Submetido para TEMA

Função de Intensidade Poisson Perturbada pelo Número de Eventos Recorrentes

MARI ROMAN[1], Departamento de Estatística, DEs - UFSCar, Universidade Federal de São Carlos - Rodovia Washington Luiz, km 235, 13565-905 - São Carlos-SP, Brasil

SABRINA LUIZA CAETANO[2], Centro Universitário da Fundação Educacional de Barretos, Avenida Professor Roberto Frade Monte nº 389 14783-226 - Barretos, SP, Brasil

FRANCISCO LOUZADA[3], Instituto de Ciências Matemáticas e de Computação, ICMC - USP, Av. Trabalhador Sancarlense, 400, 13566-590 São Carlos, SP, Brasil

JOSÉ CARLOS FOGO[4], Departamento de Estatística, - Des - UFSCar, Universidade Federal de São Carlos - Rodovia Washington Luiz, km 235, 13565-905 - São Carlos-SP, Brasil.

Resumo. Neste trabalho modela-se a função de intensidade de um processo de Poisson considerando o tempo e o total de recorrências, condicionados ao momento anterior. Adotamos um componente para o processo de Poisson e o outro para o número total de eventos ocorridos nesta mesma unidade. Estudos de simulação e testes de hipóteses empíricos da significância dos parâmetros no modelo foram realizados. A significância dos testes de hipótese de *Wald* e de razão de verossimilhança foi aproximadamente 10% para mais de 50 ocorrências. Um conjunto de dados com tempos de recorrência na aquisição de cosméticos foi modelado adequadamente, tendo parâmetros significativos e valores estimados próximos dos valores observados, justificando a utilização do modelo proposto para tempos e números de recorrências em uma unidade amostral.

Palavras-chave. Eventos Recorrentes, Processo de Poisson Perturbado, Estimação de Máxima Verossimilhança, Teste de *Wald*, Teste de Razão de Verossimilhança.

[1]mari.roman19@gmail.com
[2]sabrinalcaetano@gmail.com.
[3]louzada@icmc.usp.br.
[4]djcf@ufscar.br;

1. Introdução

Uma particularidade da análise de sobrevivência e confiabilidade se refere ao fato de existirem situações nas quais o evento de interesse pode ocorrer repetidas vezes na unidade amostral, que são denominadas dados de eventos recorrentes. Áreas como biomedicina, economia, atuária, criminologia, demografia, engenharia e confiabilidade industrial apresentam dados com essa característica. [5] trazem exemplos específicos nesta área.

Na estrutura de eventos recorrentes podem ser observados o número total de eventos em determinado período de tempo, o tempo exato de recorrência dos eventos, o tempo entre eventos, o custo relacionado a cada recorrência, entre outros [16]. E o modelo de sobrevivência pode construído considerando um componente aleatório e outro determinístico. O componente aleatório é representado pela distribuição de probabilidade vinculada ao comportamento do tempo de sobrevivência e o componente determinístico é representado pelo relacionamento entre os parâmetros da distribuição e as covariáveis. Os modelos frequentemente utilizados são baseados em processos de Poisson e/ou Renovação, denominados de modelos de intensidade. O processo de Poisson é utilizado quando o interesse é modelar o tempo total de estudo e o processo de Renovação objetiva-se modelar o tempo intervalar, ver [17].

[11] utilizam processos de Poisson para desenvolver modelos que enfocam o número esperado de eventos ocorridos em determinado intervalo de tempo, [3] sugere modificações nesse modelo. [12] consideram um processo transformado para a função de intensidade acumulada. [4] consideram situações em que, além dos eventos recorrentes, existem eventos terminais que impedem a ocorrência dos demais eventos e, portanto, induzem censuras dependentes.

[19] apresenta um estudo comparativo de modelo de regressão em processos de Poisson. Uma unificação entre processo de Poisson, de Renovação e dos modelos mistos é apresentado por [15], que o denomina de processo de Renovação e Tendência.

[9] optaram por considerar o número total de eventos de recorrência ao invés dos tempos intervalares, ao tratarem o modelo proposto por [12], que traz um modelo com dois processos. Métodos estatísticos para a análise de dados desse tipo não estão completamente disseminados e o assunto ainda é alvo de muitas discussões.

1.1. Background

Seja um sistema reparável no qual observamos o tempo $t \geq 0$ em que os eventos ocorram nos tempos $t_1, t_2, \cdots, t_m$, sendo que $t_1 < t_2 < \cdots < t_m < T$. Adotamos $x_i = t_i - t_{i-1}$ ($t_0 = 0$ e $i = 1, 2, 3, \cdots, m$) para denotar o tempo entre os eventos e $N(s,t)$ para o número de eventos no intervalo $[s,t)$, sendo $N(t) = N(0,t)$ [[12]]. O modelo de probabilidade para o processo de Poisson é especificado em termos de suas condicionais ou da função de intensidade completa (FIC). Considerando $H(t) = \{N(s) : 0 < s < t\}$ como o histórico do processo até o instante t. A FIC é

dada por

$$\lambda(t; H_t) = \lim_{\Delta t \downarrow 0} \frac{P\{N(t, t+\Delta t) = 1 | H_t\}}{\Delta t},$$

em que $H_t = \{N(s) : 0 \leq s < t\}$ corresponde ao histórico do processo no instante t.

Para este caso, devido a perda de memória, a função intensidade e a função razão são as mesmas, mas isto geralmente não acontece e a função razão não especifica o processo unicamente.

Assumindo um processo de Poisson sendo que $\rho(t)$ é função intensidade (função de risco), $N(t)$ tem distribuição de Poisson com média $P(t) = \int_0^t \rho(\mu)du$ e corresponde ao número de ocorrências relacionados com os tempos de intervalos são independentes, assim $\lambda(t, H_t) = \rho(t)$.

Usando a FIC para formular modelos incorporando os tempos de tendência (Poisson) e o número de eventos, geralmente podemos estabelecer efeitos de eventos passados dentro do modelo considerando

$$\lambda(t, H_t) = \exp\{\theta' \mathbf{z}(t)\}, \tag{1.1}$$

em que $\mathbf{z}(t)$= $(z_1(t), \cdots, z_p(t))'$ é o vetor de funções que podem depender de t e H_t, θ =$(\theta_1, \cdots, \theta_p)'$ é o vetor de parâmetros, ambos desconhecidos.

Considerando (1.1) vários modelos são casos especiais, os quais incluem processo de Poisson com função intensidade dadas por $\rho(t) = \exp(\alpha + \beta t)$, em que $z(t) = (1, t)'$ e $\theta = (\alpha, \beta)'$; $\rho(t) = \alpha t^{\beta}$, em que $z(t) = (1, \log(t))'$; e $\theta = (\log(\alpha), \beta)'$.

Métodos estatísticos para tais modelos são considerado por vários autores: [1], [2], [6], [8], [10] e [14]. E mesmo não sendo o interesse primordial do trabalho, investigamos os procedimentos de inferência para tais modelos e observamos que as variáveis explicativas podem ser incluídas em termos de $\mathbf{z}_i(t)$.

Visando acomodar a influência da quantidade de eventos ocorridos anteriormente condicionado ao momento da ocorrência do evento anterior, propomos um modelo de perturbação no processo de Poisson condicionado.

Na Seção 2 apresentamos a formulação do modelo utilizando a função intensidade, a função de verossimilhança. As estimativas de máxima verossimilhança e a validação do modelo, considerando os testes *Wald* e Razão de verossimilhança (RV) bem como as estatísticas empíricas associadas a eles, estão apresentadas na Seção 3. Na Seção 4 discutimos o ajuste do modelo proposto para dados artificiais e para dados referentes ao tempo entre compras de produtos de cosméticos. Na Seção 5 alguns comentários finais concluem o artigo.

2. Modelo condicionado para Processo de Poisson perturbado

Seja ψ um sistema no qual há ocorrência do evento W. Iniciamos o processo no tempo $t_0 = 0$ e observamos os tempos $t_i \geq 0$, $i = 1, 2, 3, \cdots, m$, tal que $t_1 < t_2 < \cdots < t_m$, nos quais ocorrem o evento de interesse, sendo os tempos t_i *i.i.d.* Seja

$N(t_i)$ o número de eventos ocorridos no intervalo $[0,\ t_{i-1}]$, assim, $N(t_i) = i-1$, valor relacionado apenas ao tempo presente t_i, sendo constante no tempo anterior, t_{i-1}.

Consideramos a FIC como apresentada em (1.1), em que $z(t_i) = (1, t_i, i-1)'$ e $\theta = (\alpha, \beta, \gamma)'$, com $\alpha, \gamma \in \Re$ e $\beta > 0$. A perturbação pelo número de eventos é obtida fazendo $g_1(t) = t$ e $g_2(N(t)) = N(t)$ em

$$z(t) = (1, g_1(t), g_2(N(t)))',$$

e a incorporação do efeito dos eventos passados é feito de forma condicional na função intensidade.

A FIC condicionada ao tempo anterior, que expressa o modelo de probabilidade de interesse, é

$$\lambda(t_i|t_i - 1) = e^{\alpha+\beta t_i+\gamma(i-1)}. \tag{2.1}$$

A expressão (2.1) corresponde ao modelo de intensidade condicional híbrido. O parâmetro α funciona como risco de base e independentemente do tempo de ocorrência do evento ele já está implícito. O parâmetro β é caracterizado como coeficiente do tempo e caracteriza o processo de Poisson. O parâmetro γ indica a influência do número de recorrência no processo e tem significância somente se a função de intensidade depender da quantidade de eventos.

Proposição 2..1. *Seja, t_i o tempo de recorrência do evento W, em que $t_{i-1} < t_i$, $i = 1, 2, \cdots, m$, com $t_0 = 0$ com função de intensidade condicional dada por* (2.1)*. A função densidade de probabilidade condicional do tempo t_i, dado t_{i-1}, é*

$$f(t_i|t_{i-1}) = e^{\alpha+\beta t_i+\gamma(i-1)} \exp\left\{\frac{-e^{\alpha+(i-1)\gamma}}{\beta}\left(e^{\beta t_i} - e^{\beta t_{i-1}}\right)\right\}. \tag{2.2}$$

Demonstração. Escrevendo $f(t_i|t_{i-1}) = S(t_i|t_{i-1}) \times \lambda(t_i|t_{i-1})$, com $\lambda(t_i|t_{i-1})$ dada em (2.1), determinamos $S(t_i|t_{i-1})$, como segue

$$\begin{aligned}
S(t_i|t_{i-1}) &= P(T \geq t_i|T \geq t_{i-1}) \\
&= \frac{P(T \geq t_i, T \geq t_{i-1})}{P(T \geq t_{i-1})} = \frac{P(T \geq t_i)}{P(T \geq t_{i-1})} = \frac{S(t_i)}{S(t_{i-1})} \\
&= \frac{\exp\left(-\int_0^{t_i} \lambda(u)du\right)}{\exp\left(-\int_0^{t_{i-1}} \lambda(u)du\right)} = \exp\left\{-\left(\int_0^{t_i} \lambda(u)du + \int_{t_{i-1}}^{0} \lambda(u)du\right)\right\},
\end{aligned}$$

portanto

$$S(t_i|t_{i-1}) = \exp\left\{-\int_{t_{i-1}}^{t_i} \lambda(u)du\right\}.$$

Para $\lambda(u)$ dada em (2.1), temos

$$S(t_i|t_{i-1}) = \exp\left\{\frac{-e^{\alpha+(i-1)\gamma}}{\beta}\left(e^{\beta t_i} - e^{\beta t_{i-1}}\right)\right\}, \tag{2.3}$$

que é uma função de sobrevivência, pois para $t_i > 0$ temos que

$$\frac{dS(t_i|t_{i-1}}{dt_i} = -exp\left\{-\frac{e^{\alpha+\gamma(i-1)}}{\beta}\left(e^{\beta t_i} - e^{\beta t_{i-1}}\right) + \alpha + \beta t_i + \gamma(i-1)\right\},$$

a qual é estritamente negativa, portanto $S(t_i|t_{i-1})$ é estritamente decrescente para $t_i > 0$. E como $\lim_{t_i \to 0} S(t_i|t_{i-1}) = 1$ e $\lim_{t_i \to \infty} S(t_i|t_{i-1}) = 0$, então $F(t_i|t_{i-1}) = 1 - S(t_i|t_{i-1})$ é uma função densidade de probabilidade acumulada.

Com o produto de (2.3) e (2.1) obtêm-se (2.2) comprovadamente uma f.d.p. □

3. Inferência

Assumimos que os tempos t_i, $i = 1, 2, \cdots, m$ e $m > 0$, são i.i.d e realizações da variável aleatória associada ao evento W. Em processos de Poisson condicionado a função de verossimilhança é dada por $L(\theta) = \prod_{i=1}^{m} \lambda(t_i|t_{i-1})S(t_i|t_{i-1})$, com $\lambda((t_i|t_{i-1})$ é dada por (2.1) e $S(t_i|t_{i-1})$ dada em (2.3). Sendo $\theta = (\alpha, \beta, \gamma)$ temos

$$L(\theta) = \prod_{i=1}^{m} e^{\alpha+\beta t_i+\gamma(i-1)} \exp\left\{\frac{-e^{\alpha+(i-1)\gamma}}{\beta}\left(e^{\beta t_i} - e^{\beta t_{i-1}}\right)\right\}. \tag{3.1}$$

O procedimento de utilização do método da máxima verossimilhança para a obtenção dos estimadores requer a função log da verossimilhança, a função escore e a matriz de informação observada de Fisher, tudo com respeito a verossimilhança (3.1).

O logaritmo da função verossimilhança (3.1), representada por $\ell(\theta) = \log(L(\theta))$, é dada por

$$\ell(\theta) = m\alpha + \beta\sum_{i=1}^{m} t_i + \gamma\sum_{i=1}^{m}(i-1) - \frac{e^{\alpha}}{\beta}\sum_{i=1}^{m} e^{(i-1)\gamma+\beta t_i} + \frac{e^{\alpha}}{\beta}\sum_{i=1}^{m} e^{(i-1)\gamma+\beta t_{i-1}}, \tag{3.2}$$

e sua correspondente função escore, obtida da derivada de (3.2) em relação ao vetor paramétrico θ, é dada por

$$U_r(\theta) = \frac{\partial l(\theta)}{\partial \theta_r} = \sum_{i=1}^{n} z_r(t_i) - \int_0^T z_r(t)\exp\{\theta' z(t)\}\, dt, \text{ para } r = 1, \cdots, p \tag{3.3}$$

para $r = 1, \cdots, p$, em que $p = 3$ corresponde a quantidade de parâmetros do vetor θ.

Os elementos da matriz escore, dos quais advêm os estimadores de máxima verossimilhança (EMVs), são

$$\begin{aligned}
\frac{\partial \log(L)}{\partial \alpha} &= m - \frac{e^{\alpha}}{\beta}\sum_{i=1}^{m} e^{(i-1)\gamma+\beta t_i} + \frac{e^{\alpha}}{\beta}\sum_{i=1}^{m} e^{(i-1)\gamma+\beta t_i},\\
\frac{\partial \log(L)}{\partial \beta} &= \sum_{i=1}^{m} t_i - \frac{e^{\alpha}}{\beta}\sum_{i=1}^{m} t_i e^{(i-1)\gamma+\beta t_i} - \frac{e^{\alpha}}{\beta^2}\sum_{i=1}^{m} e^{(i-1)\gamma+\beta t_i}\\
&\quad + \frac{e^{\alpha}}{\beta}\sum_{i=1}^{m} t_{i-1} e^{(i-1)\gamma+\beta t_{i-1}} - \frac{e^{\alpha}}{\beta^2}\sum_{i=1}^{m} e^{(i-1)\gamma+\beta t_{i-1}},\\
\frac{\partial \log(L)}{\partial \gamma} &= \sum_{i=1}^{m} (i-1) - \frac{e^{\alpha}}{\beta}\sum_{i=1}^{m} (i-1) e^{(i-1)\gamma+\beta t_i}\\
&\quad + \frac{e^{\alpha}}{\beta}\sum_{i=1}^{m} (i-1) e^{(i-1)\gamma+\beta t_{i-1}}.
\end{aligned}$$

Em princípio se o número total de eventos for grande, então para encontrar as estimativas para θ, a partir de $\widehat{\theta}$, consideramos que o último tem aproximadamente distribuição Normal com média θ e matriz de variância-covariância $I_o\left(\widehat{\theta}\right)^{-1}$. Escores ou estatísticas de verossimilhança podem ser consideradas e utilizadas de forma usual [7]. Assim, os EMVs não são expressos analiticamente e suas estimativas foram obtidas no software R, pela rotina *optim*. A obtenção numérica dos EMVs não apresentou problemas, pois a integral envolvida é de fácil resolução, diferentemente do modelo proposto por [12]. A matriz de informação observada $I_{p\times p}(\theta)$, que corresponde a derivada de (3.3) em relação ao vetor θ, com $r, s = 1, 2, \cdots, p$, é dada por

$$I_{rs}(\theta) = -\frac{\partial^2 l}{\partial\theta_r \partial\theta_s} = \int_0^T z_r(t) z_s(t) \exp\{\theta' z(t)\}\, dt. \tag{3.4}$$

Em casos especiais pode-se avaliar a integral em (3.4) analiticamente, mas em geral a integração numérica é necessária (Louzada *et al.,* 2002).

3.1. Consistência dos estimadores de máxima verossimilhança

Fixamos os vetores $\theta_1 = (2,\ 0.8,\ -0.1)$, $\theta_2 = (-2,\ 0.8,\ -1.5)$ e $\theta_3 = (1,\ 0.2,\ -1.5)$ de $\theta = (\alpha,\ \beta,\ \gamma)$ e as quantidades de recorrência $m = 10,\ 25,\ 50,\ 100,\ 150,\ 300$ e geramos amostras considerando a função inversa da função de probabilidade acumulada condicional. Para avaliar a propriedade de consistência ajustando o modelo aos dados gerados e as estimativas $\widehat{\alpha}$, $\widehat{\beta}$ e $\widehat{\gamma}$, dos respectivos parâmetros, estão apresentadas na Tabela 1. Observamos que os valores estimados se aproximam dos verdadeiros valores dos parâmetros à medida que o tamanho da amostra aumenta. Portanto, segundo [13], os estimadores são consistentes.

Tabela 1: Estimativas de máxima verossimilhança para os parâmetros.

m	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$
10	(0.993, 29.002, -3.498)	(-5.499, 2.651, -4.972)	(2.586, 0.434, -2.972)
25	(1.617, 2.976, -0.284)	(-2.264, 0.946, -1.740)	(1.497, 0.243, -1.784)
50	(2.490, 0.976, -0.141)	(-1.505, 0.774, -1.468)	(1.324, 0.190, -1.445)
100	(2.135, 0.668, -0.086)	(-1.811, 0.776, -1.458)	(1.043, 0.191, -1.438)
150	(1.963, 0.596, -0.073)	(-2.111, 0.771, -1.444)	(0.935, 0.200, -1.503)
300	(2.000, 0.803, -0.100)	(-2.000, 0.800, -1.500)	(1.000, 0.197, -1.477)

3.2. Teste de hipóteses

Utilizamos o teste de hipóteses para verificar se existe risco base na função intensidade e se existe ou não influência da quantidade de eventos recorrentes no modelo, que corresponde, respectivamente, a testar as hipóteses $H_{o1} : \alpha = 0$ e $H_{o2} : \gamma = 0$. Fixamos o nível de significância de 5% com 999 com $\theta = \theta_1 = (2,\ 0.8,\ -0.1)$ e utilizamos o teste *Wald* e o teste da RV.

Sob as hipóteses apresentadas, as estatísticas do teste *Wald* e da RV são, respectivamente, $W_\theta = \frac{\widehat{\theta}^2}{var(\widehat{\theta})}$, $RV = 2l(\widehat{\theta}) - 2l(0)$, em que $\theta = (\alpha, \gamma)$, as quais assumem assintoticamente a distribuição Qui-Quadrado com 1 grau de liberdade sob H_o, sendo $\chi^2_{1,0.05} = 3.841$. o valor crítico para ambos os testes.

A Tabela 2 dispõem as porcentagens das vezes em que o valor da estatística *Wald* (coluna 2 e 3, esquerda) e da estatística RV (colunas 4 e 5, esquerda) excederam o valor crítico. Observamos que para 10 recorrências, considerando o teste *Wald* e $\alpha = 0$, (risco base) temos que 16.9% (2^a coluna, linha 3, esquerda) dos valores são maiores que o valor crítico $\chi^2_{1,0.05}$. Contrapondo o esperado nível de significância nominal de 5%, o mesmo resultado é verificado para o parâmetro γ. Melhoras ocorrem a partir de 50 recorrências, todavia os valores encontrados estão muito acima do nível de significância nominal de 5% mesmo para amostras de tamanho 300 (ver Tabela 2). Situação pior ocorre para o teste RV, em que para 300 recorrências para a hipótese de $\gamma = 0$, temos que 16.3% (5^a coluna, linha 8, esquerda) dos valores obtidos são maiores que o valor crítico, bem maior que o valor nominal esperado (5%).

Em decorrência dos problemas apresentados acima, as distribuições assintóticas das estatísticas *Wald* e RV são obtidas via *bootstrap*.

Para utilizar a abordagem de estatísticas empíricas consideramos $\alpha = 0$, $\beta = 0.8$ e $\gamma = -0.1$ para testar a hipótese sobre α e, em seguida, $\gamma = 0$ com $\alpha = 2$ e $\beta = 0.8$ para testar a hipótese sobre γ. A obtenção das amostras, das estimativas dos parâmetros e dos valores das estatísticas empíricas dos testes *Wald* e RV foram obtidas como já descrito e armazenados para comparação futura.

Considerando as estimativas dos parâmetros não fixados pelo teste, β e γ para o primeiro teste, gerou-se nova amostra e desta calcula-se o valor das estatíticas *Wald* e RV. Repetiu-se esse procedimento 499 vezes e, dos valores calculados ao longo do processo, obtivemos o percentil 95, o qual é comparado com estatística empírica

inicialmente calculada. Esse procedimento é repetido 399 vezes e da comparação de todos com a estatística empírica resulta o percentual de valores da estatística *Wald* e RV que excederam o valor empírico, cujos resultados apresentamos na Tabela 2, nas colunas $2-5$ (direita).

Tabela 2: Valores em Porcentagem em relação ao número de vezes que a Estatística de *Wald* ou RV excedeu a $\chi^2_{1,0.05} = 3.841$ (esquerda) e em relação ao número de vezes que a estatística excedeu a Estatística de *Wald* ou RV empírica (direita)

	Wald		RV	
m	$W_{\alpha=0}$	$W_{\gamma=0}$	$W_{\alpha=0}$	$W_{\gamma=0}$
10	16.9/7.36	49.6/ 22.72	63.1/ 18.79	78.8/ 12.28
25	9.6/6.89	35.7/ 15.08	48.3/ 12.03	62.8/ 15.04
50	7.0/5.02	12.1/ 10.69	42.7/ 7.69	41.2/ 8.96
100	6.9/4.98	8.6/ 5.09	40.9/ 5.96	35.4/ 7.26
150	6.6/4.99	7.8/ 4.53	28.5/ 5.63	28.0/ 6.02
300	6.2/5.01	6.9/ 4.26	11.4/ 4.33	16.3/ 5.16

Na Tabela 2, colunas 2 e 3 (direita), apresentamos a porcentagem das estatísticas de *Wald* que excederam os valores empíricos calculados. Observa-se que para o parâmetro α os valores obtidos são próximos do nível de significância, 5% , e para γ as porcentagem se aproximam do valor esperado a partir de 50 recorrências. Para a estatística RV, a linha 7, colunas 4 e 5 (direita), apresentam-se as porcentagens dos valores da RV que excederam sua estatística empírica considerando 150 recorrências e essa porcentagem é 5.63% para o teste considerando $\alpha = 0$ e 6.02% para o teste considerando $\gamma = 0$, os quais são próximos do nível de significância estipulado, 5%.

Com base nos resultados, verificamos que com o aumento do número de recorrência, diminui a porcentagem dos valores que excedem o valor empírico, aproximando-se do valor de significância de 5% a partir de um tamanho amostral de 50 observações.

4. Análise de dados

O modelo proposto trata de tempos de recorrência do evento em um mesmo indivíduo ou de um mesmo componente. Nesta seção apresentamos duas análises utilizando o modelo híbrido proposto, inicialmente consideramos dados artificiais e em seguida dados reais referentes aos momentos (tempos) em que um cliente efetua compra de cosméticos. A necessidade de exemplificar o modelo híbrido proposto é de fundamental importância para validar o mesmo em situações em que pode ser utilizado.

4.1. Dados artificiais

Gerou-se 55 recorrências considerando $\theta = (\alpha, \beta, \gamma) = (2, 0.8, -0.1)$. As EMVs obtidas via maximização direta da função $L(\theta)$ são $\widehat{\theta} = (2.494571, 1.08182, -0.154567)$

com variância padrão de 0.07794947, 0.29474982 e 0.00551214 para $\widehat{\alpha}$, $\widehat{\beta}$ e $\widehat{\gamma}$ respectivamente.

Análise gráfica do ajuste é apresentada na A Figura 4.2. (esquerda), a qual expõe o gráfico do número de recorrência *versus* tempo de recorrência. A linha contínua central apresentada no gráfico é referente ao conjunto de dados obtidos utilizando as estimativas dos parâmetros do modelo ajustado e as demais referem-se aos limitantes de 95% de confiança e os pontos são os valores artificiais utilizados. Assim verificamos que os valores estimados são próximos dos valores artificiais considerados.

O teste de significância dos parâmetros é feita pelo testes de hipóteses *Wald*, a 5% de significância, cuja estatística é $W_{\delta}^2 = \frac{\widehat{\delta}^2}{var(\widehat{\delta})}$, em que $var(\widehat{\delta}) = \left(I_{3\times 3}\left(\widehat{\delta}\right)^{-1} \right)$ e $\widehat{\delta} = \widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}$, a qual é comparada com a distribuição Qui-Quadrado e, portanto, o limitante superior da região crítica é $\chi^2_{1,0.05} = 3.841$.

Foram montadas as hipóteses $H_0 : \delta = 0$, em que $\delta = \alpha, \beta, \gamma$ e as estatísticas calculadas são dadas por $W_{\alpha}^2 = 79.878$, $W_{\beta}^2 = 3,96$ e $W_{\gamma}^2 = 4,34$, que são maiores que o valor crítico, rejeitando as hipóteses nulas e, portanto, existe risco base e comportamento de tendência nos dados e também há um Processo de Poisson Perturbado, com 95% de confiança.

Para o caso de dados artificiais o modelo teve aplicabilidade e os testes indicam que os parâmetros são significativos ao nível de significância de 5%. Tratamos a seguir um conjunto de dados reais e nele avaliamos o comportamento do modelo.

4.2. Recorrência na compra de cosmético

Na atualidade o setor de cosméticos é um dos segmentos comerciais que mais crescem no Brasil e, nesta seção, tratamos de um exemplo real, que refere-se ao histórico de compras de um cliente de cosméticos. O evento de interesse é o tempo das compras, medido em dias. O conjunto de dados foi obtido junto à uma revendedora e corresponde aos histórico de sete anos de um cliente que teve 57 recorrências. Os dados estão apresentados em dias na Tabela 4.

Tabela 3: Conjunto de dados reais, valores em dias.

14	76	86	148	212	252	274	310	356	396	407
511	576	580	608	616	619	641	652	677	680	695
814	842	864	875	1040	1048	1231	1246	1260	1264	1321
1346	1357	1372	1490	1526	1584	1753	1757	1782	1814	1908
2016	2056	2059	2149	2167	2178	2239	2290	2300	2369	2369
2408	2430									

Para verificar a influência do número de compras anteriores no tempo da compra futura obtemos as estimativas $\widehat{\theta} = (2.472, 1.329, -0.167)$ com variância padrão respectivamente 0.071, 0.406 e 0.006, que, por serem valores baixos validam as estimativas e temos que $L\left(\widehat{\theta}\right) = 67.052$.

Com o teste de hipótese com a estatística *Wald* com nível de significância de 5%, verificamos se o risco base é significativo, se há tendência no processo e se o número de recorrência influencia o processo, utilizando as hipóteses $H_0 : \delta = 0$, *versus* $H_a : \delta \neq 0$, em que $\delta = \alpha, \beta, \gamma$.

As estatísticas calculadas são $W_\alpha^2 = 86,067$, $W_\beta^2 = 4.350$ e $W_\gamma^2 = 4.648$, sendo todas maiores que o valor crítico $\chi^2_{1,0.05} = 3.841$, dando-nos evidências para rejeitar as hipóteses nulas. Podemos afirmar, com 95% de confiança, que existe risco base e comportamento de tendência nos dados, havendo portanto um Processo de Poisson Perturbado.

Com parâmetros significativos, obtivemos o tempo estimado de recorrência, a Figura 4.2. (direita) apresenta os valores amostrados e os valores estimados.

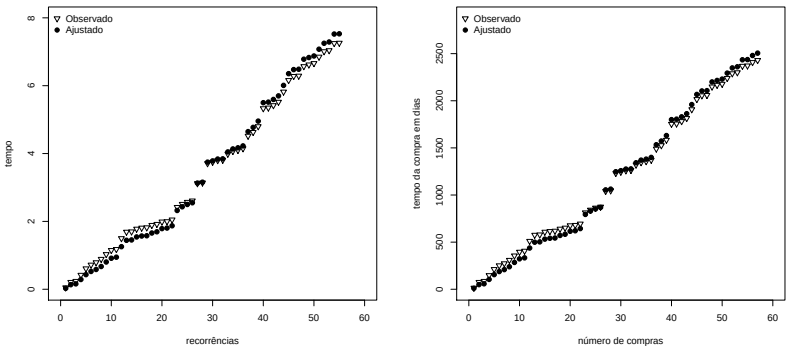

Figura 1: Painel esquerdo: Tempos estimados *versus* tempos simulado. Painel direito: Tempos de vendas, observados e estimados, pelo número de vendas.

Observa-se que o tempo estimado e o tempo observado são próximos, indiferente ao número de compras. Em alguns momentos o valor estimado superestima o valor observado, em outros subestima-o e há situação em que os valores coincidem, o que indica que o modelo proposto é adequado para a análise de eventos que consideram a variável tempo e também o número de eventos ocorridos. Desta forma, estamos satisfeitos com o modelo proposto e a seguir elencamos algumas considerações.

5. Comentários

O modelo proposto é direcionado para análise de tempos de recorrências de um evento na mesma unidade amostral, focando a importância da quantidade de eventos que podem ocorrer na mesma, referindo-se a um Processo de Poisson Perturbado, ou seja, uma componente representando o Processo de Poisson e outra o número de eventos ocorridos nesta unidade. De acordo com o nosso estudo de simulação, o modelo proposto é adequado quando trabalhamos com número de recorrências superior a 50. Em relação à consistência dos estimadores, verificamos que quanto

maior o número de recorrências mais as estimativas se aproximam do verdadeiro valor do parâmetro.

Abstract. In this paper we to model the intensity function of Poisson process considering the total number of recurrent events and the total tiem for each sample unit subject to earlier time. Being that one of the component represents the Poisson process and the other the total number of events occurring in the same unit. Simulation studies and empirical hypothesis testing for significance of the parameters were performed. Significance of the Wald and likelihood ratio hypothesis testing were was approximately 10% to more 50 recorrences. A dataset with recorrences time in the acquisition of cosmetics was modeled properly, having significant parameters and estimated and observed values are close, justifying the use of the proposed model for times and numbers of recurrences in a sample unit.

Referências

[1] Ascher, H. & Feingold, H. (1984). "Reparable Systems Reliability: Modelling, Inference, Misconceptions and Their Causes". New York, Marcel Dekker.

[2] Bain, L. J. & Engelhardt, M. (1980). Inferences on the parameters and current system reliability for a time truncated Weibull process. *Technometrics*, **22**,421-426.

[3] Cook, R. (1995). The design and analysis of randomized trials with recurrent events. *Statistics in Medicine*, **14**, 2081-2098.

[4] Cook, R. & Lawless, J. (1997). Marginal analysis of recurrent events and a terminating event, *Statistics in Medicine*, **16**, 911-924.

[5] Cook, R. J. & Lawless, J. F. (2002). Analysis of repeated events. *Statistical Methods in Medical Research*. **11**, 141-166.

[6] Cox, D. R. & Lewis, P. A. W. (1966). "The Statistical Analysis of Series of Events". London: Methuen.

[7] Cox, D. R. & Hinkley, D. V. (1974). "Theoretical Statistics". London: Chapman & Hall.

[8] Crowder, M. J., Kimber, A. C., Smith, R. L. & Sweeting, T. J. (1991). "Statistical Analysis of Reliability Data". London: Chapman and Hall.

[9] Guo, H., Zhao, W. & Mettas, A. (2006). Practical methods for modeling repairable systems with time trends and repair effects. *Proceedings of Annual Reliability and Maintainability Symposium*, California, 182-188.

[10] Lawless, J. F. (1987). Regression methods for Poisson process data. *Journal of the American Statistical Association*, **82**, 807-815.

[11] Lawless, J.F. & Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics*, **37**, 158-168.

[12] Lawless, J. F. & Thiagarajah, K. (1996). A point-process model incorporating renewals and time trends, with application to repairable systems. *Technometrics*, **38**, 131-138.

[13] Lehmann, E. L. (1999). "Elements of Large-Sample Theory". New York Springer-Verlag, New York.

[14] Lee, L., & Lee, K. (1978). Some results on inference for the Weibull process. *Technometrics*, **20**, 41-45.

[15] Lindqvist, B.H., Elvebakk, G. & Heggland K. (2003). The trend-renewal process for statistical analysis of repairable systems. *Technometrics*, **45**, 31-44.

[16] Louzada, F., Mazuchelli, J. & Achcar, J. A. (2002)."Introdução à Análise de Sobrevivência e Confiabilidade". *III Jornada Regional de Estatística*.

[17] Nelson, W. (1995). Confidence Limits for Recurrence Data - Applied to Cost or Number of Product Repair. *Technometrics*, **37**,147-157.

[18] R Development Core Team (2011). R: A Language and Environment for Statistical Computating. R Foundation for Statistical Computing, Vienna, Austria.

[19] Tomazella, V. L. D. (2003). "Modelagem de Dados de Eventos Recorrentes via Processos de Poisson com Termo de Fragilidade", 165p. Tese. (Doutorado em Ciências de Computação e Matemática Computacional)- Instituto de Ciências Matemáticas de São Carlos, Universidade de São Paulo, São Carlos.