Um olhar sobre o desenvolvimento de modelos de sobrevivência para acontecimentos recorrentes

Ivo Sousa-Ferreira[1], *ivo.ferreira@staff.uma.pt*
Ana Maria Abreu[2], *abreu@staff.uma.pt*
Cristina Rocha[1], *cmrocha@fc.ul.pt*

[1] *CEAUL* e *DEIO, Faculdade de Ciências, Universidade de Lisboa*

[2] *Departamento de Matemática, Faculdade de Ciências Exatas e da Engenharia, Universidade da Madeira e CIMA, Portugal*

1. Introdução

A partir da segunda metade do século XX, tem existido um interesse crescente em estudar os tempos até à ocorrência de acontecimentos recorrentes, os quais surgem quando se observam vários episódios de um certo acontecimento durante o período de estudo. Situações que originam dados deste tipo ocorrem, por exemplo, em estudos biomédicos, quando se pretende estudar as recidivas de uma determinada doença; em estudos atuariais, quando se procura analisar os sucessivos incumprimentos de empréstimos bancários; e em estudos de fiabilidade, quando se pretende avaliar as repetidas falhas de uma máquina. Nestas situações em que o acontecimento de interesse ocorre mais do que uma vez para um mesmo indivíduo, por vezes a multiplicidade de acontecimentos observados é ignorada e é analisado apenas o tempo até à ocorrência do primeiro acontecimento. Em consequência, parte da informação que poderia vir a contribuir para uma melhor compreensão do problema em estudo é negligenciada. A maior complexidade de fenómenos desta natureza levou a que tenham sido desenvolvidas várias abordagens para analisar este tipo de dados (Cook e Lawless, 2007).

A abordagem que mais tem sido considerada, dada a sua versatilidade, consiste em estender o modelo semiparamétrico desenvolvido por Cox (1972) ao contexto dos acontecimentos recorrentes. De facto, nas últimas quatro décadas têm sido propostas várias extensões do modelo de Cox que visam ter em conta diversos aspetos relevantes em estudos deste tipo. Sendo estes modelos também semiparamétricos, não existem quaisquer premissas relativamente à distribuição dos tempos até à ocorrência dos sucessivos acontecimentos.

Outra abordagem consiste na modelação paramétrica dos tempos de recorrência. Segundo Kwong e Hutton (2003), quando a distribuição do tempo de vida for adequadamente escolhida, os modelos paramétricos são mais eficientes que os modelos semiparamétricos, no sentido em que produzem erros padrão mais pequenos e, consequentemente, estimativas dos coeficientes de regressão mais precisas. Contudo, considerar uma abordagem paramétrica também apresenta desvantagens, pois a modelação do tempo de vida pode ficar comprometida quando há uma má especificação do modelo. Deste modo, se o objetivo do estudo for apenas obter as estimativas do risco relativo e houver incerteza sobre a distribuição do tempo, é natural ser conservador e assumir pressupostos fracos sob uma abordagem semiparamétrica. Ainda assim, Solomon (1984) afirma que há alguma robustez dos modelos no que diz respeito às covariáveis, mostrando que a importância relativa das mesmas é preservada. Outra razão que contribui para que a aplicação dos modelos paramétricos seja menos frequente, está relacionada com o facto de as distribuições disponíveis na maioria dos programas de *software* estatístico não serem, em geral, suficientemente flexíveis para captar a forma como o risco evolui ao longo do tempo. Por exemplo, ao ajustar um modelo de regressão baseado na distribuição de Weibull está-se a admitir que a função de risco pode ser somente monótona crescente, monótona decrescente ou constante. Tal pressuposto pode ser demasiado restritivo, o que tem levado ao desenvolvimento de outras abordagens paramétricas mais flexíveis.

Por estes motivos, tanto quanto é do nosso conhecimento, há pouco trabalho realizado na área da modelação paramétrica no âmbito dos acontecimentos recorrentes. A maioria dos modelos paramétricos existentes baseia-se na premissa de que os tempos podem ser modelados diretamente através de uma estrutura de riscos proporcionais. Neste contexto, Duchateau *et al.* (2003), Ip *et al.* (2015) e Sousa-Ferreira *et al.* (2020a), consideraram a distribuição de Weibull, e Sousa-Ferreira *et al.* (2019) consideraram a distribuição proposta por Chen (2000). Por outro lado, Navarro *et al.* (2012) formularam modelos de sobrevivência baseados nas distribuições de Weibull, log-normal e log-logística, mas com uma estrutura de tempo de vida acelerado. Existe ainda um tipo diferente de modelos paramétricos que têm sido desenvolvidos sob o pressuposto clássico de que o número de acontecimentos observados até um dado instante segue um processo de Poisson não homogéneo (PPNH), para o qual os tempos não são, em geral, independentes. Macera *et al.* (2015) consideraram a forma exponencial-Poisson para a função de taxa marginal subjacente derivada de um PPNH. Noutra perspetiva, Sousa-Ferreira *et al.* (2020b) propuseram modelar o logaritmo da função de taxa cumulativa subjacente proveniente de um PPNH como uma função *spline* cúbica restrita do logaritmo do tempo.

Um dos maiores constrangimentos na análise de acontecimentos recorrentes é a forte possibilidade de os tempos associados ao mesmo indivíduo estarem correlacionados entre si, isto é, ocorrer correlação intra-individual. De acordo com Box-Steffensmeier e De Boef (2006), esta correlação pode ser proveniente de duas fontes: i) dependência entre acontecimentos, ou seja, a ocorrência de um acontecimento afeta o risco de ocorrência dos acontecimentos subsequentes; ii) heterogeneidade individual não observada, a qual se deve à existência de fatores de risco desconhecidos ou não mensuráveis. Assim, os modelos de sobrevivência para acontecimentos recorrentes podem ser classificados de acordo com a forma como lidam com a correlação. A abordagem mais simples baseia-se em corrigir a estimativa usual da variância (e, por vezes, estratificar os indivíduos por acontecimento), originando um modelo de variância corrigida. Outra abordagem consiste em incorporar um efeito aleatório, designado por variável fragilidade, o qual permite modelar a associação entre os tempos correspondentes ao mesmo indivíduo, dando origem a modelos com fragilidade. Por último, existe ainda a abordagem que recorre à obtenção da distribuição condicional dos tempos de recorrência, fazendo com que a relação existente entre os sucessivos acontecimentos deixe de ser um problema (Cook e Lawless, 2007).

O principal propósito deste trabalho é dar a conhecer dois tipos de abordagem para a modelação dos tempos relativos a acontecimentos recorrentes: semiparamétrica *versus* paramétrica. Para tal, começa-se por referir a abordagem semiparamétrica, onde serão consideradas duas extensões do modelo de Cox e, em seguida, a abordagem paramétrica, onde serão considerados dois modelos baseados na distribuição de Weibull e dois modelos baseados na distribuição de Chen. De forma a ilustrar a metodologia considerada, será apresentado um exemplo de aplicação a um conjunto de dados reais. Por fim, serão feitos alguns comentários a respeito destas duas abordagens.

2. Metodologia

Admita-se que existem n indivíduos em estudo e que cada um deles pode sofrer no máximo K recorrências de um certo acontecimento. Seja T_{ik} a variável aleatória (v.a.) que representa o tempo desde o início do estudo até à ocorrência do k-ésimo acontecimento ($i = 1, \ldots, n$ e $k = 1, \ldots, K$). Define-se a v.a. $Y_{ik} = T_{ik} - T_{i,k-1}$ como sendo a duração do intervalo de tempo entre dois acontecimentos consecutivos (*gap time*), com $0 \equiv T_{i0} < T_{i1} < \cdots < T_{iK}$. Por último, denote-se por $\boldsymbol{z}_{ik} = (z_{ik1}, \ldots, z_{ikp})'$ o vetor de p covariáveis referente ao k-ésimo acontecimento, associado ao i-ésimo indivíduo, e por $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ o correspondente vetor de coeficientes de regressão.

Entre as várias extensões do modelo de Cox para analisar acontecimentos recorrentes, os dois modelos que mais têm sido aplicados foram propostos por Andersen e Gill (AG) e Prentice, Williams e Peterson (PWP) (Andersen e Gill, 1982; Prentice *et al.*, 1981). No que diz respeito à formulação dos modelos AG e PWP, para o i-ésimo indivíduo em estudo, as respetivas funções de risco são dadas por

$$h(y; \boldsymbol{z}_{ik}) = h_0(y)\exp(\boldsymbol{\beta}'\boldsymbol{z}_{ik}), \qquad y \geq 0 \tag{1}$$

e

$$h(y;\boldsymbol{z}_{ik}) = h_{0k}(y)\exp(\boldsymbol{\beta}'\boldsymbol{z}_{ik}), \qquad y \geq 0, \tag{2}$$

onde $h_0(\cdot)$ representa a função de risco subjacente comum a todos os acontecimentos e $h_{0k}(\cdot)$ é a função de risco subjacente específica do acontecimento k. Estas extensões são classificadas como modelos semiparamétricos pois, tal como no modelo de Cox, a forma da função de risco subjacente não é especificada, isto é, $h_0(\cdot)$ e $h_{0k}(\cdot)$ são funções arbitrárias. Além disso, repare-se que os modelos AG (1) e PWP (2) são formulados segundo uma estrutura de riscos proporcionais, onde se assume que as covariáveis exercem um efeito multiplicativo sobre a função de risco.

Em termos gerais, o modelo AG (2) foi proposto para o caso em que os acontecimentos ocorrem de forma ordenada e apresentam igual risco de ocorrência, pelo que se considera uma função de risco subjacente comum a todos os acontecimentos. Neste modelo, considera-se que todos os indivíduos em estudo contribuem para o conjunto de risco de qualquer acontecimento, seja qual for o número de acontecimentos observados para cada indivíduo. Por essa razão, diz-se que o conjunto de risco é não restritivo.

O modelo PWP (3) também surgiu para analisar acontecimentos ordenados, mas pressupõe que o risco de ocorrência de um acontecimento é afetado pela ocorrência do acontecimento que o antecede. Consequentemente, é necessário estratificar os indivíduos segundo a ordem pela qual os acontecimentos ocorrem. Assim, se for possível observar k acontecimentos, existirão k estratos ordenados, sendo que a cada um deles estará associada a função de risco subjacente $h_{0k}(\cdot)$, $k = 1, \ldots, K$. Note-se que neste modelo tanto pode ser obtida uma estimativa global dos coeficientes de regressão $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ (alternativa que adotamos de modo a poder ser comparada com a obtida pelo modelo AG), como as estimativas específicas associadas a cada acontecimento k, $\boldsymbol{\beta_k} = (\beta_{k1}, \ldots, \beta_{kp})'$. Relativamente ao conjunto de indivíduos em risco, considera-se que estão em risco para o k-ésimo acontecimento apenas os indivíduos aos quais já foi observado o acontecimento $k-1$, o que se traduz num conjunto de risco restritivo.

Neste trabalho considera-se que, em ambos os modelos AG (1) e PWP (2), a construção do intervalo de risco é feita segundo o tempo por intervalos (*gap time*), onde o relógio reinicia a contagem do tempo voltando ao instante zero após a ocorrência de cada acontecimento. Assim, a escala de tempo diz respeito ao tempo desde o último acontecimento. Interessa salientar que o modelo AG foi originalmente proposto para analisar o tempo desde o início do estudo, onde a construção do intervalo de risco é feita através dos processos de contagem. A possibilidade de o modelo AG ser formulado para analisar o tempo entre dois acontecimentos consecutivos foi explorada por Kelly e Lim (2000). Para mais detalhes sobre este assunto consulte-se o referido trabalho, onde os autores também examinaram minuciosamente as particularidades que as diferentes extensões do modelo de Cox apresentam.

Tendo por base a formulação dos modelos AG (1) e PWP (2), quando se especifica uma determinada distribuição para o tempo os modelos resultantes são totalmente paramétricos. Vários autores têm seguido esta estratégia para desenvolver novos modelos paramétricos (Duchateau *et al.*, 2003; Ip *et al*., 2015; Sousa-Ferreira *et al*., 2019; Sousa-Ferreira *et al*., 2020a). Nesse sentido, a distribuição de Weibull tem sido a escolha preferida para especificar a forma da função de risco subjacente, provavelmente devido ao facto desta distribuição ocupar um lugar de referência na análise de dados de sobrevivência. Assim sendo, também se optou por utilizar a distribuição de Weibull com parâmetro de escala $\lambda > 0$ e parâmetro de forma $\alpha > 0$, cuja função de risco é $h(y) = \lambda\alpha y^{\alpha-1}$, $y \geq 0$. Por conseguinte, para o i-ésimo indivíduo em estudo, as correspondentes funções de risco dos modelos Weibull não estratificado (WNE) e Weibull estratificado (WE) são definidas por

$$h(y;\boldsymbol{z}_{ik}) = \lambda\alpha y^{\alpha-1}\exp(\boldsymbol{\beta}'\boldsymbol{z}_{ik}), \qquad y \geq 0 \tag{3}$$

e

$$h(y;\boldsymbol{z}_{ik}) = \lambda_k\alpha_k y^{\alpha_k-1}\exp(\boldsymbol{\beta}'\boldsymbol{z}_{ik}), \qquad y \geq 0, \tag{4}$$

em que $\lambda > 0$ e $\alpha > 0$ denotam os parâmetros de escala e de forma comuns a todos os acontecimentos, enquanto que $\lambda_k > 0$ e $\alpha_k > 0$ representam os parâmetros de escala e de forma específicos do acontecimento k, $k = 1, \ldots, K$, respetivamente. Apesar destes dois modelos já permitirem obter

estimativas da função de risco, são inadequados para situações em que a função de risco não seja monótona.

Para fenómenos complexos associados ao ciclo de vida de um indivíduo, é admissível considerar uma função de risco *bathtub-shaped*. Neste contexto, a distribuição proposta por Chen (2000) é uma boa opção, visto que a sua função de risco é dada por $h(y) = \varphi\theta y^{\theta-1}\exp(y^{\theta}),\ \varphi,\theta > 0$ e $y \geq 0$, onde θ é o único parâmetro que influencia a forma da distribuição. De facto, tem-se que esta função de risco é: *bathtub-shaped* quando $\theta < 1$ (atingindo o valor mínimo em $y_{min} = (1/\theta - 1)^{1/\theta}$); e monótona crescente quando $\theta \geq 1$. Sousa-Ferreira *et al.* (2019) verificaram que esta distribuição pode ser usada para formular um modelo de regressão pertencente à classe de modelos de riscos proporcionais. Deste modo, considerando a distribuição de Chen para especificar a forma da função de risco subjacente em (1) e (2), tem-se que as funções de risco dos modelos Chen não estratificado (CNE) e Chen estratificado (CE) são, respetivamente, definidas por

$$h(y; \boldsymbol{z}_{ik}) = \varphi\theta y^{\theta-1}\exp(y^{\theta} + \boldsymbol{\beta}'\boldsymbol{z}_{ik}), \qquad y \geq 0 \tag{5}$$

e

$$h(y; \boldsymbol{z}_{ik}) = \varphi_k\theta_k y^{\theta_k-1}\exp(y^{\theta_k} + \boldsymbol{\beta}'\boldsymbol{z}_{ik}), \qquad y \geq 0, \tag{6}$$

onde $\varphi, \theta > 0$ são parâmetros comuns a todos os acontecimentos e $\varphi_k, \theta_k > 0$ são parâmetros específicos do acontecimento k, $k = 1, \ldots, K$, respetivamente.

Em suma, os modelos paramétricos WNE (3) e CNE (5) permitem analisar situações em que o risco de ocorrência dos acontecimentos não se altera. Por outro lado, como nos modelos WE (4) e CE (6) são consideradas distribuições com parâmetros específicos para modelar os *gap times* de cada estrato/acontecimento, estes modelos abrangem situações em que os acontecimentos têm riscos de ocorrência distintos.

O método de inferência é baseado na teoria assintótica de máxima verosimilhança, assumindo que os *gap times* são independentes. Considere-se uma amostra em que a cada indivíduo i corresponde o vetor $(y_{ik}, \delta_{ik}, \boldsymbol{z}_{ik})$, onde y_{ik} é o *gap time* observado e δ_{ik} denota a variável indicatriz que caracteriza o estado do i-ésimo indivíduo relativamente ao acontecimento k. Além disso, admita-se que os indivíduos estão sujeitos a um mecanismo de censura à direita e que a censura é não informativa. No caso dos modelos semiparamétricos AG (1) e PWP (2), para estimar o vetor de coeficientes de regressão $\boldsymbol{\beta}$ é necessário adaptar a função de verosimilhança parcial do modelo de Cox ao contexto dos acontecimentos recorrentes (Cox, 1975; Kelly e Lim, 2000). Quanto aos modelos paramétricos WNE (3), WE (4), CNE (5) e CE (6), os estimadores de máxima verosimilhança dos vários parâmetros de cada modelo podem ser obtidos maximizando o logaritmo da função de verosimilhança dado por $\ell = \log\sum_{i=1}^{n}\sum_{k=1}^{K}\{\delta_{ik}\log h(y_{ik}; \boldsymbol{z}_{ik}) + \log S(y_{ik}; \boldsymbol{z}_{ik})\}$, onde $h(y_{ik}; \boldsymbol{z}_{ik})$ e $S(y_{ik}; \boldsymbol{z}_{ik}) = \exp\left[-\int_0^{y_{ik}} h(u; \boldsymbol{z}_{ik})\right]du$ são, respetivamente, as funções de risco e de sobrevivência do modelo que estiver a ser ajustado.

Embora o método de máxima verosimilhança assegure a obtenção de estimadores consistentes e assintoticamente normais, a existência de correlação intra-individual faz com que o estimador usual da matriz de covariância (baseado na matriz de informação de Fisher) não seja válido para realizar inferência. Na verdade, esta é uma abordagem *naïve* que usualmente deflaciona o erro padrão, originando resultados demasiado otimistas. Deve-se então considerar um estimador mais robusto para a variância. Lin e Wei (1989) recorreram aos estimadores *sandwich*, com o objetivo de desenvolverem um estimador robusto para o modelo de Cox, o qual foi generalizado por Wei *et al.* (1989) de modo a ser utilizado nas extensões do modelo de Cox para acontecimentos recorrentes. No caso dos modelos paramétricos, optou-se por aplicar o estimador *jackknife "one-step"* que foi desenvolvido para dados de sobrevivência agrupados (Lipsitz *et al.*, 1994), tal como acontece no âmbito dos acontecimentos recorrentes em que as observações estão agrupadas por indivíduo. Este estimador é assintoticamente equivalente ao estimador *sandwich* e tem a vantagem de ser mais fácil de programar computacionalmente.

A implementação dos modelos foi realizada no *software* estatístico **R** (R Core Team, 2020), versão 4.0.3, sendo que para os modelos semiparamétricos (AG e PWP) foi usado o *package* **survival** (Therneau, 2020) e para os modelos paramétricos (WNE, WE, CNE e CE) desenvolveu-se a sua programação utilizando o método de otimização de Broyden-Fletcher-Goldfarb-Shanno disponível no *package* **maxLik** (Toomet e Henningsen, 2020).

3. Aplicação a um conjunto de dados reais

Com o propósito de exemplificar a aplicação dos modelos formulados anteriormente, considerou-se os dados relativos à doença granulomatosa crónica (DGC) reportados em Fleming e Harrington (1991). A DGC caracteriza-se por uma imunodeficiência, o que faz aumentar a suscetibilidade a infeções recorrentes por microrganismos patogénicos como, por exemplo, determinadas bactérias e fungos. Em termos clínicos, é uma condição que se caracteriza pelo desenvolvimento de áreas de inflamação, designadas por granulomas, em diversos tecidos biológicos.

Os dados considerados representam o tempo (em anos) até uma infeção grave em 128 pacientes, dos quais 63 receberam *interferon gamma* (rIFN-g) e 65 receberam um placebo. O tempo máximo de *follow-up* foi de 1.2 anos (439 dias). Na Tabela 1, sumarizou-se a informação relevante sobre a constituição e evolução do conjunto de indivíduos em risco por acontecimento. Como o conjunto de risco diminui gradualmente, nos modelos com estratificação (PWP, WE e CE) os últimos estratos foram agrupados a partir do 4.º acontecimento, de forma a evitar a obtenção de estimativas pouco fiáveis.

Tabela 1: Resumo da informação sobre os dados da DGC

	Número do acontecimento							
	1	2	3	4	5	6	7	8
N.º de indivíduos em risco	128	44	16	8	3	2	1	1
N.º de acont. observados								
rIFN-g	14	5	1	0	0	0	0	0
Placebo	30	12	7	3	2	1	1	0
Total	44	17	8	3	2	1	1	0
Percentagem de censura (%)	65.6	61.4	50	62.5	33.3	50	0	100

A título ilustrativo, foram incorporadas duas covariáveis nos modelos: o tratamento (0: placebo ou 1: rIFN-g) e a idade (em anos) à entrada no estudo. Os resultados obtidos no ajustamento de cada modelo, relativamente às estimativas dos coeficientes de regressão, encontram-se compilados na Tabela 2.

Tabela 2: Estimativas dos coeficientes de regressão para cada modelo

Covariável/Modelo	$\hat{\beta}_j$	$\exp(\hat{\beta}_j)$	$EP(\hat{\beta}_j)$	$EP_r(\hat{\beta}_j)$	Valor-p
Tratamento					
AG	-1.1143	0.3281	0.2683	0.3208	0.0005
WNE	-1.0522	0.3492	0.2619	0.3225	0.0011
CNE	-1.0775	0.3404	0.2623	0.3285	0.0010
PWP	-0.9069	0.4038	0.2796	0.2881	0.0016
WE	-0.8245	0.4389	0.2730	0.3013	0.0063
CE	-0.8382	0.4325	0.2731	0.3063	0.0062
Idade					
AG	-0.0294	0.9711	0.0131	0.0139	0.0352
WNE	-0.0286	0.9718	0.0130	0.0148	0.0529
CNE	-0.0292	0.9712	0.0130	0.0151	0.0529
PWP	-0.0236	0.9767	0.0134	0.0110	0.0314
WE	-0.0225	0.9777	0.0134	0.0122	0.0644
CE	-0.0228	0.9774	0.0129	0.0125	0.0672

Comparando as estimativas usual e robusta do erro padrão, observa-se que $EP_r(\hat{\beta}_j)$ é, em geral, superior a $EP(\hat{\beta}_j)$, advertindo para a existência de correlação intra-individual, como antecipado. Quando $EP_r(\hat{\beta}_j)$ é muito inflacionado comparativamente a $EP(\hat{\beta}_j)$, significa que há menor variação entre as observações pertencentes a um mesmo indivíduo do que entre as observações de indivíduos diferentes.

Reciprocamente, a obtenção de um $EP_r(\hat{\beta}_j)$ muito deflacionado sugere que há menor variação entre os indivíduos do que em cada um deles. Ambas as situações constituem indícios de violação do pressuposto de independência entre as observações. A segunda situação apenas ocorreu nos modelos com estratificação, mais precisamente nas estimativas do erro padrão do efeito estimado da covariável idade. Este aspeto não é surpreendente visto que o número de indivíduos em risco diminui com as recorrências, fazendo com que o conjunto de risco de torne cada vez menos heterogéneo.

No que concerne ao efeito estimado de cada covariável, $\hat{\beta}_j$, e correspondente estimativa do risco relativo, $\exp(\hat{\beta}_j)$, observa-se que nos modelos sem estratificação (AG, WNE e CNE) os resultados obtidos são semelhantes e o mesmo sucede nos modelos com estratificação (PWP, WE e CE). Aplicando o teste de Wald robusto[1], verifica-se que em todos os modelos apenas o tratamento tem influência significativa sobre o tempo até à ocorrência do acontecimento, ao nível de significância de 0.01.

Para selecionar o modelo que melhor se ajusta a estes dados, obteve-se o valor do critério de informação de Akaike (AIC) para cada modelo (ver Tabela 3). Qualquer que seja a abordagem considerada, semiparamétrica ou paramétrica, os modelos com estratificação são aqueles que têm um valor do AIC mais baixo, indicando assim um melhor ajustamento aos dados da DGC, de entre os modelos considerados. Em particular, na abordagem paramétrica constata-se que o menor valor do AIC está associado ao modelo com estratificação baseado na distribuição de Chen (modelo CE). Então, para este conjunto de dados, a distribuição de Chen proporciona uma qualidade de ajustamento superior à distribuição de Weibull.

Tabela 3: Critério de informação de Akaike dos modelos ajustados aos dados da DGC

Modelos Semiparamétricos		Modelos Paramétricos			
AG	PWP	WNE	WE	CNE	CE
705.184	541.691	183.072	178.291	181.329	174.771

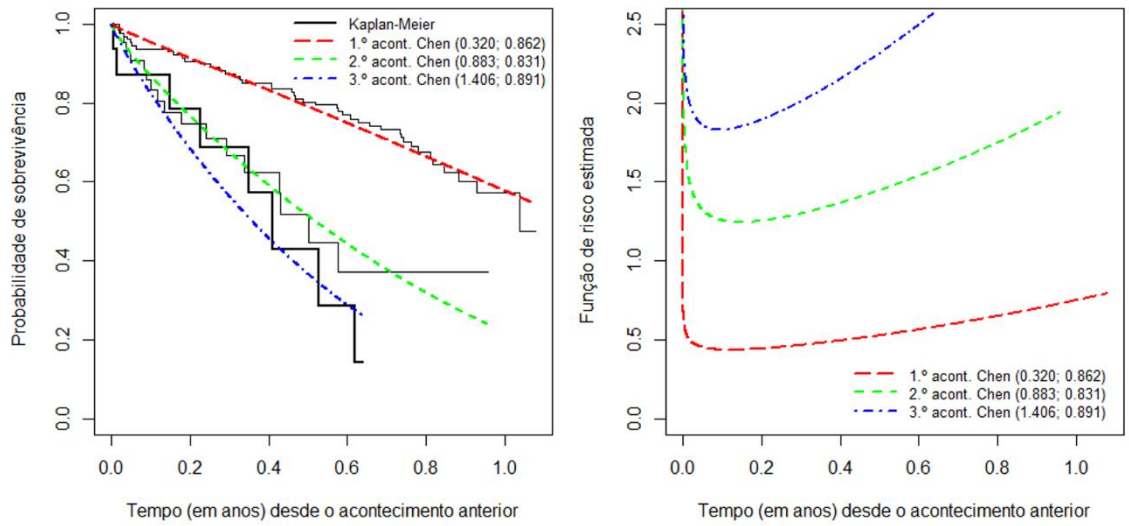


Figura 1: Estimativas da função de sobrevivência, pelo método de Kaplan-Meier e usando a distribuição de Chen, referentes aos três primeiros acontecimentos (à esquerda); e correspondentes estimativas da distribuição de Chen para a função de risco (à direita).

A adequabilidade do modelo CE pode ainda ser avaliada, de modo informal, representando graficamente as estimativas da função de sobrevivência obtidas pelo estimador de Kaplan-Meier e pelo modelo CE nulo (sem incluir as covariáveis), tendo-se obtido o gráfico apresentado à esquerda na Figura 1. Através deste gráfico, verifica-se que as estimativas baseadas na distribuição de Chen são, em geral, bastante próximas das estimativas de Kaplan-Meier, indicando que esta distribuição é uma alternativa paramétrica adequada para modelar o tempo até à recorrência de infeções causadas pela DGC. Além

[1] O teste de Wald robusto é assim designado por considerar a estimativa robusta do erro padrão.

disso, as estimativas específicas do parâmetro de forma da distribuição de Chen para o 1.º, 2.º e 3.º acontecimentos são 0.862, 0.831 e 0.891, respetivamente. Então, como estas estimativas são todas inferiores a 1, as correspondentes funções de risco estimadas são *bathtub-shaped*, tal como é possível observar no gráfico à direita. Assim sendo, a melhoria observada na qualidade do ajustamento dos modelos baseados na distribuição de Chen, pode ser atribuída à capacidade que esta distribuição tem para modelar uma forma mais complexa da função de risco.

4. Considerações finais

Em termos gerais, quando a função de risco subjacente é corretamente especificada, sabe-se que os modelos paramétricos evidenciam maior eficiência que os modelos semiparamétricos. Embora a aplicação dos modelos paramétricos baseados nas distribuições de Weibull e de Chen aos dados da DGC não tenha melhorado a precisão das estimativas dos coeficientes de regressão, estes modelos têm a vantagem de permitir estimar de forma suave a função de risco. Além disso, as estimativas robustas inflacionadas do erro padrão salientaram a importância de ter em conta a correlação intra-individual. Repare-se que os modelos considerados neste trabalho são na verdade classificados como modelos de variância corrigida, pois os parâmetros de cada modelo foram estimados ignorando a correlação, corrigindo posteriormente a estimativa usual da variância.

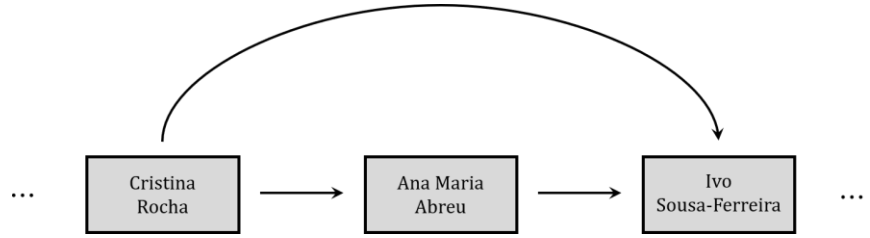
Nos modelos para acontecimentos recorrentes que são extensões do modelo de Cox, a função de risco subjacente não é especificada, o que pode constituir uma limitação.

Com efeito, em certas circunstâncias, a estimação desta função é de extrema importância, em particular na área da saúde, pois contribui para o estudo da evolução de uma doença ao longo do tempo. Na análise de acontecimentos recorrentes, a modelação paramétrica dos tempos é uma abordagem menos frequente, logo propícia a novos desenvolvimentos. Como os modelos paramétricos têm a vantagem de permitir a estimação da função de risco subjacente, atualmente o nosso interesse encontra-se direcionado para o desenvolvimento de novos modelos paramétricos que sejam capazes de descrever estes cenários complexos de forma útil, contribuindo para uma melhor compreensão do processo de recorrência.

Neste âmbito, pretende-se desenvolver modelos de sobrevivência paramétricos capazes de abranger situações em que a heterogeneidade individual não seja totalmente quantificável através das covariáveis observadas. Nesse sentido, será estudada a inclusão de um efeito aleatório (variável fragilidade), com o intuito de modelar essa heterogeneidade. No contexto dos acontecimentos recorrentes, a existência de heterogeneidade não observada é uma fonte de correlação intra-individual que pode fazer com que os indivíduos tenham uma maior (ou menor) propensão para sofrer os acontecimentos (Box-Steffensmeier e De Boef, 2006). Assim, a variável fragilidade irá representar a estrutura de dependência entre os tempos de um mesmo indivíduo, permitindo explicar alguns resultados inesperados ou até mesmo fornecer uma explicação alternativa em certas situações (ver, por exemplo, Duchateau *et al.*, 2003; Ip *et al.*, 2015). Por fim, também se pretende desenvolver modelos para lidar com situações em que haja suspeita da existência de indivíduos curados/imunes na população. Assim sendo, futuramente será investigada a inclusão de uma fração de cura nos modelos desenvolvidos, pois é plausível considerar que, também no contexto dos acontecimentos recorrentes, existam indivíduos para os quais nunca será observado qualquer episódio do acontecimento de interesse (ver, por exemplo, Tawiah *et al*., 2020).

Sobre os autores

A maioria dos docentes da área de Probabilidades e Estatística da Universidade da Madeira (UMa) deve a sua formação estatística a docentes da FCUL. Essa formação tem sido profícua e continua a dar frutos. E porque o tema do presente boletim é "*De onde viemos? Onde estamos? Para onde vamos?",* apresentamos-vos um esquema do nosso percurso em termos de orientação científica:



Agradecimentos

O presente trabalho foi desenvolvido na vigência da bolsa de doutoramento de Ivo Sousa-Ferreira, atribuída pela Universidade de Lisboa, e foi parcialmente financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia, no âmbito dos projetos UIDB/00006/2020 (CEAUL – Centro de Estatística e Aplicações) e UIDB/04674/2020 (CIMA – Centro de Investigação em Matemática e Aplicações, grupo de Estatística, Processos Estocásticos e Aplicações).

Referências

Andersen, P. K. e Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics, 10*(4), 1100-1120.

Box-Steffensmeier, J. M. e De Boef, S. (2006). Repeated events survival models: the conditional frailty model. *Statistics in Medicine, 25*(20), 3518-3533.

Chen, Z. (2000). A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. *Statistics and Probability Letters, 49*(2), 155-161.

Cook, R. J. e Lawless, J. F. (2007). *The Statistical Analysis of Recurrent Events.* New York: Springer Science & Business Media.

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

Cox, D. R. (1975). Partial likelihood. *Biometrika, 62*(2), 269-276.

Duchateau, L., Janssen, P., Kezic, I. e Fortpied, C. (2003). Evolution of recurrent asthma event rate over time in frailty models. *Journal of the Royal Statistical Society (Series C), 52*(3), 355-363.

Fleming, T. R. e Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

Ip, E. H., Efendi, A., Molenberghs, G. e Bertoni, A. G. (2015). Comparison of risks of cardiovascular events in the elderly using standard survival analysis and multiple-events and recurrent-events methods. *BMC Medical Research Methodology, 15*(15), 1-7.

Kelly, P. J. e Lim, L. L-Y. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in Medicine, 19*(1), 13-33.

Kwong, G. P. S. e Hutton, J. L. (2003). Choice of parametric models in survival analysis: applications to monotherapy for epilepsy and cerebral palsy. *Journal of the Royal Statistical Society (Series C), 52*(2), 153-168.

Lin, D. Y. e Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association, 84*(408), 1074-1078.

Lipsitz, S. R., Dear, K. B. G. e Zhao, L. (1994). Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. *Biometrics, 50*(3), 842-846.

Macera, M. A. C., Louzada, F., Cancho, V. G. e Fontes, C. J. F. (2015). The exponential-Poisson model for recurrent event data: an application to a set of data on malaria in Brazil. *Biometrical Journal, 57*(2), 201–214.

Navarro, A., Moriña, D., Reis, R., Nedel, F. B., Martín, M. e Alvarado, S. (2012). Hazard functions to describe patterns of new and recurrent sick leave episodes for different diagnoses. *Scandinavian Journal of Work, Environment & Health, 38*(5), 447-455.

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

R Core Team. (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria.

Solomon, P. J. (1984). Effect of misspecification of regression models in the analysis of survival data. *Biometrika, 71*(2), 291-298.

Sousa-Ferreira, I., Abreu, A. M. e Rocha, C. (2020a). Modelos de sobrevivência aplicados à análise de acontecimentos múltiplos. Em M. F. Salgueiro, P. Vicente, T. Calapez, C. Marques, & M. E. Silva, *Atas do XXIII Congresso da SPE* (pp. 145-159). Lisboa, Portugal: Sociedade Portuguesa de Estatística.

Sousa-Ferreira, I., Abreu, M. A. e Rocha, C. (2019). Parametric regression models for recurrent events analysis based on Chen distribution. Em M. Meira-Machado, & G. Soutinho, *Proceedings of the 34th International Workshop on Statistical Modelling* (Vol. 2, pp. 200-205). Guimarães, Portugal.

Sousa-Ferreira, I., Rocha, C. e Abreu, A. M. (2020b). A flexible marginal rate model for recurrent events with a zero-recurrence proportion. Em I. Irigoien, D.-J. Lee, J. Martínez-Minaya, & M. X. Rodríguez-Álvarez, *Proceedings of the 35th International Workshop on Statistical Modelling* (pp. 417-420). Bilbao, Espanha.

Tawiah, R., McLachlan, G. J., e Ng, S. K. (2020). Mixture cure models with time-varying and multilevel frailties for recurrent event data. *Statistical Methods in Medical Research, 29*(5), 1368-1385.

Therneau, T. M. (2020). *survival: A package for survival analysis in S*. Obtido de Package do R versão 3.2-7: *https://CRAN.R-project.org/package=survival*.

Toomet, O. e Henningsen, A. (2020). Title maximum likelihood estimation and related tools. *Package do R versão 1.4-4*. Obtido de *https://CRAN.R-project.org/package=maxLik*.

Wei, L. J., Lin, D. Y. e Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association, 84*(408), 1065-1073.