

# Função de intensidade Poisson Perturbada pelo Número de Ocorrências para dados de eventos Recorrentes

Sabrina Luzia Caetano

Orientador: Prof. Dr. Francisco Louzada Neto

Defesa apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

São Carlos  
Junho de 2007

**Ficha catalográfica elaborada pelo DePT da  
Biblioteca Comunitária da UFSCar**

C128fi

Caetano, Sabrina Luzia.

Função de intensidade Poisson perturbada pelo número de ocorrências para dados de eventos recorrentes / Sabrina Luzia Caetano. -- São Carlos : UFSCar, 2008.  
63 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2007.

1. Distribuição de Poisson perturbada. 2. Função de intensidade híbrida. 3. Análise de sobrevivência. 4. Distribuição de Poisson. 5. Método de estimação. 6. Método bootstrap. I. Título.

CDD: 519.54 (20ª)

Dedico esta tese, com muito amor e carinho, aos meus pais Luiz Antônio Caetano e Rosa Maria dos Santos Caetano, que sempre me incentivaram e me ajudaram a lutar pelas coisas em que eu acredito: e ao meu namorado Samuel Maria, que sempre acreditou que eu fosse capaz.

"O débil, acovardado, indeciso e servil não conhece nem pode conhecer, o generoso impulso que guia aquele que confia em si mesmo, e cujo prazer não é o de ter conseguido a vitória, senão de sentir capaz de conquistá-la."

William Shakespeare

## *Agradecimentos*

Agradeço...

Primeiramente a Deus, que sempre me iluminou e me guiou para mais esta realização em minha vida.

Aos meus familiares, Luíz Antônio, Rosa Maria, Vanderlei, Rogério, Ivanete, Rosemari, Isadora, Laura, Leonardo, Luizinho e Samuel que sempre estiveram do meu lado.

Ao meu orientador Prof. Dr. Francisco Louzada Neto, por ter acreditado em minha capacidade e ter me dado a oportunidade de iniciar este trabalho, ao Prof. Dr. José Carlos Fogo, pela admirável dedicação, pelo esforço e sugestões durante todo o processo de elaboração e desenvolvimento deste trabalho.

A Prof. Dra. Vera Lucia Damasceno Tomazella pelas sugestões e comentários dados quando da realização do meu exame de qualificação.

A todos os professores e funcionários do Departamento de Estatística da UFSCar - DEs - pelo apoio às minhas dificuldades e pelo companheirismo.

A todos os amigos que, de alguma maneira contribuíram para a realização deste trabalho em especial aos amigos, Juliana, Fausto, Beliza, Daniela, Glaucy, Adriano, Juninho, Tássia, Gabriela, Vera, Valdemar, Silvia, Maihara e Leandra, pela amizade e carinho e pela ajuda no decorrer do trabalho.



# *Resumo*

Uma particularidade da análise de sobrevivência e confiabilidade se refere ao fato de que existem situações em que um evento de interesse pode ocorrer várias vezes para uma unidade amostral. Nesta dissertação estudamos o modelo de intensidade Poisson Perturbada para dados de eventos recorrentes.

Primeiramente uma revisão da literatura sobre os processos de Poisson e Renovação é realizada. Um modelo de intensidade Poisson Perturbada para dados de eventos recorrentes é apresentado, sendo que a estimação dos parâmetros de interesse é realizada para sistemas reparáveis simples, retratando apenas uma unidade, podendo ser esta caracterizada por um indivíduo ou componente.

Os exemplos considerados neste estudo, referem-se a dados artificiais e também a um exemplo real, tal que o evento de interesse é o tempo de reincidência da variável considerada, podendo ser esta variável o tempo até a falha de um componente, tempo até uma nova crise de uma determinada doença ou até mesmo tempo de compras de um determinado produto.

As estimativas dos parâmetros do modelo são obtidas via máxima verossimilhança. Procedimentos de estimação intervalar são apresentados, bem como testes baseados nas estatísticas de Wald e da razão de verossimilhança são realizados para os parâmetros de interesse.

Palavras-chave: Eventos Recorrentes, Função de Intensidade Híbrida, Processo de Poisson, Estimação de Máxima Verossimilhança, Teste de Wald e Teste de Razão de Verossimilhança.

# *Abstract*

A particularity of the survival analysis refers to the fact of that exist some situations where an event of interest can occur several times for the same sample element. In this dissertation we study a hybrid Poisson/Number of events intensity model recurrence events.

First a revision of the literature about the trials of Poisson and Renewal is carried out. A hybrid Poisson/Number of events intensity model recurrence events is presented, being that the estimation of the parameters of interest is carried out for simple repairable systems, portraying barely a unique sample element, which can be characterized by an individual or component.

The examples considered in this study refer to an artificial and a real example, such that the event of interest is the time of relapse of the considered variable, which can represent the time up to default of a component, the time to a new crisis of a determined disease or to the time of purchases of a determined product.

The estimates of the parameters are obtained via maximum likelihood. Procedures of interval estimation are presented, as well as tests based in the statistics of Wald and the maximum likelihood ratio are carried out for the parameters of interest.

**Keywords:** recurrence events, Hybrid Intensity Function, Poisson Process, Maximum Likelihood Estimation, Wald Test, Likelihood Ratio.

# Sumário

<b>1</b>	<b>Introdução</b>	<b>1</b>
1.1	Conceitos Básicos . . . . .	2
1.1.1	Descrição do Comportamento do tempo de Sobrevivência . . . . .	2
1.1.2	Peculiaridades em Sobrevivência . . . . .	4
<b>2</b>	<b>Sistemas Reparáveis com função de intensidade Híbrida Poisson/Número de eventos</b>	<b>11</b>
2.1	Formulação do Modelo . . . . .	12
2.2	Métodos de Máxima Verossimilhança . . . . .	14
2.3	Exemplo (dados artificiais) . . . . .	17
2.4	Conclusões e Observações . . . . .	20
<b>3</b>	<b>Simulação</b>	<b>21</b>
3.1	Consistência dos Estimadores . . . . .	21
3.2	Testes de Hipóteses . . . . .	22
3.2.1	Estatística de Wald . . . . .	23
3.2.2	Estatística de Razão de Verossimilhança . . . . .	25
3.3	QQ- Plot . . . . .	28
3.4	Conclusões e Observações . . . . .	29
<b>4</b>	<b>Exemplo- Dados Reais</b>	<b>31</b>
4.1	Estimativas . . . . .	33
4.2	Conclusões e Observações . . . . .	35
	<b>Apêndices</b>	<b>36</b>

---

<b>A</b>	<b>Processos de Contagem</b>	<b>37</b>
A.1	Conceitos Básicos . . . . .	38
A.2	Processo de Poisson . . . . .	39
A.3	Processo de Renovação Ordinário . . . . .	42
A.4	Estimação dos parâmetros do modelo . . . . .	45
A.4.1	Intervalo Assintótico . . . . .	45
A.4.2	Intervalo Via Bootstrap . . . . .	46
<b>B</b>	<b>Códigos</b>	<b>48</b>
B.1	Utilizado no Capítulo 2 . . . . .	48
B.2	Utilizado no Capítulo 3 . . . . .	50
B.2.1	Teste de Wald Empírico . . . . .	50
B.2.2	Teste de Razão de Verossimilhança Empírico . . . . .	54
B.3	Utilizado no Capítulo 4 . . . . .	58
	<b>Referências Bibliográficas</b>	<b>60</b>

# Capítulo 1

## Introdução

Uma particularidade da análise de sobrevivência e confiabilidade se refere ao fato de que existem situações nas quais um evento de interesse do pesquisador pode ocorrer várias vezes em uma unidade amostral. Na literatura, dados referentes a situações como estas são chamados de dados de eventos recorrentes e aparecem freqüentemente em áreas como biomedicina, economia, engenharia e confiabilidade industrial, atuária, criminologia e demografia. Alguns exemplos específicos nesta área podem ser encontrados em Cook & Lawless (2002).

Na área biomédica, podemos citar a recorrência de um tipo de tumor em pacientes com câncer, recorrências de epilepsia e episódios de pneumonia, entre outros. Em engenharia e confiabilidade industrial a estrutura de dados de eventos recorrentes pode aparecer no estudo de sistemas mecânicos ou eletrônicos, na análise de defeitos em um software e na verificação da presença de rachaduras em estruturas de concreto, dentre outros.

Nesta estrutura de dados, para cada unidade amostral é observado o número de eventos ocorridos em um determinado período de tempo, os tempos de ocorrência de cada um destes eventos e, além disso, podemos ter disponível um vetor de covariáveis e um vetor de indicadores de censura.

O objetivo deste trabalho, reside em entender e caracterizar o processo de recorrência de eventos nas unidades amostrais e possivelmente a comparação de tratamentos, com base no tempo para cada evento distinto, o número de eventos, os tipos de eventos e as interdependências entre eventos, visando explicar a natureza da variação entre unidades amostrais em termos de tratamento, covariáveis e até mesmo em situações de censura.

Uma breve descrição em relação ao comportamento do tempo de sobrevivência será apresentado na Seção 1, deste capítulo, assim como as suas possíveis peculiaridades (eventos recorrentes, censura e presença de covariáveis). No capítulo 2, o modelo proposto será estudado, tal que este incorpora o Processo de Poisson e o número de eventos concomitantemente, considerando um único sistema recorrente, isto é, um único indivíduo ou componente. sendo que no capítulo 3 é realizada uma simulação e no último capítulo, trabalhamos com um exemplo real.

## 1.1 Conceitos Básicos

Nesta seção iremos apresentar alguns conceitos básicos em relação ao comportamento do tempo de sobrevivência e suas peculiaridades.

### 1.1.1 Descrição do Comportamento do tempo de Sobrevivência

O comportamento do tempo de sobrevivência pode ser descrito por várias funções matematicamente equivalentes, Nelson (1982). A função de densidade de probabilidade é definida como o limite da probabilidade do evento de interesse acontecer no intervalo de tempo  $[t, t + \Delta t]$ , e é expressa por:

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t},$$

em que  $f(t) \geq 0$  para todo  $t$ .

A função de sobrevivência é a probabilidade de um indivíduo sobreviver mais do que um determinado tempo  $t$  e é dada a seguir:

$$S(t) = P(T \geq t) = 1 - F(t), \tag{1.1}$$

em que  $F(t)$  é a função acumulada da função densidade  $f(t)$ .

A função de risco é definida por:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t / T \geq t)}{\Delta t}, \tag{1.2}$$

A função de risco (1.2) é interpretada como a taxa instantânea de falha que se modifica com o passar do tempo.

Podemos observar que algumas distribuições conhecidas podem ser descritas de acordo com o seu comportamento. A distribuição exponencial acomoda funções de risco constantes, a distribuição Weibull, funções monotonicamente decrescente ou crescente em  $t$ ; enquanto a log-logística e a log-normal acomodam funções de risco unimodais.

Além das relações mencionadas anteriormente, descrevemos a seguir algumas outras relações importantes.

Considere a função de densidade de uma variável aleatória  $t$ , dada por:

$$f(t) = \frac{\partial F(t)}{\partial t}.$$

A função de distribuição acumulada de  $t$  é dada por:

$$F(t) = 1 - S(t), \quad (1.3)$$

em que  $S(t)$  é definida em (1.1), enquanto sua função de densidade é a derivada da equação (1.3), ou seja,

$$f(t) = \frac{\partial(1 - S(t))}{\partial t} = -S'(t). \quad (1.4)$$

Além disso, a função de risco (1.2) pode ser escrita como

$$h(t) = \frac{f(t)}{S(t)}. \quad (1.5)$$

Então, substituindo (1.4) em (1.5) temos:

$$h(t) = \frac{-S'(t)}{S(t)} = \frac{-\partial[\log S(t)]}{\partial t}$$

em que o logaritmo de  $S(t)$  assume a relação

$$\log S(t) = - \int_0^t h(u) du. \quad (1.6)$$

Aplicando exponencial em (1.6) temos,

$$S(t) = \exp\left(-\int_0^t h(u)du\right), \quad (1.7)$$

em que  $S(0)=1$ .

A função de risco acumulada é obtida por meio da integração de  $h(t)$ , ou seja,

$$H(t) = \int_0^t h(u)du. \quad (1.8)$$

Substituindo a função de risco acumulada (1.8) em (1.7) temos:

$$S(t) = \exp[-H(t)],$$

como  $S(\infty) = 0$  então  $H(\infty) = \lim_{t \rightarrow \infty} H(t) = \infty$ , e

$$f(t) = h(t)S(t). \quad (1.9)$$

Substituindo também (1.7) em (1.9) temos,

$$f(t) = h(t) \exp\left(-\int_0^t h(u)du\right). \quad (1.10)$$

Estas relações são muito utilizadas, afinal, obtendo qualquer uma dessas funções, podemos chegar nas demais e assim obter quantidades de interesse, como média, probabilidades, sobrevivência e taxas de risco.

Algo que deve ser levado em consideração, e que mais adiante será abordado, é o fato de não trabalharmos necessariamente com a função de risco, isto ocorre porque estamos trabalhando com eventos recorrentes, então desta forma a função de intensidade é a mais conveniente.

### 1.1.2 Peculiaridades em Sobrevivência

A análise de sobrevivência e confiabilidade se faz peculiar devido às características especiais, inerentes aos tipos de dados que são normalmente disponíveis para análise (Louzada-Neto, *et al.*, 2002).

Aqui são discutidas algumas das principais características dos dados de sobrevivência e confiabilidade, que estão diretamente ligadas a este trabalho: a presença de censuras, eventos recorrentes e a presença de covariáveis.

## Censura

Em dados de análise de sobrevivência e confiabilidade, é usual a presença de observações parciais ou incompletas. Isso decorre do fato que o tempo até a ocorrência de um evento não é medido, sendo que valores grandes desta variável necessitam de mais tempo para serem observados. Logo, existe a possibilidade de o evento de interesse não ocorrer até o final do estudo. Existe também a situação na qual não é possível mais obter informações sobre uma determinada unidade amostral sem que o evento de interesse tenha ocorrido. Por exemplo, em um estudo sobre o tempo até a ocorrência de uma doença em um grupo de pacientes, existe a possibilidade de óbito de qualquer unidade da amostra dentro do período estudado. Quando as informações relativas a uma ou mais unidades, por algum motivo, ficam inacessíveis, diz-se, em análise de sobrevivência e confiabilidade, que uma ou mais censuras ocorreram.

Em geral, as censuras podem ocorrer de várias formas, de acordo com diferentes mecanismos, dentre os quais se podem citar: censuras do tipo I, censuras do tipo II e censuras aleatórias (Lawless, 1982). Por exemplo, em análise de sobrevivência, se o estudo termina em um instante pré-estabelecido e alguns dos tempos até a ocorrência de um evento não são observados para algumas unidades amostrais, tem-se a censura do tipo I.

Caso o estudo termine após a ocorrência de uma determinada quantidade pré-estabelecida de censuras, dentre as unidades em estudo, têm-se a censura do tipo II. O tipo de censura aleatória é mais comum na prática médica. Isto acontece, por exemplo, quando um paciente é retirado no decorrer do estudo sem ter ocorrido à falha, ou então, se o paciente morrer por uma razão diferente da estudada.

Geralmente, as censuras de tipo I e aleatórias são observadas com maior frequência em estudos biomédicos, enquanto em experimentos industriais, as censuras do tipo II são predominantes. Além disso, as censuras podem ainda ser classificadas como: censura à direita, à esquerda e intervalar.

A censura à direita ocorre quando a unidade está em observação e, em algum tempo

dentro do período de estudo, seja pela sua exclusão da amostra ou pelo final do estudo, as respectivas informações tornam-se inacessíveis após este tempo.

A censura à esquerda acontece na situação na qual o evento de interesse já ocorreu quando uma unidade é incluída no grupo de unidades em estudo.

A censura intervalar é um tipo mais geral de censura que acontece quando se sabe somente que a censura ocorreu dentro de um certo intervalo de tempo, digamos  $t$  e  $t_0$ .

## Eventos Recorrentes

Experimentos relacionados com dados de eventos recorrentes são comuns em análise de sobrevivência e confiabilidade, sendo que, neste tipo de experimento, a definição do evento de interesse deve ser feita de forma clara e concisa. Como exemplo de experimentos com dados de eventos recorrentes podemos citar:

1. Uma empresa do ramo automobilístico está lançando um novo veículo no mercado e pretende negociá-lo com, no mínimo, um ano de garantia total. Para oferecer esta garantia ao consumidor, ela necessita de uma estimativa do número e custo médio acumulado de reparos deste veículo no período de um ano.

2. Uma cooperativa que oferece serviços hospitalares possui um grupo de pacientes que se submeteram a uma cirurgia, e tem-se o interesse em estimar o número de vezes que esses pacientes reutilizarão os serviços oferecidos pela cooperativa.

3. Uma administradora de cartões de crédito tem o interesse em estimar qual será o número de vezes que um grupo de clientes deixará de pagar a fatura, após um determinado tempo  $t$ .

4. Uma empresa prestadora de serviços oferece uma garantia de seis meses pelos serviços prestados e deseja expandir essa garantia para doze meses. Para isto, ela necessita de uma estimativa do número e custo médio de revisões de seus serviços em um ano.

Existem algumas situações nas quais não se sabe exatamente quais foram os tempos em que os eventos de interesse ocorreram nas unidades amostrais, sendo possível apenas observar intervalos de tempo, nos quais se sabe que os eventos ocorreram. Aqui é feita a suposição de que os tempos de recorrência de eventos são exatamente conhecidos.

Em experimentos com dados de eventos recorrentes existe um conjunto de variáveis que podem ser observado em cada unidade amostral tais como: número total de eventos

em um determinado período de tempo, tempos de recorrência dos eventos, tempos entre os eventos, e, em muitas situações, o custo relacionado a cada recorrência. A Figura 1.1 ilustra as possíveis variáveis observáveis em um experimento na presença de três eventos recorrentes para uma unidade amostral (Louzada-Neto, *et al.*, 2002).

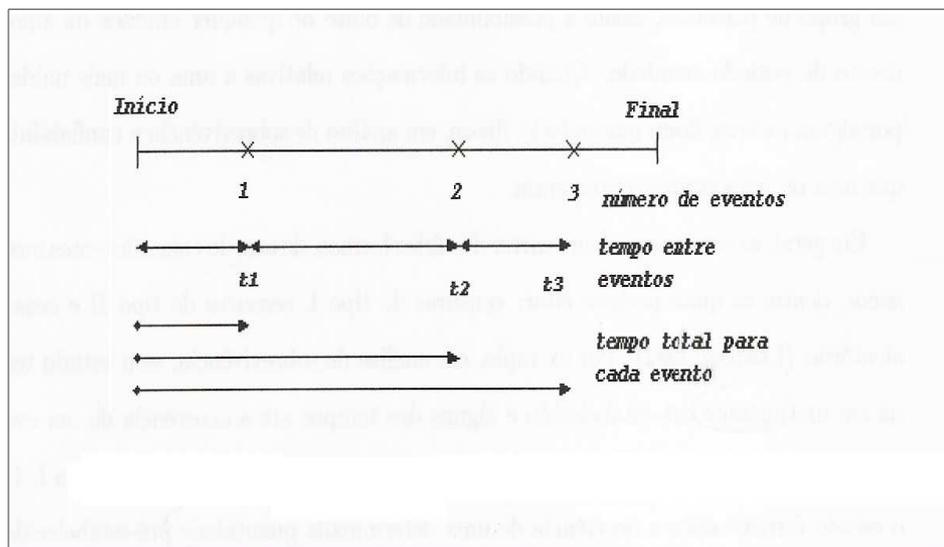


Figura 1.1. Variáveis observáveis em uma unidade amostral com três eventos recorrentes.

Um exemplo que pode ser citado, é o número de crises de bronquite de uma criança de 10 anos ao longo de sua vida, desta forma teríamos o tempo total para cada evento e também o tempo entre eventos.

### Presença de Covariáveis

Em várias situações práticas podemos observar nos dados, além do tempo de sobrevivência e da variável indicadora de censura, a presença de variáveis explicativas ou covariáveis que representam a heterogeneidade existente na população, tais como, idade, sexo, entre outras, e também possíveis tratamentos ou testes, aos quais os indivíduos ou componentes são submetidos.

Neste contexto o objetivo da análise de sobrevivência está centrado na relação entre os tempos de sobrevivência e as variáveis explicativas de interesse.

Aqui consideramos situações em que as quantidades observadas em cada indivíduo

são representadas pela tripla  $(t_i, \delta_i, \mathbf{z}_i)$ , para  $i = 1, 2, \dots, n$ , em que  $t_i$  denota a variável tempo de sobrevivência,  $\delta_i$  denota a variável de censura, tal que  $\delta_i = 1$  representa uma observação sem censura e  $\delta_i = 0$  uma observação censurada e  $\mathbf{z}_i^t = (z_{i1}, \dots, z_{ik})$  denota o **vetor** de  $k$  covariáveis.

Os modelos de sobrevivência para este tipo de dados são construídos a partir de dois componentes, um aleatório e outro determinístico, isto no ponto de vista paramétrico. Assim o componente aleatório é representado por uma distribuição de probabilidade vinculada ao comportamento do tempo de sobrevivência, enquanto que o componente determinístico é representado pelo relacionamento entre os parâmetros desta distribuição e as covariáveis.

Considerando o modelo exponencial dada sua função de probabilidade, podemos escrevê-la da seguinte forma:

$$f(t, \varphi_i) = \varphi_i \exp\{-\varphi_i t\},$$

em que  $\varphi_i > 0$  representa a taxa constante de falha no  $i$ -ésimo nível da covariável  $z_i$ ,  $i = 1, \dots, k$ .

O componente determinístico da formulação do modelo para dados de sobrevivência na presença de covariáveis é representado por uma função que estabelece o relacionamento entre o parâmetro da função de distribuição do tempo de sobrevivência e as covariáveis. Este relacionamento pode ser representado genericamente por:

$$\varphi_i = g(\boldsymbol{\beta} \mathbf{z}_i), \tag{1.11}$$

em que  $g(\cdot)$  é uma função positiva e contínua,  $\boldsymbol{\beta}$  é um vetor de coeficientes a serem estimados e  $\mathbf{z}_i$  é um vetor de covariáveis.

Quando somente uma covariável é considerada, reescrevemos (1.11) na forma:

$$\varphi_i = \exp\{-(\beta_0 + \beta_1 z_i)\}, \tag{1.12}$$

em que  $\beta_0$  e  $\beta_1$  são coeficientes a serem estimados, tal que  $-\infty < \beta_0, \beta_1 < \infty$  para  $i = 1, \dots, k$ .

Considerando uma amostra aleatória  $T_1, \dots, T_n$ , e que associado a  $T_i$  exista um vetor

de covariáveis  $z_i$  e uma variável indicadora de censura  $\delta_i$ , a função de verossimilhança, baseada nas informações disponíveis,  $(t_1, \delta_1, z_1), \dots, (t_n, \delta_n, z_n)$ , pode ser escrita de (1.11) e de (1.12), na forma (Lawless, 1982):

$$L(\beta|z) = \prod_{i=1}^n f(t_i|z_i)^{\delta_i} S^{1-\delta_i}(t_i|z_i),$$

$$L(\beta|z) = \prod_{i=1}^n \exp\{-\delta_i(\beta_0 + \beta_1 \sum_{j=1}^K \beta_j z_{ij})\} \exp\{-t_i \exp\{-(\beta_0 + \beta_1 \sum_{j=1}^k \beta_j z_{ij})\}\}.$$

E assim podemos calcular os estimadores de máxima verossimilhança (*EMV*), que pode ser feito via métodos computacionais.

Quando temos grandes amostras, as inferências sobre  $\beta$  podem ser baseadas na aproximação assintótica dos estimadores de máxima verossimilhança (Cox e Hinkley, 1974).

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1) \sim N(\beta, I_o^{-1}(\hat{\beta})),$$

em que  $I_o(\hat{\beta})$  é a matriz de informação observada com elementos definidos por:

$$I_{pq}(\beta) = \frac{\partial^2 \log(L(\beta|z))}{\partial \beta_{p-1} \partial \beta_{q-1}} \Big|_{\hat{\beta}_0, \hat{\beta}_1},$$

sendo  $p, q = 1, 2$ . Assim,

$$p = 1, q = 1 \rightarrow I_{11} = \frac{\partial^2 \log(L(\beta|z))}{\partial \beta_0 \partial \beta_0} = \frac{\partial^2 \log(L(\beta|z))}{\partial \beta_0^2}$$

$$p = 1, q = 2 \rightarrow I_{12} = \frac{\partial^2 \log(L(\beta|z))}{\partial \beta_0 \partial \beta_1}$$

$$p = 2, q = 1 \rightarrow I_{21} = \frac{\partial^2 \log(L(\beta|z))}{\partial \beta_1 \partial \beta_0}$$

$$p = 2, q = 2 \rightarrow I_{22} = \frac{\partial^2 \log(L(\beta|z))}{\partial \beta_1 \partial \beta_1} = \frac{\partial^2 \log(L(\beta|z))}{\partial \beta_1^2}$$

A parte de Processos de contagem que conta como a base deste trabalho, foi estudada

e encontra-se no apêndice deste trabalho, caso haja o interesse de consulta.

## Capítulo 2

# Sistemas Reparáveis com função de intensidade Híbrida Poisson/Número de eventos

Existem vários modelos para a análise de dados de eventos recorrentes. Metodologias baseadas no processo de Poisson e no processo de renovação são freqüentemente utilizadas. O processo de Poisson é utilizado quando o interesse é modelar o tempo total de estudo, enquanto que o processo de Renovação é utilizado para modelar o tempo intervalar, ou seja, o tempo entre eventos, como já tratado no capítulo anterior, ver Nelson, (1995).

Todavia o modelo proposto neste trabalho, envolve o número de eventos recorrentes e não o tempo intervalar, isto foi realizado em virtude de que em muitas situações a quantidade de eventos influencia os próximos tempos de recorrência, ou seja, a partir daí teremos um Processo de Poisson Perturbado.

A Seção 2.1 trata da formulação do modelo, a Seção 2.2 descreve a estimação de máxima verossimilhança e teste de hipótese. A Seção 2.3, um exemplo com dados artificiais, incluindo as estimativas dos parâmetros, e a Seção 2.4 conclui com algumas observações à extensão deste capítulo.

## 2.1 Formulação do Modelo

Processos envolvendo eventos recorrentes são comuns em confiabilidade e várias outras áreas, desta forma considere um sistema reparável, observando um tempo  $t \geq 0$ , e suponha que os eventos ocorram nos tempos  $(t_1, t_2, \dots)$ , tal que  $t_1 < t_2$ . Dado que  $x_i = t_i - t_{i-1}$  (com  $t_0 = 0$  e  $i = 1, 2, \dots$ ), assim  $x_i$  denota o tempo entre os eventos, e  $N(s, t)$  denota o número de eventos no intervalo  $[s, t)$ , a notação  $N(t)$  também pode ser usada para representar  $N(0, t)$ , ver Lawless e Thiagarajah (1996).

Geralmente, um modelo de probabilidade para um determinado ponto no processo pode ser especificado em termos de uma condicional ou uma função intensidade completa (FIC).

Definimos  $H_t = \{N(s) : 0 \leq s < t\}$  como o histórico do processo acima para o tempo  $t$ . Então a FIC é dada por:

$$\lambda(t; H_t) = \lim_{\Delta t \downarrow 0} \frac{P\{N(t, t + \Delta t) = 1 | H_t\}}{\Delta t} \quad (2.1)$$

Nós observamos que a função média para o processo é definida como  $P(t) = E\{N(t)\}$ , e que  $p(t) = P'(t)$  é chamado de razão de ocorrência, ou apenas de função razão. Agora para processos em que os eventos não ocorrem simultaneamente, a função razão é dada por  $\lim_{\Delta t \downarrow 0} P\{N(t, t + \Delta t) = 1\} / \Delta t$ .

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

Processos de Poisson são processos em que (2.1) é da seguinte forma

$$\lambda(t, H_t) = \rho(t), \quad (2.2)$$

neste caso  $\rho(t)$  é a intensidade e a função razão. Para processos de Poisson, bem se sabe, que  $N(t)$  tem uma distribuição de Poisson com média  $P(t) = \int_0^t \rho(\mu) d\mu$  e que o número de eventos em seqüência relacionados com os tempos de intervalos são independentes.

Processos de renovação, por outro lado, são modelos os quais (2.1) é da forma:

$$\lambda(t, H_t) = h(t - t_{N(t-)}), \quad (2.3)$$

em que  $t_{N(t-)}$  é o último evento antes de  $t$ . Então (2.3) implica que os tempos  $x_i$  entre os sucessivos eventos são independentes e identicamente distribuídos (iid), com função de risco  $h(x)$ , que é a maneira que os processos de renovação são normalmente definidos.

Como já se sabe, o processo de Poisson homogêneo, no qual  $\rho(t)$  é constante, são também processos de renovação com  $x'_i$ s seguindo uma distribuição exponencial, todavia, as famílias dos processos de Poisson e renovação são distintas, ver Cox e Isham (1980) ou Ross (1983).

Nós podemos usar (2.1) para formular modelos incorporando os tempos de tendência (Poisson) e número de eventos. Geralmente, podemos estabelecer efeitos de eventos passados dentro do modelo.

Desta forma considerando a equação em (2.1), temos:

$$\lambda(t, H_t) = e^{\theta' \mathbf{z}(t)}, \quad (2.4)$$

em que  $\mathbf{z}(t) = (z_1(t), \dots, z_p(t))'$  é um **vetor** de funções que pode depender de  $t$  e  $H_t$ , e  $\theta = (\theta_1, \dots, \theta_p)'$  é um vetor de parâmetros desconhecidos. Todos os vetores são considerados vetor coluna.

Vários modelos comuns são casos especiais de (2.4). Estes incluem processo de Poisson com função intensidade: a)  $\rho_1(t) = \exp(\alpha + \beta t)$  e b)  $\rho_2(t) = \alpha t^\beta$ , dados por a)  $\mathbf{z}(t) = (1, t)'$ ,  $\theta = (\alpha, \beta)'$  e b)  $\mathbf{z}(t) = (1, \log t)'$ ,  $\theta = (\log \alpha, \beta)'$  respectivamente. Métodos estatísticos para estes modelos tem sido considerado por vários autores (e.g., Ascher e Feingold, 1984; Bain e Engelhardt, 1980; Cox e Lewis, 1966; Crowder *et al.*, 1991; Lawless, 1987; Lee e Lee, 1978).

Assumindo que tempos de recorrência dos eventos são observados, inferência paramétrica para modelos da forma (2.4) mostram-se diretos e rápidos através de desenvolvimento computacional. O restante deste capítulo desenvolve e investiga procedimentos de inferência, observando-se que variáveis explicativas podem ser incluídas nos termos  $\mathbf{z}_i(t)$  em (2.4), ainda que este trabalho não inclua este tipo de situação. Todavia o modelo que iremos tratar tem como  $\mathbf{z}(t) = (1, t, N(t))'$  e  $\theta = (\alpha, \beta, \gamma)'$ , assim a função intensidade

pode ser escrita como:

$$\lambda(t, H_t) = e^{\alpha + \beta t + \gamma N(t)}, \quad (2.5)$$

O que podemos verificar no modelo (2.5), é que o parâmetro  $\alpha$  funciona como risco de base, ou seja, independentemente do tempo de ocorrência do evento ele já estará implícito, agora o parâmetro  $\beta$  é caracterizado como o coeficiente do tempo, ou seja, ele será significativo se o tempo interferir na função intensidade. O último parâmetro que é dado por  $\gamma$ , também só terá significância se a função de risco, no caso a intensidade por se tratar de eventos recorrentes depender da quantidade de eventos.

## 2.2 Métodos de Máxima Verossimilhança

Nós consideramos a função de verossimilhança para um simples processo observado no intervalo de tempo  $(0, T]$ . A verossimilhança é proporcional para densidade de probabilidade para dados observados, na qual é da forma  $\{n \text{ eventos, no tempo } t_1 < \dots < t_n \leq T\}$ ; em que  $n \geq 0$ . Geralmente, a verossimilhança para um processo com intensidade (2.1) é dado por:

$$L(\theta) = \prod_{i=1}^n \lambda(t_i, H_{t_i}) \exp\left\{-\int_0^T \lambda(t, H_t) dt\right\}. \quad (2.6)$$

Neste caso  $\lambda(t, H_t)$  é dado por (2.4), tal que  $i = 1, 2, \dots, n$ , representa as recorrências.

Em particular  $T$  não está pré-especificado, mas pode ser considerado como o último tempo, sendo assim  $T = t_n$ .

Para o modelo da forma (2.4), a função do log da verossimilhança  $l(\theta) = \log L(\theta)$  é dada por:

$$l(\theta) = \sum_{i=1}^n \theta' z(t_i) - \int_0^T e^{\theta' z(t)} dt. \quad (2.7)$$

A função escore é dada por:

$$U_r(\theta) = \frac{\partial l(\theta)}{\partial \theta_r} = \sum_{i=1}^n z_r(t_i) - \int_0^T z_r(t) e^{\theta' z(t)} dt, \text{ para } r = 1, \dots, p, \quad (2.8)$$

e a matriz de informação observada  $p \times p$   $I(\theta)$  dada por:

$$I_{rs}(\theta) = -\frac{\partial^2 l}{\partial \theta_r \partial \theta_s} = \int_0^T z_r(t) z_s(t) e^{\theta' z(t)} dt. \quad (2.9)$$

Em casos especiais nós podemos avaliar as integrais em (2.7) – (2.9) analiticamente, mas em geral a integração numérica é necessária.

Se o número total esperado de eventos é grande e apropriadas condições no modelo são asseguradas, então, para encontrar as estimativas ou testes para  $\theta$  nós podemos tratar  $\hat{\theta}$  como tendo aproximadamente uma distribuição normal com média  $\theta$  e matriz de variância-covariância  $I_o(\theta)^{-1}$ . Escores ou estatísticas de verossimilhança podem também serem usadas de forma usual.

Considerando a função dada em (2.4), dado que o modelo proposto por este estudo tem que  $\mathbf{z}(t) = (1, t, N(t))'$  e  $\boldsymbol{\theta} = (\alpha, \beta, \gamma)'$ , a função intensidade é dada por (2.5).

Desta forma a função de verossimilhança (2.6) considerada, é dada por:

$$L(\theta) = \prod_{i=1}^n e^{\alpha + \beta t + \gamma N(t)} \cdot \exp\left\{-\int_0^T e^{\alpha + \beta t + \gamma N(t)} dt\right\}. \quad (2.10)$$

As equações de máximo da verossimilhança ou equação score e a matriz de informação observada  $I_{p \times p}(\theta)$  são encontradas através das devidas substituições de  $\mathbf{z}(t) = (1, t, N(t))'$  e  $\boldsymbol{\theta} = (\alpha, \beta, \gamma)'$ , nas equações (2.8) e (2.9). Considerando inicialmente a função de distribuição condicional acumulada, dado que  $T = t_n$  e  $N(t_i) = (i - 1)$ , temos:

$$F(t_i|t_{i-1}) = P(t \leq t_i | t > t_{i-1}) = \frac{F(t_i) - F(t_{i-1})}{S(t_{i-1})} = 1 - \frac{S(t_i)}{S(t_{i-1})} \quad (2.11a)$$

$$S(t_i|t_{i-1}) = 1 - F(t_i|t_{i-1})$$

$$S(t_i|t_{i-1}) = \exp\left(-\int_{t_{i-1}}^{t_i} \lambda(t) dt\right) \quad (2.11b)$$

$$S(t_i|t_{i-1}) = \exp\left(-\int_{t_{i-1}}^{t_i} \exp(\alpha + \beta t + (i-1)\gamma) dt\right)$$

$$S(t_i|t_{i-1}) = e^{\frac{-e^{(\alpha+(i-1)\gamma)}(e^{\beta t_i}) - e^{(\beta t_{i-1})}}{\beta}} \quad (2.12)$$

Através de (2.12), podemos encontrar a função densidade utilizando a fórmula dada em (1.3).

$$f(t_i|t_{i-1}) = e^{\alpha + \beta t + \gamma(i-1)} e^{\frac{-e^{(\alpha+(i-1)\gamma)}(e^{\beta t_i}) - e^{(\beta t_{i-1})}}{\beta}}$$

Sabendo que a função de verossimilhança é dada por (2.6), através da função densidade e sobrevivência temos:

$$L(\alpha, \beta, \gamma | \text{dados}) = e^{n\alpha} e^{\beta \sum_{i=1}^n t_i} e^{\gamma \sum_{i=1}^n (i-1)} e^{\frac{-e^\alpha}{\beta} \sum_{i=1}^n e^{(i-1)\gamma + \beta t_n}} e^{\frac{e^\alpha}{\beta} \sum_{i=1}^n e^{(i-1)\gamma + \beta t_{n-1}}} \quad (2.13)$$

Aplicando agora o *logaritmo* na função (2.13), temos:

$$\ln(L) = n\alpha + \beta \sum_{i=1}^n t_i + \gamma \sum_{i=1}^n (i-1) - \frac{e^\alpha}{\beta} \sum_{i=1}^n e^{(i-1)\gamma + \beta t_i} + \frac{e^\alpha}{\beta} \sum_{i=1}^n e^{(i-1)\gamma + \beta t_{i-1}} \quad (2.14)$$

Também encontramos as expressões da primeira derivada, desta forma elas são dadas a seguir.

$$\frac{\partial \ln(L)}{\partial \alpha} = n - \frac{e^\alpha}{\beta} \sum_{i=1}^n e^{(i-1)\gamma + \beta t_i} + \frac{e^\alpha}{\beta} \sum_{i=1}^n e^{(i-1)\gamma + \beta t_i}$$

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

$$\begin{aligned} \frac{\partial \ln(L)}{\partial \gamma} &= \sum_{i=1}^n (i-1) - \frac{e^\alpha \sum_{i=1}^n e^{(i-1)\gamma + \beta t_i} (i-1)}{\beta} \\ &\quad + \frac{e^\alpha \sum_{i=1}^n e^{(i-1)\gamma + \beta t_{i-1}} (i-1)}{\beta} \end{aligned}$$

Como o principal interesse é encontrar os estimadores de máxima verossimilhança, neste modelo não encontraremos problemas para estimá-los, uma vez que a integral envolvida foi de fácil resolução, diferentemente do modelo proposto por Lawless e Thiagarajah (1996). O encontro das estimativas é feito através do software R, através da rotina *optim*.

### 2.3 Exemplo (dados artificiais)

Para ilustrar o uso do modelo (2.10), e associá-lo com processo de inferência, nós consideramos o subconjunto de dados gerados pela inversa da função densidade condicional, sendo esta dada por:

$$f(t_i|t_{i-1}) = e^{\alpha + \beta t + \gamma(i-1) - \frac{e^{\alpha + (i-1)\gamma} (e^{\beta t_i} - e^{\beta t_{i-1}})}{\beta}}. \quad (2.15)$$

Considerando os parâmetros para geração do modelo sendo  $\alpha = 2$ ,  $\beta = 0.8$  e  $\gamma = -0.1$ , foi gerado um conjunto de dados correspondente à 55 recorrência de tal forma que arredondamos para apenas 2 casas decimais, assim os dados são encontrados na Tabela 2.1.

Tabela 2.1. Conjunto de dados simulado através da função dada em (2.15) (55 recorrências)

0.05	0.21	0.24	0.42	0.61	0.72	0.79	0.89	1.03	1.15	1.18
1.50	1.69	1.70	1.78	1.81	1.82	1.89	1.92	1.99	2.00	2.05
2.42	2.51	2.57	2.61	3.12	3.14	3.71	3.75	3.80	3.81	3.99
4.07	4.10	4.15	4.52	4.63	4.80	5.33	5.35	5.42	5.52	5.82
6.16	6.28	6.29	6.57	6.62	6.66	6.85	7.01	7.04	7.25	7.26

Com o conjunto de dados gerado, consideramos o modelo (2.5) e encontramos as estimativas dos parâmetros e a matriz de Informação Observada respectivamente, através da utilização do software R, assim temos:

$$\hat{\theta} = (2.4945709, 1.0818178, -0.1545671),$$

e

$$I_o(\hat{\theta}) = \begin{pmatrix} 55.00013 & 192.1218 & 1485.004 \\ 190.57057 & 948.4795 & 7240.210 \\ 1485.00438 & 7240.2103 & 55504.390 \end{pmatrix}$$

A estimativa da matriz de covariância  $I_o(\hat{\theta})^{-1}$  dá a variância padrão (tirando a raiz obtemos os desvios das estimativas padronizadas) para  $\hat{\alpha}$ ,  $\hat{\beta}$  e  $\hat{\gamma}$  são 0.077949475, 0.29474982 e 0.005512137 respectivamente.

A Figura 2.1 expõe o gráfico do tempo acumulativo de recorrência *versus* número acumulativo de recorrências no tempo em relação ao conjunto de dados gerados a linha contínua apresentada no gráfico é referente ao conjunto de dados obtidos através das estimativas dos parâmetros do modelo ajustado. Assim podemos verificar que os valores estimados são muito próximos dos valores originais (gerados inicialmente).

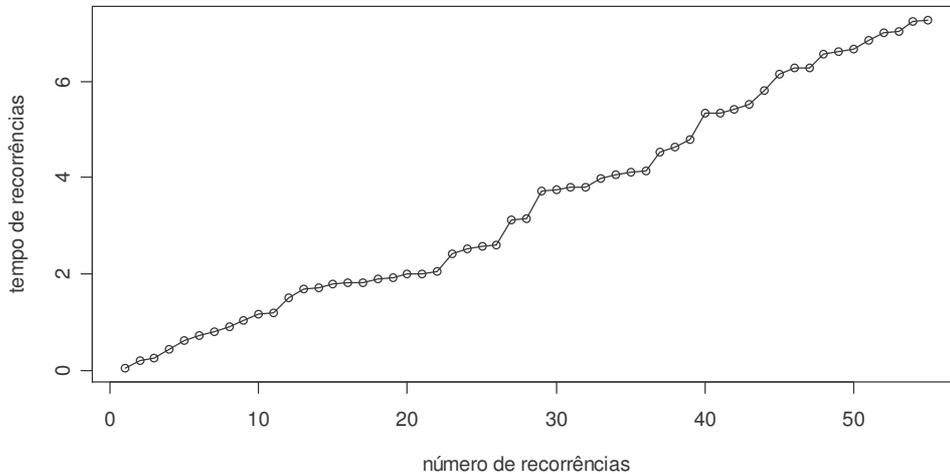


Figura 2.1. Tempo acumulado de recorrência versus número acumulado de recorrências no tempo

Apesar das estimativas terem sido verificadas, temos que analisar se os parâmetros são realmente significativos, desta forma verificaremos se existe ou não o comportamento de tendência, referente ao processo de Poisson e também se o número de recorrência influencia ou não este processo, ou seja, se podemos dizer que realmente temos um Processo de Poisson Perturbado. Então iremos trabalhar com o teste de hipótese de Wald.

A hipótese a ser testada inicialmente, será em relação ao comportamento de tendência, desta forma a hipótese nula é dada por  $H_0 : \beta = 0$ . O teste de Wald é utilizado para verificação de não rejeição ou rejeição da hipótese  $H_0$ . Esta estatística é baseada em  $W^2 = \hat{\beta}^2 / S^2(\hat{\beta})$ , em que  $S^2(\hat{\beta}) = (I_{3 \times 3}(\hat{\beta}))^{-1}$ , desta maneira a estatística  $W^2$  é comparada com uma distribuição qui-quadrado com nível de significância de 5%, assim  $\chi_{1,0.05}^2 = 3.841$ .

Realizando os cálculos, temos:

$$W_{\beta}^2 = 1.0818178^2 / 0.29474982 = 3.970587.$$

Através do valor obtido pela estatística de Wald em relação ao parâmetro  $\beta$ , podemos verificar que este valor é maior que 3.841, assim podemos rejeitar a hipótese nula, ou seja,

temos evidência que existe comportamento de tendência.

Agora analisando o parâmetro referente ao número de ocorrência, iremos testar a hipótese nula que é dada por  $H_0 : \gamma = 0$ , sendo a estatística de teste  $W^2 = \hat{\gamma}^2/S^2(\hat{\gamma})$ , em que  $S^2(\hat{\gamma}) = (I_{3 \times 3}(\hat{\gamma}))^{-1}$ , desta maneira a estatística  $W^2$  é comparada com uma distribuição  $\chi_{1,0.05}^2$ , tal que  $\chi_{1,0.05}^2 = 3.841$ .

Realizando os cálculos, temos:

$$W_\gamma^2 = (-0.1545671)^2/0.005512137 = 4.33425.$$

O valor obtido foi de 4.33425, que também é maior que  $\chi_{1,0.05}^2 = 3.841$ , desta maneira rejeitamos a hipótese nula, ou seja, estes dados tem influência em relação ao número de eventos que já ocorreram.

## 2.4 Conclusões e Observações

Modelos da forma (2.5) são úteis para explorar uma série de eventos, particularmente quando pode-se ter um comportamento de tendência relacionado com o número de eventos, gerando assim um Processo de Poisson Perturbado. Como ilustração, foi realizado um exemplo, no qual verificamos as estimativas dos parâmetros através do software R, verificamos se estas estimativas eram significantes ou não para o modelo.

Desta forma, concluímos através do exemplo que os dados comportam um Processo de Poisson Perturbado, ou seja, o tempo de recorrência depende do número de ocorrências.

Em suma, observamos que existem várias considerações a serem trabalhadas, então desta forma no Capítulo 3, será realizado um estudo mais aprofundado em relação ao modelo proposto, na tentativa de avaliar melhor as condições que este poderá ser utilizado.

# Capítulo 3

## Simulação

Neste capítulo, desenvolvemos análises em relação ao modelo proposto em (2.5), isto foi realizado em virtude de verificar as condições nas quais o modelo é viável ou não, assim analisaremos qual a consistência da função inversa apresentada em (2.15), de gerar dados de diferentes tamanhos e também serão realizados testes de hipóteses para validar o modelo.

Primeiramente utilizando novamente a inversa da função densidade condicional (2.15), foram gerados conjuntos de dados com 10, 25, 50, 100, 150 e 300 recorrência para realização da verificação da consistência dos estimadores e na seção 4.2 a consistência dos testes de hipóteses.

### 3.1 Consistência dos Estimadores

Um estimador é consistente quando seu valor se aproxima do verdadeiro valor do parâmetro à medida que aumenta-se o tamanho da amostra, Lehmann (1999).

Considerando a Tabela 3.1, apresentada a seguir, temos na primeira coluna os números de recorrências consideradas, na segunda coluna o valor das estimativas do parâmetro  $\alpha$ , na terceira as estimativas de  $\beta$  e na quarta as de  $\gamma$ .

Tabela 3.1. Estimativas dos Parâmetros, utilizando a função (2.15)

n	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$
10	0.9911541	29.0343849	-3.5022875
25	1.6173930	2.9772930	-0.2847253
50	2.4921083	0.9808284	-0.1418630
100	2.1268794	0.6932792	-0.0892514
150	1.94619481	0.69247182	-0.08491928
300	2.0000001	0.7999990	-0.1000242

Os valores dos parâmetros utilizados para encontrar as estimativas foram:  $\alpha = 2$ ,  $\beta = 0.8$  e  $\gamma = -0.1$ , sendo que a quantidade de recorrências consideradas foram 10, 25, 50, 100, 150 e 300.

O que podemos visualizar através da Tabela 3.1, é que as estimativas dos parâmetros não são interessantes para menos de 50 recorrências, todavia quando o número de recorrência aumenta as estimativas vão se aproximando do verdadeiro valor do parâmetro, desta forma podemos afirmar que os estimadores dos parâmetros estabelecidos são consistentes e desta forma podemos estar assegurados em relação a forma de geração da função inversa apresentada em (2.15).

## 3.2 Testes de Hipóteses

Fundamentado nos conceitos estatísticos, um dos problemas provenientes da modelagem é o de testar hipóteses, para verificação se os parâmetros envolvidos no mesmo são, ou não significativos. Então é realizado uma determinada afirmação sobre um modelo, usualmente sobre o parâmetro testado, e assim desejamos saber se os resultados experimentais provenientes de uma amostra contrariam ou não tal afirmação.

O objetivo do teste estatístico de hipótese é, então, fornecer uma metodologia que nos permita verificar se os dados amostrais (no caso do estudo em questão, dados gerados) trazem evidências que apóiam ou não uma hipótese estatística formulada.

### 3.2.1 Estatística de Wald

Como tratado anteriormente, o teste de hipótese é realizado em virtude de avaliar se a hipótese testada, no caso a hipótese nula, é ou não rejeitada. Todavia fundamentado neste conceito, realizamos o Teste de Wald que é dado por:

$$W_{\alpha} = \frac{\hat{\alpha}^2}{s^2(\hat{\alpha})} \quad (3.1)$$

quando o parâmetro analisado é dado por  $\alpha$ , que representa o risco base da função intensidade dada por (2.5), e

$$W_{\beta} = \frac{\hat{\gamma}^2}{s^2(\hat{\gamma})} \quad (3.2)$$

quando o parâmetro analisado é dado por  $\gamma$ , que caracteriza a influência da quantidade de eventos recorrentes no modelo.

Desta forma os testes de hipóteses foram fundamentados em  $H_0 : \alpha = 0$  e  $H_0 : \gamma = 0$ , utilizando um nível de significância de 5%, sendo que foram gerados valores de tal sorte que inicialmente o risco base era igual a zero ( $\alpha = 0$ ), para 10, 25, 50, 100, 150 e 300 recorrências em uma mesma unidade amostral, obtendo-se então a segunda coluna de valores da Tabela 4.2. Em seguida, realizamos o teste para o parâmetro  $\gamma=0$ , sendo este relacionado com a influência do número de ocorrência do evento, em que os valores obtidos se encontram na terceira coluna.

Para realização do teste, a estatística de Wald foi comparada com o valor  $\chi_{1,0.05}^2 = 3.841$ , assim verificamos a quantidade de vezes, em 999 reamostras, que o valor calculado através da simulação excedia o valor da qui-quadrado estabelecida. Através da Tabela 4.2, podemos verificar os resultados obtidos, que são apresentados em porcentagem.

Tabela 3.2. Valores em Porcentagem em relação ao número de vezes que a Estatística de Wald excedeu a  $\chi_{1,0.05}^2 = 3.841$

n	$W_{\alpha=0}$	$W_{\gamma=0}$
10	16,9	49,6
25	9,6	35,7
50	7,0	12,1
100	6,9	8,6
150	6,6	7,8
300	6,2	6,9

Analisando a Tabela 3.2, verificamos que para 10 ocorrências, considerando apenas o  $\alpha = 0$ , (risco base) 16,9% dos valores são maiores que  $\chi_{1,0.05}^2 = 3.841$ , sendo que o esperado seria apenas 5% (dado que o modelo de geração tem  $\alpha = 0$ ), isto também é verificado para o parâmetro  $\gamma$ , porém isto começa a melhorar a partir de 50 recorrências. Todavia os valores encontrados não satisfazem as suposições, já que o nível de significância adotado é de 5%.

Devido aos valores obtidos na simulação apresentada anteriormente, verificamos que a teoria assintótica não é conveniente, pois mesmo para grande reamostras não conseguimos obter os resultados necessários para justificá-la, em virtude deste fato o procedimento apresentado a seguir é referente à valores empíricos para a estatística de Wald, desconsiderando desta forma a distribuição  $\chi_{1,0.05}^2$ , que é referente a teoria assintótica.

Assim os valores apresentados na Tabela 3.3, referem-se à porcentagem dos números de estatísticas de teste, no caso a de Wald, que excederam o valor empírico. Desta forma, segue o procedimento utilizado para realização dos cálculos dos valores apresentados:

1. Geramos dados com  $\alpha = 0$ ,  $\beta = 0.8$  e  $\gamma = -0.1$ , e calculamos a Estatística de Wald Empírica;
2. Os valores estimados serão os parâmetros para nova geração de valores, porém  $\alpha = 0$ , calcula-se as estatística de Wald (Este processo repete-se 399 vezes);
3. As estatísticas obtidas são ordenadas e calcula-se o 95<sup>o</sup> percentil, e compara-se com a Estatística de Wald Empírica calculada no passo 1.

Os 3 passos são repetidos 499 vezes.

O mesmo procedimento é realizado para o parâmetro  $\gamma$ .

Tabela 3.3. Valores em Porcentagem em relação ao número de vezes que a Estatística de Wald excedeu a Estatística de Wald Empírica

n	$W_{\alpha=0}$	$W_{\gamma=0}$
10	7,36	22,72
25	6,89	23,08
50	5,02	10,69
100	4,98	5,09
150	4,99	4,53
300	5,01	4,26

Através da Tabela 3.3, observamos que para o parâmetro  $\alpha$  os valores obtidos são coerentes com o nível de significância de 5%, desta forma podemos dizer que existe uma confiabilidade de 95%. Agora realizando a mesma simulação, porém com os parâmetros  $\alpha = 2, \beta = 0.8$  e  $\gamma = 0$ , não obtivemos valores muito satisfatórios quando se trata de uma pequena quantidade de ocorrências, todavia isto se torna mais interessante após 50 recorrências, com uma significância empírica não maior que 10% e na medida que o número de recorrências aumentam no caso 100, 150 e 300 podemos afirmar que a confiabilidade de 95% foi atendida. Os códigos utilizados nesta simulação encontram-se no Apêndice.

### 3.2.2 Estatística de Razão de Verossimilhança

Outro teste de hipótese também considerado foi o da Razão de Verossimilhança ( $RV$ ), sendo este baseado na comparação entre os máximos com restrição e sem restrição do logaritmo da função de verossimilhança e tem uma distribuição assintótica qui-quadrado com  $k$  graus de liberdade, tal que  $k$  será considerado igual a 1 e o nível de significância igual a 5%, assim  $\chi_{1,0.05}^2 = 3.841$ .

A estatística de razão de verossimilhança é dada por:

$$RV = 2l(\hat{\theta}) - 2l(\tilde{\theta}) \quad (3.3)$$

em que  $\hat{l}(\theta)$  é dado pelo logaritmo da verossimilhança calculado com todos os parâmetros e  $l(\theta)$  o valor do logaritmo da verossimilhança considerando um dos parâmetros iguais a 0, isto é, o teste da razão de verossimilhança é obtido através da comparação entre os modelos com e sem os parâmetros sendo testadas.

Como segue na Tabela 3.4, consideramos primeiramente a hipótese nula  $H_0 : \alpha = 0$  e desta forma geramos os dados em relação ao tempo de recorrência, tal que o número de recorrências foram: 10, 25, 50, 100, 150 e 300, os valores dos parâmetros considerados foram:  $\alpha = 0, \beta = 0.8, \gamma = -0.1$  e em cada situação calculamos a estatística de  $RV$  999 vezes, e em seguida comparamos-as com a  $\chi_{1,0.05}^2 = 3.841$ . Este procedimento foi realizado tanto para  $\alpha$ , quanto para  $\gamma$  ( $H_0 : \gamma = 0$ ), porém aqui utilizamos os parâmetros de acordo com o teste,  $\alpha = 2, \beta = 0.8, \gamma = 0$ , e os resultados referente ao número de vezes que os valores da  $RV$  excederam o valor de 3.841 considerado, encontram-se em porcentagem na Tabela 3.4.

Tabela 3.4. Valores em Porcentagem em relação ao número de vezes que a Estatística de  $RV$  excedeu a  $\chi_{1,0.05}^2 = 3.841$

n	$W_{\alpha=0}$	$W_{\gamma=0}$
10	63,1	78,8
25	48,3	62,8
50	42,7	41,2
100	40,9	35,4
150	28,5	28,0
300	11,4	16,3

Os valores apresentados na segunda coluna foram calculados levando em conta  $\alpha = 0, \beta = 0.8, \gamma = -0.1$ , e através da função de verossimilhança, calculamos  $\hat{l}(\theta)$  o valor da log-verossimilhança para esta situação, em seguida estimamos os valores dos parâmetros e assim obtemos o valor de  $l(\theta)$ , e desta maneira realizamos os cálculos da Razão de Verossimilhança (3.3), isto foi realizado 999 vezes como já foi dito anteriormente. Portanto, através dos valores apresentados pela Tabela 4.4, verificamos que apesar do valor de geração do parâmetro  $\alpha$  ser igual 0, os valores das estatísticas excederam o valor da  $\chi_{1,0.05}^2 = 3.841$  em 63,1% das vezes quando o número de ocorrências é 10, e na medida

que o número de ocorrências foi aumentando este número vai diminuindo tanto para o parâmetro  $\alpha$ , quanto para o parâmetro  $\gamma$ , porém estes valores não se mostraram satisfatórios em relação ao nível de significância esperado.

Em virtude dos valores encontrados não serem aceitáveis em relação à teoria assintótica, realizamos a simulação considerando o valor empírico da função de Razão de Verossimilhança, assim como já foi feito para a estatística de Wald. A seguir, segue o procedimento utilizado para realização dos cálculos dos valores apresentados na Tabela 3.5:

1. Primeiramente consideramos  $\alpha = 0$ ,  $\beta = 0.8$ ,  $\gamma = -0.1$  e desta forma geramos os dados, referentes ao tempo de recorrência dos eventos;
2. Em seguida, as estimativas dos parâmetros, de acordo com os dados gerados em (1) foram encontradas;
3. Calculamos então, a estatística de Razão de Verossimilhança Empírica, dada por (3.3);
4. Agora, considerando os parâmetros estimados anteriormente, porém  $\alpha = 0$ , geramos novos dados;
5. Através dos dados gerados, calculamos as estimativas dos parâmetros e desta forma calculamos a estatística de  $RV$  novamente;
6. Os procedimentos quatro e cinco são realizados 499 vezes, ou seja, teremos 499 estatísticas de  $RV$ , todavia apenas o valor que representa o 95º percentil é utilizado para ser comparado com o valor da estatística de  $RV$  empírica, obtida em (3);

Todos estes procedimentos citados anteriormente foram realizados 399, e assim conseguimos os valores que se encontram em porcentagem na Tabela 3.5.

Os passos apresentados referem-se ao parâmetro  $\alpha$ , sendo este o parâmetro que retrata o risco base da função intensidade apresentada em (2.5). Para o parâmetro  $\gamma$ , a simulação foi realizada de forma similar, todavia nos passos (1) e (4),  $\gamma = 0$ , os valores que iremos apresentar na Tabela 3.5, referem-se ao número de vezes em porcentagem que a  $RV$  do 95º percentil foi maior do que  $RV$  empírica.

Tabela 3.5. Valores em Porcentagem em relação ao número de vezes que a Estatística de RV excedeu a Estatística de RV Empírica

n	$W_{\alpha=0}$	$W_{\gamma=0}$
10	18,79	12,28
25	12,03	15,04
50	7,69	8,96
100	5,96	7,26
150	5,63	6,02
300	4,33	5,16

Através da Tabela 4.5, verificamos que à medida que o número de recorrência em uma mesma unidade amostral aumenta, a porcentagem dos valores que excedem o valor empírico diminui, aproximando-se do valor de significância de 5%, porém o valor nominal não foi atingido de forma precisa, mas podemos considerar que o modelo proposto é interessante para um grande número de recorrências

Ainda podemos considerar que acima de 50 recorrências a significância empírica não é maior que 10% tanto para o parâmetro  $\alpha$ , quanto para o parâmetro  $\gamma$ . Os códigos utilizados nesta simulação encontram-se no Apêndice.

### 3.3 QQ- Plot

Podemos utilizar recursos computacionais e em particular simulações para inferir distribuições amostrais de quantidades de interesse. Na teoria de estatística existem vários resultados que podem ser ilustrados via simulação, o que ajuda na compreensão e visualização dos conceitos e resultados, como já vimos na seção apresentada anteriormente.

Este uso de simulações é apenas um ponto de partida, pois estas são especialmente úteis para explorar situações onde resultados teóricos não são conhecidos ou não podem ser obtidos.

Uma forma eficaz de comparar distribuições empíricas e teóricas é comparar os quantis das distribuições e para isto utilizamos o qq-plot. O qq-plot é um gráfico dos dados ordenados contra os quantis esperados de uma certa distribuição. Quanto mais próximo

os pontos estiverem da bissetriz do primeiro quadrante mais próximos os dados observados estão da distribuição considerada.

A Figura 3.6 ilustra os qq-plots da Estatística de Wald de acordo com o número de observações, isto é, de acordo com o número de recorrências. A estatística de Wald foi escolhida em virtude de seus resultados referentes a simulação apresentada na seção anterior.

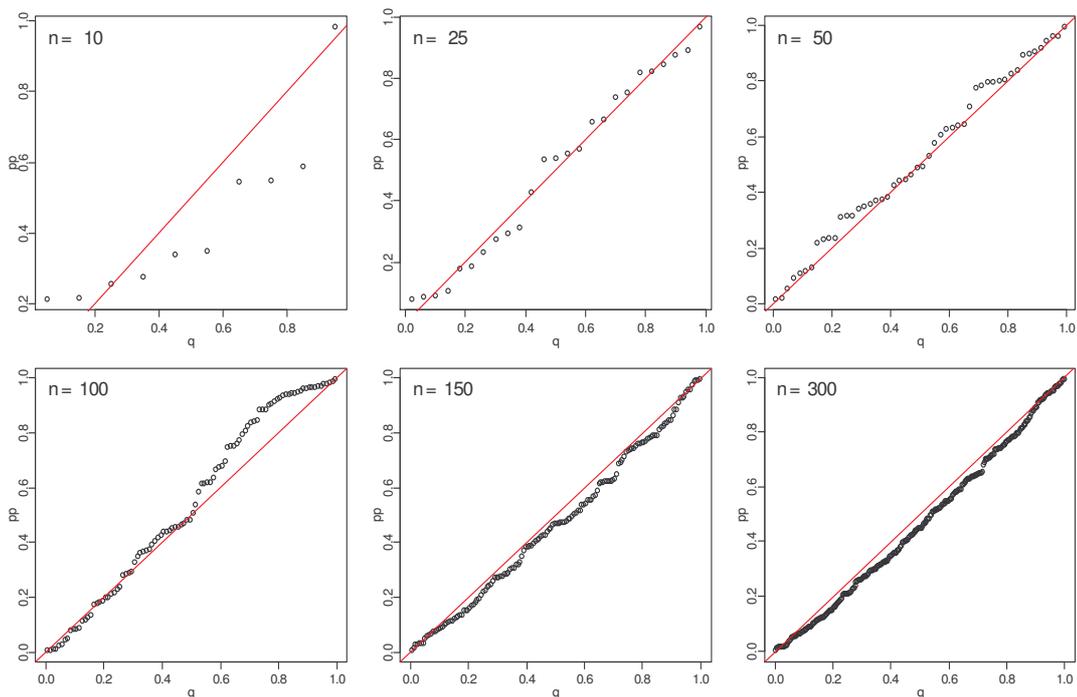


Figura 3.6. QQ-Plots em relação ao número de recorrência considerado

O que podemos concluir é que quanto maior o número de recorrências, mais os pontos se aproximam da bissetriz do primeiro quadrante, desta forma podemos dizer que mais próximos os dados observados estão da distribuição considerada, considerando a estatística de Wald.

### 3.4 Conclusões e Observações

O modelo proposto sugere a análise de tempos de recorrências de uma determinada variável de interesse em uma mesma unidade amostral, focando a importância da quan-

tidade de eventos que podem estar ocorrendo nesta unidade, referindo-se a um Processo de Poisson Perturbado, ou seja, uma componente representando o Processo de Poisson e a outra o Número Eventos ocorridos nesta unidade.

Como podemos verificar através dos testes de hipóteses a significância nominal foi atingida, entretanto acima de 50 recorrências a significância empírica se torna mais interessante, porém não maior que 5%, considerando o teste de Wald e não maior que 10% considerando a estatística de Razão de Verossimilhança. Em relação à consistência dos estimadores, verificamos que quanto maior o número de recorrência as estimativas se aproximam cada vez mais do verdadeiro valor do parâmetro e através dos qq-plots verificamos também que os dados gerados aproximam-se da distribuição, no caso do modelo considerado, levando em consideração a estatística de Wald.

Um exemplo com dados reais será muito interessante para validar o modelo e será realizado no capítulo 4.

# Capítulo 4

## Exemplo- Dados Reais

Devido ao modelo proposto requerer tempos de recorrência de um mesmo indivíduo ou de um mesmo componente, neste capítulo iremos apresentar um exemplo com dados reais com as condições estabelecidas. A necessidade de exemplificar o modelo híbrido proposto é de fundamental importância para validar o mesmo, em situações problemas nas quais ele poderá ser utilizado.

O exemplo apresentado refere-se a um indivíduo que faz compras de uma revendedora da empresa de cosméticos, sendo que o evento de interesse é referente ao momento no qual o indivíduo adquire produtos dessa revendedora, representado pela variável tempo. A coleta de dados foi possível, graças a um conjunto de dados de sete anos cedido por uma consultora, tendo este mais de 50 recorrências.

A procura por um conjunto de dados com mais de 50 observações se fez necessária, devido às simulações e resultados obtidos no capítulo 3, sendo que estes apontavam que o modelo é adequado para um grande número de recorrências.

O objetivo deste modelo na situação problema envolvida relaciona-se com a verificação da influência do número de compras anteriores no tempo das compras futuras, desta forma o conjunto de dados encontrados encontra-se na Tabela 4.1.

Tabela 4.1. Conjunto de dados reais, valores em dias (57 recorrências)

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									

O conjunto de dados apresentado na Tabela 4.1, refere-se ao tempo em dias de uma nova compra feita pela cliente considerada, todavia o modelo proposto trabalha com dados representados em números decimais, então decidimos trabalhar com os dados não em dias, mas sim em anos. Considerando o que foi dito até o momento o conjunto de dados da Tabela 4.1, foi dividido por 360 que representa o ano comercial.

Tabela 4.2. Conjunto de dados reais, valores em anos (57 recorrências)

0.04	0.21	0.24	0.41	0.59	0.70	0.76	0.86	0.99	1.10	1.13
1.42	1.60	1.61	1.69	1.71	1.72	1.78	1.81	1.88	1.89	1.93
2.26	2.34	2.40	2.43	2.89	2.91	3.42	3.46	3.50	3.51	3.67
3.74	3.77	3.81	4.14	4.24	4.40	4.87	4.88	4.95	5.04	5.30
5.60	5.71	5.72	5.97	6.02	6.05	6.22	6.36	6.39	6.58	6.59
6.69	6.75									

Através da Tabela 4.2, podemos observar que os dados coletados referem-se a um período de 7 anos, como já tratado previamente, assim sendo, verificamos que esta cliente efetua compras e é fiel a esta revendedora no período de 2001 até o ano de 2007, considerado.

## 4.1 Estimativas

Considerando agora o conjunto de dados já em decimais apresentado na Tabela 4.2, podemos estimar os parâmetros utilizando o modelo proposto em (2.5), tal que os códigos para realização deste encontram-se no Apêndice referente ao capítulo 4.

As estimativas dos parâmetros são  $\hat{\theta} = (2.47004, 1.315204, -0.1659582)$ , tal que a matriz de Informação observada de Fisher é dada por:

$$I_o(\hat{\theta}) = \begin{pmatrix} 56.98301 & 188.6643 & 1596.459 \\ 188.66434 & 878.7202 & 7332.701 \\ 1596.45930 & 7332.7011 & 61380.727 \end{pmatrix}$$

A estimativa da matriz de covariância  $I_o(\hat{\theta})^{-1}$  dá a variância padrão (tirando a raiz obtemos os desvios das estimativas padronizadas) para  $\hat{\alpha}$ ,  $\hat{\beta}$  e  $\hat{\gamma}$  são 0.07183600, 0.40583449 e 0.006191541 respectivamente e a função de máxima verossimilhança  $L(\hat{\theta}) = 67.05319$ .

Apesar das estimativas terem sido verificadas, temos que analisar se os parâmetros são realmente significativos para validar o modelo, desta forma verificaremos se existe ou não o comportamento de tendência, referente ao processo de Poisson e também se o número de recorrência influencia ou não este processo, ou seja, se podemos dizer que realmente temos um Processo de Poisson Perturbado. O teste de hipótese de Wald será utilizado.

A hipótese a ser testada inicialmente, será em relação ao comportamento de tendência, desta forma a hipótese nula é dada por  $H_0 : \beta = 0$ . O teste de Wald é utilizado para verificação de não rejeição ou rejeição da hipótese  $H_0$ . Esta estatística é baseada em  $W_{\beta}^2 = \hat{\beta}^2 / S^2(\hat{\beta})$ , em que  $S^2(\hat{\beta}) = (I_{3 \times 3}(\hat{\beta}))^{-1}$ , desta maneira a estatística  $W_{\beta}^2$  é comparada com uma distribuição qui-quadrado com nível de significância de 5%, assim  $\chi_{1,0.05}^2 = 3.841$ .

Realizando os cálculos, temos:

$$W_{\beta}^2 = 2.47004^2 / 0.40583449 = 15.0335.$$

Através do valor obtido pela estatística de Wald em relação ao parâmetro  $\beta$ , podemos verificar que este valor é maior que 3.841, assim podemos rejeitar a hipótese nula, ou seja, temos evidência que existe o comportamento de tendência.

Agora analisando o parâmetro referente ao número de ocorrência, iremos testar a hipótese nula que é dada por  $H_0 : \gamma = 0$ , sendo a estatística de teste  $W_\gamma^2 = \hat{\gamma}^2 / S^2(\hat{\gamma})$ , em que  $S^2(\hat{\gamma}) = (I_{3 \times 3}(\hat{\gamma}))^{-1}$ , desta maneira a estatística  $W^2$  é comparada com uma distribuição  $\chi_{1,0.05}^2$ , tal que  $\chi_{1,0.05}^2 = 3.841$ .

Realizando os cálculos, temos:

$$W_\gamma^2 = (-0.1659582)^2 / 0.006191541 = 4.4483.$$

O valor obtido foi de 4.4483, que também é maior que  $\chi_{1,0.05}^2 = 3.841$ , desta maneira rejeitamos a hipótese nula, ou seja, estes dados tem influência em relação ao número de eventos que já ocorreram.

Após a verificação da significância dos parâmetros, iremos verificar os tempos estimados de acordo com os valores dos parâmetros estimados, desta maneira a Figura 4.1, apresenta através dos pontos os verdadeiros valores, ou seja o tempo de recorrência das compras, enquanto que a linha contínua apresenta os valores dos tempos de compras estimados.

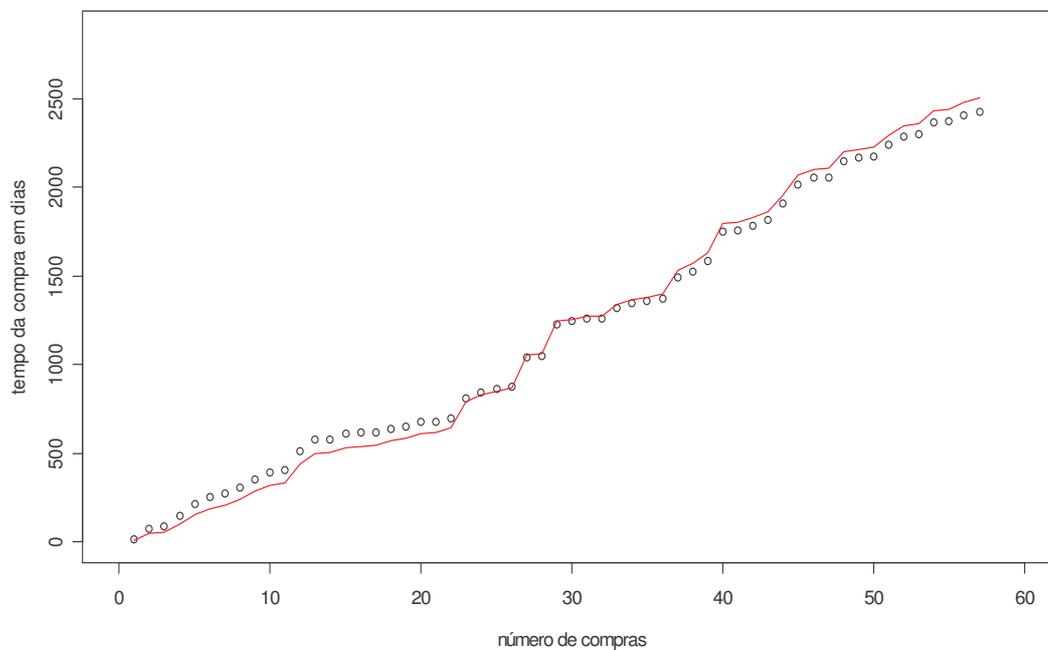


Figura 4.1. Gráfico das estimativas dos tempos de vendas em relação ao número de vendas

O que podemos observar é que os tempos estimados de compra são muito próximos dos tempos reais, desta forma podemos dizer que o modelo é adequado.

Este modelo se faz interessante devido estarmos trabalhando não apenas com a variável tempo, mas também com a variável número de eventos e como já vimos no capítulo 4, este modelo se faz interessante a partir de 50 recorrências, não obstante podemos utilizá-los em várias situações.

## 4.2 Conclusões e Observações

O que podemos concluir em relação a todo o trabalho apresentado, é que o modelo proposto em (2.5) é adequado quanto estamos trabalhando com um número superior a 50 recorrência, desta forma iremos verificar um processo de Poisson Perturbando, em que o número de ocorrência influenciam no tempo das recorrências.

O exemplo com dados reais, ajudou-nos a verificar melhor as situações nas quais o modelo proposto poderá ser utilizado, e como pudemos visualizar através da Figura 4.1, os tempos estimados encontram-se muito próximos dos tempos observados, portanto possíveis previsões para verificar os futuros tempos de compras são interessantes.

Contudo, verificamos também os possíveis valores nos quais os parâmetros devem assumir para que os valores gerados sejam significativas, assim analisando o parâmetro  $\alpha$ , referente ao risco base da função dada por (2.5), caso ele tenha valores negativos o primeiro valor de geração aumento, e se o valor dele for positivo, diminui o primeiro valor, desta forma verificamos que o parâmetro  $\alpha$  é responsável pelo início dos tempos de recorrência dos dados. O parâmetro  $\beta$ , também foi analisado, e assim observamos que os valores atribuídos a ele não poderão ser negativos, todavia se os valores estiverem entre 0 e 1, verificamos o aumento da amplitude dos dados, e se ele for maior do que 1, existe uma diminuição da amplitude.

O parâmetro  $\gamma$ , foi o mais interessante de ser analisado, já que ele pode assumir valores positivos e negativos, porém sua modificação em relação a décimos modifica demais a estrutura da geração, sendo que se ele for negativo aumenta muito a amplitude e se ele for positivo ele diminui demais a amplitude. Após esta análise fica fácil visualizar possíveis conjuntos de dados que serão aptos a se adequar ao modelo proposto, não esquecendo

porém que estes devem estar na estrutura de decimais.

Portanto fica a proposta do modelo de intensidade Poisson Perturbada pelo Número de Ocorrências para dados de eventos Recorrentes com suas validações em caráter clássico.

# Apêndice A

## Processos de Contagem

O desenvolvimento de modelos estatísticos baseados em processos de contagem para analisar dados de eventos recorrentes foi originalmente introduzido por Aalen (1978). Esse trabalho de Aalen serviu de inspiração a vários modelos de sobrevivência definidos a partir de intensidades de processo de contagem, quantificando ocorrências de um evento de interesse ao longo do tempo. O caráter temporal desses modelos traduz a característica dinâmica das unidades em análise de sobrevivência que se movem entre um número finito de estados. Por exemplo, mudança de um indivíduo do estado vivo para o estado morto, que é uma situação usual neste tipo de dados.

Existe uma extensa literatura sobre modelos de processo de contagem. Esta abordagem nos proporciona ferramentas bastante poderosas, capazes de generalizar diversas situações. Os modelos freqüentemente utilizados para dados de eventos recorrentes, que proporcionam o aprendizado sobre o processo individual, são aqueles baseados no processo de Poisson e processo de renovação. Esses modelos são conhecidos como modelos de intensidade. Cox e Isham (1980) e Cox e Lewis (1966) são alguns exemplos para situações nas quais covariáveis não estão presentes. Covariáveis são, entretanto, tratadas por Andersen *et al.* (1993), Follmann & Goldberg (1988), Prentice *et al.* (1981) e Cox (1972b), entre outros.

Com relação à interpretação dos modelos, pode ser dito que, se for empregado uma representação de Poisson para os processos individuais, o tempo total de estudo está sendo modelado, enquanto que, considerando o processo de renovação, é o tempo intervalar, ou seja, o tempo entre eventos que está sendo modelado. Como discutido por Lipschutz e

Snapinn (1997), modelar o tempo total parece natural se os eventos podem ser vistos como independentes; também é razoável assumir que eles podem se desenvolver simultaneamente, com o risco para a ocorrência de cada evento iniciando no mesmo tempo. Nesse contexto, os eventos ordenados refletem a ordem de suas ocorrências no tempo. Por outro lado, modelando o intervalo de tempo, considera-se situações em que o risco para o próximo evento inicia depois de o evento prévio ter ocorrido.

Alguns conceitos de processos de contagem aplicados à análise de sobrevivência são relatados neste apêndice.

## A.1 Conceitos Básicos

Um processo estocástico  $\{N(t), t \geq 0\}$  é dito um processo de contagem se  $N(t)$  representa o número total de “eventos” que ocorreram até o tempo  $t$ . Propriedades importantes são dadas a seguir:

1.  $N(t) \geq 0$ .
2.  $N(t)$  é um valor inteiro.
3. Se  $s < t$  então  $N(s) \leq N(t)$ .
4. Se  $s < t$  então  $N(t) - N(s)$  representa o número de eventos que ocorreram no intervalo  $(s, t]$

Alguns exemplos que envolvem o processo de contagem são apresentados.

(a) Se considerarmos  $N(t)$  o número de pessoas que entraram em um shopping até o tempo  $t$ , então  $\{N(t), t \geq 0\}$  é um processo de contagem, no qual um evento corresponde a entrada de uma pessoa no shopping.

(b) Se considerarmos que um evento ocorre sempre que uma criança nasce, então  $\{N(t), t \geq 0\}$  é um processo de contagem, em que  $N(t)$  é o número total de pessoas que nasceram até o tempo  $t$ .

(c) Seja  $N(t)$  o número de ligações telefônicas que chegam a uma central telefônica até o tempo  $t$ , então  $\{N(t), t \geq 0\}$  é um processo de contagem.

Dizemos que um processo de contagem possui incrementos independentes se as variáveis aleatórias que representam o número de eventos que ocorrem em intervalos disjuntos de tempo são independentes. Isto é, para quaisquer  $t_0 < t_1 < \dots < t_n$ , teremos

que as variáveis aleatórias  $N(t_0), N(t_1) - N(t_0), N(t_2) - N(t_1), \dots, N(t_n) - N(t_{n-1})$  são independentes. Intuitivamente, podemos dizer que num processo com incrementos independentes, os “acontecimentos” do passado não influenciam o futuro. É razoável supor que os exemplos (a) e (c) possuem esta característica, no entanto o mesmo não ocorre em (b).

Dizemos também que um processo de contagem possui incrementos estacionários se a distribuição de qualquer variável aleatória que representa o número de eventos que ocorrem em um intervalo de tempo depende somente do comprimento do intervalo. Isto é, para todo  $t$ , se  $h > 0$ , temos que  $P\{N(t+h) - N(t) = n\} = P\{N(s+h) - N(s) = n\}$ . No exemplo (c) esta hipótese não é razoável, uma vez que a ocorrência de ligações entre 1 e 2 da tarde é muito mais provável que a ocorrência de ligações entre 1 e 2 da manhã.

## A.2 Processo de Poisson

Um dos mais importantes tipos de processo de contagem é o processo de Poisson, que é tratado em detalhes na maioria dos livros de processos estocásticos (ver Karlin & Taylor, 1975). Um processo de contagem  $\{N(t), t \geq 0\}$  é um processo de Poisson com taxa  $\lambda > 0$ , se:

- i)  $N(0) = 0$ .
- ii) O processo tem incrementos independentes.
- iii) Para todo  $s, t \geq 0$ ,

$$P\{N(t+s) - N(s) = n\} = \frac{e^{-\varphi t} (\varphi t)^n}{n!}, n = 0, 1, \dots$$

Pelo item (iii), podemos concluir que  $N(t+s) - N(s)$  é uma variável aleatória que tem distribuição de Poisson com taxa  $\varphi t$  e o processo tem incrementos estacionários, pois a probabilidade é função apenas de  $t$  (comprimento do intervalo). Disto, sabemos que a taxa  $\varphi$  do processo representa o número médio de chegadas em qualquer intervalo de tempo de comprimento unitário.

Para determinar se um processo de contagem é um processo de Poisson, devemos verificar as condições (i), (ii) e (iii). As condições (i) e (ii), em geral, seguem do nosso conhecimento prévio sobre o processo, mas a condição (iii) não parece de fácil verifi-

cação. Por isso, apresentaremos uma outra definição, que é equivalente a esta. Antes, precisaremos definir o seguinte conceito:

Uma função  $f(\cdot)$  é dita ser  $o(h)$  se converge “mais rápido” para zero que a função  $h$ , quando  $h \rightarrow 0$ . Alguns exemplos de funções deste tipo são  $x^2$  e  $x^3$ .

$$o(h) = \lim_{h \rightarrow 0} \frac{f(h)}{h} = 0.$$

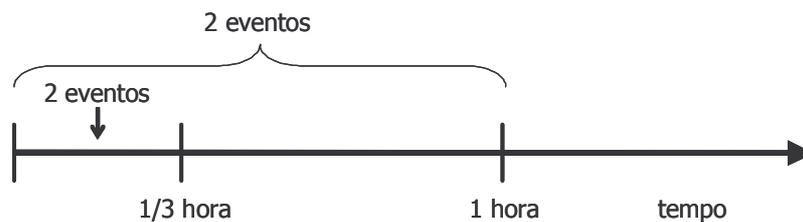
Definição: Um processo de contagem  $\{N(t), t \geq 0\}$  é um processo de Poisson com taxa  $\varphi > 0$ , se:

- i)  $N(0) = 0$
- ii) O processo tem incrementos independentes e estacionários
- iii)  $P\{N(h) = 1\} = \varphi h + o(h)$
- iv)  $P\{N(h) \geq 2\} = o(h)$

Os itens (i) e (ii) da definição, em geral, seguem da natureza do processo. O item (iii) significa que a probabilidade de ocorrer um evento em um intervalo de tamanho  $h$  ( $h$  pequeno) é aproximadamente  $\lambda h$ . O item (iv) garante que dois eventos não ocorrem simultaneamente no mesmo intervalo.

O exemplo a seguir ilustra melhor a idéia.

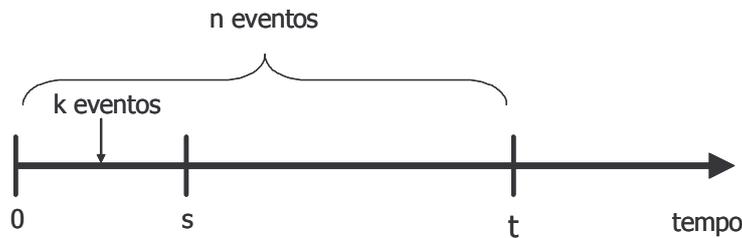
Clientes chegam a um banco segundo um processo de Poisson com taxa  $\lambda$ . Suponha que dois clientes chegam na primeira hora. Qual a probabilidade que ambos tenham chegado durante os 20 primeiros minutos?



$$\begin{aligned}
P\{N(\frac{1}{3}) = 2 | N(1) = 2\} &= \frac{P\{N(\frac{1}{3})=2, N(1)=2\}}{P\{N(1)=2\}} = \\
\frac{P\{N(\frac{1}{3})=2, N(1)-N(\frac{1}{3})=0\}}{P\{N(1)=2\}} &= \frac{P\{N(\frac{1}{3})=2\}P\{N(1)-N(\frac{1}{3})=0\}}{P\{N(1)=2\}} = \\
= \frac{(\frac{1}{3}\lambda)^2 e^{-\frac{1}{3}\lambda} (\frac{2}{3}\lambda)^0 e^{\frac{2}{3}\lambda}}{\frac{2!}{(\lambda)^2 e^{-\lambda}}} &= \frac{1}{9}
\end{aligned}$$

Generalizando para  $s < t$ , temos que  $N(s) | N(t) = n \sim binomial(n, s/t)$ , isto é:

$$P\{N(s) = k | N(t) = n\} = \binom{n}{k} \left(\frac{s}{t}\right)^k \left(1 - \frac{s}{t}\right)^{n-k}$$



O Processo de Poisson é uma abordagem freqüentemente utilizada para modelar eventos recorrentes, que permite conhecer o processo individual. O desenvolvimento de modelos estatísticos baseados no processo de contagem para analisar dados de eventos recorrentes foi originalmente introduzido por Aalen (1978).

Lawless & Nadeau (1995) utilizaram o processo de Poisson para desenvolver modelos que enfocam o número esperado de eventos ocorridos em um determinado intervalo de tempo, desta forma o interesse pode ser caracterizar o processo de ocorrência de eventos para um indivíduo ou verificar a comparação de tratamentos com base no tempo de cada evento distinto, no número de evento, no tipo de evento e na interdependência entre eles, pelo fato deste ser utilizado quando modelamos o tempo total em estudo.

Como discutido por Lipschutz e Snapinn (1997), modelar tempo total parece natural se os eventos, em algumas situações especiais, podem ser vistos como independentes, assim é razoável assumir que eles podem ser desenvolvidos simultaneamente, com o risco para a ocorrência de cada evento começando ao mesmo tempo. Neste contexto, os eventos

ordenados simplesmente refletem a ordem de suas ocorrências no tempo. Por exemplo, pacientes com síndrome de imunodeficiência humana podem experimentar a recorrência oportunista de infecções. Eles podem estar em risco para um tipo de infecção, tal como candidíase superficial, em várias localizações no começo do estudo, com o primeiro e segundo eventos subseqüentes refletindo a ordem de suas ocorrências em tempo progressivo.

Observe que a base para modelar dados de eventos recorrentes, com suposição de que o risco de experimentar um evento é independente do estado corrente é claramente um modelo de Markov. A distribuição do número e eventos em um intervalo  $(s, t]$  é a distribuição de Poisson com média  $\varphi$ , por exemplo.

Considerando situações em que  $n$  indivíduos são observados, tendo cada um deles  $m_i$  eventos ( $i = 1, \dots, n$  e  $j = 1, \dots, m_i$ ) e  $m_i$  representa a ocorrência de um determinado evento, assim os dados do  $i$ -ésimo indivíduo consistem no número total  $m_i$  de eventos (tempo de vida) observados por um período de tempo  $(0, T_i]$  e a época ordenada dos  $m_i$ , isto é, os tempos de vida no tempo,  $0 \leq t_{i1} < t_{i2} < \dots < t_{im_i} \leq T_i$ .

Baseados nos modelos do processo de Poisson um indivíduo com um vetor de covariável  $\mathbf{z}$  tem ocorrência de eventos com função intensidade definida como segue:

$$f(t, \varphi) = \varphi(t) \exp(\mathbf{z}'\beta).$$

em que  $\varphi(t)$  é a taxa de falha e  $\mathbf{z}'\beta$  é a incorporação de uma regressão, tal que os  $\beta$  sejam os parâmetros e  $\mathbf{z}$  as covariáveis.

### A.3 Processo de Renovação Ordinário

O processo de Renovação foi desenvolvido para modelar tempo entre eventos, ver Cox (1962). Suponha que temos uma população de componentes com os seus respectivos tempos de falhas,  $X$ , tal que esta é uma variável aleatória contínua com f.d.p.  $f(x)$ . É interessante notar que a falha do componente pode ser dada por várias interpretações.

Vamos supor que analisaremos um componente novo no tempo 0. Este componente falha no tempo  $X_1$ , ele é repostado imediatamente por um componente novo com tempo de falha  $X_2$ , assim a segunda falha ocorrerá depois de um tempo total de  $X_1 + X_2$ .

Este processo é contínuo, ou seja, um componente é repostado imediatamente após a

falha, por um componente novo.

O tempo do  $r$ -ésimo componente utilizado é  $X_r$ , e a  $r$ -ésima ocorrência de falha é dada pelo tempo  $S_r$ , em que

$$S_r = X_1 + \dots + X_r. \quad (\text{A.1})$$

Se  $\{X_1, X_2, \dots\}$  são variáveis aleatórias independentes e identicamente distribuídas, todas com f.d.p.  $f(x)$ , nós chamamos o sistema de um processo de renovação (ordinário).

Se o tempo de falha seguir uma distribuição exponencial com f.d.p.  $\varphi e^{-\varphi x}$ , nós chamamos o processo de renovação ordinário de processo de Poisson com taxa  $\varphi$  constante.

Nós agora definiremos algumas situações utilizadas para um processo de renovação ordinário.

i) Para alguns valores fixados de  $r$  nós podemos desejar saber o tempo em que a  $r$ -ésima renovação ocorrerá, isto é, precisamente a variável aleatória  $S_r$ , definida em (A.1).

ii) Fixado um tempo  $t$ , definimos a variável aleatória  $N(t)$ , como o número de renovações ocorridas em  $(0, t)$ , o componente novo que começa a funcionar no tempo  $t = 0$  não deve ser contado como uma renovação. A variável aleatória  $N(t)$ , é frequentemente a variável no processo de mais interesse e geralmente podemos selecionar dois tempos fixos distintos para calcularmos o número médio de renovações, como por exemplo,  $t_1, t_2$  ( $t_1 < t_2$ ) e definir

$$N(t_1, t_2) = N(t_2) - N(t_1), \quad (\text{A.2})$$

como sendo o número de renovações em  $(t_1, t_2)$ .

Podemos verificar que, formalmente, a propriedade de  $N(t_1, t_2)$  pode ser obtida do mesmo modo que  $N(t)$ .

iii) Nosso interesse pode estar em encontrar o valor médio de  $N(t)$ . Desta forma a função renovação é expressa por

$$P(t) = E(N(t)). \quad (\text{A.3})$$

Note que, para a variável aleatória  $N(t_1, t_2)$  nós temos de (A.2) que

$$E(N(t_1, t_2)) = P(t_2) - P(t_1). \quad (\text{A.4})$$

iv) Considere para algum tempo  $t$ , a função  $p(t)$ , definida por

$$p(t) = \lim_{\Delta t \rightarrow 0^+} \frac{E(N(t, t + \Delta t))}{\Delta t} = P'(t). \quad (\text{A.5})$$

esta função é chamada de renovação densidade e especifica o número médio de renovações esperado em um intervalo próximo de  $t$ . Dado as variáveis aleatórias  $\{X_r\}$  serem contínuas, sem concentração de probabilidade no tempo de falha zero, a probabilidade de mais que uma falha no intervalo de tamanho  $\Delta t$  é dado por  $o(\Delta t)^2$ . Segue que

$$p(t) = \lim_{\Delta t \rightarrow 0^+} \frac{\text{prob}\{\text{uma, ou mais ,renovação em } (t, t + \Delta t)\}}{\Delta t}. \quad (\text{A.6})$$

Uma alternativa na interpretação de  $p(t)$  é que se temos um número muito grande  $m$  de processos de renovação independentes em operação simultaneamente,  $mp(t)$  é o número de renovações no intervalo de tempo  $(t, t + \Delta t)$ .

v) Dado  $t$  um tempo fixado, definimos a variável aleatória  $U_t$  como o comprimento do tempo medido antes de  $t$  (*backward*), até a última renovação ou até o tempo estabelecido  $t$ . Se não houver nenhuma renovação até  $t$ ,  $U_t$  é definida como  $t$ . Então  $U_t$  é a idade do componente em uso até o tempo  $t$ .

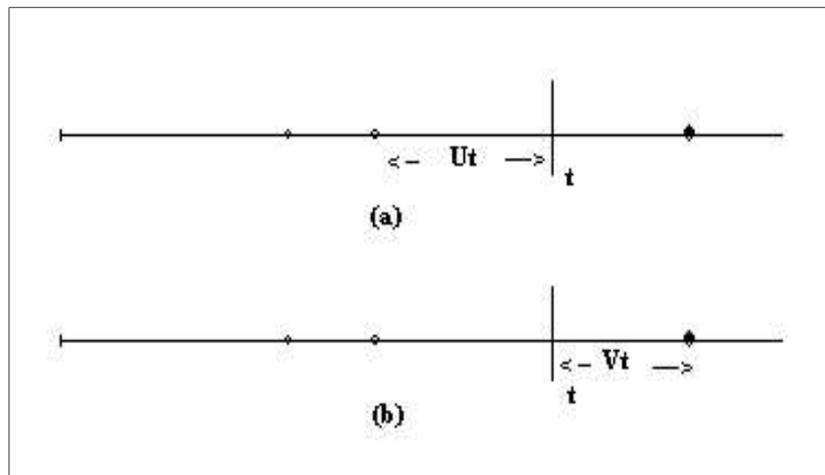


Figura A: Definição de tempo recorrente  $U_t, V_t$ .

vi) Similarmente, dado  $V_t$  como o tempo medido de  $t$  até a próxima renovação ocorrer após o tempo fixado  $t$ .

Esta variável aleatória é chamada de tempo recorrente *forward* (para frente), outro nome utilizado é tempo de vida residual, para  $V_t$  a vida do componente permanece em uso até o tempo  $t$ .

## A.4 Estimação dos parâmetros do modelo

Nesta seção, serão apresentados dois tipos de intervalos de confiança sendo estes o assintótico e o *Bootstrap*, estes serão utilizados nos próximos capítulos.

### A.4.1 Intervalo Assintótico

Intervalos de confiança para os parâmetros podem ser baseados na distribuição normal assintótica.

O intervalo assintótico pode ser construído utilizando os estimadores de máxima verossimilhança (EMV) e suas variâncias estimadas. Sejam  $\theta' = (\alpha, \beta, \gamma)$  e  $L(\alpha, \beta, \gamma)$  a função de verossimilhança correspondente, e  $l(\alpha, \beta, \gamma) = \log L(\alpha, \beta, \gamma)$ . Seja  $\hat{\theta}$  o vetor de EMV's para  $\theta$ . Para obtermos inferências para  $\theta$  podemos utilizar a normalidade assintótica dos EMV (Cox e Hinkley, 1974),

$$\begin{aligned}\hat{\alpha} &\sim N(\alpha, I_o^{-1}(\hat{\alpha})), \\ \hat{\beta} &\sim N(\beta, I_o^{-1}(\hat{\beta})), \\ \hat{\gamma} &\sim N(\gamma, I_o^{-1}(\hat{\gamma})).\end{aligned}\tag{A.7}$$

em que  $\alpha, \beta$  e  $\gamma$  são os parâmetros utilizado no modelo que serão propostos a partir do capítulo 3, assim  $I(\hat{\alpha}), I(\hat{\beta})$  e  $I(\hat{\gamma})$  são os elementos da diagonal principal da matriz de informação de Fisher observada.

Desta forma, o cálculo dos intervalos de confiança de 95% são dados por:

$$\begin{aligned}\hat{\alpha} &\pm 1.96\sqrt{I_o^{-1}(\hat{\alpha})}, \\ \hat{\beta} &\pm 1.96\sqrt{I_o^{-1}(\hat{\beta})}, \\ \hat{\gamma} &\pm 1.96\sqrt{I_o^{-1}(\hat{\gamma})}.\end{aligned}$$

A utilização de (A.7) é direcionada pelo tamanho da amostra, que deve ser suficientemente grande. Alternativamente, uma vez que muitas situações só podemos dispor de amostras pequenas ou moderadas, podemos utilizar a técnica *Bootstrap* paramétrico e/ou não-paramétrico (Davison e Hinkley, 1997) na construção de intervalos de confiança e testes de hipóteses.

#### A.4.2 Intervalo Via Bootstrap

O termo *Bootstrap* é utilizado para denominar uma técnica de simulação que visa à obtenção de estimativas intervalares empíricas para os estimadores dos parâmetros de interesse, através da reamostragem do conjunto de dados original. Essa técnica foi introduzida por Efron (1979).

Sejam  $\alpha$ ,  $\beta$  e  $\gamma$  os parâmetros de interesse. Para cada amostra, calculamos os EMV's para  $\alpha$ ,  $\beta$  e  $\gamma$  e temos no final de  $R$  reamostras,  $\hat{\alpha}_1^* < \dots < \hat{\alpha}_R^*$ ,  $\hat{\beta}_1^* < \dots < \hat{\beta}_R^*$  e  $\hat{\gamma}_1^* < \dots < \hat{\gamma}_R^*$  valores dos EMV ordenados. Utilizamos:

$$\begin{aligned}\hat{\alpha}_{(R+1)(\frac{\alpha}{2})}^* &\text{ e } \hat{\alpha}_{(R+1)(1-\frac{\alpha}{2})}^* \\ \hat{\beta}_{(R+1)(\frac{\alpha}{2})}^* &\text{ e } \hat{\beta}_{(R+1)(1-\frac{\alpha}{2})}^* \\ \hat{\gamma}_{(R+1)(\frac{\alpha}{2})}^* &\text{ e } \hat{\gamma}_{(R+1)(1-\frac{\alpha}{2})}^*\end{aligned}\tag{A.8}$$

como sendo os limites inferiores e superiores do intervalo  $100(1 - \alpha)\%$  de confiança para  $\alpha$ ,  $\beta$  e  $\gamma$  respectivamente. Estipulamos, o número de reamostragens  $R$  que é fixado aqui em 999.

Dessa forma, através de (A.8) podemos obter intervalos de confiança percentil *Boot-*

*strap*  $100(1 - \alpha)\%$  para os parâmetros de interesse.

Existem dois tipos básicos de *Bootstrap*. O *Bootstrap* paramétrico em que os EMV's são obtidos através do modelo ajustado, isto é, geramos dados do modelo ajustado com os valores dos parâmetros fixados nos EMV obtidos da amostra original, e o *Bootstrap* não-paramétrico que será utilizado neste estudo, em que os EMVs são baseadas em  $R$  reamostras com reposição obtidas da amostra original. Maiores detalhes sobre a técnica podem ser obtidos em Efron e Tibishirani (1993), Davison e Hinkley (1997) e Hall (1986).

# Apêndice B

## Códigos

### B.1 Utilizado no Capítulo 2

```
# Programa referente- Modelo Híbrido Poisson-Número de Eventos
# Processo de Poisson Perturbado
#
t=numeric()
#-----Geração dos Valores-----#
n=55
t=numeric()
alpha=2
beta=0.8
gama=-0.1
a=0
for(i in 1:n){
set.seed(123*i)
u=runif(1,0,1)
x=exp(alpha+beta*a+(i-1)*gama)
y=beta*log(1-u)
z=exp(alpha+(i-1)*gama)
t[i]=log((x-y)/z)/beta
a=t[i]
```

```
}
t=round(t,2)
#-----#
n=length(t)
t0=numeric()
param=numeric()
x=numeric()
y=numeric()
t0[1]=0
for(i in 2:n){
t0[i]=t[i-1]
}
funcao=function(param){
alpha <- param[1]
beta <- param[2]
gama <- param[3]
for(i in 1:n){
x[i]=exp(alpha+beta*t[i]+gama*(i-1))
y[i]=exp(alpha+beta*t0[i]+gama*(i-1))
z[i]=(i-1)
}
w=sum(x-y)/beta
-(n*alpha+beta*sum(t)+gama*sum(z)-w)
}
po=c(alpha,beta,gama)
optm3 <- optim(par=po, fn=funcao, method="BFGS",hessian=T)
optm3
a=solve(optm3$hessian)
```

## B.2 Utilizado no Capítulo 3

### B.2.1 Teste de Wald Empírico

```
# Programa referente- Modelo Híbrido Poisson-Número de Eventos
# Teste de Wald Empírico
#
t1=numeric()
t=numeric()
b=numeric()
#-----Geração dos Valores-----#
for(j in 1:399){
n=300
alpha=0 #(com alpha=0)
beta=0.8
gama=-0.1
a1=0
p1=0
x1=numeric()
y1=numeric()
z1=numeric()
u1=numeric()
for(i in 1:n){
set.seed(1*i+2*j)
u1[i]=runif(1,0,1)
x1[i]=exp(alpha+beta*a1+(i-1)*gama)
y1[i]=beta*log(1-u1[i])
if(y1[i]>x1[i]){
y1[i]=p1}
z1[i]=exp(alpha+(i-1)*gama)
t[i]=log((x1[i]-y1[i])/z1[i])/beta
a1=t[i]
```

```
p1=min(y1)
}
#-----#
n1=length(t)
t0=numeric()
param=numeric()
x=numeric()
y=numeric()
z=numeric()
t0[1]=0
for(i in 2:n1){
t0[i]=t[i-1]}
funcao=function(param){
alpha <- param[1]
beta <- param[2]
gama <- param[3]
for(i in 1:n1){
x[i]=exp(alpha+beta*t[i]+gama*(i-1))
y[i]=exp(alpha+beta*t0[i]+gama*(i-1))
z[i]=(i-1)
}
w=sum(x-y)/beta
-(n*alpha+beta*sum(t)+gama*sum(z)-w)
}# fecha função
po1=c(alpha,beta,gama)
optm1 <- optim(par=po1, fn=funcao, method="BFGS",hessian=T)
alpha2=optm1$par[1]
beta2=optm1$par[2]
gama2=optm1$par[3]
varcov=solve(optm1$hessian)
w1.novo[j]=alpha2^2/varcov[1,1]
```

```
# ----- Reamostrando ----- #
reamostras=499
t1=numeric()
t=numeric()
w1=numeric()
w2=numeric()
w3=numeric()
cont.alpha=0
cont.beta=0
cont.gama=0
# ----- Geração dos Valores ----- #
# ----- Reamostragem ----- #
for(s in 1:reamostras){
  alpha=0
  beta=beta2
  gama=gama2
  a2=0
  p2=0
  x2=numeric()
  y2=numeric()
  z2=numeric()
  u2=numeric()
  for(i in 1:n){
    set.seed(i+s+2*j)
    u2[i]=runif(1,0,1)
    x2[i]=exp(alpha+beta*a2+(i-1)*gama)
    y2[i]=beta*log(1-u2[i])
    if(y2[i]>x2[i]){
      y2[i]=p2}
    z2[i]=exp(alpha+(i-1)*gama)
    t[i]=log(abs((x2[i]-y2[i])/z2[i]))/beta
```

```
a2=t[i]
p2=min(y2)
}
#-----#
n1=length(t)
t0=numeric()
param=numeric()
x=numeric()
y=numeric()
z=numeric()
t0[1]=0
for(i in 2:n1){
t0[i]=t[i-1]}
funcao=function(param){
alpha <- param[1]
beta <- param[2]
gama <- param[3]
for(i in 1:n1){
x[i]=exp(alpha+beta*t[i]+gama*(i-1))
y[i]=exp(alpha+beta*t0[i]+gama*(i-1))
z[i]=(i-1)
}
w=sum(x-y)/beta
-(n*alpha+beta*sum(t)+gama*sum(z)-w)
}# fecha função
po=c(alpha2,beta2,0)
optm <- optim(par=po, fn=funcao, method="BFGS",hessian=T)
alpha1=optm$par[1]
beta1=optm$par[2]
gama1=optm$par[3]
varcov=solve(optm$hessian)
```

```
w1[s]=alpha1^2/varcov[1,1]
cat("\n", "reamostragem=", s, " loop=", j) # fecha a reamostragem
a=sort(w1)
b[j]=a[475]
}
for(i in 1:399){
if (w1.novo[i]>b[i])
cont.alpha=cont.alpha+1}
(cont.alpha/399)*100
```

## B.2.2 Teste de Razão de Verossimilhança Empírico

```
# Programa referente- Modelo Híbrido Poisson-Número de Eventos
# Teste de Razão de Verossimilhança Empírico
#
t=numeric()
cont.rv=0
RV=numeric()
cont.alpha=0
RV.novo=numeric()
b=numeric()
#-----Geração dos Valores-----#
for(j in 1:399){
n=300
alpha=0
beta=0.8
gama=-0.1
a1=0
p1=0
x1=numeric()
y1=numeric()
z1=numeric()
```

```
u1=numeric()
for(i in 1:n){
  set.seed(1*i+2*j)
  u1[i]=runif(1,0,1)
  x1[i]=exp(alpha+beta*a1+(i-1)*gama)
  y1[i]=beta*log(1-u1[i])
  if(y1[i]>x1[i]){
    y1[i]=p1}
  z1[i]=exp(alpha+(i-1)*gama)
  t[i]=log((x1[i]-y1[i])/z1[i])/beta
  a1=t[i]
  p1=min(y1)
}
#-----#
n1=length(t)
t0=numeric()
param=numeric()
x=numeric()
y=numeric()
z=numeric()
t0[1]=0
for(i in 2:n1){
  t0[i]=t[i-1]}
funcao=function(param){
  alpha <- param[1]
  beta <- param[2]
  gama <- param[3]
  for(i in 1:n1){
    x[i]=exp(alpha+beta*t[i]+gama*(i-1))
    y[i]=exp(alpha+beta*t0[i]+gama*(i-1))
    z[i]=(i-1)
```

```

}
w=sum(x-y)/beta
-(n*alpha+beta*sum(t)+gama*sum(z)-w)
}# fecha função
po1=c(alpha,beta,gama)
optm1 <- optim(par=po1, fn=funçao, method="BFGS",hessian=T)
p=c(alpha,beta,gama)
optm.verd1=funçao(p)
alpha2=optm1$par[1]
beta2=optm1$par[2]
gama2=optm1$par[3]
RV.novo[j]=2*optm.verd1-2*optm1$value
#-----Reamostrando-----#
reamostras=499
#-----Geração dos Valores-----#
for(s in 1:reamostras){
alpha=0
beta=beta2
gama=gama2
a2=0
p2=0
x2=numeric()
y2=numeric()
z2=numeric()
u2=numeric()
for(i in 1:n){
set.seed(i+s+2*j)
u2[i]=runif(1,0,1)
x2[i]=exp(alpha+beta*a2+(i-1)*gama)
y2[i]=beta*log(1-u2[i])
if(y2[i]>x2[i]){

```

```

y2[i]=p2}
z2[i]=exp(alpha+(i-1)*gama)
t[i]=log(abs((x2[i]-y2[i])/z2[i]))/beta
a2=t[i]
p2=min(y2)
}
#-----#
n1=length(t)
t0=numeric()
param=numeric()
x=numeric()
y=numeric()
t0[1]=0
for(i in 2:n1){
t0[i]=t[i-1]}
funcao=function(param){
alpha <- param[1]
beta <- param[2]
gama <- param[3]
for(i in 1:n1){
x[i]=exp(alpha+beta*t[i]+gama*(i-1))
y[i]=exp(alpha+beta*t0[i]+gama*(i-1))
z[i]=(i-1)
}
w=sum(x-y)/beta
-(n*alpha+beta*sum(t)+gama*sum(z)-w)
}# fecha função
po=c(0,beta2,gama2)
optm <- optim(par=po, fn=funcao, method="BFGS",hessian=T)
p=c(0,beta2,gama2)
optm.verd=funcao(p)

```

```

alpha1=optm$par[1]
beta1=optm$par[2]
gama1=optm$par[3]
RV[s]=2*optm.verd-2*optm$value
cat("\n", "reamostragem=",s,"loop=",j)} # fecha a reamostragem
a=sort(RV)
b[j]=a[475]
}
for(i in 1:399){
if (RV.novo[i]>b[i])
cont.alpha=cont.alpha+1}
(cont.alpha/399)*100

```

### B.3 Utilizado no Capítulo 4

```

#Programa para o cálculo das estimativas, referente ao exemplo com dados reais
#
t=c(0.04,0.21,0.24,0.41,0.59,0.70,0.76,0.86,0.99,1.10,1.13,1.42,1.60,1.61,1.69,
1.71,1.72,1.78,1.81,1.88,1.89,1.93,2.26,2.34,2.40,2.43,2.89,2.91,3.42,3.46,3.50,
3.51,3.67,3.74,3.77,3.81,4.14,4.24,4.40,4.87,4.88,4.95,5.04,5.30,5.60,5.71,5.72,
5.97,6.02,6.05,6.22,6.36,6.39,6.58,6.59,6.69,6.75)
n=length(t)
t0=numeric()
param=numeric()
x=numeric()
y=numeric()
z=numeric()
t0[1]=0
for(i in 2:n){
t0[i]=t[i-1]
}

```

```
funcao=function(param){
alpha <- param[1]
beta <- param[2]
gama <- param[3]
for(i in 1:n){
x[i]=exp(alpha+beta*t[i]+gama*(i-1))
y[i]=exp(alpha+beta*t0[i]+gama*(i-1))
z[i]=(i-1)
}
w=sum(x-y)/beta
-(n*alpha+beta*sum(t)+gama*sum(z)-w)
}
po=c(2,1,-0.2)
optm <- optim(par=po, fn=funcao, method="BFGS",hessian=T)
alpha1=optm$par[1]
beta1=optm$par[2]
gama1=optm$par[3]
#Verificação
t.novo=numeric()
n=57
alpha=alpha1
beta=beta1
gama=gama1
a=0
for(i in 1:n){
set.seed(123*i)
u=runif(1,0,1)
x=exp(alpha+beta*a+(i-1)*gama)
y=beta*log(1-u)
z=exp(alpha+(i-1)*gama)
t.novo[i]=log((x-y)/z)/beta
```

```
a=t.novo[i]
}
plot(t*360, xlab='número de compras', ylab='tempo da compra em dias',xlim=c(0,60),ylim=c(0,8*360))
lines(t.novo*360,col='red')
```

# Referências Bibliográficas

- [1] Aalen, O. O. (1978). Nonparametric inference for a family of counting processes, *The Annals of Statistics*, 6, 701-726.
- [2] Andersen, P. K., Borgan, O. , Gill, R. D. and Keiding, N. N. (1993). *Statistical models based on counting processes*. New York: Springer-Verlag.
- [3] Ascher, H., and Feingold, H.(1984), *Reparaible Systems Reliability: Modelling, Inference, Misconceptions and Their Causes*, New York: Marcel Dekker.
- [4] Bain, L.J., and Engelhardt, M.(1980), "Inferences on the Parameters and Current System Reliability for a Time Truncated Weibull Process", *Technometrics*, 22,421-426.
- [5] Cook, R.J. and Lawless, J.F. (2002). Analysis of repeated events. *Statistical Methods in Medical Research* 11, 141-166.
- [6] Cox, D. R. *Renewal Theory*, (1962).
- [7] Cox, D. R., and Lewis, P. A. W.(1966), *The Statistical Analysis of Series of Events*, London: Methuen.
- [8] Cox, D. R. (1972b). The statistical analysis of dependences in point processes. In *Stochastic point Processes*, Ed. P.A.W. Lewis, 55-66. New York: Wiley.
- [9] Cox, D. R., and Isham, V. (1980). *Point Processes*, London: Chapman & Hall.
- [10] Cox, DR., Hinkley, DV. (1974). *Theoretical Statistics*, London: Chapman and Hall.
- [11] Crowder, M. J., Kimber, A. C., Smith, R. L., and Sweeting, T. J. (1991), *Statistical Analysis of Reliability Data*, London: Chapman and Hall.

- [12] Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.
- [13] Efron, B. (1979). *Bootstrap methods: Another look at the jackknife*. The Annals of Statistics, 7, 1-26.
- [14] Efron, B., and Tibshirani, R. J. (1993). *An introduction to the Bootstrap*. New York: Chapman & Hall.
- [15] Follmann, D. R. and Goldberg, M. S. (1988). Distinguishing heterogeneity from decreasing hazard rates. *Technometrics*, 1, 30, 389-96.
- [16] Hall, P. (1986). On the number of *Bootstrap* simulations required to construct a confidence interval. *Annals of Statistics*, 14, 1453-1462.
- [17] Karlin, S., and Taylor, H. M. (1975). *A First Course in Stochastic Processes*. New York: Academic.
- [18] Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. John Wiley and Sons, New York, New York.
- [19] Lawless, J. F. (1987), "Regression Methods for Poisson Process Data", *Journal of the American Statistical Association*, 82, 807-815.
- [20] Lawless, J. F. and Nadeaus, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics*, 20, 41-45.
- [21] Lawless, J.F. & Thiagarajah, K. (1996). A point-process model incorporating renewals and time trends, with application to repairable systems. *Technometrics* 38, 131-38.
- [22] Lehmann, E. L. (1999). *Elements of Large-sample Theory*. New York Springer-Verlag, New York.
- [23] Lee, L., and Lee, K. (1978), "Some Results on Inference for the Weibull Process", *Technometrics*, 20, 41-45.
- [24] Lipschutz, K. H. and Snapinn, S. M. (1997). Discussion of the paper by Wei, L. J. and Glidden, D. V. *Statistics in Medicine*, 16, 846-48.

- 
- [25] Louzada-Neto, F.; Mazuchelli, J.; Achcar, J.A. (2002). Introdução à Análise de Sobrevivência e Confiabilidade. III Jornada Regional de Estatística.
- [26] Nelson, W. (1982). Applied Life Data Analysis. John Wiley and Sons, New York, NY.
- [27] Nelson, W. (1995). Confidence Limits for Recurrence Data - Applied to Cost or Number of Product Repair. *Technometrics*, 37, p.147-157.
- [28] Prentice, R.L., Williams, B. J. and Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 68, 373-79.
- [29] Ross, S. M. (1983), *Stochastic Processes*, New York: John Wiley.