

# 1

## Introdução

Na análise de dados de séries temporais se assume, freqüentemente, que a forma funcional da associação das covariáveis com a variável resposta é linear. Contudo, se esta relação exibe um padrão complexo, esta abordagem torna a modelagem inadequada.

Em algumas aplicações, como a análise de séries temporais epidemiológicas, o que se pretende usualmente é estimar o efeito de uma ou mais covariáveis que representam uma exposição sobre a variável resposta que corresponde a um desfecho de saúde. Entretanto, é necessário remover o efeito de covariáveis que constituem fatores de confusão. As formas paramétricas podem não ser suficientemente flexíveis para proporcionar o ajuste necessário para corrigir este efeito.

Atualmente, os modelos aditivos generalizados (MAG) constituem a metodologia padrão para a análise de séries temporais epidemiológicas em estudos que investigam a associação de poluição do ar com eventos de saúde, como morbidade e mortalidade, devidos a doenças respiratórias e cardiovasculares. Os modelos aditivos generalizados permitem que funções genéricas das covariáveis sejam introduzidas no preditor aditivo. No caso trivial, estas funções são lineares. A associação de cada covariável com a variável resposta pode ser definida por uma função suave determinada pelos próprios dados. Estas funções são denominadas suavizadores.

Na literatura referente à regressão não-paramétrica, há um grande número de funções suavizadoras. Entretanto, a família de funções mais utilizada é a *spline*. A *spline* cúbica natural é o suavizador mais frequentemente utilizado devido a características matemáticas que são atrativas para a análise estatística de dados e para diagnósticos de adequação do modelo.

Não obstante constituir a metodologia padrão dos estudos de séries temporais em epidemiologia ambiental, os modelos aditivos generalizados não são capazes de ajustar adequadamente estruturas intrínsecas das séries temporais. Por exemplo, a autocorrelação precisa ser ajustada através de métodos *ad hoc*. A solução natural para o problema é a formulação de

uma classe de modelos capazes de lidar com os componentes das séries temporais e que permitam associações não-lineares das covariáveis com a variável resposta.

A classe de modelos Poisson-gama para séries temporais proposta por Harvey e Fernandes [20] constitui uma abordagem para a análise de dados de contagem, na qual as estruturas intrínsecas das séries temporais são devidamente ajustadas. Entretanto, estes modelos não permitem que associações mais complexas das covariáveis com a variável resposta sejam modeladas de forma não-paramétrica.

Considere-se, então, uma classe de modelos capazes de incorporar no seu preditor formas não-paramétricas das covariáveis e que, ainda, sejam capazes de ajustar devidamente estruturas inerentes à dinâmica das séries temporais. Estes modelos possuiriam a adequação de um modelo próprio para séries temporais e a flexibilidade de um modelo aditivo generalizado.

Neste trabalho é proposta a extensão dos modelos Poisson-gama para uma formulação mais geral na qual o preditor linear das covariáveis é substituído por um preditor aditivo de funções genéricas destas covariáveis. Como nos modelos aditivos generalizados, as funções lineares das covariáveis constituem um caso particular. As funções suavizadoras utilizadas neste trabalho são as *splines* cúbicas naturais, embora outros suavizadores possam ser facilmente incorporados ao modelo.

O processo de estimação da partição paramétrica do modelo híbrido ou semi-paramétrico, como será referido a partir de agora, é obtida pelo método de maximização numérica da verossimilhança. A partição não-paramétrica é estimada pelo algoritmo *backfitting*. Os dois algoritmos de estimação combinados são iterados até que o valor da verossimilhança atinja um critério de convergência. Alguns procedimentos inferenciais são adaptados dos modelos lineares generalizados na falta de uma teoria distribucional adequada aos modelos semi-paramétricos.

Os algoritmos de estimação dos modelos Poisson-gama semi-paramétricos bem como ferramentas para diagnósticos da adequação do ajuste de modelos, foram implementados na forma de uma biblioteca nas linguagens *R* e *C*. Apesar de poderem ser executados em outros sistemas operacionais suportados pelo *R*, os códigos foram otimizados para execução sobre o sistema operacional *Linux*. A escolha da linguagem de programação e do sistema operacional concorda com a filosofia de *software* livre e código aberto.

A extensão do modelo Poisson-gama para uma forma mais geral na qual seja possível introduzir funções não-paramétricas, sobretudo funções

suaves dos dados como as *splines* cúbicas naturais, de forma aditiva no preditor linear permite ampliar o campo de aplicação desta classe de modelos.

O trabalho está organizado da seguinte forma: No capítulo 2 é realizada uma revisão detalhada de teorias e métodos dos modelos Poisson-gama paramétricos e de regressão não-paramétrica com ênfase em *splines* cúbicas naturais e no algoritmo *backfitting*. No capítulo 3, os modelos Poisson-gama semi-paramétricos são especificados teoricamente, são apresentados alguns procedimentos de inferência estatística e detalhados os aspectos computacionais e algoritmo de estimação. O capítulo 4 consiste da aplicação do modelo semi-paramétrico a um problema de epidemiologia ambiental e comparação do modelo Poisson-gama semi-paramétrico com um modelo Poisson-gama paramétrico e um modelo aditivo generalizado na análise do mesmo problema de associação de poluição do ar com o número diário de internações de crianças por doenças respiratórias. O capítulo 5 conclui o trabalho apontando resultados e dificuldades. O código fonte dos programas encontra-se no anexo A. O anexo B, é um CD-ROM contendo o código fonte da biblioteca *pgam* e também sua versão compilada para o sistema operacional *Windows*, bem como as versões atuais do sistema *R*.