ANÁLISE DE SOBREVIVÊNCIA APLICADA A DADOS DE REINCIDÊNCIA AO CRIME

Matheus Henrique Pimenta ZANON[1]
Emílio Augusto Coelho BARROS[1]

- RESUMO: O sistema carcerário brasileiro é ponto de grandes discussões em nosso atual momento político e social. O último relatório do (DEPEN, 2014) aponta que o Brasil é o quarto pais com maior número de indivíduos privados de liberdade no mundo, estimasse ainda que o índice de reincidência ao crime para indivíduos do sexo masculino é em torno de 50% e para indivíduos do sexo feminino é de 35%. Este trabalho foi realizado com informações do banco de dados do SECAT da comarca de Primeiro de Maio-PR, contendo 356 indivíduos, abrangendo os períodos de dezembro de 2009 à dezembro de 2015. Nosso evento de interesse é a reincidência ao crime de cada indivíduo durante o período de observação. Para realizar este estudo será proposto a utilização de técnicas de análise de sobrevivência, devido ao fato de conter observações censuradas. Neste trabalho é proposto a utilização da distribuição Burr XII com e sem fração de cura, além da distribuição de Weibull com e sem fração de cura para a modelagem e análise dos dados, visto que há evidências que muitos indivíduos nunca enfrentaram o evento de interesse. Para a obtenção dos estimadores dos modelos propostos são consideradas as inferências frequentista e Bayesiana.

- PALAVRAS-CHAVE: Análise de sobrevivência; tempo de reincidência ao crime; fração de cura; distribuição Weibull; distribuição Burr XII.

1 Introdução

Em análise de sobrevivência, um modelo de mistura de longa duração, também conhecido como modelo de fração de cura, assume que a população em estudo é uma mistura de indivíduos suscetíveis a um evento de interesse e indivíduos não

[1]Universidade Tecnológica Federal do Paraná - UTFPR, Departamento Acadêmico de Matemática, CEP: 86300-000, Cornélio Procópio, PR, Brasil. E-mail: *matheus.pimenta@outlook.com; eabarros@utfpr.edu.br*

suscetíveis, em que nunca é observado o evento de interesse. Esses indivíduos que não estão em risco a respeito ao evento de interesse e são considerados imunes, não suscetíveis ou curados (MALLER e ZHOU, 1996), dependendo do contexto do estudo que geram os dados.

Diferentes metodologias, paramétricas e não paramétricas, podem ser consideradas para modelar a proporção de imunes. Nesse sentido, vários autores podem ser citados, como por exemplo, Boag (1949), Berkson e Gage (1952), Haybittle (1965), Farewell (1982, 1986), Meeker (1987), Dunsmuir *et al.* (1989), Gamel *et al.* (1990), Ghitany e Maller (1992), Taylor (1995), Copas e Heydari (1997), Ng e McLachlan (1998), Angelis *et al.* (1999), Peng e Dear (2000), Sy e Taylor (2000) e Yu *et al.* (2004). Wienke *et al.* (2006) propõem um modelo de fração de cura para dados de sobrevivência bivariados.

De acordo com Maller e Zhou (1996), em um modelo de fração de cura assume-se que uma certa fração p de indivíduos na população é curada ou nunca experimenta o evento de interesse (são imunes). Logo $1-p$ é a fração de indivíduos não curados. A função de sobrevivência, nesse caso, pode ser escrita considerando a seguinte mistura (BERKSON e GAGE, 1952),

$$S(t) = p + (1-p)S_0(t), \tag{1}$$

em que $p \in (0,1)$ é o parâmetro de mistura (proporção de imunes) e $S_0(t)$ é a função de sobrevivência basal para a população de indivíduos não curados (indivíduos suscetíveis). Considerando uma amostra aleatória de tempos de vida (t_i, δ_i), $i = 1, \ldots, n$, a contribuição do *i-ésimo* indivíduo para a função de verossimilhança é dada por,

$$L_i = \left[f\left(t_i\right)\right]^{\delta_i} \left[S\left(t_i\right)\right]^{1-\delta_i}, \tag{2}$$

em que δ_i é a variável indicadora de censura, ou seja, $\delta_i = 1$ quando o tempo de sobrevivência é exatamente observado e $\delta_i = 0$ quando o tempo é censurado (não observado) para o *i-ésimo* indivíduo.

A partir da função de sobrevivência definida em (1), é possível obter a função densidade de probabilidade, utilizando o resultado $f\left(t_i\right) = -\frac{d}{dt}S(t_i)$, dada por,

$$f\left(t_i\right) = \left(1-p\right) f_0\left(t_i\right), \tag{3}$$

em que $f_0\left(t_i\right)$ é a função densidade de probabilidade para os indivíduos suscetíveis. Substituindo a função densidade (3) e a função de sobrevivência (1) na função de verossimilhança (2) obtêm-se a seguinte função de verossimilhança para o modelo de mistura de longa duração,

$$L_i = \left[\left(1-p\right) f_0\left(t_i\right)\right]^{\delta_i} \left[p + \left(1-p\right) S_0\left(t_i\right)\right]^{1-\delta_i}. \tag{4}$$

Portanto a função log-verossimilhança considerando todas as observações é dada por,

$$l = r\log\left(1-p\right) + \sum_{i=1}^{n} \delta_i \log f_0\left(t_i\right) + \sum_{i=1}^{n} \left(1-\delta_i\right) \log\left[p + \left(1-p\right) S_0\left(t_i\right)\right], \tag{5}$$

em que, $r = \sum_{i=1}^{n} \delta_i$ é o número de observações não censuradas.

Usualmente, assume-se que a função de sobrevivência $S_0(t)$, em (1), é a função de sobrevivência das distribuições exponencial ou Weibull. Peng *et al.* (1998) considerou a função de sobrevivência da distribuição Fisher-Snedecor generalizada. A distribuição Fisher-Snedecor generalizada inclui os modelos mais usuais de sobrevivência como casos particulares, como à exemplo, as distribuições exponencial, Weibull e log-normal. Yamaguchi (1992) considera a distribuição log-gama generalizada para modelos de fração de cura no contexto de regressão com tempos de falha acelerados. A distribuição de Gompertz é considerada em Gieser *et al.* (1998), enquanto que as distribuições Weibull exponenciada e exponencial exponenciada são consideradas, respectivamente, por Cancho e Bolfarine (2001) e Kannan *et al.* (2010). Um modelo de fração de cura utilizando a distribuição Conway-Maxwell Poisson é proposto por Rodrigues *et al.* (2009) como alternativa ao modelo discutido por Yin e Ibrahim (2005). Shao e Zhou (2004) propõem um modelo de mistura de longa duração considerando a distribuição Burr XII.

A situação carcerária é tema de grandes debates no cenário político e social. Conhecer alguns aspectos deste cenário proporciona novos horizontes, estudos de diversas áreas estão sendo realizados com este intuito: ampliar o debate e entender o que esta relacionado ao tipo de perfil dos indivíduos e se há correlações com alguma covariável em comum. Este estudo, buscou utilizar-se de um recorte adequado para o estudo, onde é considerado "reincidente" o indivíduo que por algum motivo teve mais que uma passagem pela DP da comarca de Primeiro de Maio, buscando utilizar técnicas de análise de sobrevivência para acrescentar informações dos indivíduos que passam pela DP e esboçar um possível perfil do indivíduo reincidente da comarca de Primeiro de Maio.

2 Distribuição Burr XII na presença de fração de cura

Uma distribuição que vem se mostrando ser mais flexível em relação as distribuições usualmente utilizadas na aplicação em análise de sobrevivência é proposta por Burr (1942), esta distribuição consegue assumir diversos casos particulares, englobando funções como Weibull, gama, exponencial, log-normal entre outras, fazendo com que seja muito flexível no ponto de vista da modelagem, assumindo diversas formas. A distribuição Burr XII com três parâmetros tem função densidade de probabilidade, função de sobrevivência e função de risco, dadas,

respectivamente, por,

$$f(t) = \frac{\alpha}{\mu^{\alpha}} t^{\alpha-1} \left[1+\lambda\left(\frac{t}{\mu}\right)^{\alpha}\right]^{-\left(1+\frac{1}{\lambda}\right)}; \quad (6)$$

$$S(t) = \left[1+\lambda\left(\frac{t}{\mu}\right)^{\alpha}\right]^{-\frac{1}{\lambda}}; \quad (7)$$

$$h(t) = \frac{\alpha\left(\frac{1}{\mu}\right)^{\alpha} t^{\alpha-1}}{1+\lambda\left(\frac{t}{\mu}\right)^{\alpha}}. \quad (8)$$

em que $\mu > 0$ é o parâmetro de locação e $\alpha > 0$ e $\lambda > 0$ são parâmetros de forma. Observar que, para $\alpha \to 0^+$ tem-se a distribuição Weibull como caso particular. A função de risco para a distribuição Burr XII é decrescente se $\alpha \leq 1$ e unimodal com moda em $t = \frac{(\alpha-1)^{1/\alpha}}{\mu^{-1}\lambda^{1/\alpha}}$ quando $\alpha > 1$.

De (7), o modelo Burr XII na presença de fração de cura tem função densidade de probabilidade, função distribuição e função de sobrevivência definidas, respectivamente, por,

$$f(t \mid \boldsymbol{\theta}) = (1-p)\frac{\alpha}{\mu^{\alpha}} t^{\alpha-1} \left[1+\lambda\left(\frac{t}{\mu}\right)^{\alpha}\right]^{-\left(1+\frac{1}{\lambda}\right)}; \quad (9)$$

$$F(t \mid \boldsymbol{\theta}) = (1-p)\left\{1-\left[1+\lambda\left(\frac{t}{\mu}\right)^{\alpha}\right]^{-\frac{1}{\lambda}}\right\}; \quad (10)$$

$$S(t \mid \boldsymbol{\theta}) = p+(1-p)\left[1+\lambda\left(\frac{t}{\mu}\right)^{\alpha}\right]^{-\frac{1}{\lambda}}, \quad (11)$$

em que $\boldsymbol{\theta} = (\mu, \alpha, \lambda, p)$, μ é o parâmetro de escala, α e λ são os parâmetros de forma e p é a proporção de indivíduos imunes ou não suscetíveis. Supor os dados na forma (t_i, δ_i), $i = 1, \ldots, n$, em que $\delta_i = 1$ se t_i não é censurado e $\delta_i = 0$ caso contrário e que $f(t_i)$ é dado por (9). Assumindo censuras à direita para os tempos de sobrevivência, a função de verossimilhança é dada por,

$$L(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}) = L_1(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}) \times L_2(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}). \quad (12)$$

Dessa forma, as funções log-verossimilhança, $l_j(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}) = \log\left[L_j(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta})\right]$, $j = 1, 2$, são dadas, respectivamente, por,

$$l_1(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}) = r\log(1-p) + r\log(\alpha) - r\alpha\log(\mu) + (\alpha-1)\tilde{t} - \left(1+\frac{1}{\lambda}\right)\sum_{i=1}^{n}\delta_i \log(A_i); \quad (13)$$

$$l_2(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}) = \sum_{i=1}^{n}(1-\delta_i)\log\left\{p+(1-p)A_i^{-\frac{1}{\lambda}}\right\}, \quad (14)$$

em que $r = \sum_{i=1}^{n} \delta_i$, $\tilde{t} = \sum_{i=1}^{n} \delta_i \log(t_i)$, $A_i = 1 + B_i$ e $B_i = \lambda \left(\frac{t_i}{\mu}\right)^{\alpha}$.

Sejam os tempos de sobrevivência observados, (t_i, δ_i), $i = 1, \ldots, n$, definindo $l(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}) = \log L(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta})$; então, os estimadores de máxima verossimilhança para $\boldsymbol{\theta} = (\mu, \alpha, \lambda, p)$, denotados por $\widehat{\boldsymbol{\theta}} = \left(\widehat{\mu}, \widehat{\alpha}, \widehat{\lambda}, \widehat{p}\right)$, são obtidos resolvendo, por algum método numérico, o seguinte sistema de equações,

$$\begin{cases} \frac{\partial}{\partial \mu} l(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}) = -\frac{r\alpha}{\mu} + \frac{\alpha\left(1+\frac{1}{\lambda}\right)}{\mu} \sum\limits_{i=1}^{n} \frac{\delta_i B_i}{A_i^{\alpha}} + \frac{(1-p)\alpha}{\lambda\mu} \sum\limits_{i=1}^{n} \frac{A_i^{-\left(1+\frac{1}{\lambda}\right)} B_i}{p+(1-p)A_i^{-\frac{1}{\lambda}}} = 0 \\ \frac{\partial}{\partial \alpha} l(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}) = \frac{r}{\alpha} - r\log(\mu) + \tilde{t} - \left(1+\frac{1}{\lambda}\right) \sum\limits_{i=1}^{n} \frac{\delta_i B_i \log\left(\frac{t_i}{\mu}\right)}{A_i} - \frac{(1-p)}{\lambda} \sum\limits_{i=1}^{n} \frac{A_i^{-\left(1+\frac{1}{\lambda}\right)} B_i \log\left(\frac{t_i}{\mu}\right)}{p+(1-p)A_i^{-\frac{1}{\lambda}}} = 0 \\ \frac{\partial}{\partial \lambda} l(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}) = \frac{1}{\lambda^2} \sum\limits_{i=1}^{n} \delta_i \log(A_i) - \frac{\lambda+1}{\lambda^2} \sum\limits_{i=1}^{n} \frac{\delta_i B_i}{A_i} + \frac{(1-p)}{\lambda^2} \sum\limits_{i=1}^{n} \frac{A_i^{-\frac{1}{\lambda}} \left[\log(A_i) - B_i A_i^{-1}\right]}{p+(1-p)A_i^{-\frac{1}{\lambda}}} = 0 \\ \frac{\partial}{\partial p} l(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}) = -\frac{r}{1-p} + \sum\limits_{i=1}^{n} \frac{1-A_i^{-\frac{1}{\lambda}}}{p+(1-p)A_i^{-\frac{1}{\lambda}}} = 0 \end{cases} \tag{15}$$

Intervalos de confiança e testes de hipóteses de interesse podem ser obtidos usando métodos assintóticos, como a normalidade assintótica dos estimadores de máxima verossimilhança (EMV) ou usando testes de razão de verossimilhança (LAWLESS, 1982).

3 Distribuição de Weibull na presença de fração de cura

Uma das distribuições mais utilizadas na área de análise de sobrevivência é a distribuição de Weibull, pelo fato de sua função de risco pode assumir diversas formas, sendo todas monótonas, isto é, crescente, decrescente ou constante. Sua função densidade de probabilidade para uma variável aleatória T é dada por,

$$f(t) = \frac{\beta}{\mu^{\beta}} t^{\beta-1} \exp\left[-\left(\frac{t}{\mu}\right)^{\beta}\right], \; t \geq 0, \tag{16}$$

em que $\beta > 0$ e $\mu > 0$ são os parâmetros de forma e escala, respectivamente. O parâmetro μ tem a mesma unidade do tempo de falha e o parâmetro β não possui unidade. Suas funções de sobrevivência e risco, são dadas, respectivamente, por,

$$S(t) = \exp\left[-\left(\frac{t}{\mu}\right)^{\beta}\right], \tag{17}$$

$$h(t) = \frac{\beta}{\mu^{\beta}} t^{\beta-1} \tag{18}$$

Vale observar que, se $\beta = 1$ a expressão (16) se torna a função densidade de probabilidade da distribuição exponencial, fazendo com que a distribuição exponencial seja um caso particular da distribuição de Weibull.

Assumindo a parametrização $\lambda = \frac{1}{\mu^\beta}$ e o modelo de mistura (1), a função log-verossimilhança para β, λ e p (ver (5)), é dada por,

$$l\left(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}\right) = r\log\left(1-p\right) + r\log\left(\beta\right) + r\log\left(\lambda\right) + \left(\beta-1\right)v - \lambda A_1\left(\beta\right) + A_2\left(\beta,\lambda,p\right), \quad (19)$$

em que $\theta = (\beta, \lambda, p)$, $r = \sum_{i=1}^{n} \delta_i$, $v = \sum_{i=1}^{n} \delta_i \log\left(t_i\right)$, $A_1\left(\beta\right) = \sum_{i=1}^{n} \delta_i t_i^\beta$ e $A_2\left(\beta,\lambda,p\right) = \sum_{i=1}^{n} \left(1-\delta_i\right) \log\left[p + \left(1-p\right)\exp\left(-\lambda t_i^\beta\right)\right]$.

As primeiras derivadas de $l\left(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}\right)$ em relação a β, λ e p, são dadas, respectivamente, por,

$$\begin{aligned}
\frac{\partial}{\partial\beta} l\left(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}\right) &= \frac{r}{\beta} + v - \lambda\frac{\partial}{\partial\beta}A_1\left(\beta\right) + \frac{\partial}{\partial\beta}A_2\left(\beta,\lambda,p\right), \\
\frac{\partial}{\partial\lambda} l\left(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}\right) &= \frac{r}{\lambda} - A_1\left(\beta\right) + \frac{\partial}{\partial\lambda}A_2\left(\beta,\lambda,p\right), \\
\frac{\partial}{\partial p} l\left(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}\right) &= -\frac{r}{1-p} + \frac{\partial}{\partial p}A_2\left(\beta,\lambda,p\right),
\end{aligned} \quad (20)$$

em que,

$$\begin{aligned}
&\tfrac{\partial}{\partial\beta}A_1\left(\beta\right) = \sum_{i=1}^{n} \delta_i t_i^\beta \log\left(t_i\right); \\
&\tfrac{\partial}{\partial\beta}A_2\left(\beta,\lambda,p\right) = -\lambda\left(1-p\right)\sum_{i=1}^{n} \tfrac{\left(1-\delta_i\right)t_i^\beta \exp\left(-\lambda t_i^\beta\right)\log(t_i)}{p+\left(1-p\right)\exp\left(-\lambda t_i^\beta\right)}; \\
&\tfrac{\partial}{\partial\lambda}A_2\left(\beta,\lambda,p\right) = -\left(1-p\right)\sum_{i=1}^{n} \tfrac{\left(1-\delta_i\right)t_i^\beta \exp\left(-\lambda t_i^\beta\right)}{p+\left(1-p\right)\exp\left(-\lambda t_i^\beta\right)}; \\
&\tfrac{\partial}{\partial p}A_2\left(\beta,\lambda,p\right) = \sum_{i=1}^{n} \tfrac{\left(1-\delta_i\right)\left[1-\exp\left(-\lambda t_i^\beta\right)\right]}{p+\left(1-p\right)\exp\left(-\lambda t_i^\beta\right)}.
\end{aligned} \quad (21)$$

Igualando a zero as equações obtidas em (20) e resolvendo o sistema resultante por método numérico, tem-se os estimadores de máxima verossimilhança para β, λ e p, denotados por $\widehat{\beta}$, $\widehat{\lambda}$ e $\widehat{p}$. Os $100\times\left(1-\psi\right)\%$ intervalos de confiança para β, λ e p podem ser obtidos a partir da normalidade assintótica dos estimadores de máxima verossimilhança.

4 Análise bayesiana

Na resolução de problemas reais, muitas vezes a solução mais viável é utilizar a metodologia Bayesiana, dada a possível dificuldade na obtenção de estimadores de máxima verossimilhança e resultados assintóticos confiáveis. Recentemente, as técnicas Bayesianas vêm sendo utilizadas em larga escala (ACHCAR *et al.*, 2016; COELHO-BARROS *et al.*, 2016; MOORE *et al.*, 2016). Este fato se verifica, pois, o uso de métodos Bayesianos levam a resultados que não são dependentes de resultados assintóticos e ainda permitem a incorporação de informações de especialistas a partir da escolha de distribuições *a priori* para os parâmetros do modelo.

Para os modelos utilizados nesse artigo, não é possível obter a distribuição a posteriori por meio de uma solução analítica. Portanto, amostras da distribuição conjunta a posteriori de interesse são simuladas utilizando métodos MCMC, como o popular algoritmo de Gibbs (GELFAND e SMITH, 1990; CASELLA e GEORGE, 1992) e o algoritmo Metropolis-Hastings (CHIB e GREENBERG, 1995).

4.1 Análise bayesiana para a distribuição Burr XII

Para a análise Bayesiana considerando a distribuição Burr XII na presença ou não de fração de cura, assume-se distribuição a priori não informativa uniforme $U(0;1)$ definida no intervalo $(0,1)$ para a probabilidade de cura p (para o caso de fração de cura), e distribuição a priori não informativa uniforme $U(0;10000)$ definida no intervalo $(0,10000)$ para os parâmetros de locação μ e forma α e λ. Assume-se independência a priori entre p, μ, α e λ.

Na presença de um vetor de covariáveis $\mathbf{x} = (x_1, \ldots, x_k)^{'}$ afetando o parâmetro de locação μ, o seguinte modelo de regressão é proposto,

$$\mu_i = \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}). \tag{22}$$

Considerando os modelos de mistura e não mistura apresentados na Seção 2, assume também, independência a priori entre os parâmetros. Uma distribuição a priori não informativa normal $N(0;100)$ é considerada para os parâmetros β_l, $l = 0, \ldots, k$, em que $N(\mu;\sigma^2)$ denota a distribuição normal com média μ e variância σ^2.

4.2 Análise bayesiana para a distribuição Weibull

Para a análise bayesiana considerando a distribuição Weibull na presença ou não de fração de cura, assume-se distribuição a priori não informativa uniforme $U(0;1)$ definida no intervalo $(0,1)$ para a probabilidade de cura p (para o caso de fração de cura) e distribuição a priori não informativa gama $Gama(0,001;0,001)$ para os parâmetros de forma β e escala μ. Assume-se, também, independência a priori entre os parâmetros β, μ e p.

É importante observar que para os modelos introduzidos aqui, é utilizado a procedure `MCMC` (SAS, 2010a) do `software SAS`, que só requer a introdução da distribuição para os dados e as distribuições a priori para os parâmetros do modelo proposto. Assim, não são introduzidas as distribuições condicionais a posteriori necessárias para a geração de amostras da posteriori conjunta de interesse usando o amostrador de Gibbs ou o algoritmo de Metropolis-Hastings.

5 Aplicação aos dados de reincidência ao crime de detentos do SECAT da comarca de Primeiro de Maio-PR

Para ilustrar a metodologia proposta e verificar o desempenho dos modelos propostos, é considerado um conjunto de dados composto por todos os indivíduos

detidos por algum motivo no SECAT da comarca de Primeiro de Maio (356 indivíduos) entre dezembro de 2009 até dezembro de 2015 (final do período observacional). Os dados de entrada e saída, tipo de delito cometido, motivo da prisão e tipo de benefício que o fez ser solto foram obtidos através do livro de registro de presos do SECAT, que é preenchido manualmente. Duas informações adicionais foram obtidas através do Sistema de Registro Policial da Polícia Civil do Paraná são elas, o sexo do indivíduo e a idade em que o mesmo cometeu o primeiro delito. A variável dependente de interesse é o tempo em dias da soltura até o detento reincidir ao crime.

O objetivo da análise desses dados é estudar o comportamento da variável tempo até a reincidência ao crime utilizando a distribuição de Weibull e Burr XII. Existem evidências que sugerem que uma grande proporção de detentos não voltam a reincidir ao crime após cumprir a pena, portanto um modelo de mistura de longa duração será proposto.

Considerando o objetivo de ajustar o tempo de reincidência ao crime aos modelos Weibull e Burr XII na presença ou não de fração de cura, os parâmetros de interesse foram estimados através das metodologias clássica (Estimadores de Máxima Verossimilhança) e Bayesiana.

Os estimadores de máxima verossimilhança foram obtidos utilizando a procedure `NLMIXED` do software `SAS` (SAS, 2010b), pelo algoritmo de Newton-Raphson. Para obter as estimativas Bayesianas foi utilizado o método MCMC disponível no software `SAS 9.3` na procedure `MCMC` (SAS, 2010a). Uma única cadeia é utilizada para ambos os modelos considerando 200.000 simulações para cada parâmetro com um *burn-in* de tamanho 15.000 para eliminar os possíveis efeitos dos valores iniciais da simulação. Os valores simulados foram selecionados de 100 em 100, para se ter amostras aproximadamente não correlacionadas, no que resulta em uma amostra final de tamanho 2.000. Assume-se distribuições a priori não informativas para cada parâmetro dos modelos. Diagnósticos usuais de convergência observados na literatura estão avaliados na procedure `MCMC` do software `SAS`; nesse caso a indicação de convergência para todos os parâmetros foi observada.

Seja T a variável aleatória que representa o tempo até a reincidência do detento ao crime, as estimativas de máxima verossimilhança para os parâmetros dos modelos Weibull e Burr XII na presença ou não de fração de cura são apresentadas na Tabela 1, as inferências considerando a análise Bayesiana são apresentadas na Tabela 2. Nas Tabelas 1 e 2 também é possível observar o AIC (*Akaike Information Criterion*) e as estimativas de Monte Carlo para o DIC (*Deviance Information Criterion*), utilizados como critério de discriminação de modelos. Menores valores de AIC e DIC indicam melhores modelos, esses critérios são detalhados no Apêndice A.

Na Figura 1, tem-se os gráficos das curvas de sobrevivência estimadas pelo método de máxima verossimilhança e Bayesiano considerando os modelos Weibull e Burr XII na presença ou não de fração de cura e as curvas de Kaplan-Meier estimadas não parametricamente (KAPLAN e MEIER, 1958).

Observado o ajuste dos dados aos modelos de sobrevivência (ver Figura 1), conclui-se que todos os modelos se ajustam bem aos tempos de reincidência,

Tabela 1 - Estimadores de Máxima Verossimilhança

Distribuição	Parâmetro	Estimativa	Erro Padrão	AIC
Weibull (sem mistura)	μ	15997	5931, 23	1283, 3
	β	0, 5577	0, 06193	
Weibull (mistura)	μ	1222, 15	1028, 68	1283, 7
	β	0, 6701	0, 1044	
	p	0, 6606	0, 1133	
Burr XII (sem mistura)	μ	3805, 83	3641, 57	1283, 4
	α	0, 7312	0, 1611	
	λ	4, 5583	3, 7153	
Burr XII (mistura)	μ	983, 10	622, 48	1285, 7
	α	0, 7233	0, 1633	
	λ	0, 4046	2, 1019	
	p	0, 6513	0, 2154	

Tabela 2 - Médias a posteriori

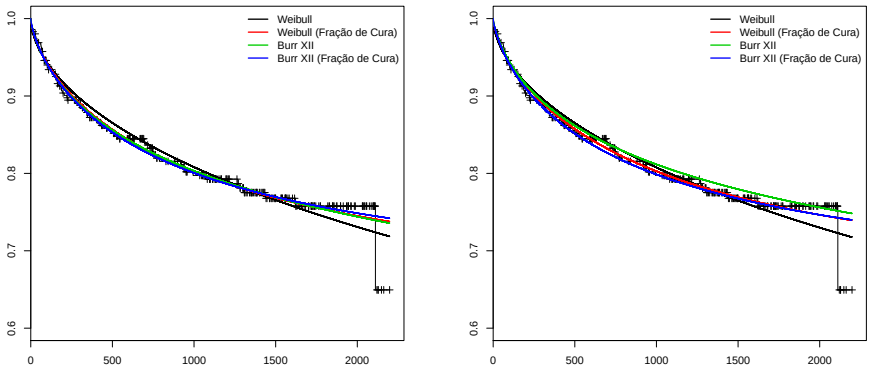
Distribuição	Parâmetro	Média	Desvio Padrão	DIC
Weibull (sem mistura)	μ	15778, 3	127, 4	1281, 299
	β	0, 5595	0, 0364	
Weibull (mistura)	μ	1135, 5	511, 0	1282, 726
	β	0, 6880	0, 0795	
	p	0, 6713	0, 0622	
Burr XII (sem mistura)	μ	3524, 6	697, 8	1281, 507
	α	0, 7564	0, 0687	
	λ	5, 3817	1, 6932	
Burr XII (mistura)	μ	961, 2	8, 9180	1281, 692
	α	0, 7392	0, 0774	
	λ	0, 5219	0, 2897	
	p	0, 6413	0, 0438	

porém, os modelos de mistura aparentemente apresentam melhor ajuste. É possível perceber, também, que o modelo Burr XII sem fração de cura se aproxima de ambos os modelos com fração de cura, mostrando sua maior flexibilidade em relação ao modelo Weibull.

Considerando o modelo Burr XII na presença das covariáveis sexo e idade do primeiro delito afetando, o parâmetro de locação μ, tem-se o seguinte modelo de regressão,

$$\mu_i = \exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\right), \tag{23}$$

em que, x_1 é uma variável binária que assume valor 1, se o indivíduo é do sexo masculino, e valor 0, se do sexo feminino, e x_2 é a variável que representa a idade em anos completos em que o indivíduo cometeu seu primeiro delito. Os parâmetros β_1 e β_2 medem, respectivamente, o efeito das covariáveis sexo e idade do primeiro



(a) EMV (b) Bayesiano

Figura 1 - Ajuste dos modelos para o tempo até a reincidência do detento ao crime, considerando as distribuições Burr XII e Weibull, com e sem fração de cura.

delito no tempo de reincidência ao crime. A escolha pelo modelo Burr XII sem a presença de fração de cura é devido ao seu melhor ajuste aos tempos de reincidência (ver Figura 1 e Tabelas 1 e 2) e por ser um modelo mais simples se comparado com o modelo Burr XII na presença de fração de cura.

Nas Tabelas 3 e 4 tem-se, respectivamente, os estimadores de Máxima Verossimilhança e os resultados inferenciais considerando a análise Bayesiana para os parâmetros do modelo de regressão proposto em (23).

Tabela 3 - Estimadores de Máxima Verossimilhança para os parâmetros do modelo de regressão

Parâmetro	Estimativa	Erro Padrão	Intervalo de Confiança 95%
β_0	$8,1246$	$1,5514$	$(5,0735; 11,1757)$
β_1	$-2,2136$	$0,9041$	$(-3,9917; -0,4355)$
β_2	$0,0903$	$0,0258$	$(0,0395; 0,1411)$
α	$0,6571$	$0,1392$	$(0,3834; 0,9308)$
λ	$1,8223$	$2,5565$	$(-3,2055; 6,8502)$

A partir dos resultados das Tabelas 3 e 4, é possível concluir que as covariáveis sexo e idade do primeiro crime afetam o tempo de reincidência ao crime, visto que o valor zero não está incluído nos intervalos de confiança e credibilidade para os parâmetros β_1 e β_2.

Por fim, pode-se concluir ao analisar os resultados obtidos pelas estimativas no modelo de regressão que indivíduos do sexo masculino tendem a reincidir com

Tabela 4 - Médias a posteriori para os parâmetros do modelo de regressão

Parâmetro	Média	Desvio Padrão	Intervalo de Credibilidade 95%
β_0	$8,1075$	$1,1604$	$(5,7250; 10,1352)$
β_1	$-2,4465$	$0,8439$	$(-4,2572; -0,9710)$
β_2	$0,0931$	$0,0265$	$(0,0438; 0,1464)$
α	$0,6866$	$0,1137$	$(0,4963; 0,9186)$
λ	$2,8561$	$1,8951$	$(0,1570; 6,3236)$

um tempo menor se comparado com os indivíduos do sexo feminino, devido ao valor negativo do parâmetro β_1. E ainda que, quanto mais velho for o indivíduo, maior o tempo de reincidência ao crime, devido ao valor positivo da estimativa do parâmetro β_2, estas estimativas são possíveis devido a incorporação das covariáveis no parâmetro de locação μ do modelo de Burr XII.

Conclusões

Usualmente na análise de dados de sobrevivência tem-se a presença de fração de cura quando uma certa proporção de indivíduos não experimentam o evento de interesse. Para a análise desse tipo de dados, pode-se utilizar diferentes formulações paramétricas, como por exemplo, o modelos de mistura. Essas formulações usualmente assumem uma distribuição paramétrica, assim como, Weibull, log-normal ou exponencial para os indivíduos susceptíveis. Como visto na Seção 5, o uso da distribuição Burr XII pode ser de grande interesse prático, pois esse modelo apresenta uma grande flexibilidade no ajuste aos dados se comparado com outras distribuições mais usuais. Além disso, a distribuição Burr XII ainda foi pouco explorada na análise de dados de sobrevivência, principalmente na presença de fração de cura.

A partir dos resultados dados nas Tabelas 1, 2, 3 e 4 observa-se, considerando a presença ou não de fração de cura, que as estimativas pontuais para os parâmetros dos modelos propostos são muito similares, mas os erros-padrão são bem menores considerando a metodologia Bayesiana. Esta comparação pode ser feita pois as distribuições a priori utilizadas são não informativas e isto implica em estimativas bem mais precisas. É importante salientar que os resultados clássicos são obtidos utilizando métodos assintóticos nem sempre bem precisos e dependentes do tamanho amostral e a proporção de dados censurados. Pela Tabela 3, é possível verificar que o intervalo de confiança para o parâmetro λ se inicia em -3.2055, o que não ocorre com a estimativa Bayesiana observada na Tabela 4, este é um problema recorrente em estimativas via métodos frequentista, visto que $\lambda > 0$. Observa-se, também, que os critérios DIC e AIC para todos os modelos propostos são muito próximos.

6 Agradecimentos

Agradecemos à Universidade Tecnológica Federal do Paraná pelo auxílio no desenvolvimento deste trabalho. Os autores agradecem aos dois revisores e editores pelas sugestões.

PIMENTA-ZANON, M. H.; COELHO-BARROS, E. A.; Survival analysis applied to data of recidivism to crime. *Rev. Bras. Biom.,* Lavras, v.36, n.3, p.527-542, 2018.

- *ABSTRACT: The Brazilian prison system is a point of great discussion in our current political and social moment. The last report (DEPEN, 2014) points out that Brazil is the fourth country with the largest number of individuals deprived of their freedom in the world, esteemed although the rate of recidivism to crime for males is around* 50% *and for females is* 35%. *This work was based on the SECAT database information of the city of Primeiro de Maio-PR, Brazil, containing* 356 *individuals, covering the period from December* 2009 *to December* 2015. *Our event of interest is recidivism crime of each individual during the observation period. To conduct this study, we are suggesting to use survival analysis techniques due to the fact that it contains censored observations. This paper advocate the use of distribution Burr XII with and without fraction of cure, besides the Weibull distribution with and without cure rate for modeling and analysing the data, since there is evidence that many individuals have never faced the event of interest. To obtain the estimators of the proposed models, we have considered the frequentist and Bayesian inference approaches.*

- *KEYWORDS: Survival analysis; time of crime to recidivism; cure fraction; Weibull distribution; Burr XII distribution.*

Referências

ACHCAR, J. A. ; COELHO-BARROS, E. A. ; DE SOUZA, R. M. Use of non-homogeneous Poisson process (NHPP) in presence of change-points to analyze drought periods: a case study in Brazil. *Environmental and Ecological Statistics*, v.23, p.405-419, 2016.

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. *Proceedings of the* 2^{nd} *International Symposium on Information Theory.* N. Petrov and F. Caski (eds), p.176-723, 1973.

AKAIKE, H. A new look at statistical model identification. *IEEE Transactions Automatic Control*, v.19, p.716-722, 1974.

ANGELIS, R. D.; CAPOCACCIA, R.; HAKULINEN, T; SODERMAN, B.; VERDECCHIA, A. Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in Medicine*, v.18, n.4, p.441-454, 1999.

BERG, A.; MEYER, R; YU, J., Deviance information criterion for comparing stochastic volatility models, *Journal of Business and Economic Statistics*, v.22, p.107-120, 2004.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, v.47, p.501-515, 1952.

BOAG, J. Maximum likelihood estimation of the proportion of patients cured by Cancer therapy. *Journal of the Royal Statistical Society*, B, v.11, p.15-53, 1949.

BURR, I. W. Cumulative frequency functions. *Annals of Mathematical Statistics*, v.13, p.215-232, 1942.

CANCHO, V. G.; BOLFARINE, H. Modeling the presence of immunes by using the exponentiated-Weibull model. *Journal of Applied Statistics*, v.28, n.6, p.659-671, 2001.

CASELLA, G.; GEORGE, E. I. Explaining the Gibbs sampler, *The American Statistician*, v.46, n.3, p.167-174, 1992.

CHIB, S.; GREENBERG, E. Undestanding the Metropolis-Hastings Algorithm, *The American Statistician*, v.49, n.4, p.327-335, 1995.

COELHO-BARROS, E. A.; ACHCAR, J. A.; MAZUCHELI, J. Bivariate Weibull Distributions Derived From Copula Functions In The Presence Of Cure Fraction And Censored Data. *Journal of Data Science*, v.14, p.295-316, 2016.

COPAS, J. B.; HEYDARI, F. Estimating the risk of reoffending by using exponential mixture models. *Journal of the Royal Statistical Society*, A, v.160, n.2, p.237-252, 1997.

DEPARTAMENTO PENITENCIÁRIO NACIONAL. Levantamento Nacional de Informações Penitenciárias. *INFOPEN*, DEPEN, 2014.

DUNSMUIR, W.; TWEEDIE, R.; FLACK. L.; MERGERSEN, K. Modeling of transitions between employement states for young australians. *Australian Journal of Statistics*, v.31, n.A, p.165-196, 1989.

FAREWELL, V. T. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, v.38, p.1041-1046, 1982.

FAREWELL, V. T. Mixture models in survival analysis: Are they worth the risk?. *The Canadian Journal of Statistics*, v.14, n.3, p.257-262, 1986.

GAMEL, J. W.; MCLEAN, I. W.; ROSENBERG, S. H. Proportion cured and mean log survival time as functions of tumor size. *Statistics in Medicine*, v.9, p.999-1006, 1990.

GAMEL, J. W.; MCLEAN, I. W.; ROSENBERG, S. H. Proportion cured and mean log survival time as functions of tumor size. *Statistics in Medicine*, v.9, p.999-1006, 1990.

GELFAND, A. E.; SMITH, A. F. M. Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, v.85, n.410, p.398-409, 1990.

GIESER, P. W.; CHANG, M. N.; RAO, P. V.; SHUSTER, J. J.; PULLEN, J. Modelling cure rates using the gompertz model with covariate information. *Statistics in Medicine*, n.17, p.831-839, 1998.

HAYBITTLE, J. L. A two parameter model for the survival curve of treated cancer patients. *Journal of the American Statistical Association*, v.53, p.16-26, 1965.

KANNAN, N.; KUNDU, D.; NAIR, P.; TRIPATHI, R. C. The generalized exponential cure rate model with covariates. *Journal of Applied Statistics*, v.37, n.9-10, p.1625-1636, 2010.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, v.53, p.457-481, 1958.

LAWLESS, J. F. *Statistical models and methods for lifetime data*. New York: John Wiley and Sons, 1982, 580p.

MALLER, R. A.; ZHOU, X. *Survival analysis with long-term survivors*. Wiley Series in Probability and Statistics: Applied Probability and Statistics, Chichester: John Wiley & Sons Ltd., 1996, 278 p.

MEEKER, W. Q. Limited failure population life tests: Application to integrated circuit reliability. *Technometrics*, v.29, n.1, p.51-65, 1987.

MOORE, B. R.; HÖHNA, S.; MAY, M. R.; RANNALA, B.;HUELSENBECK, J. P. Critically evaluating the theory and performance of Bayesian analysis of macroevolutionary mixtures. *Proceedings of the National Academy of Sciences*, v.113, n.34, p.9569-9574, 2016.

NG, S. K.; MCLACHLAN, G. J. On modifications to the long-term survival mixture model in the presence of competing risks. *Journal of Statistical Computation and Simulation*, v.61, p.77-96, 1998.

PENG, Y.; DEAR, K. B. G. A nonparametric mixture model for cure rate estimation. *Biometrics*, v.56, p.237-243, 2000.

PENG, Y.; DEAR, K. B. G.; DENHAM, J. W. A generalized F mixture model for cure rate estimation. *Statistics in Medicine*, v.17, n.8, p.813-830, 1998.

RODRIGUES, J.; CASTRO, M.; CANCHO. V. G.; BALAKRISHNAN, N. COM-Poisson cure rate survival models and an application to a cutaneous melanoma data. *J. Statist. Plann. Inference*, v.139, n.10, p.3605-3611, 2009.

SAS *SAS/STAT User's Guide, Version 9.22*. Cary, NC: SAS Institute Inc.: , 2010, 5062p.

SHAO, Q.; ZHOU, X. A new parametric model for survival data with long-term survivors. *Stat Med*, v.23, n.22, p.3525-43, 2004.

SPIEGELHALTER, D.; BEST, N.; CARLIN, B.;VAN DER LINDE, A. Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society*, Series B, Methodological, v.64, p.583-639, 2002.

SY, J. P.; TAYLOR, J. M. G. Estimation in a Cox proportional hazards cure model. *Biometrics*, v.56, p.227-236, 2000.

TAYLOR, J. M. G. Semiparametric estimation in failure time mixture models. *Biometrics*, v.51, p.899-907, 1995.

WIENKE, A.; LOCATELLI, I.; YASHIN, A. I. The modelling of a cure fraction in bivariate time-to-event data. *Austrian Journal of Statistics*, v.35, n.1, p.67-76, 2006.

YAMAGUCHI, K. Accelerated failure-time regression model with a regression model for the survining fraction: an application to the analysis of a permanent employment ain japan, *Journal of the American Statistical Association*, v.87, p.284-292, 1992.

YIN, G.; IBRAHIM, J. G. Cure rate models: a unified approach. *The Canadian Journal of Statistics*, v.33, n.4, p.559-570, 2005.

YU, B.; TIWARI, R. C.; CRONIN, K. Z. Cure fraction estimation from the mixture cure models for grouped survival times. *Statistics in Medicine*, v.23, p.1733-1747, 2004.

Recebido em 11.11.2016.

Aprovado após revisão em 21.02.2018.

APÊNDICE A: Critérios de Discriminação de Modelos

Diferentes métodos de discriminação de modelos podem ser escolhidos para verificar o modelo mais adequado considerando o paradigma Bayesiano (BERG *et al.*, 2004). Aqui é considerado o critério DIC. O *deviance* pode ser expresso por,

$$D(\boldsymbol{\theta}) = -2 \log L(\boldsymbol{\theta} \mid \mathbf{y}) + c, \tag{24}$$

em que $L(\boldsymbol{\theta} \mid \mathbf{y})$ é a função de verossimilhança para os parâmetros desconhecidos $\boldsymbol{\theta}$ dado os dados observados $\mathbf{y}$ e c é uma constante, que não é considerada quando o interesse é comparar modelos.

Spiegelhalter *et al.* (2002) define o critério DIC como segue,

$$DIC = D(\widehat{\boldsymbol{\theta}}) + 2n_D, \tag{25}$$

em que $D(\widehat{\boldsymbol{\theta}})$ é o *deviance* encontrado a partir da média a posteriori $\widehat{\boldsymbol{\theta}}$ e n_D é o número de parâmetros do modelo, definido por $n_D = \bar{D} - D(\widehat{\boldsymbol{\theta}})$, em que $\bar{D} = E[D(\boldsymbol{\theta})]$ é o *deviance* a posterior que mede a qualidade do ajuste do modelo aos dados observados. Menores valores de DIC indicam melhores modelos, notar que esses valores podem ser negativos.

Outro método comum para discriminar modelos é o chamado AIC (AKAIKE, 1973; AKAIKE, 1974) dado por,

$$AIC = -2 \log L(\widehat{\theta} \mid \mathbf{y}) + 2p, \tag{26}$$

em que $L(\widehat{\theta} \mid \mathbf{y})$ é o valor da máxima verossimilhança e p é o número de parâmetros do modelo. Menores valores de AIC indicam melhores modelos.