definidas em (3.20) e (3.21), respectivamente.

Substituindo em (3.42) as funções densidade de probabilidade e sobrevivência por forma a incorporar uma proporção de imunes, π, obtém-se a seguinte função de verosimilhança para um modelo de duração de mistura:

$$L = \prod_{i=1}^{n} [(1-\pi) f_u(t_i|\boldsymbol{x}_i)]^{d_i} [\pi + (1-\pi) S_u(t_i|\boldsymbol{x}_i)]^{(1-d_i)}, \tag{3.47}$$

e logaritmo da função de verosimilhança é dado por:

$$\begin{aligned} l(\boldsymbol{\varphi}) = \ln L = \sum_{i=1}^{n} \{ & d_i \ln[(1-\pi) f_u(t_i|\boldsymbol{x}_i)] \\ & + (1-d_i) \ln[\pi + (1-\pi) S_u(t_i|\boldsymbol{x}_i)]\}. \end{aligned} \tag{3.48}$$

Ao incluir uma proporção de imunes em (3.45), a contribuição individual para o valor do logaritmo da função de verosimilhança para a distribuição log-normal resulta na seguinte fórmula:

$$\begin{aligned} \ln L_i = {} & d_i \ln(1-\pi) - d_i \left\{ \ln t_i + \lambda + \ln \sqrt{2\pi} + \frac{(\ln t_i - \mu)^2}{2e^{2\lambda}} \right\} + \\ & + (1-d_i) ln \left\{ \pi + (1-\pi) \left[1 - \Phi \left(\frac{\ln t_i - \mu}{e^{\lambda}} \right) \right] \right\}, \end{aligned} \tag{3.49}$$

em que π pode ser modelado, por exemplo, através de uma função de ligação do tipo logit, como apresentado em (3.34).

3.5. Diagnóstico do modelo

Os vários modelos podem ser comparados através da análise do pseudo-R^2, normalmente utilizado em modelos não lineares, calculado da seguinte forma:

$$pseudo - R^2 = 1 - \frac{l(\hat{\boldsymbol{\varphi}})}{l(\hat{\boldsymbol{\varphi}}_0)}, \tag{3.50}$$

em que $l(\hat{\boldsymbol{\varphi}}_0)$ é o valor da log-verosimilhança do modelo composto apenas por constantes e $l(\hat{\boldsymbol{\varphi}})$ o valor da log-verosimilhança do modelo avaliado com parâmetros $\boldsymbol{\varphi}$. Outro modo de seleccionar o modelo mais apropriado de entre um conjunto de modelos candidatos é através dos critérios de informação, AIC e BIC, que penalizam o acréscimo de parâmetros a estimar,

$$AIC = -2l(\hat{\boldsymbol{\varphi}}) + 2k \tag{3.51}$$

$$BIC = -2l(\hat{\boldsymbol{\varphi}}) + \ln(n)\, k \tag{3.52}$$

com *k* parâmetros estimados e *n* observações.

Estes critérios podem ser úteis não só para avaliar a inclusão de novas variáveis, mas também para decidir sobre qual a distribuição a utilizar na modelação da duração (*t*), uma vez que o número de parâmetros a estimar varia consoante a distribuição.

Porém, escolher o melhor modelo com base nos critérios acima apresentados não garante por si só que o modelo seleccionado seja o mais adequado, existindo a possibilidade de não serem respeitados os pressupostos do modelo específico. Uma abordagem gráfica para avaliar se os pressupostos do modelo são respeitados, consiste em representar a função de sobrevivência obtida por um modelo não-paramétrico (KM) e a função de sobrevivência obtida pelo modelo que se pretende avaliar. Através de uma transformação adequada da função de sobrevivência obtida pelo modelo paramétrico que se pretende avaliar, de modo a que o resultado seja uma função linear que possa ser avaliada visualmente. Por exemplo, considerando a função de sobrevivência de um modelo baseado numa distribuição log-normal, definida em (3.22) é possível obter a seguinte transformação linear:

$$\Phi^{-1}(1 - S(t)) = -\frac{\mu}{\sigma} + \frac{1}{\sigma}\ln(t) \tag{3.53}$$

em que Φ^{-1} representa a inversa da função de distribuição de uma normal-padrão.

Se o modelo baseado numa distribuição log-normal for adequado, então a relação entre $\Phi^{-1}(1 - S(t))$ e $\ln(t)$ deve ser aproximadamente linear.

Outro modo de avaliar se os pressupostos do modelo estimado são respeitados é através da análise de resíduos, permitindo analisar desvios do modelo estimado e perceber de que forma o modelo pode ser melhorado. Blossfeld *et al.* (2007) referem uma análise de resíduos com base nos resíduos generalizados de Cox e Snell (1968) definidos da seguinte forma para o indivíduo *i*:

$$\hat{e}_i = \int_o^{t_i} \hat{h}(u|\boldsymbol{x}_i)du, \tag{3.54}$$

equivalente a[6]:

[6] Ver (3.3) e (3.4).

$$\hat{e}_i = -\ln \hat{S}(t_i|\boldsymbol{x}_i). \tag{3.55}$$

Assim, os resíduos generalizados, ou pseudo-resíduos, são definidos como uma taxa de risco acumulada. Se o modelo especificado for adequado, os resíduos devem seguir aproximadamente uma distribuição exponencial, à qual está associada uma taxa de risco constante, independente da duração *t*. Recorrendo ao estimador Kaplan-Meier é possível estimar a função de sobrevivência dos resíduos $S_{\hat{e}}(e)$ e comparar graficamente $-\ln[S_{\hat{e}}(e)]$ com $\hat{e}_i$ como apresentado na Figura 3.5. Se o modelo se ajustar aos dados, então o gráfico deverá apresentar uma linha recta com declive = 1.

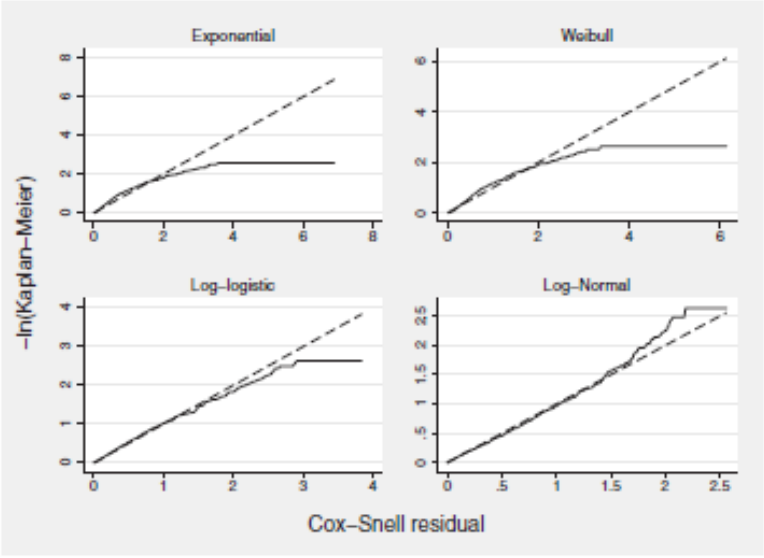

Figura 3.5. Exemplo de análise gráfica dos pseudo-resíduos
(Blossfeld *et al.*, 2007, p.222)

A Tabela 3.1 resume os principais modelos anteriormente apresentados, bem como outros modelos estimados no Capítulo 5, para o caso específico de uma distribuição log-normal. Estes modelos foram estimados recorrendo ao *software* Stata 10.

A terminologia apresentada na Tabela 3.1, que será utilizada neste estudo para identificar o tipo de modelo e os parâmetros modelados, é *M(.)[K1(.) K2(.) Z(.)]*, em que M_A = modelo agregado, M_I = modelo com imunidade e M_S = modelo de mistura com *S* segmentos. O parâmetro *K1* é parâmetro de localização e *K2* o parâmetro de dispersão da distribuição. A variável *Z* indica as variáveis usadas na função de ligação logit. *K1(.)* identifica o parâmetro *K1* como sendo uma constante e *K1(X)* como sendo modelado. O mesmo se aplica para *K2* e *Z*. Por exemplo, $M_I[\mu(.)\ \sigma(.)\ Z(X)]$, respeita a um modelo

com imunidade com modelação da proporção de imunes em que a duração dos não imunes segue uma distribuição log-normal.

Tabela 3.1. Síntese de modelos de análise de sobrevivência com distribuição log-normal

Modelação	**Tipo de Modelo**		
	M_A - Agregado $S = 1$	**M_I - Com imunes** $S = 2$ $\pi_u = \frac{1}{1+e^{-z_u}}$	**M_S - Com Mistura** $S \geq 2$ $\pi^s = \frac{e^{z_s}}{1+\sum_{s=1}^{S-1} e^{z_s}}, z_S = 0$
*μ(.) lnσ(.) [Z(.)]**	$\mu=\beta_0$ $ln\ \sigma=\alpha_0$	$\mu_u=\beta_0$ $ln\ \sigma_u=\alpha_0$ $z_u=\gamma_0$	$\mu_s=\beta_{0,s}$, s = 1, … , S $ln\ \sigma_s=\alpha_{0,s}$, s=1,…,S $z_s=\gamma_{0,s}$, s=1,…,S-1
μ(.) lnσ(.) Z(X)	N.A.	$\mu_u=\beta_0$ $ln\ \sigma_u=\alpha_0$ $z_u = \boldsymbol{\gamma_u w}$	$\mu_s=\beta_{0,s}$, s = 1, … , S $ln\ \sigma_s=\alpha_{0,s}$, s=1,…,S $z_s = \boldsymbol{\gamma_s w}$, s=1,…, S-1
*μ(X) lnσ(.) [Z(.)]**	$\mu = \boldsymbol{\beta x}$ $ln\ \sigma=\alpha_0$	$\mu_u = \boldsymbol{\beta_u x}$ $ln\ \sigma_u=\alpha_0$ $z_u=\gamma_0$	$\mu_s = \boldsymbol{\beta_s x}$, s = 1, … , S $ln\ \sigma_s=\alpha_{0,s}$, s=1,…,S $z_s=\gamma_{0,s}$, s=1,…,S-1
*μ(X) lnσ(X) [Z(.)]**	$\mu = \boldsymbol{\beta x}$ $ln\ \sigma = \boldsymbol{\alpha x}$	$\mu_u = \boldsymbol{\beta_u x}$ $ln\ \sigma_u = \boldsymbol{\alpha_u x}$ $z_u=\gamma_0$	$\mu_s = \boldsymbol{\beta_s x}$, s = 1, … , S $ln\ \sigma_s = \boldsymbol{\alpha_s x}$, s = 1, … , S $z_s=\gamma_{0,s}$, s=1,…,S-1
*μ(X) lnσ() [Z(X)]**	N.A.	$\mu_u = \boldsymbol{\beta_u x}$ $ln\ \sigma_u=\alpha_0$ $z_u = \boldsymbol{\gamma_u w}$	$\mu_s = \boldsymbol{\beta_s x}$, s = 1, … , S $ln\ \sigma_s=\alpha_{0,s}$, s=1,…,S $z_s = \boldsymbol{\gamma_s w}$, s=1,…, S-1
*μ(X) lnσ(X) [Z(X)]**	N.A.	$\mu_u = \boldsymbol{\beta_u x}$ $ln\ \sigma_u = \boldsymbol{\alpha_u x}$ $z_u = \boldsymbol{\gamma_u w}$	$\mu_s = \boldsymbol{\beta_s x}$, s = 1, … , S $ln\ \sigma_s = \boldsymbol{\alpha_s x}$, s = 1, … , S $z_s = \boldsymbol{\gamma_s w}$, s=1,…, S-1
Comandos STATA	*streg*	*cureregr / strsmix*	N.d.

* Não aplicável nos modelos agregados.

Os modelos agregados, *M_A[K1(X) K2(X)]*, podem ser estimados com o comando *streg* do Stata. Este comando além da distribuição log-normal, permite considerar outras funções na distribuição da duração t, exponencial, Gompertz, log-logística, Weibull e a

gama generalizada, com a introdução de covariáveis para cada parâmetro da função considerada. Porém, o *streg* não permite considerar mistura de distribuições.

No Stata os modelos de sobrevivência com imunes, dependendo da distribuição desejada, podem ser estimados utilizando os comandos *cureregr* (Buxton, 2007) e *strsmix* (Lambert, 2007)[7]. O comando *cureregr* tem como opções para a distribuição de *T* as funções Weibull, log-normal, logística, gama e exponencial. O comando *strsmix* permite considerar as funções Weibull, log-normal, gama, mistura de uma Weibull com uma exponencial e mistura de duas Weibull, para a distribuição de *T*. Em ambos os comandos, a proporção de imunes pode ser modelada através de uma função de ligação tipo logit.

Nenhum dos modelos descritos na coluna da direita da Tabela 3.1 pode ser estimado através dos comandos disponíveis no Stata, pelo que foi necessário programar o respectivo estimador de máxima verosimilhança no Stata. É necessário especificar a função de log-verosimilhança via o modelo, $M(.)$, a distribuição da variável dependente e as variáveis explicativas.

3.6. Ensaio com dados simulados

Para melhor compreender e validar os procedimentos nos modelos estimados na aplicação empírica, foi efectuado um ensaio com base em dados simulados, considerando um modelo base em que a duração *t* segue uma distribuição log-normal, com 5 covariáveis ($x1 - x5$), como especificado na Tabela 3.2.

Tabela 3.2. Modelo log-normal gerado

Parâmetro	**Coeficiente**
μ	
x1	0.00
x2	-0.75
x3	1.25
x4	-0.50
x5	-0.50
_cons	1.00
σ	
_cons	1.00

[7] Estes comandos podem ser encontrados através do comando *findit.*

A amostra utilizada nos ensaios tem dimensão *n*=10000 pelo que as estimativas são muito próximas dos valores dos parâmetros reais. Assim, foram feitos 4 ensaios, considerando 4 bases de dados diferentes:

i) M_A **[μ(X) lnσ(.)] sem dados censurados**. Foi utilizada uma amostra retirada de uma população gerada de acordo com o modelo especificado na Tabela 3.2;

ii) M_A **[μ(X) lnσ(.)]**. Foram acrescentados dados censurados à amostra utilizada para o modelo i);

iii) M_I **[μ(X) lnσ(.) Z(.)]**. Com censura e proporção de imunes π. Foi acrescentada uma proporção (20%)[8] de dados imunes à base de dados utilizada para estimar o modelo ii);

iv) M_I **[μ(X) lnσ(.) Z(X)]**. Com censura e proporção de imunes π=*f(w)*. Semelhante ao modelo iii), mas π é função logit de uma covariável (w) e de uma constante.

Na Tabela 3.3 apresentam-se os coeficientes dos modelos estimados, mostrando que as estimativas estão próximas dos verdadeiros valores dos parâmetros. Comparando as estimativas dos modelos i) e ii) é possível concluir que o efeito da censura sobre os coeficientes das covariáveis é mínimo. Os coeficientes apresentados pelos modelos iii) e iv) mostram que mesmo adicionando uma proporção de imunes, os coeficientes associados aos não imunes estão próximos dos reais.

Tabela 3.3. Coeficiente dos modelos ensaiados

Parâmetros	**Modelo verdadeiro**	**Modelo i)**	**Modelo ii)**	**Modelo iii)**	**Modelo iv)**
$ln[\pi/(1-\pi)]$					
w	-1.00	n.a.	n.a.	n.a.	-0.9755
_cons	0.10	n.a.	n.a.	-1.2319	0.0942
μ					
x1	0.50	0.4976	0.4975	0.4840	0.4999
x2	-0.75	-0.7454	-0.7342	-0.7713	-0.7437
x3	1.25	1.2375	1.2512	1.2604	1.2289
x4	-0.50	-0.5008	-0.5151	-0.4791	-0.5029
x5	-0.50	-0.4972	-0.4983	-0.4987	-0.4822
_cons	1.00	0.9946	1.1319	1.0221	1.0011
σ					
_cons	1.00	1.0006	1.0166	0.9945	1.0020

[8] $z = \ln\left(\frac{\pi}{1-\pi}\right) \approx -1.3863$

4. Dados

Os dados consistem num conjunto de 8680 clientes de uma instituição financeira de crédito aos quais foi concedido crédito entre Janeiro de 2007 e Julho de 2009. Os dados recolhidos fornecem informação, não só sobre as condições inicialmente acordadas e segundo as quais o contrato irá decorrer, nomeadamente o prazo da operação, o valor da prestação e a taxa, mas também informação de carácter demográfico relativa aos intervenientes (cliente e avalistas), nomeadamente idade, estado civil, tipo de habitação, rendimento, entre outros. Além de dados relativos ao momento da concessão do crédito foi também recolhida informação mensal sobre o desempenho do cliente, como a data em que o cliente registou o primeiro atraso no pagamento da prestação e a data em que o cliente entrou em situação de *default*. A última "fotografia" recolhida respeita a Agosto de 2009.

Os dados utilizados na construção do modelo de sobrevivência são apurados com base neste último conjunto de dados longitudinais tendo sido definidos dois acontecimentos. O primeiro acontecimento está relacionado com a transição de um estado em que o cliente não regista qualquer atraso para um estado em que regista um primeiro incidente. O segundo acontecimento regista-se quando o cliente transita de um estado de atraso até 90 dias para um atraso superior a 90 dias, i.e., entra em *default*. Outras situações podem ser também classificadas em *default* sem que a antiguidade da dívida ultrapasse os 90 dias, por exemplo casos em que haja uma manifesta incapacidade por parte do cliente.

Por questões de confidencialidade, a amostra foi seleccionada de forma a distorcer os índices de incumprimento associados à carteira de crédito da instituição financeira. Optou-se, assim, por construir uma amostra composta por uma proporção idêntica de clientes que entram ou não em situação de *default*. Este procedimento de emparelhamento é comum por exemplo nos estudos das causas de falências de empresas (Hensher e Jones, 2004).

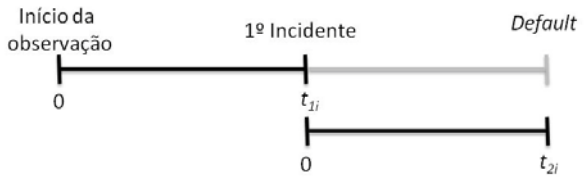

Figura 4.1. Abordagem condicional tipo II

Uma vez que para atingir uma situação de *default* o cliente teve obrigatoriamente de passar por uma situação de registo de incidente, está implícita uma ordem entre acontecimentos, impondo-se assim uma abordagem que permita a introdução de eventos recorrentes, nomeadamente, uma abordagem condicional do tipo II (Prentice *et al.*, 1981) que se centra na análise da duração entre os dois acontecimento, outra alternativa seria uma abordagem marginal (Wei *et al.*, 1989) que se centra na duração total entre o início do período de observação e o acontecimento, mas que ignora a característica recorrente destes acontecimentos. A Figura 4.1 mostra a dinâmica inerente a uma abordagem condicional tipo II, para o cliente *i*, em que t_{1i} representa a duração até o primeiro incidente e t_{2i} a duração entre o registo do primeiro incidente e a transição para uma situação de *default*. Deste modo, a estrutura dos dados utilizada no modelo de duração encontra-se apresentada na Tabela 4.1.

Tabela 4.1. Estrutura de dados utilizada para o modelo de duração

ID	**Intervalo**	**Duração**	**Acontecimento**	$\mathbf{X_1}$	$\mathbf{X_2}$	**…**	$\mathbf{X_j}$
1	1	t_{11}	1	x_{11}	x_{21}	…	x_{j1}
1	2	t_{21}	1	x_{11}	x_{21}	…	x_{j1}
2	1	t_{12}	1	x_{12}	x_{22}	…	x_{j2}
2	2	t_{22}	0	x_{12}	x_{22}	…	x_{j2}
3	1	t_{13}	0	x_{13}	x_{23}	…	x_{j3}
…	…	…	…	…	…	…	…
n	1	t_{1n}	0	x_{1n}	x_{2n}	…	x_{jn}

Em que *X* representa as variáveis explicativas utilizadas para modelar o tempo até ao acontecimento ou momento de censura dos dados. A Tabela 4.2 descreve as variáveis consideradas como candidatas a entrar no modelo como variáveis explicativas.

A selecção de variáveis teve por base estudos anteriores que consideram a idade do cliente, o montante financiado, o prazo da operação, o tipo de produto, o número de dependentes, um número de telefone fixo, profissão, existência de avalistas, rendimento anual, estado civil e tipo de habitação (Stepanova e Thomas, 2000; Andreeva *et al*., 2006; Sarlija *et al*., 2009; Till e Hand, 2003). Outros autores introduziram, entre outras variáveis, um *scoring* interno como variável explicativa da duração até o cliente entrar em *default* (Beran e Djaïdja, 2006; Malik e Thomas, 2010; Cao *et al*., 2009). Além das utilizadas em estudos anteriores foram recolhidas outras variáveis, com base na experiência do autor, que além de poderem melhorar a capacidade explicativa dos modelos, podem também servir de indicação para estudos futuros. Estas variáveis são (*vide* Tabela 4.2): *Idadebem*, *eurotax*, *loantovalue*, *taxa*, *prestacao* e *taxaesforco*.

De notar que as variáveis *loantovalue*[9], *taxaesforço*[10] e *prazoremincidente*[11] são calculadas com base em variáveis apresentadas na Tabela 4.2, pelo que estamos na presença de colinearidade perfeita. Porém, optou-se por incluir para já estas variáveis no estudo, deixando para depois de uma análise univarida, através de uma regressão de Cox, a decisão sobre as variáveis a excluir.

Tabela 4.2. Descrição de variáveis candidatas a covariáveis

Variável	**Tipo**	**Descrição**
Idadebem	Métrica	Antiguidade (anos) do bem financiado no início do contrato
eurotax	Métrica	Valor (u.m.) do bem no início do contrato
vf	Métrica	Valor (u.m.) do crédito concedido
loantovalue	Métrica	Relação entre o valor do crédito concedido e o valor do colateral
taxa	Métrica	Taxa nominal acordada no início do contrato de crédito
prazo	Métrica	Duração do contrato em meses acordada no início do crédito
prestacao	Métrica	Valor (u.m.) da prestação mensal do contrato de crédito
produto	Nominal	Tipo de produto (1=A, 2=B, 3=C)
scoring	Ordinal	Nível de scoring (*applicant scoring*) global atribuído à operação
numinterv	Ordinal	Número de intervenientes no contrato (cliente, cotitular e avalistas)
Idadecliente	Métrica	Idade do cliente no início do contrato de crédito
dependentescli	Ordinal	Número de dependentes do cliente no início do contrato de crédito
telefempregofixo	Nominal	Indica se o cliente comunicou telefone fixo da entidade patronal (1=Telefone fixo da entidade patronal)
telefcontactofixo	Nominal	Indica se o cliente possui telefone fixo (1=Possui telefone fixo)
habitacao	Nominal	Tipo de habitação do cliente (1=Arrendada, 2=Outra, 3=Própria com hipoteca, 4= Própria sem hipoteca)
estadocivil	Nominal	Estado civil do cliente (1=Casado, 2=Divorciado, 3=Solteiro, 4=Viúvo)
rendimensal	Métrica	Rendimento mensal do cliente (u.m.)
tipo2interv	Nominal	Tipo de segundo interveniente (1=Avalista, 2=Cotitular, 3=Único titular)
taxaesforco	Métrica	Taxa de esforço (*prestacao* / *rendimensal*)
tincidente	Métrica	Número de meses decorridos até o cliente registar o 1º incidente.
prazoremincidente	Métrica	Prazo remanescente quando o cliente registou o 1º incidente.

As variáveis *numinterv* e *dependentescli* foram recodificadas de modo a agregar na mesma categoria observações com valores mais elevados. Assim, a variável *numinterv*

[9] $loantovalue = vf/eurotax$
[10] $taxaesforco = prestacao\ /\ rendimensal$
[11] $prazoremincidente = prazo - tincidente$

assume valor 3 em operações de crédito com 3 ou mais intervenientes e a variável *dependentescli* assume valor 2 para clientes com 2 ou mais dependentes (e passando a variável ordinal).

A Tabela 4.3 resume a distribuição de clientes por intervalo (1: incidente; 2: *default*) e que registaram o acontecimento.

Tabela 4.3. Distribuição de clientes por tipo de acontecimento

Intervalo	Acontecimento		Total
	0	1	
1 – Incidente	3278	5402	8680
2 - *Default*	1749	3653	5402

Durante o período analisado, 38% dos clientes não registaram qualquer incumprimento e 58% dos clientes não registaram atrasos superiores a 90 dias. De notar que as observações do intervalo 2 – *default* – dizem respeito a clientes que registaram o acontecimento no intervalo 1 – Incidente.

As Figuras 4.2 e 4.4 mostram a função de sobrevivência segundo o estimador de Kaplan-Meier para cada um dos acontecimentos em estudo e as Figuras 4.3 e 4.5 a respectiva taxa de risco.

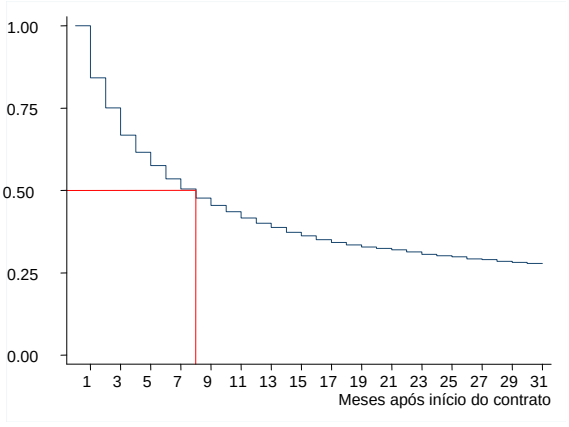

Figura 4.2. Função sobrevivência KM (1º Incidente)

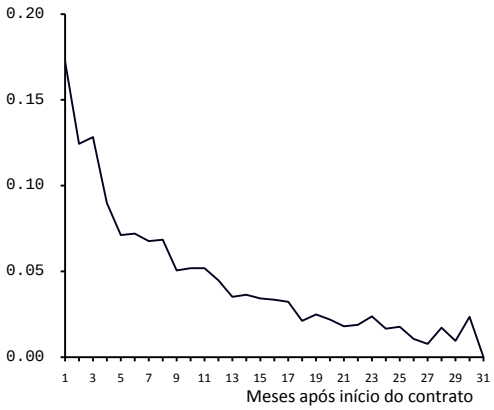

Figura 4.3. Taxa de risco (1º Incidente)

As Figuras 4.3 e 4.5 mostram uma clara diferença no comportamento da taxa de risco associada a cada um dos acontecimentos. A taxa de risco associada ao primeiro atraso é mais elevada nos meses iniciais do contrato de crédito diminuindo progressivamente com o decorrer do contrato atingindo valores muito próximos de 0, que resulta num decréscimo cada vez menos acentuado da função de sobrevivência. Na Figura 4.2 é

possível observar que a função sobrevivência associada ao primeiro acontecimento parece estabilizar ligeiramente acima de 0.25, o que sugere a existência de um grupo de clientes que cumpre o contrato sem registar qualquer atraso no pagamento das prestações, ou seja, um grupo de "imunes" ao risco de crédito.

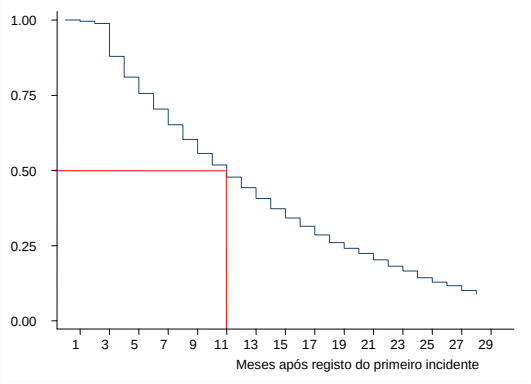

Figura 4.4. Função sobrevivência KM (*default*)

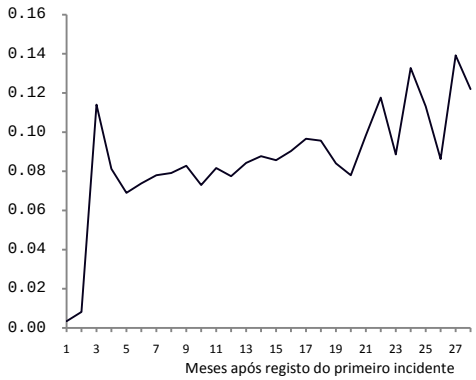

Figura 4.5. Taxa de risco (*default*)

Contrariamente, a taxa de risco associada ao segundo acontecimento apresenta um comportamento ligeiramente crescente ao longo da vida do contrato. De notar que, a taxa de risco para os primeiros 2 meses após o registo do primeiro incidente é muito próxima de zero, registando um forte acréscimo para 3 meses. Este comportamento da taxa de risco sugere a existência de um grupo de clientes que com uma transição muito rápida para um situação de *default*, deixando de pagar as prestações precedentes após um primeiro incumprimento, entrando em *default* em apenas 3 meses após o registo do primeiro incidente, como exemplificado na Figura 4.6.

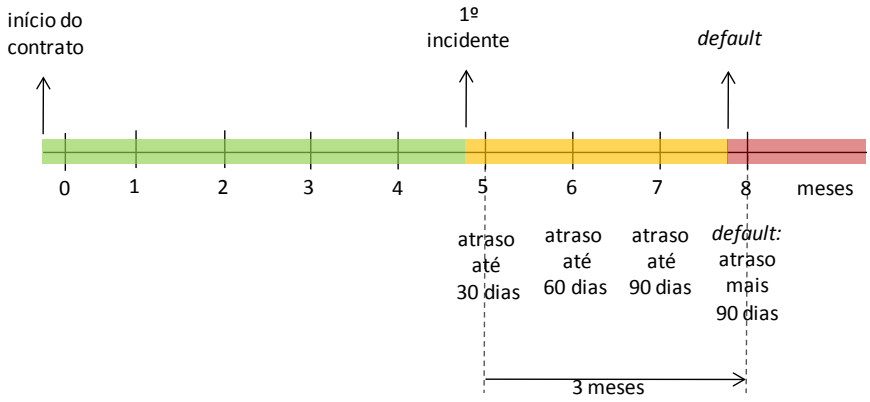

Figura 4.6. Exemplo: duração até *default*

De qualquer modo, em termos globais a transição para um estado em que o cliente já registou atrasos é mais rápida, comparativamente com uma transição posterior para uma situação de *default*, como se pode concluir através da análise da mediana do tempo até ao acontecimento (linha vermelha nos gráficos das Figuras 4.2 e 4.4), em 50% dos clientes da amostra regista um primeiro incidente durante os primeiros 8 meses do contrato e 50% dos clientes que registam incidentes entram em *default* 11 meses após o registo do primeiro incidente.

Dado o conjunto de variáveis candidatas apresentadas na Tabela 4.2 e para avaliar a capacidade explicativa de cada uma nos modelos de sobrevivência, para ambos os acontecimentos começou-se por analisar as variáveis não métricas, sendo posteriormente analisadas as variáveis métricas aplicando uma regressão univariada de Cox.

Assim, para avaliar se existem diferenças significativas nas funções de sobrevivência associadas a cada uma das categorias das variáveis não métricas apresentadas na Tabela 4.2 recorreu-se ao teste de igualdade *Log-Rank* e *Wilcoxon*, cuja hipótese nula estabelece que a função de sobrevivência é comum a todas as categorias da variável. Os resultados dos testes são apresentados nas Tabela 4.4 e 4.5 para o 1º e 2º acontecimento, respectivamente.

Tabela 4.4. Resultados do teste de igualdade *Log-Rank* e *Wilcoxon-Breslow-Gehan* (1º incidente)

Variável (não métrica)	**Log-Rank**		**Wilcoxon-Breslow-Gehan**	
scoring	$\chi^2(6)=$	1110.65 ***	$\chi2(6)=$	1024.08 ***
numinterv	$\chi^2(2)=$	134.66 ***	$\chi2(2)=$	150.91 ***
dependentescli	$\chi^2(2)=$	7.04 *	$\chi2(2)=$	14.83 ***
telefempregofixo	$\chi^2(1)=$	68.59 ***	$\chi2(1)=$	54.40 ***
telefcontactofixo	$\chi^2(1)=$	81.84 ***	$\chi2(1)=$	83.54 ***
habitacao	$\chi^2(3)=$	41.33 ***	$\chi2(3)=$	47.22 ***
estadocivil	$\chi^2(3)=$	110.08 ***	$\chi2(3)=$	118.32 ***
tipo2interv	$\chi^2(2)=$	160.46 ***	$\chi2(2)=$	179.99 ***
produto	$\chi^2(2)=$	172.08 ***	$\chi2(2)=$	179.08 ***

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Tabela 4.5. Resultados do teste de igualdade *Log-Rank* e *Wilcoxon-Breslow-Gehan (default)*

Variável (não métrica)	**Log-Rank**	**Wilcoxon-Breslow-Gehan**
scoring	$\chi^2(6)=$ 317.03 ***	$\chi2(6)=$ 290.25 ***
numinterv	$\chi^2(2)=$ 168.59 ***	$\chi2(2)=$ 172.71 ***
dependentescli	$\chi^2(2)=$ 19.99 ***	$\chi2(2)=$ 26.99 ***
telefempregofixo	$\chi^2(1)=$ 25.43 ***	$\chi2(1)=$ 19.76 ***
telefcontactofixo	$\chi^2(1)=$ 39.91 ***	$\chi2(1)=$ 40.94 ***
habitacao	$\chi^2(3)=$ 46.46 ***	$\chi2(3)=$ 60.30 ***
estadocivil	$\chi^2(3)=$ 64.64 ***	$\chi2(3)=$ 64.68 ***
tipo2interv	$\chi^2(2)=$ 162.47 ***	$\chi2(2)=$ 170.33 ***
produto	$\chi^2(2)=$ 35.22 ***	$\chi2(2)=$ 33.05 ***

* p < 0.05; ** p < 0.01; *** p < 0.001.

Os valores dos testes são significativos para todas as variáveis não métricas, pelo que nesta fase nenhuma variável é descartada, continuando como possíveis variáveis explicativas ou de estratificação na análise aqui apresentada.

Ainda relativamente às variáveis não métricas, mais concretamente às ordinais, interessa também analisar se o comportamento da taxa de risco para cada categoria é coerente com a ordem implícita nas categorias. Nas Figuras 4.7 a 4.9 são apresentadas as taxas de risco associadas a ambos os acontecimentos para as variáveis *scoring*, *numinterv* e *dependentescli*.

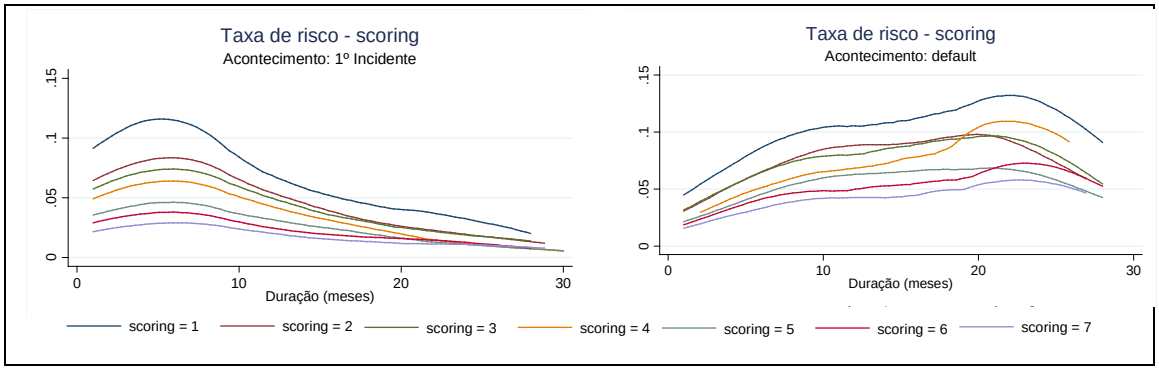

Figura 4.7. Taxa de risco por nível de *scoring*.

As taxas de risco associadas a cada nível de *scoring* são globalmente coerentes com a ordem implícita nos níveis de *scoring*, ou seja, taxas de risco mais elevadas estão associadas a níveis de *scoring* mais baixos, verificam-se apenas algumas incoerências para durações mais prolongadas, provavelmente devido a um número reduzido de observações.

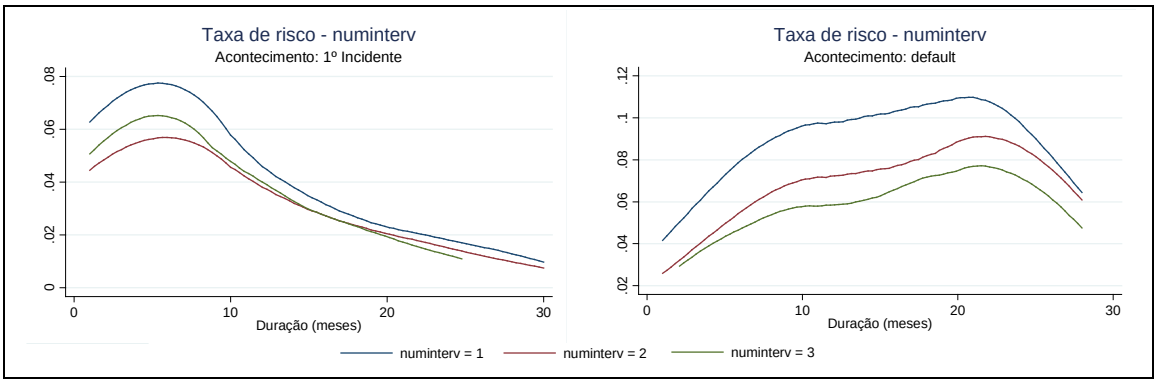

Figura 4.8. Taxa de risco por número de intervenientes.

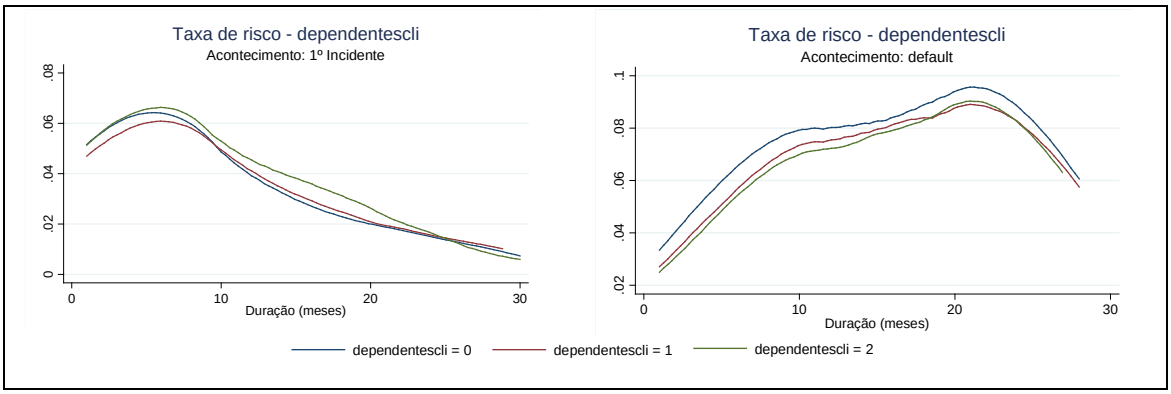

Figura 4.9. Taxa de risco por número de dependentes.

Relativamente à variável ordinal *numinterv*, parece não haver grande diferença nas taxas de risco entre operações com 2 ou mais intervenientes para o primeiro acontecimento, o que sugere que o principal factor de diferenciação para este acontecimento é o facto de o cliente ser ou não o único interveniente. Esta proximidade das taxas de risco para 2 ou mais intervenientes deixa de ser evidente para o segundo acontecimento – *default*. Na Figura 4.10 são apresentadas as taxas de risco por tipo de segundo interveniente, para cada acontecimento. No que respeita ao primeiro acontecimento, as taxas associadas a cada categoria de *tipo2interv* apresentam-se mais destacadas, quando comparadas com as categorias de *numinterv*, pelo que esta variável não será considerada nos modelos relativos ao primeiro acontecimento. De notar que também a variável *tipo2interv*[12] fornece informação sobre se o cliente é ou não o único interveniente e ao mesmo tempo, caso não seja, fornece informação se o segundo interveniente é um mero avalista ou cotitular do contrato de crédito. Optar entre estas duas variáveis, no caso do segundo acontecimento, equivale a optar entre a quantidade de intervenientes (*numinterv*) na proposta e a qualidade dos segundos intervenientes (*tipo2interv*).

[12] Tipo de segundo interveniente (1=Avalista, 2=Cotitular, 3=Único titular).

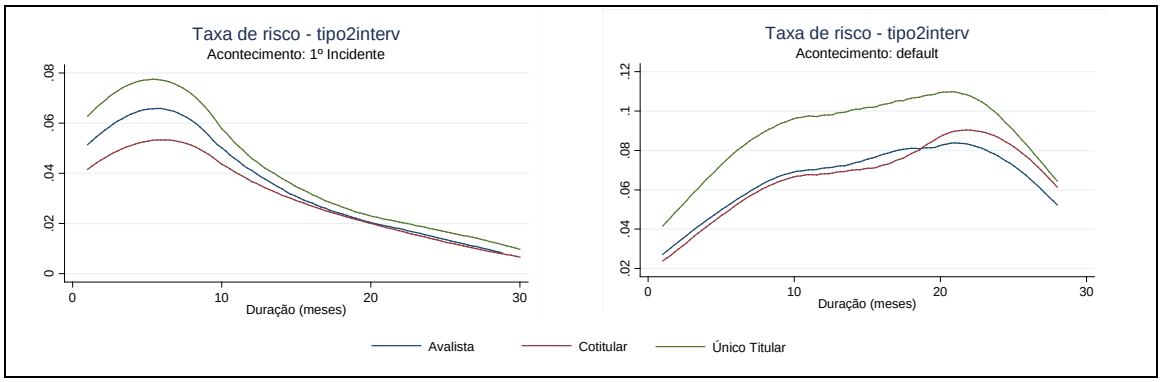

Figura 4.10. Taxa de risco por tipo de 2º interveniente.

As taxas de risco associadas às categorias de *dependentescli* apresentam uma maior proximidade comparativamente com as apresentadas para *scoring* e *numinterv*. Além disso, não apresentam um comportamento coerente com a ordem das categorias para o primeiro acontecimento. Apesar de clientes com 2 ou mais dependentes apresentarem taxas de risco globalmente mais elevadas para o primeiro acontecimento, esta situação altera-se para a taxa de risco relativa ao *default*, em que clientes sem dependentes apresentam globalmente uma taxa de risco mais elevada.

Relativamente às variáveis métricas, recorreu-se ao teste de rácio de verosimilhança (LRT - *Likelihood Ratio Test*)[13] para testar se a sua inclusão num modelo de regressão univariada (*Cox proportional hazard*) resultaria numa melhoria significativa relativamente ao modelo nulo. A Tabela 4.6 mostra os resultados do teste aplicando uma regressão univariada de Cox para cada um dos acontecimentos: o tempo até ao primeiro incidente e o tempo até à situação de *default*.

As variáveis *vf* e *prestacao* não são estatisticamente significativas na explicação de ambos os acontecimentos, pelo que não serão utilizadas como explicativas nos modelos estimados no Capítulo 5. Deste modo, evita-se também o problema de colinearidade perfeita com as variáveis *loantovalue* e *taxaesforco*, como atrás referido.

Por outro lado, as variáveis *prazo*, *rendimensal* e *taxa de esforço* surgem como possíveis variáveis explicativas apenas para um modelo relativo ao primeiro acontecimento, registo do primeiro atraso. A variável *tincidente*, que representa o tempo

[13] O valor do teste (G) segue uma distribuição do tipo χ^2_{g-r}, considerando $l(\hat{\boldsymbol{\varphi}}_g)$ o máximo da função de log-verosimilhança do modelo geral e $l(\hat{\boldsymbol{\varphi}}_r)$ o máximo da função de log-verosimilhança do modelo restrito:
$G = -2\ln\left[\frac{l(\hat{\boldsymbol{\varphi}}_r)}{l(\hat{\boldsymbol{\varphi}}_g)}\right]$, em que g e r são o número de parâmetros dos modelos respectivos.

em meses até ao momento em que o cliente registar um primeiro atraso e a variável *prazoremincidente* que representa o prazo remanescente até ao final do contrato quando o cliente regista o primeiro incidente. Assim, ambas as variáveis são potenciais variáveis explicativas da duração até à situação de *default*. Também aqui é evitado o problema de colinearidade perfeita, uma vez que a variável *prazo* não será considerada no modelo relativo ao 2º acontecimento. Salienta-se porém que o facto de *tincidente* ser um variável endógena, uma vez que esta é a variável que se pretende modelar no modelo relativo ao primeiro acontecimento, apesar de significativa numa regressão univariada, deixa de fazer sentido a sua inclusão numa modelo multivariado.

Tabela 4.6. Regressão de Cox univariada – LR $\chi^2(1)$

Variável	**Tempo até:**	
	1º Incidente	***Default***
idadebem	228.86***	56.58***
eurotax	67.67***	9.79*
loantovalue	245.52***	27.41***
vf	2.40	0.40
taxa	345.33***	105.05***
prazo	94.58***	1.38
idadecliente	29.11***	17.73***
prestacao	0.95	0.45
rendimensal	19.21***	0.45
taxaesforco	4.64*	1.41
tincidente	n.a.	30.25***
prazoremincidente	n.a.	5.31*

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Assim, as variáveis explicativas candidatas para os modelos associados a cada acontecimento são:

- Primeiro incidente: *scoring*, *tetelefcontactofixo*, *telefempregofixo*, *habitacao*, *estadocivil*, *tipo2interv*, *produto*, *idadebem*, *eurotax*, *loantovalue*, *taxa*, *prazo*, *idadecliente*, *rendimensal* e *taxaesforco*;
- Situação de *default*: *scoring*, *numinterv*, *tetelefcontactofixo*, *telefempregofixo*, *habitacao*, *estadocivil*, *tipo2interv*, *produto*, *idadebem*, *eurotax*, *loantovalue*, *taxa*, *idadecliente tincidente* e *prazoremincidente*.

5. Modelos estimados

Como referido no Capítulo 4, neste estudo foram considerados dois acontecimentos. Assim, a primeira parte deste Capítulo pretende estimar a duração até ao registo de um primeiro atraso – primeiro incidente – e ao mesmo tempo avaliar a existência de uma proporção de clientes imunes ao risco de crédito através de um modelo de imunidade.

Na segunda parte serão ensaiados modelos com o objectivo de estimar a duração entre o primeiro incidente e uma situação com atraso superior a 90 dias – *default*. Será analisada a heterogeneidade de comportamentos na transição do primeiro para o segundo acontecimento, investigando a existência de um grupo de clientes com uma transição muito rápida, apenas 3 meses, entre estes dois acontecimentos.

5.1. Duração até ao 1º incidente

Os modelos de imunidade são um tipo de modelos de análise de sobrevivência que assumem a presença de uma proporção de indivíduos para os quais o acontecimento em estudo nunca irá ocorrer. Como referido anteriormente, a taxa de risco associada à ocorrência do primeiro incidente tende a aproximar-se de 0 à medida que o contrato avança no tempo, resultando numa função de sobrevivência quase paralela ao eixo das abcissas. Neste estudo foram testados modelos admitindo que *t* segue uma distribuição exponencial, Weibull, log-normal e log-logística. A Tabela 5.1 resume os parâmetros estimados em cada uma das distribuições analisadas. Considerou-se a função de ligação logit[14].

Tabela 5.1. Parâmetros a modelar para o modelo de imunidade

Distribuição	**Parâmetros**	
	K1	**K2**
Exponencial	$-ln\,\lambda$	-
Weibull	$-ln\,\lambda$	$ln\,p$
Log-normal	μ	$ln\,\sigma$
Log- logística	$-ln\,\lambda$	$-\,ln\,\alpha$

Começou-se por estimar um conjunto de modelos agregados considerando as distribuições da Tabela 5.1, para posteriormente estimar modelos com imunes, avaliando se estes são mais adequados comparativamente com os modelos agregados. As covariáveis para *K1* foram seleccionadas por *stepwise*[15], disponível no comando

[14] $\pi = (1 + \exp(-z))^{-1}$.
[15] Com recurso ao *streg* do Stata: *stepwise, pe(0.05) pr(0.1) forward:streg lista_cov, dist(distrition)*

streg, de entre as variáveis identificadas no capítulo anterior, não sendo para já considerada a variável *scoring* uma vez que praticamente todas as outras variáveis são consideradas no seu cálculo. Para *K2*, e como o *stepwise* no *streg*, analisa apenas o parâmetro de localização – *K1*, foram adicionadas todas as variáveis, sendo posteriormente retiradas uma a uma, caso não apresentassem coeficientes significativos ($\alpha = 0.05$).

A Tabela 5.2 resume os modelos agregados estimados. Destes modelos, o modelo log-normal apresenta melhores resultados (AIC e BIC), com modelação de ambos os parâmetros da distribuição, $M_A[\mu(X)\ ln\sigma(X)]$.

Tabela 5.2. Critérios de informação – Modelos agregados

Modelo	$f(t)$	**Exponential**	**Weibull**	**Log-normal**	**Log-logística**
$M_A[K1(.)\ K2(.)]$	LL	-19896.97	-19658.35	-19054.76	-19250.03
	df	1	2	2	2
	AIC	39795.95	39320.7	38113.53	38504.06
	BIC	39803.02	39334.84	38127.66	38518.19
$M_A[K1(X)\ K2(.)]$	LL	-19101.8	-18975.16	-18452.44	-18597.29
	df	14	14	15	14
	AIC	38231.6	37978.33	36934.89	37222.59
	BIC	38330.56	38077.29	37040.92	37321.55
$M_A[K1(X)\ K2(X)]$	LL	-	-18965.61	-18372.2	-18534.4
	df	-	17	20	20
	AIC	-	37965.22	36784.4	37108.81
	BIC	-	38085.39	36925.78	37250.18

Uma vez que nenhum dos comandos disponibilizados no Stata para estimar modelos de imunidade suporta o método *stepwise*, a selecção das covariáveis para os modelos de imunidade foi efectuada de modo idêntico ao parâmetro *K2* nos modelos agregados, ou seja, foram adicionadas todas as variáveis e posteriormente retiradas uma a uma, caso não apresentassem coeficientes significativos ($\alpha = 0.05$). A proporção de mistura foi modelada apenas pela variável *scoring*, uma vez que esta variável pretende precisamente estimar a probabilidade de um cliente entrar em incumprimento durante o prazo do contrato.

Para cada uma das distribuições foram testados 5 tipos de modelos, excepto para a exponencial que foram testados apenas 4, como apresentado na Tabela 5.3.

Tabela 5.3. Critérios de informação – Modelos de Imunidade

Tipo de modelo	$f_u(t)$	Exponential	Weibull	Log-normal	Log-logística
*M*I*[K1(.) K2(.) Z(.)]*	LL	-19200.13	-19178.78	-18766.93	-18932.43
	df	2	3	3	3
	AIC	38404.26	38363.57	37539.86	37870.85
	BIC	38418.4	38384.77	37561.06	37892.06
*M*I*[K1(X) K2(.) Z(.)]*	LL	-18754.15	-18752.29	-18369.79	-18502.59
	df	12	13	15	14
	AIC	37532.29	37530.57	36769.58	37033.18
	BIC	37617.12	37622.47	36875.61	37132.14
*M*I*[K1(X) K2(X) Z(.)]*	LL	-	-18740.08	-18257.31	-18414.71
	df	-	16	19	18
	AIC	-	37512.15	36552.63	36865.42
	BIC	-	37625.26	36686.93	36992.66
*M*I*[K1(.) K2(.) Z(X)]*	LL	-18753.7	-18732.02	-18313.95	-18476.77
	df	8	9	9	9
	AIC	37523.41	37482.04	36645.9	36971.55
	BIC	37579.96	37545.66	36709.52	37035.17
*M*I*[K1(X) K2(X) Z(X)]*	LL	-18551.73	-18521.35	-18090.47	-18240.82
	df	18	22	24	23
	AIC	37139.46	37086.71	36228.94	36527.64
	BIC	37266.7	37242.22	36398.59	36690.23

O modelo escolhido, de entre os modelos agregados e de imunidade, de acordo com os critérios de informação (AIC e BIC) é o modelo com imunidade em que o tempo até ao acontecimento segue uma distribuição log-normal, modelando tanto a proporção de imunes (π) como ambos os parâmetros da distribuição (μ e σ).

A Figura 5.1 permite comparar os pseudo-resíduos[16] das 4 distribuições consideradas, para os modelos tipo *M*I*[K1(X) K2(X) Z(X)]*. A análise sugere que o pressuposto do modelo log-normal escolhido é respeitado, uma vez que o gráfico com $-ln[S_{KM}(\hat{e}_i)]$ vs $\hat{e}_i$ se aproxima bastante de uma recta com declive igual a um, ou seja, os resíduos seguem aproximadamente uma distribuição exponencial, à qual está associada uma taxa de risco constante, independente da duração *t*.

Comparativamente com o modelo com distribuição log-logística essa diferença é menos evidente, uma vez que a distribuição dos resíduos do modelo com distribuição log-logística seguem também aproximadamente uma distribuição exponencial.

[16] Secção 3.5 - Diagnóstico do modelo.

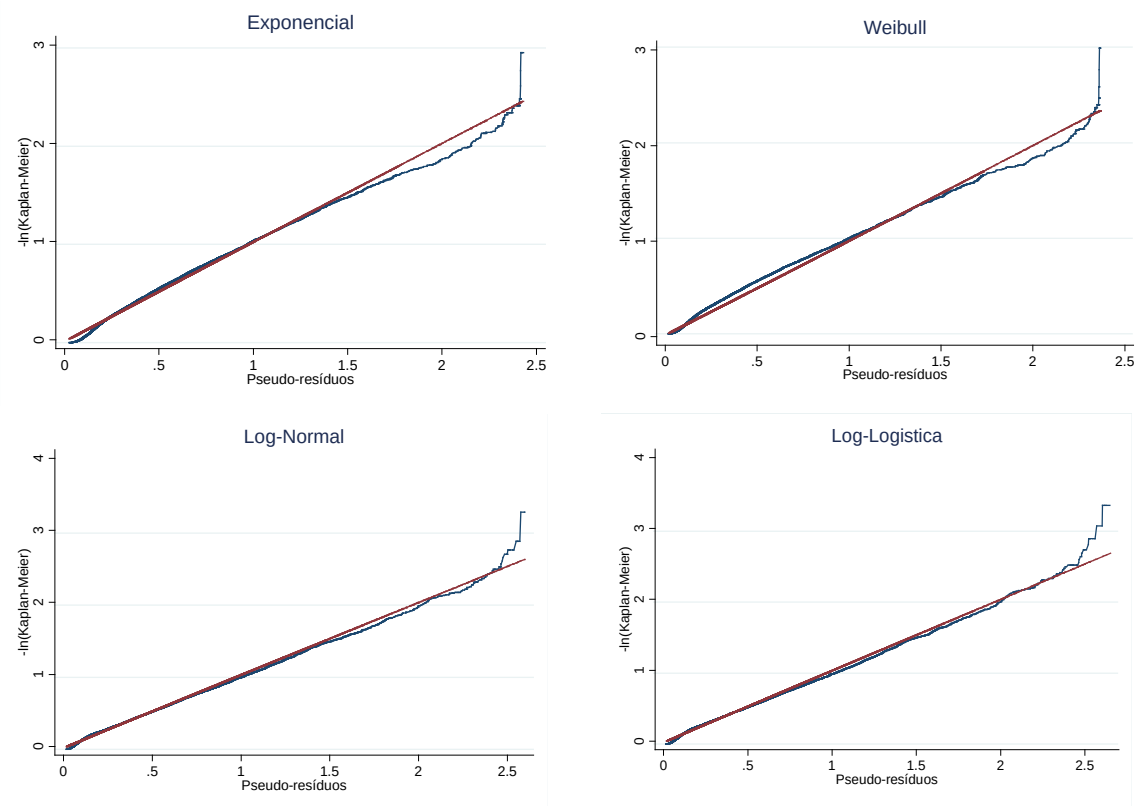

Figura 5.1. Pseudo-resíduos – Modelo com imunidade

A Tabela 5.4 apresenta o modelo seleccionado com uma distribuição log-normal.

Tabela 5.4. Modelo de imunidade – Distribuição log-normal

Covariável	**z**			**μ**			**ln (σ)**		
	coef.		**S.E.**	**coef.**		**S.E.**	**coef.**		**S.E.**
score7	2.431	***	0.1576						
score6	2.145	***	0.1505						
score5	1.922	***	0.1461						
score4	1.378	***	0.1508						
score3	1.084	***	0.1574						
score2	0.819	***	0.1531						
taxa				-0.041	***	0.005	-0.016	***	0.003
prazo				-0.014	***	0.001	-0.005	***	0.001
idadebem				-0.037	***	0.010	-0.013	*	0.006
idadecliente				0.005	***	0.002	0.002	*	0.001
telefempregofixo				0.181	***	0.036	0.081	**	0.026
prodCNS				0.213	***	0.050			
telefcontactofixo				0.067	*	0.033			
unicoInterv				-0.159	***	0.040			
loantovalue				-0.001	**	0.0004			
cotitular				0.107	*	0.041			
_cons	-2.542	***	0.125	2.996	***	0.151	0.556	***	0.093

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Analisando os coeficientes das covariáveis utilizadas para modelar o parâmetro de localização (μ) da log-normal é possível perceber se uma determinada característica do cliente/operação se traduz numa diminuição ou aumento da duração *t*, ou seja, se a taxa de risco apresenta um máximo para valores de *t* mais próximos ou mais afastados do início do contrato. Assim, incrementos na taxa, no prazo, na antiguidade do bem e no rácio (valor financiado / valor do colateral) têm um impacto na função estimada, log-normal, deslocando o intervalo de tempo até ao acontecimento onde a taxa de risco é mais elevada para a esquerda, ou seja, um incremento na taxa de risco mais próximo do início do contrato. O mesmo sucede quando o cliente não apresente avalistas. As restantes variáveis estão positivamente relacionadas com a duração *t*.

O parâmetro σ fornece-nos informação, tal como numa distribuição normal, sobre a dispersão. Na prática, um incremento em σ resulta num decréscimo menos acentuado da taxa de risco ao longo da vida do contrato. No modelo apresentado, as variáveis *idadecliente* e *telefempregofixo* estão positivamente relacionadas com σ, ao passo que a taxa, o prazo e a antiguidade do bem estão inversamente relacionadas com σ, o que implica maior concentração de valores em torno do momentos mais próximos da originação, tanto pelo efeito em μ como pelo efeito sobre σ, apresentando a função de risco uma "cauda" mais curta.

5.2. Duração até *default*

Para modelar o tempo entre o registo do primeiro incidente e a entrada de um cliente em situação de *default* foram ensaiados, primeiro, modelos de sobrevivência agregados e com proporção de imunes, testando as distribuições: exponencial, Weibull, log-normal e log-logística, à semelhança do ponto anterior. Posteriormente foram estimados modelos de sobrevivência com mistura. A Tabela 5.5 resume o primeiro conjunto de modelos estimados.

A selecção de variáveis para cada parâmetro seguiu o mesmo processo dos modelos agregados para o 1º acontecimento, ou seja, as covariáveis para *K1* foram seleccionadas por *stepwise*, com o comando *streg*, e para *K2* analisaram-se os coeficientes de cada variável, sendo retirados aqueles que não fossem significativos ($\alpha = 0.05$).

Tabela 5.5. Critérios de informação – Modelos agregados e com imunidade

Modelo	$f(t)$	**Exponential**	**Weibull**	**Log-normal**	**Log-logística**
	K1	$-ln\ \lambda$	$-ln\ \lambda$	μ	$-ln\ \lambda$
	K2	-	$ln\ p$	$ln\ \sigma$	$-ln\ \alpha$
$M_A[K1(.)\ K2(.)]$	LL	-13568.97	-13178.44	-12987.9	-13080.55
	df	1	2	2	2
	AIC	27139.93	26360.87	25979.8	26165.1
	BIC	27146.53	26374.06	25992.99	26178.29
$M_A[K1(X)\ K2(.)]$	LL	-13388.43	-12943.19	-12764.21	-12843.59
	df	12	13	13	13
	AIC	26800.86	25912.39	25554.41	25713.17
	BIC	26879.99	25998.12	25640.14	25798.9
$M_A[K1(X)\ K2(X)]$	LL	-	-12939.11	-12749.67	-12837.69
	df	-	15	19	16
	AIC	-	25908.23	25537.34	25707.37
	BIC	-	26007.15	25662.64	25812.89
$M_I[K1(X)\ K2(X)$	π	0.000000	0.041091	0.000000	0.000000
$Z(.)]$	LL	-13388.43	-12923.77	-12764.21	-12843.59
	df	13	14	14	14
	AIC	26802.86	25875.55	25556.41	25715.17
	BIC	26888.59	25967.87	25648.73	25807.5

De acordo com os critérios de informação AIC e BIC, o modelo seleccionado inclui uma distribuição log-normal, tal como no modelo estimado no ponto anterior. Este modelo, $M_A[\mu(X)\ ln\sigma(.)]$, que se apresenta por enquanto como o modelo mais adequado para estimar a duração *t* desde o registo do primeiro incidente até à entrada do cliente em situação de *default*, inclui covariáveis apenas para o cálculo do parâmetro μ da log-normal, não se verificando vantagem, segundo o AIC e o BIC, em modelar σ. De salientar ainda o facto do modelo estimado com proporção de imunes, apresentar uma estimativa de π muito próxima de zero para praticamente todas as distribuições consideradas, como tinha aliás sido sugerido no Capítulo 4.

Os gráficos na Figura 5.2 dizem respeito aos modelos tipo $M_A[K1(X)K2(.)]$ apresentados na Tabela 5.5 Tal como no modelo com imunes para estimar a duração *t* até ao primeiro acontecimento, apresentado na Secção 5.1, os modelos com função log-normal e log-logística são aqueles que apresentam resíduos mais bem comportados, mais próximos de uma recta com declive igual a um (diagonal).

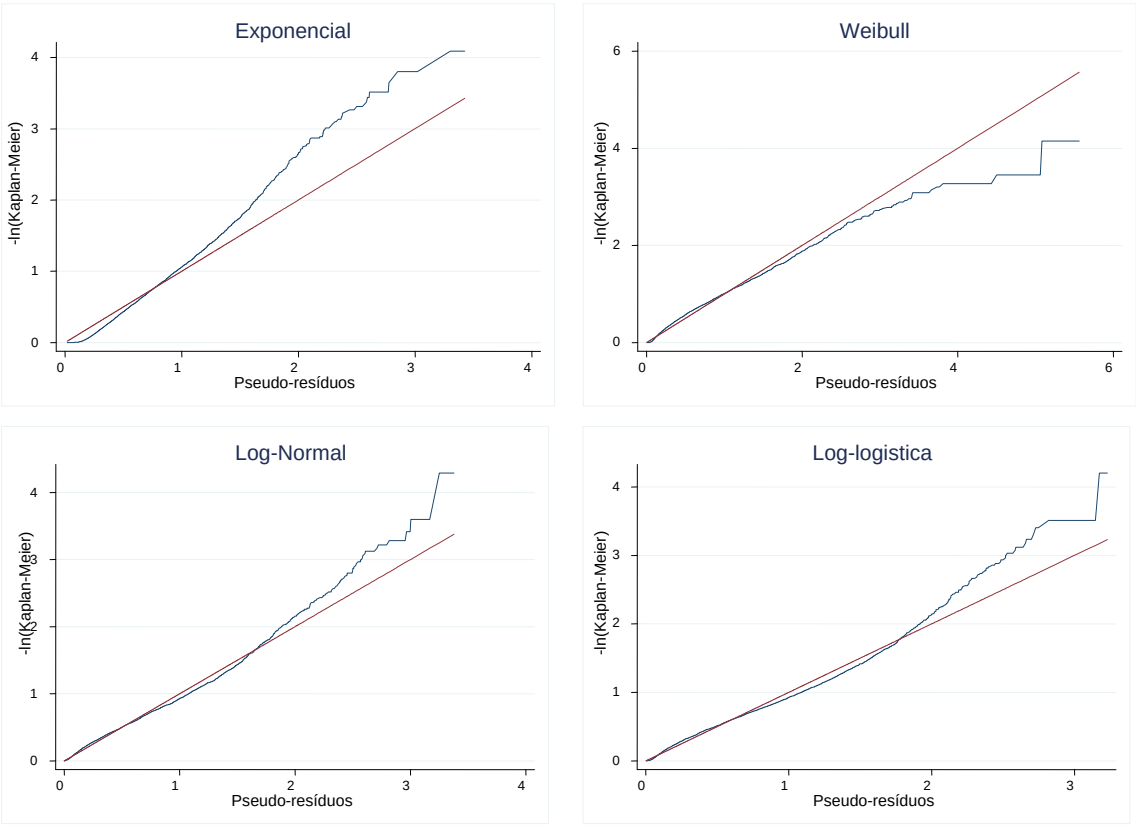

Figura 5.2. Pseudo-resíduos $-M_A[K1(X)\ K2(.)]$ (*default*)

A Tabela 5.6 apresenta as estimativas dos parâmetros do modelo com distribuição log-normal, $M_A[\mu(X)\ ln\sigma(.)]$. Nesta, a coluna da direita recorda os coeficientes de variáveis que também foram utilizadas no modelo com imunes do ponto anterior para o parâmetro μ da log-normal.

Tal como no modelo relativo à duração até ao primeiro acontecimento, o rácio (valor financiado / valor do colateral), a antiguidade do bem, o prazo remanescente do contrato, a taxa e o facto de o cliente ser o único interveniente no contrato têm um impacto negativo na duração *t* (primeiro incidente até à entrada em *default*). O mesmo se verificando para a idade do cliente, o que sugere que clientes de um escalão etário mais avançado tendem a transitar mais depressa para uma situação de *default* após terem registado um primeiro incidente. Este comportamento da variável idade do cliente pode significar uma de duas situações: (i) situações em que indivíduos pertencentes a um escalão etário mais avançado se predispõem a assumir a titularidade do contrato de crédito apesar de ser um terceiro a usufruir do crédito, que num cenário em que este deixe de pagar ao titular do contrato, este automaticamente deixa de cumprir com o

contrato de crédito; ou (ii) situação em que um acontecimento inesperado o coloca em dificuldades financeiras (*e.g.*, desemprego) difíceis de ultrapassar.

Tabela 5.6. Coeficientes para μ - Modelo agregado (*default*) vs Modelo com imunes (1º incidente)

Covariável	**Modelo agregado** *Default*		**Parâmetro μ do Modelo de imunidade** 1º incidente	
	coef	**S.E.**	**coef.**[17]	
μ				
loantovalue	-0.001 **	0.0003	-0.001 **	
eurotax	-0.002 **	0.000002	n.a.	
cotitular	0.087 **	0.034	0.107 *	
idadebem	-0.017 **	0.007	-0.037 ***	
telefcontatofixo	0.095 ***	0.026	n.a.	
idadecliente	-0.006 ***	0.001	0.005 ***	
habPropHipotec	0.135 ***	0.030	n.a.	
telefempregofixo	0.126 ***	0.027	0.181 ***	
prazoRemIncidente[18]	-0.006 ***	0.001	-0.014 ***	
taxa	-0.028 ***	0.004	-0.041 ***	
unicoInterv	-0.233 ***	0.031	-0.159 ***	
_cons	3.502 ***	0.099		
$ln\ \sigma$				
_cons	-0.237 ***	0.012		

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

As variáveis *telefcontactofixo* e *habPropHipotec*, apesar de não serem significativas para prever a duração até ao 1º acontecimento, apresentam coeficientes significativos no modelo ATF aqui apresentado, com impacto positivo sobre a duração do 1º até ao 2º acontecimento, ou seja, um cliente que possua habitação própria com hipoteca ou apresente um número de contacto de rede fixa tende a transitar mais tarde para o estado de *default*.

Uma abordagem alternativa ao modelo agregado apresentado, passaria por estimar um modelo com mistura de log-normais - $M_S[\mu(.)\ ln\sigma(.)\ Z(.)]$, ou seja, pressupor que na população de clientes que registaram o 1º incidente existe heterogeneidade na transição entre estados, representada por S grupos ($s = 1,2,\dots,S$).

[17] Os erros-padrão (S.E.) encontram-se na Tabela 5.4.
[18] O coeficiente apresentado para o modelo com imunes respeita à variável prazo que representa também o prazo remanescente do contrato no início do episódio.

A Tabela 5.7 resume os modelos com mistura estimados. Nestes modelos, tal como nos anteriores (com imunes) a proporção de mistura resulta de uma função de ligação logit. Nos modelos com modelação de μ e/ou de π, foi considerado o conjunto de variáveis incluídas no modelo agregado, apresentado na Tabela 5.6.

Tabela 5.7. Avaliação dos modelos de duração com mistura

Modelo	*f(t)*	**Log-normal** *S=2*	**Log-normal** *S=3*	**Log-normal** *S=4*
$M_S[\mu(.)ln\sigma(.)\ Z(.)]$	LL	-3631.13	-3614.36	-3611.19
	df	5	8	11
	AIC	7272.261	7244.71	7244.38
	BIC	7305.233	7297.466	7316.92
$M_S[\mu(.)ln\sigma(.)\ Z(X)]$	LL	-3565.98	-3405.33	-3351.00
	df	16	30	44
	AIC	7163.964	6870.653	6790.01
	BIC	7269.477	7068.489	7080.16
$M_S[\mu(X)ln\sigma(.)\ Z(.)]$	LL	-3466.35	-3423.3	-
	df	21	35	-
	AIC	6974.692	6916.597	-
	BIC	7113.177	7147.405	-
$M_S[\mu(X)ln\sigma(.)\ Z(X)]$	LL	-3484.18	-3421.94	-3393.99
	df	18	32	46
	AIC	7004.351	6907.87	6879.982
	BIC	7123.053	7118.895	7183.331

Com base nos critérios apresentados na Tabela 5.7, o modelo eleito é $M_3[\mu(.)\ ln\sigma(.)\ Z(X)]$, ou seja, com 3 segmentos e apenas com modelação dos pesos de cada grupo.

Apesar do maior número de parâmetros estimados, o modelo com mistura apresenta AIC e BIC mais favoráveis. Também através da análise de resíduos se conclui que o modelo de mistura apresenta um melhor ajustamento (Figura 5.3).

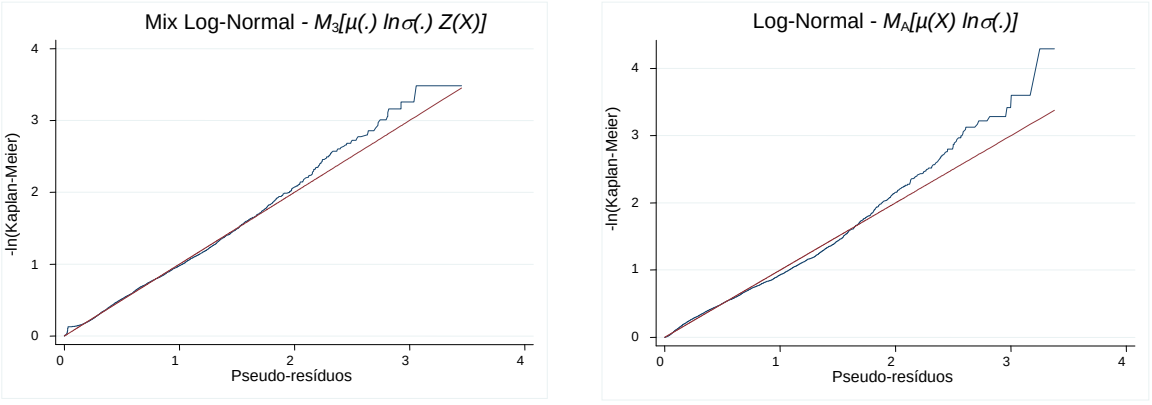

Figura 5.3. Pseudo-resíduos – Agregado vs Mistura (*default)*

As estimativas do modelo escolhido encontram-se na Tabela 5.8, que apresenta também os coeficientes estimados no modelo agregado com base numa distribuição log-normal, apresentado na Tabela 5.6.

Tabela 5.8. Coeficientes do Modelo com mistura vs Modelo agregado

Parâmetro / Covariáveis	**Modelo Mistura** $M_3[\mu(.)\ ln\sigma(.)\ Z(X)]$						***Modelo Agregado*** $M_A[\mu(X)ln\sigma(.)]$
	s=1		s=2		s=3		
	coef	**S.E.**	**coef**	**S.E.**	**coef**	**S.E.**	**coef**
z_s							
loantovalue			0.000	0.001	-0.006*	0.003	-0.000*
eurotax			0.000	0.00001	-0.000*	0.00001	-0.002*
cotitular			-0.144	0.145	0.231	0.146	0.087*
idadebem			-0.013	0.027	-0.047	0.029	-0.017*
telefcontatofixo			-0.057	0.107	0.431***	0.117	0.095***
idadecliente			0.004	0.004	-0.028***	0.006	-0.006***
habPropHipotec			-0.341*	0.137	0.345*	0.132	0.135***
telefempregofixo			-0.050	0.108	0.503*	0.145	0.126***
prazoRemIncidente			0.008*	0.003	-0.019***	0.003	-0.006***
taxa			0.045*	0.015	-0.124***	0.026	-0.028***
unicoInterv			0.209	0.126	-0.921***	0.175	-0.233***
_cons			-2.918***	0.442	3.810***	0.537	3.502***
$f_s(t)$							
μ_s	2.194***	0.037	1.099***	0.00001	3.120***	0.034	
ln_σ_s	-0.514***	0.027	-53.165	26285.5	-0.991***	0.097	
σ_s	0.598		0.000		0.371		

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Além dos aspectos positivos relacionados com a avaliação dos modelos, este modelo com mistura permite uma interpretação mais objectiva. No modelo agregado, com modelação do parâmetro de posicionamento (μ), a interpretação dos coeficientes é algo limitada uma vez que, de uma forma directa, pouco mais se pode concluir para além do sentido do impacto de uma determinada característica do cliente/operação na duração *t*. Este modelo com mistura permite apresentar uma função sobrevivência para cada grupo e posteriormente quantificar o impacto no *odds ratio*, através do modelo *logit*, de cada covariável.

A Figura 5.4 apresenta a função sobrevivência para cada grupo.

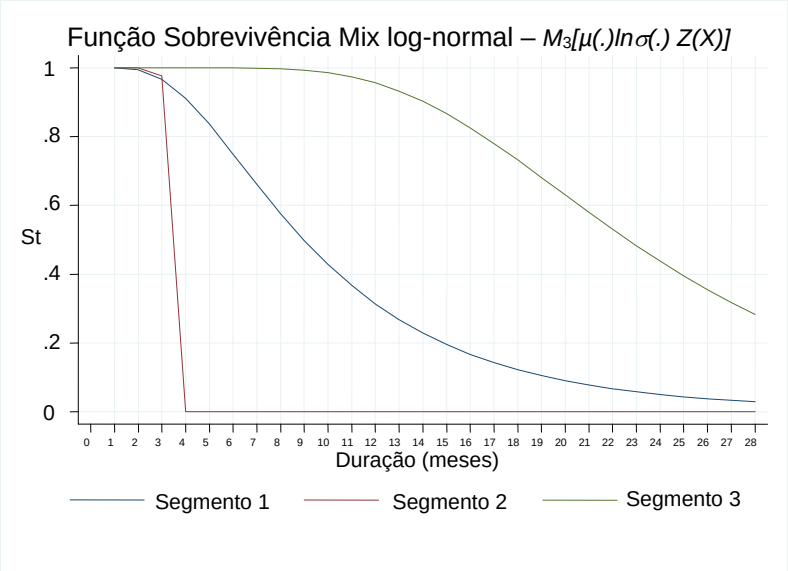

Figura 5.4. Funções sobrevivência para cada grupo

Em termos de duração até ao momento do cliente entrar numa situação de *default*, o grupo 3 representa aqueles que demoram mais tempo a transitar de uma situação registo de incidente para uma situação de *default*, o grupo 2 por seu lado representa aqueles que transitam mais rapidamente para situação de *default* e o grupo 1 representa aqueles clientes com um comportamento intermédio. Através da análise do parâmetro de posicionamento μ da log-normal em cada grupo é também possível tirar a mesma conclusão, com um valor mais elevado para o grupo 3, $\widehat{\mu}_3$=3.12 e um valor mais reduzido para o grupo 2, $\widehat{\mu}_2$=1.09. O grupo 2 caracteriza-se por uma variabilidade praticamente nula $\widehat{\sigma}_2 \approx 0$, com os clientes pertencentes a este grupo a entrarem em situação de *default* 3 meses após o registo do primeiro incidente[19], confirmando-se assim a existência de um grupo de indivíduos com uma transição muito rápida para uma situação de *default*, não pagando quaisquer das mensalidades após registarem o 1º atraso, como sugerido no Capítulo 4.

Comparativamente com o modelo agregado, os coeficientes associados à probabilidade de um cliente pertencer ao grupo 3, *i.e.*, com uma transição mais lenta para uma situação de *default*, apresentam o mesmo sentido dos coeficientes do modelo AFT agregado. No modelo agregado, as covariáveis com coeficientes negativos associados tinham um impacto negativo sobre a duração *t*, ou seja, a transição mais rápida entre estados. No modelo de mistura, essas covariáveis têm um impacto negativo sobre a

[19] $e^{1.09861} \approx 3$ meses

probabilidade do cliente pertencer ao grupo 3. Por exemplo, um cliente que se apresente numa operação como sendo o único interveniente (*unicointerv*) terá uma diminuição de 60.19%[20] no *odds ratio*, ao contrário de um cliente que possui habitação própria com hipoteca que terá associado um aumento de 41.17%[21] no *odds ratio*.

As covariáveis *cotitular* e *idadebem* que apresentaram capacidade explicativa no modelo agregado, deixam de ter um efeito significativo na probabilidade de pertença a cada grupo do modelo de mistura. A probabilidade de pertença ao grupo 2, que representa os clientes com uma transição mais rápida para uma situação de *default*, é explicada pelas covariáveis *habPropHipotec*, *prazoRemIncidente* e *taxa*. Para um cliente que possua habitação própria hipotecada, o *odds ratio* para o grupo 2 sofre uma redução de 28.9%. Uma taxa mais elevada e um período mais longo para o final do contrato no momento em que o cliente regista o primeiro incidente têm um impacto positivo na probabilidade de pertença ao grupo 2, mais concretamente, um impacto de 4.62% e 0.77% no *odds ratio*, por cada 100 p.b. ou por cada mês, respectivamente.

[20] $e^{-0.92117} - 1 \approx -0.6019$
[21] $e^{0.34480} - 1 \approx 0.4117$

6. Conclusão

Com o objectivo de criar modelos que apoiem a gestão de cobranças de clientes numa instituição financeira de crédito, foram estimados modelos de sobrevivência heterogéneos, para prever a duração até dois acontecimentos: (i) registo do primeiro atraso no pagamento das mensalidade do contrato de crédito, que resulta na transição do cliente para gestão de cobranças; (ii) registo de atrasos superiores a 90 dias – *default*, situação muito próxima de um incumprimento definitivo, que resulta normalmente na transferência do processo de cobranças para uma departamento de contencioso. Seguiu-se uma abordagem condicional tipo II, utilizando todos os clientes da amostra para estimar a duração entre o início do contrato e o registo do primeiro atraso. Foi criado um segundo modelo, considerando apenas os clientes da amostra que registaram o primeiro atraso, para estimar a duração entre o registo do primeiro incidente e a transição para uma situação de *default*.

Os modelos heterogéneos de sobrevivência apresentados neste estudo, para ambos os acontecimentos, apresentam-se como sendo mais parcimoniosos (AIC e BIC) quando comparados com os modelos agregados, reflexo da heterogeneidade de comportamentos verificada para clientes de crédito. Talvez a principal desvantagem deste tipo de modelos seja a sua complexidade, não sendo possível, por vezes, estimá-los através dos comandos mais usuais de um software estatístico. Neste estudo, por exemplo, foi necessário programar o estimador de máxima verosimilhança no Stata.

É crença geral que o nível de rendimentos e a taxa de esforço do cliente estão relacionados com o seu desempenho num contrato de crédito, porém, para os dados utilizados neste estudo conclui-se que o rendimento, a taxa de esforço não têm capacidade explicativa da duração em ambos os acontecimentos. De notar que a variável *taxaesforco*, utilizada neste estudo, representa apenas o impacto na taxa de esforço global de um cliente, na medida em que, por falta de informação, é calculada apenas com base nos encargos provenientes do crédito com a instituição financeira, ignorando outras responsabilidades que o cliente já tenha assumido. Outras variáveis como o número de dependentes a cargo do cliente e o montante de crédito, também não apresentaram capacidade explicativa. Se por um lado um maior número de dependentes pode significar maiores gastos fixos mensais, por outro pode ser também um factor de estabilidade. Verificou-se também que, para ambos os acontecimentos, solicitar um terceiro interveniente não tem impacto significativo sobre a duração.

Foram ensaiados modelos com base nas funções exponencial, Weibull, log-normal e log-logística, concluindo-se que a log-normal é função que melhor se ajusta para representar a distribuição da duração para ambos os acontecimentos.

O primeiro modelo estimado para a duração até ao momento do primeiro incidente foi um modelo com imunidade, existindo um grupo de clientes que cumprem sem registarem atrasos no pagamento das mensalidades do contrato de crédito, utilizando o *scoring* aplicacional interno para estimar a probabilidade do cliente pertencer ao grupo dos imunes. Relativamente ao grupo dos não imunes, foi utilizada uma distribuição log-normal, com modelação de ambos os parâmetros, μ e σ, tendo sido consideradas covariáveis relacionadas com a estabilidade do cliente (idade e o facto de possuir um número de telefone de rede fixa em casa e no emprego), com as garantias apresentadas (como outros intervenientes, idade do bem financiado e o rácio entre o montante do crédito e o valor do colateral) e características da operação (tipo de produto, prazo e taxa).

Na análise ao segundo acontecimento, verificou-se que os clientes que registam um primeiro incidente transitam para uma situação de *default*, não se verificando por isso uma proporção de imunes. No entanto, através de um modelo de sobrevivência de mistura com modelação da probabilidade de pertença a cada grupo, foram identificados 3 padrões de comportamento na transição para uma situação de *default*: (i) clientes com uma duração mais longa; (ii) clientes com uma duração intermédia; e (iii) clientes com duração mais curta, ou seja, com uma transição muito rápida, não pagando consecutivamente as mensalidades após registarem o 1º atraso, transitando para uma situação de *default*, 3 meses após o primeiro acontecimento. A probabilidade de um cliente pertencer ao grupo com duração mais longa aumenta nos casos em que este apresente sinais de estabilidade (um contacto telefónico fixo e habitação própria), melhores garantias (relação montante financiado e valor do colateral mais favorável/reduzido e um segundo interveniente). A este grupo de clientes estão ainda associadas condições de crédito mais competitivas, *i.e.*, taxas mais reduzidas, na medida que a probabilidade de pertencer ao grupo diminui com incrementos na taxa. Ao contrário do que seria de esperar, clientes pertencentes a um escalão etário mais elevado têm uma menor probabilidade de pertencerem ao grupo com duração mais longa. Para o grupo com duração mais curta, a probabilidade associada aumenta nos casos com

contratos com taxas mais elevadas e maior prazo remanescente no momento em que o cliente regista o primeiro incidente.

A análise de sobrevivência como técnica aplicada em modelos de *scoring* comportamental permite projectar um filme do desempenho futuro de um cliente de crédito, que permite a sua utilização em outros tipos de modelos, com o *profit scoring*. Os modelos de sobrevivência apresentados neste estudo ao incluir na sua grande maioria informação recolhida no momento da originação, permitem também a sua aplicação em modelos de *profit scoring*, estimando o envolvimento no momento em que o cliente entra em situação de *default*, calculando desta forma uma perda esperada que deve ser considerada na avaliação da rendibilidade de uma operação.

Outra técnica que também permite projectar um filme do desempenho futuro de um cliente de crédito são as Cadeias de Markov. Além desta sugestão para investigação futura aplicada aos dados da instituição financeira, a introdução de outras variáveis relativas ao histórico do cliente, como o número de atrasos, exposição máxima ou montantes vencido, pode resultar num melhor desempenho dos modelos para estimar o momento em que um cliente transita para uma situação de *default*¸ não sendo possível porém, a sua aplicação em modelos de *profit scoring* em novos clientes.

Bibliografia

Abreu, A.M. e C.S Rocha (2006), Um novo modelo de cura paramétrico, *Ciência Estatística*, Edições SPE, 151-162.

Andreeva, G., J. Ansel e J. Crook (2006), Modelling profitability using survival combination scores, *European Journal of Operational Research*, 183(3), 1537-1549.

Bellotit, T. e J. Crook (2007), Credit scoring with macroeconomic variables using survival analysis, *Journal of the Operational Research Society*, 60, 1699-1707.

Beran, J. e A.K. Djaïdja (2006), Credit risk modeling based on survival analysis with immunes, *Statistical Methodology*, 4(3), 251-276.

Blossfeld, H., Golsh, K. e G. Rohwer. (2007), *Event History Analysis with Stata*, Lawrence Earlbaum Associates, New York / London.

Buxton, A. (2007), *CUREREGR: Stata module to estimate parametric cure regression*, Statistical Software Components S446901, Boston College Department of Economics, revised 25 Sep 2007.

Cao, R., J.M. Vilar e A. Devia (2009), Modelling consumer credit risk via survival analysis, *Sort: Statistics and Operations Research Transactions*, 33(1), 3-30.

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

Cox, D.R. e E.J. Snell (1968), A general definition of residuals, *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 30(2), 248-275.

De Angelis, R., R. Capocaccia, T. Hakulinen, B. Soderman, e A. Verdecchia (1999), Mixture models for cancer survival analysis: application to population-based data with covariates, *Statistics in Medicine*,18(4), 441-454.

Dias, J.G. (2004), *Finite Mixture Models – Review, Applications, and Computer-intensive Methods*, SOM, Groningen, Holanda.

Gutierrez, R.G. (2002), Parametric frailty and shared frailty survival models, *The Stata Journal*, 2(1), 22-44.

Hensher, D. e S. Jones (2004), Predicting firm financial distress: A mixed logit model, *The Accounting Review*, 79(4), 1011-1038.

Hosmer, D.W. e S. Lemeshow (1999), *Applied Survival Analysis – Regression modeling of time to event data*, John Wiley & Sons, New York.

Jiang, R e D.N.P. Murthy (1998), Mixture of Weibull distributions - parametric characterization of failure rate function, *Applied Stochastic Models and Data Analysis*, 14, 47-65.

Kaplan E.L e P. Meier (1958), Nonparametric-estimation from incomplete observations, *Journal of the American Statistical Association*, 53(282), 457-481.

Keidel, A. (2008), China's Economic Rise – Fact and Fiction, *Carnegie Endowment for International Peace*, Policy Brief 61.

Khanna, T. (2007), China + India: the power of two, *Harvard Business Review*, http://hbr.org/product/china-india-the-power-of-two/an/R0712D-PDF-ENG.

Kleinbaum, D.G. e M. Klein (2005), *Survival Analysis: A self-learning text*, Second Edition, Springer, New York.

Kuk, A.Y.C e C.H. Chen (1992), A mixture model combining logistic-regression with proportional hazards regression, *Biometrika*, 79(3), 531-541.

Lambert, P.C. (2007), Modeling of the cure fraction in survival studies, *The Stata Journal*, 7(3), 351-375.

Lee, B. e H.J.P. Timmermans (2007), A latent class accelerated hazard model of activity episode durations, *Transportation Research Part B-Methodological*, 41(4), 426-447.

Lee, E.T. e J.W. Wang (2003), *Statistical Methods for Survival Data Anlysis*, Third edition, John Wiley & Sons, New Jersey.

Malik, M. e L.C. Thomas (2010), Modeling credit risk of portfolio of consumer loans, *Journal of the Operational Research Society*, 61(3), 411-420.

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

Ortega, E.M.M., V.G. Cancho e G.A. Paula (2009), Generalized log-gamma regression models with cure fraction, *Lifetime Data Analysis*, 15(1), 79-106.

Prentice, R.L., B.J. Williams e A.V. Peterson (1981), On the regression analysis of multivariate failure data, *Biometrika*, 68(2), 373-379.

Sarlija, N., M. Bensic e M. Zekic-Susac (2009), Comparison procedure of predicting the time to default in behavioural scoring, *Expert Systems with Applications*, 36(5), 8778-8788.

Stepanova M. e L.C Thomas (2000), Survival analysis methods for personal loan data, *Operations Research*, 50(2), 277-289.

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

Thomas, L.C. (2000), A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers, *International Journal of Forecasting*, 16(2), 149-172.

Thomas, L.C. (2009), *Consumer Credit Models: Pricing, Profit, and Portfolios*, Oxford University Press, New York.

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

Thomas, L.C., J. Ho e W.T. Scherer (2001*)*, Time will tell: behavioural scoring and the dynamics of consumer credit assessment, *Institute of Mathematics and its Applications*, 12(1), 89-103.

Till, R.J. e D.J. Hand (2003), Behavioural models of credit card usage, *Journal of Applied Statistics*, 30(10), 1201-1220.

Vakratsas, D. (1998), Consumer heterogeneity and the shape of purchase rate functions, *Applied Stochastic Models and Data* Analysis, 14, 85-44.

Vakratsas, D. e F.M. Bass (2002), A segment-level hazard approach to studying household purchase timing decisions, *Journal of Applied Econometrics*, 17(1), 49-59.

Wei, L.J., D.Y. Lin e L. Weissfeld (1989), Regression-analysis of multivariate incomplete failure time data by modeling marginal distributions, *Journal of American Statistical Association*, 84, 1065-1073.

Yildirim, Y. (2008), Estimating default probabilities of CMBS loans with clustering and heavy censoring, *Journal of Real Estate Finance and Economics*, 37(2), 93-111.