NAYARA FRANCINE DE MOURA GONÇALVES

BOOTSTRAP EM MODELOS AUTO-REGRESSIVOS ADITIVOS GENERALIZADOS

UNIVERSIDADE FEDERAL DE MINAS GERAIS
MAIO 2009

BOOTSTRAP EM MODELOS AUTO-REGRESSIVOS ADITIVOS GENERALIZADOS

NAYARA FRANCINE DE MOURA GONÇALVES

Orientadora: GLAURA DA CONCEIÇÃO FRANCO - UFMG
Co-Orientador: VALDÉRIO ANSELMO REISEN - UFES

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial à obtenção do título de MESTRE em ESTATÍSTICA.

Maio 2009

Ah, se o mundo inteiro me
pudesse ouvir! Tenho muito pra contar,
dizer que aprendi...

Azul da cor do mar - Tim Maia

Agradecimentos

Aos meus pais e irmão

Desde o começo vocês estiveram comigo. Inúmeras foram as vezes que sacrificaram seus objetivos em favor dos meus. Também não foram raros os momentos em que buscaram meu sorriso e me encontraram tão cheia de pressa. Eu agradeço por todo amor que recebi, pois em todas as lições de vida vocês estavam presentes e sempre que foi preciso decidir vocês acreditaram em mim!

À Glaura

Você me convidou a voar na sua sabedoria e o que eu aprendi foi que este voar sempre dependeu das minhas próprias asas. Agradeço pela amizade, paciência e pela tão dedicada orientação.

Ao Prof. Valdério e à minha amiga Fabiana

Pelas fundamentais contribuições.

Ao Prof. Paulo Sérgio Lúcio (Departamento de Estatística da UFRN)

Por gentilmente ceder o banco de dados reais utilizado neste trabalho.

Resumo

A classe dos Modelos Aditivos Generalizados (MAG), considerados uma extensão dos Modelos Lineares Generalizados, vem atraindo a atenção de pesquisadores principalmente em função de sua flexibilidade. Apesar de construído sob a hipótese de independência dos dados, os MAG's são muito aplicados em estudos de séries temporais, sobretudo como alternativa para modelagem de variáveis de confusão tais como tendência e sazonalidade. Recentemente, modelos mais gerais, que consideram a estrutura de correlação entre os dados, como os modelos GLARMA (*autoregressive moving average generalized linear models*), têm sido utilizados. Este trabalho estende os modelos GLARMA para uma classe de modelos auto-regressivos aditivos generalizados para séries de contagem cuja distribuição condicional, dadas as observações passadas e as variáveis explicativas, segue uma distribuição de Poisson.

Além de apresentar uma conceituação desses modelos bem como procedimentos de ajustes, este trabalho emprega, em um estudo empírico, o procedimento *bootstrap* em três formas (*bootstrap* nas observações, *bootstrap* condicional e *bootstrap* nos resíduos) na inferência pontual dos parâmetros do modelo e compara dois métodos de construção de intervalos de confiança *bootstrap* - *bootstrap* percentílico e *bootstrap* com correção do vício – na estimação intervalar.

Os resultados mostram que, em geral, os procedimentos e os intervalos de confiança *bootstrap* apresentam um bom desempenho quando utilizados na classe de MAG's que por sua vez, quando auxiliados pela modelagem GLARMA, modelam bem dados de contagem com estrutura auto-regressiva de ordem 1, apresentando estimativas próximas dos valores verdadeiros dos parâmetros.

Abstract

The class of Generalized Additive Models (GAM), considered an extension of the Generalized Linear Models (GLM), is attracting the attention of researchers mainly due to the flexibility of these procedures. In spite of being built under the hypothesis of independency of the data, the GAM is widely applied to time series data, as an alternative to model variables such as trend and seasonality. Recently, more general models, which consider the correlation structure among the data, like the GLARMA models (autoregressive moving average generalized linear models), are being used. This work extends the GLARMA models to a class of autoregressive generalized additive models of count series whose conditional distribution, given the past observations and the independent variables, follows a Poisson distribution.

Besides presenting the definition of the model, as well as the fitting procedures, this work employs, in a empirical study, the bootstrap procedure in three different ways (bootstrap in the observations, conditional bootstrap and the bootstrap in the residuals) in the interval inference of the parameters, comparing two bootstrap methods of building confidence intervals – percentile bootstrap and bootstrap with bias correction.

The results show that, in general, the procedures and the bootstrap confidence intervals present a satisfactory performance when used in the GAM models with the GLARMA structure, modeling count data with an autoregressive structure of order 1, and presenting estimates close to the true values of the parameters.

Sumário

1. INTRODUÇÃO

1.1 Revisão de Literatura

Durante muitos anos os modelos lineares, sob a suposição de normalidade, foram utilizados para descrever fenômenos aleatórios. Se o fenômeno sob estudo não apresentasse uma resposta para a qual fosse razoável supor a distribuição Gaussiana, algum tipo de transformação da variável era utilizado, com o propósito de se alcançar a normalidade.

Mais tarde, pesquisadores abriram o leque de opções para a distribuição da variável resposta permitindo que a mesma pertencesse à família exponencial de distribuição. Em 1972, Nelder e Wedderburn unificaram estes procedimentos introduzindo a classe dos Modelos Lineares Generalizados (MLG). Muitas distribuições conhecidas pertencem à família exponencial, como a Normal, a Poisson, a Binomial e a Gama. Uma característica dos MLG's é que a forma da relação funcional entre a média da variável resposta e as variáveis preditoras é completamente linear e especificada por termos paramétricos.

Como a suposição de linearidade pode ser irrealista em situações práticas, Hastie e Tibshirani (1990) propuseram os Modelos Aditivos Generalizados (MAG), cuja principal diferença com os MLG's é a substituição da usual forma linear das covariáveis por funções suavizadoras não paramétricas que sumarizam a associação entre a variável resposta e as variáveis explicativas.

Existem na literatura várias técnicas de suavização – Hastie e Tibshirani (1990) apresentam algumas delas. Em qualquer uma das técnicas, a suavização é obtida ajustando-se uma curva aos dados de tal forma que a cada ponto, a curva dependa somente das observações naquele ponto (x_0) e em uma vizinhança.

Entre as formas mais simples de suavização estão as técnicas *running-mean* e a *running-line* (Hastie e Tibshirani, 1896). A primeira, também conhecida como *moving-average*, é popular por ser utilizada em dados temporais igualmente espaçados, entretanto, apesar da simplicidade dos cálculos, esta técnica tende a ser viciada por suavizar a tendência nos pontos iniciais e finais da série.

O suavizador r*unning-line*, uma simples generalização do suavizador *running-mean*, reduz o problema do vício utilizando os mínimos quadrados lineares ao invés da média em cada vizinhança. Por outro lado, a técnica tende a gerar curvas bastante irregulares então um segundo estágio de suavização geralmente é necessário. Uma maneira de melhorar os resultados em um suavizador *running-line* é usar o ajuste de mínimos quadrados ponderados (MQP) em cada vizinhança. O suavizador *running-line* pode produzir resultados irregulares porque pontos em uma determinada vizinhança têm pesos iguais (não nulos), enquanto pontos fora desta região têm peso zero.

Em 1977, Cleveland abranda este problema propondo uma técnica mais robusta de suavização, *locally weighted running line smoother* (*loess*). O procedimento, que pode ser usado para suavizar dados com configurações mais gerais que não sejam necessariamente para dados de séries temporais igualmente espaçados, é uma adaptação do ajuste sucessivo de modelos de regressão pelo método de mínimos quadrados ponderados (Beaton and Tukey, 1974; Andrews, 1974). A proposta é dar maior peso para x_0 e para os pontos da vizinhança e pesos que decrescem suavemente para pontos mais afastados desta região. Visando tornar o ajuste *loess* ainda mais robusto, Cleveland propõe, em 1979, o suavizador *robust locally weighted running line smoother* introduzindo um novo conjunto de pesos a serem utilizados no ajuste de MQP.

Outras alternativas de técnicas de suavização são propostas na literatura. Introduzida inicialmente por Whittaker (1923), a suavização por *spline* é uma técnica bastante utilizada no ajuste do MAG; autores como Dominici *et al.* (2002) e Ramsay *et al.* (2003), descrevem e utilizam as funções splines e algumas de suas extensões – penalized splines e parametric splines – como forma de suavização; a técnica Kernel é destacada, por exemplo, em Silverman (1986), Buja *et al.* (1989) e Härdle(1990).

Nenhuma das técnicas de suavização faz alguma suposição paramétrica sobre a curva a ser estimada; cada suavizador tem um parâmetro que determina o "quanto" a função estará suavizada. Para a função *loess*, por exemplo, este parâmetro é chamado *span*. Para *splines* e *parametric splines* o grau de suavização é especificado através do número de graus de liberdade.

Também fazem parte da classe de MAG's os modelos semiparamétricos – segundo Buja *et al.* (1989) pode-se entender estes modelos como uma generalização dos MAG's – , constituídos pela soma de termos paramétricos de algumas variáveis

preditoras e funções suavizadoras de outras. Discutido por Stone (1985), Hastie e Tibshirani (1990) e Lee (1990), este modelo tem sido bastante explorado na literatura devido a esta peculiaridade de compor a flexibilidade do modelo aditivo não paramétrico com um componente paramétrico. Uma contribuição de interesse é feita por Buja *et al.* (1989) ao apresentar diversas técnicas de estimação aplicáveis a este tipo de modelo. Além desses trabalhos destacam-se outros como Schick (1986, 1993,1996) e Bhattacharya e Zhao (1997) que enriqueceram a teoria assintótica acerca do modelo semiparamétrico apresentando assim uma ponte entre os resultados assintóticos para o modelo puramente não paramétrico já discutidos por Stone (1977), Cox (1983) ou Rice e Rosenblatt (1983).

Um grande número de pesquisadores vem utilizando os MAG's e sua extensão semiparamétrica em séries temporais, principalmente em estudos cujo objetivo é avaliar os efeitos da poluição atmosférica sobre a saúde de seres humanos (Schwartz *et al.*, 1993). Alguns fatores como condições meteorológicas e os dias da semana influenciam os dados (Díez, 1999) e confundem a associação entre a exposição de interesse e o desfecho. Além disso, ainda existem as componentes da própria série temporal como tendência, sazonalidade e autocorrelação. O MAG tem sido uma técnica alternativa facilitadora no controle desses fatores de confusão já que esta modelagem elimina a necessidade de especificar uma forma paramétrica para a associação entre covariáveis e preditor.

Nestes estudos epidemiológicos a variável resposta é, geralmente, alguma contagem de eventos que representam danos à saúde, como o número de óbitos ou o número de internações por determinada causa respiratória. Na maioria das vezes, estes desfechos são modelados com o pressuposto de que as contagens dos eventos seguem uma distribuição Poisson – como exemplos de aplicações desta técnica ver estudos de Schwartz *et al.* (1992), Conceição *et al.* (2001) e Lima *et al.* (2001).

Em 1999, Davis *et al.* faz uma revisão dos modelos já propostos na literatura para séries temporais de contagem que seguem a distribuição Poisson. Em particular, uma nova classe de modelo, GLARMA (*autoregressive moving average generalized linear models*), é introduzida e suas propriedades desenvolvidas em parte. Por serem capazes de capturar uma gama de estruturas de dependência entre as observações das séries temporais, os modelos GLARMA também vêm sendo utilizados nestes estudos epidemiológicos nos casos, diferentemente do MAG, em que a relação entre a variável

resposta e as covariáveis assume a forma linear e os dados apresentam uma estrutura de dependência, por exemplo a auto-regressiva.

Um problema comumente encontrado na estimação dos termos dos modelos semiparamétricos é a baixa freqüência de dados. Pequenas amostras e a dificuldade em se determinar a distribuição assintótica dos dados fazem com que os métodos de estimação percam um pouco de sua eficiência. Nestes casos o *bootstrap*, sugerido por Efron (1979), pode ser utilizado para a melhoria das inferências intervalares.

O método *bootstrap* consiste em uma técnica de reamostragem que permite aproximar a distribuição de uma função das observações pela distribuição empírica dos dados, baseando-se em uma amostra finita (Efron e Tibshirani, 1993). Em série temporal essa técnica tem sido bastante empregada, porém, o fato dos dados em série não serem independentes torna a aplicação do método bastante criteriosa – as observações não devem ser reamostradas diretamente, pois sua estrutura original pode ser perdida. A utilização do *bootstrap* em modelos MAG ainda é pouco discutida na literatura estatística, mas já considerada em alguns trabalhos recentes – Härdle *et al.* (2004) mostram como o procedimento pode ser utilizado na correção do vício das estimativas paramétricas e não-paramétricas, em testes de hipótese e na construção de bandas de confiança; Figueiras *et al.* (2005), utilizam um *bootstrap* condicional para corrigir problemas de concurvidade em modelos MAG. Entretanto, não foram encontrados trabalhos utilizando o *bootstrap* em modelos GLARMA.

Existem vários métodos *bootstrap* de construção de intervalos de confiança e entre eles estão os intervalos *bootstrap* percentílico (Efron e Tibshirani, 1986) e o *bootstrap* com correção do vício (Efron e Tibshirani, 1986 e Hall, 1988) utilizados neste estudo.

1.2 Objetivos

Este trabalho tem por objetivos (a) propor uma modelagem para dados de contagem que substitua a usual forma linear do GLARMA pela estrutura semiparamétrica dos MAG's – neste caso, além dos termos lineares serão consideradas também funções não lineares de variáveis explicativas na construção dos modelos; (b) comparar a estimação dos parâmetros lineares da modelagem proposta com a estimação do MAG em estudos simulados de dados independentes e de séries

temporais, ambos com respostas que seguem a distribuição Poisson e (c) utilizar a técnica *bootstrap* com o intuito de fazer inferências sobre os parâmetros lineares dos modelos.

1.3 Organização do Trabalho

O abundante uso da técnica MAG em séries temporais, apesar de não existirem resultados teóricos sobre sua utilização em dados dependentes, e a presença de estruturas de correlação nas séries, motivaram a proposta deste trabalho que é estender os modelos GLARMA para uma classe de modelos auto-regressivos aditivos generalizados, MAG-AR, ainda não abordada na literatura.

No decorrer do texto, são apresentados os modelos generalizados MAG e GLARMA bem como algumas de suas propriedades. Alguns resultados gerais sobre suavizadores lineares também serão apresentados, no entanto o foco maior é dado ao suavizador *loess* – escolhido por apresentar diferentes propriedades estatísticas e também por ser bastante utilizado em estudos temporais sobre os efeitos da poluição atmosférica na saúde dos seres humanos; a citada técnica será utilizada no ajuste da parte não-paramétrica do MAG e MAG-AR. Três diferentes abordagens *bootstrap* – *bootstrap* nas observações, *bootstrap* condicional e *bootstrap* nos resíduos – são avaliadas.

Para atender aos objetivos acima, estudos de simulação Monte Carlo serão realizados para que se compare, segundo o vício e o erro quadrático médio, o comportamento dos estimadores do parâmetro linear dos modelos. Além disto, intervalos de confiança obtidos através da técnica *bootstrap*, como o intervalo *bootstrap* percentílico e o *bootstrap* com correção do vício, serão comparados ao intervalo de confiança assintótico quanto ao percentual de cobertura e tamanho dos intervalos.

Este trabalho está organizado da seguinte maneira: o Capítulo 2 apresenta as técnicas de suavização descrevendo com maiores detalhes o suavizador *loess*. No Capítulo 3 apresenta-se o modelo MAG dissertando sobre sua estimação e propriedades. A descrição do modelo GLARMA e a proposta do novo modelo MAG-AR são feitas no Capítulo 4. No Capítulo 5 o procedimento *bootstrap* é abordado.

Resultados das simulações e da análise de dados reais são discutidos, respectivamente, nos Capítulos 6 e 7 e o Capítulo 8 conclui o trabalho.

2. TÉCNICAS DE SUAVIZAÇÃO

Um suavizador é uma ferramenta utilizada para descrever a dependência de uma variável resposta Y como uma função de uma ou mais variáveis preditoras X. Uma importante propriedade de um suavizador é sua natureza não paramétrica: a forma da curva estimada é determinada pelos próprios dados e, de fato, não é necessário nem mesmo conhecer previamente a forma dessa relação para estimá-la, por esse motivo os procedimentos de suavização são também denominados técnicas de regressão não paramétrica. Os valores destas funções devem ser mais "suaves" do que os valores de Y, isto porque a estimativa obtida de um procedimento de suavização tem variabilidade menor que a de Y, daí a razão do nome suavizador ou alisador.

2.1 Definição e propriedades dos suavizadores

Formalmente, um suavizador é definido como uma função de $x=(x_1,...,x_n)$ e $y=(y_1,...,y_n)$, $f=\delta(y\mid x)$, com mesmo domínio de x. Para alguns suavizadores, a função $f=\delta(y\mid x)$ calculada em x_0, $f(x_0)$, é definida para todo x_0. Outras vezes ela é definida apenas para os valores observados $x_1,...,x_n$ e, neste caso, algum tipo de interpolação é necessário para obter estimativas para outros valores de X.

As curvas suavizadas podem ser utilizadas com diferentes objetivos; tipicamente são empregadas no ajuste do modelo

$$E(Y\mid X=x_i)=f(x_i), \tag{2.1}$$

uma generalização dos modelos de regressão linear simples, onde f é uma função arbitrária desconhecida não especificada a priori.

Os suavizadores são classificados como lineares ou não lineares. Um suavizador é linear se $\delta(ay_1+by_2\mid x)=a\delta(y_1\mid x)+b\delta(y_2\mid x)$ (Hastie e Tibshirani, 1990).

Focando no vetor de valores estimados $\hat{y} = (\hat{y}_1,...,\hat{y}_n)' = (\hat{f}(x_1),...,\hat{f}(x_n))' = \hat{f}$, um suavizador linear pode ser escrito como nas equações (2.2) e (2.3)

$$\hat{f} = Sy \tag{2.2}$$

$$\hat{f}(x_i) = s_{x_i,1}y_1 + s_{x_i,2}y_2 + ... + s_{x_i,n}y_n \ , \ i = 1,...,n \tag{2.3}$$

onde $S = \{s_{x_i j}\} = \{s_{ij}\}$ é uma matriz de dimensão $n \times n$ chamada matriz suavizadora que não depende de y, mas apenas de X e da técnica de suavização adotada, e $s_{x_i,1}, s_{x_i,2},...,s_{x_i,n}$ é a i-ésima linha da matriz S. Pelo fato de suas matrizes não dependerem da variável resposta, a análise por meio destes suavizadores é relativamente simples (Hastie e Tibshirani 1989).

Dado um algoritmo de um suavizador linear pode-se produzir a correspondente matriz suavizadora S suavizando o vetor de base unitária: o resultado de suavização do i-ésimo vetor unitário é a i-ésima coluna de S (Buja *et al.* 1989). *Running-mean*, *running-line*, *spline* e *loess* são exemplos de suavizadores lineares.

Um simples exemplo de suavizador não linear para o qual a matriz suavizadora não pode ser construída é o suavizador *running median*. A diferença entre esta técnica e o suavizador *running mean* é que o valor ajustado através do *running median* para x_0 e vizinhança é dado pela mediana destas observações. Neste caso, para variáveis aleatórias X e Y, mediana($X + Y$) $\neq$ mediana (X) + mediana (Y), assim as estimativas dependem de Y de um modo não linear.

A matriz suavizadora desempenha papel semelhante ao da matriz *hat*[1] no método de estimação de mínimos quadrados e algumas de suas propriedades são demonstradas por Hoaglin e Welsch (1978).

i. $0 \leq s_{ii} \leq 1$;

ii $-1 \leq s_{ij} \leq 1$ para $i \neq j$;

[1] No modelo de regressão normal linear, a estimativa do vetor de médias $\hat{\mu} = X\hat{\beta}$, onde $\hat{\beta} = (X'X)^{-1}X'y$, pode ser reescrita como $\hat{\mu} = Hy$, com $H = X(X'X)^{-1}X'$. A matriz H é a matriz de projeção ortogonal dos vetores de R^n no subespaço gerado pelas colunas da matriz X.

iii $s_{ii} = 1$ se e somente se $s_{ij} = 0$ para todo $i \neq j$;

iv $\sum_{j=1}^{n} s_{ij} = 1.$

Porém, para algumas técnicas de suavização como *running line* e *loess* S não é simétrica e nem idempotente, ou seja, não é um operador de projeção como ocorre com a matriz *hat*. Para os suavizadores *Bin*, *least-square line*, *polynomial regression* e *splines*, por exemplo, a matriz suavizadora é simétrica e seus autovalores são sempre reais.

Se o suavizador é linear, o estimador $\hat{f}$ para uma função arbitrária f é sempre viciado

$$E(f - \hat{f}) = f - E(\hat{f}) = f - E(SY) = f - Sf \tag{2.5}$$

e a matriz de covariância de $\hat{f}$ é dada por

$$Var(\hat{f}) = Var(SY) = SVar(Y)S' \tag{2.6}$$

e, sob a suposição de que os Y_i's são independentes com $Var(Y_i) = \sigma^2$, a expressão (2.6) pode ser reescrita

$$Var(\hat{f}) = SS'\sigma^2. \tag{2.7}$$

Considerando estas informações, o erro quadrático médio

$$EQM = \frac{\sum_{i=1}^{n} E[f(x_i) - \hat{f}(x_i)]^2}{n} \tag{2.8}$$

assume a forma

$$EQM = \frac{1}{n}\sum_{i=1}^{n} Var\left(\hat{f}(x_i)\right) + \frac{1}{n}\sum_{i=1}^{n} b_i^2$$

$$= \frac{tr(SS')}{n}\sigma^2 + \frac{b'b}{n}, \tag{2.9}$$

onde b é o vetor de vício definido em (2.5). O parâmetro σ^2 em (2.7) e em (2.9) geralmente é desconhecido. Um estimador para este parâmetro, assumindo que f é não viciado é dado por

$$\hat{\sigma}^2 = \frac{\sum\left(y_i - \hat{f}(x_i)\right)^2}{n - tr(2S - SS')} = \frac{\hat{e}'\hat{e}}{n - tr(2S - SS')}, \tag{2.10}$$

onde $\hat{e} = y - \hat{f} = y - Sy = (I - S)y$, com I representando uma matriz identidade $n \times n$. No caso em que S é idempotente, $tr(SS') = tr(S) = rank(S)$ e $tr(2S - SS') = \sum_{i=1}^{n}(2\theta_i - \theta_i^2)$, sendo θ_i, $i = 1,...,n$, os autovalores de S.

O resultado (2.10) é apresentado sob a suposição de ausência de viés de $\hat{f}$. Porém o viés é nulo apenas para uma classe restrita de funções. O suavizador *loess*, por exemplo, fornece funções $\hat{f}$ viciadas para uma f arbitrária. Uma solução para esse problema é considerar o comportamento assintótico discutidos amplamente por Stone (1977), Cox (1983) ou Rice e Rosenblatt (1983).

2.2 Seleção do parâmetro de suavização

Na maioria das técnicas lineares de suavização, o valor alisado é obtido segundo o comportamento de uma vizinhança. Diferentes formas de cálculos nesta vizinhança definem as diferentes técnicas de suavização. A escolha do tamanho da vizinhança está associada a um parâmetro λ, denominado parâmetro de suavização. Sendo mais importante até do que a escolha da técnica de suavização, a definição dos valores para λ é um passo importante no processo, isto porque o parâmetro está diretamente relacionado à relação de ganho e perda entre o viés e a variância da curva estimada: aumentar λ implica aumentar a suavização da curva, logo, diminui-se a variância, por

outro lado, perde-se informação no ajuste implicando no aumento do viés. A Figura 2.1 ilustra este efeito.

É interessante observar também a influência de λ nos termos do EQM (2.9); de forma geral, aumentando λ, o $tr(SS')$ tende a diminuir e o vício tende a aumentar.

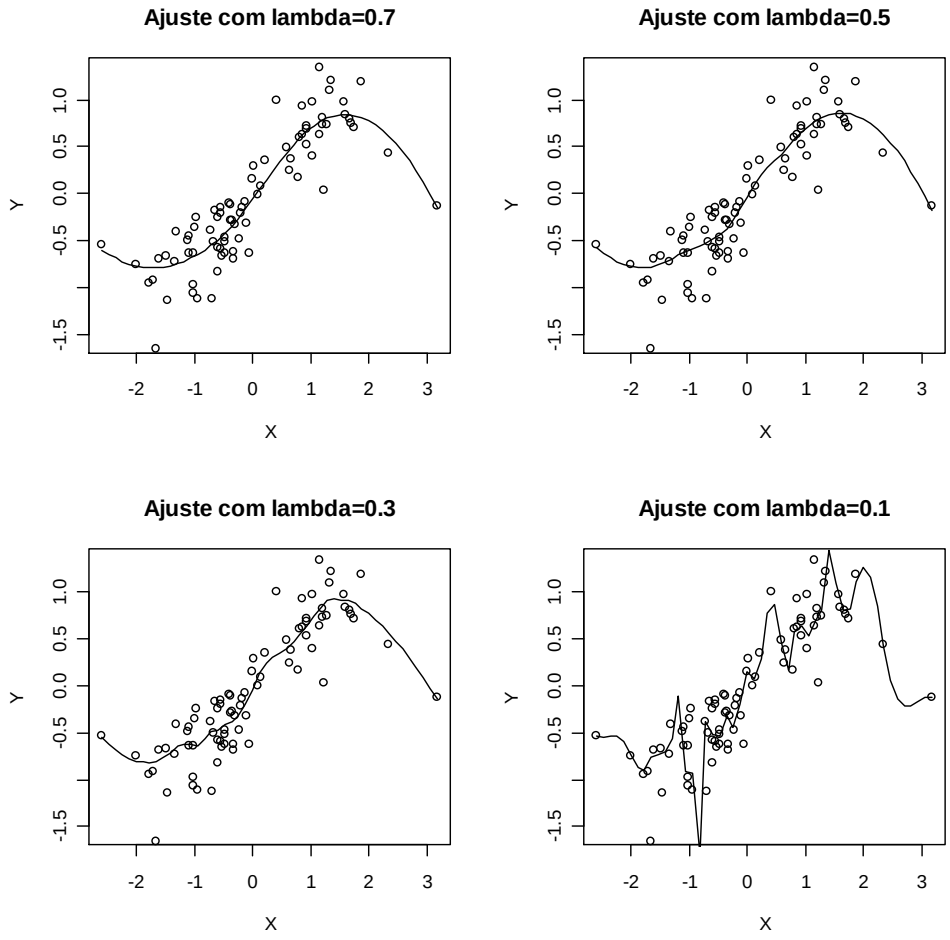

Figura 2.1 – Diagrama de dispersão de X e Y com curva suavizada pelo método *loess* com diferentes valores de λ.

Não existe um critério rígido para a escolha do valor de λ. Na prática estes valores são escolhidos a *priori* através da inspeção visual da curva ou através de um método automático que geralmente testa vários valores de λ para um mesmo conjunto de dados. A Validação Cruzada (*cross-validation*), por exemplo, é uma técnica automática de seleção do parâmetro suavizador; o método consiste em retirar o ponto

(x_i, y_i) da base de dados e calcular $\hat{f}(x_i)$ apenas com os $n-1$ pontos restantes, $i = 1,...,n$. A estatística de validação cruzada é dada por

$$VC(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_{\lambda}^{-i}(x_i))^2 \text{ , } i = 1,...,n \qquad (2.4)$$

onde $\hat{f}_{\lambda}^{-i}(x_i)$ indica o valor estimado para Y quando o ponto (x_i, y_i) é eliminado. A expressão (2.4) é calculada considerando-se um conjunto de valores de λ fixados a *priori*, sendo selecionado o valor de λ que minimize esta expressão. Alguns detalhes podem ser encontrados em Silverman (1985) e Craven e Wahba (1979).

Apesar da validação cruzada e de outras formas de seleção automática para o parâmetro de suavização parecerem bem fundamentadas, suas performances são questionáveis. Hastie e Tibshirani (1990) mostraram, em um estudo simulado, que os valores de λ assim obtidos apresentam grande variabilidade, mesmo para dados gerados a partir de modelos simples, com pequena variância para os erros. Os autores sugerem que a escolha deste parâmetro seja feita com métodos gráficos auxiliados por medidas dos graus de liberdade dos suavizadores.

Dada uma matriz suavizadora S de um suavizador linear com um λ fixado, o número dos graus de liberdade pode ser definido como

$$gl = tr(S). \qquad (2.12)$$

Quanto maior o número de graus de liberdade, menor o valor de λ e, por conseqüência, menor a quantidade de suavização. Existem ainda pelo menos outras duas definições de graus de liberdade – ver em Hastie e Tibshirani (1990). Estas definições, assim como (2.12) são derivadas da analogia dos modelos de regressão linear e podem ser utilizadas com vários propósitos, entre eles comparar os suavizadores descritos nas seções anteriores levando em conta à "quantidade" de suavização ou comparar duas técnicas com base no valor esperado da soma de quadrado residual (ver maiores detalhes em Buja *et al.* 1989); além disso, qualquer uma das expressões para graus de liberdade pode ser usada para auxiliar na escolha de um valor para o parâmetro de suavização.

2.3 O SUAVIZADOR *loess*

Proposto por Cleveland em 1977, o *loess* (l*ocally weighted running line smoother*) é um método de suavização que se baseia no ajuste sucessivo de n modelos de regressão pelo método de mínimos quadrados ponderados (MQP).

Para cada ponto (x_i, y_i) define-se uma vizinhança e aos pontos (x_k, y_k) nessa vizinhança são atribuídos pesos através de uma função U. Esses pesos são utilizados no ajuste de um polinômio por MQP.

A vizinhança de cada (x_i, y_i) é constituída por l pares de observações (x_k, y_k) que possuem as coordenadas x_k mais próximas a x_i. A quantidade l a ser considerada é dada por

$$l = \lambda n \tag{2.15}$$

onde λ, pertencente ao intervalo $(0, 1]$, é o parâmetro de suavização correspondente à proporção do número total de observações a ser utilizado em cada ajuste local.

O grau do polinômio deve ser fixado com base no padrão apresentado pelos dados num diagrama de dispersão. De uma forma geral, se a nuvem de pontos sugere uma tendência sem máximos ou mínimos locais, então um ajuste linear é adequado. Se existem regiões com máximos ou mínimos locais, então um ajuste quadrático geralmente produz uma curva que descreve localmente melhor o padrão dos dados.

A função U que atribui pesos em cada ajuste local do polinômio, tendo (x_i, y_i) como ponto alvo tem a forma

$$u_{x_i,k} = U(h_i^{-1}(x_i - x_k)),\ i = 1,...,n,\ k = 1,...,n \tag{2.16}$$

onde h_i é a distância entre x_i e seu l-ésimo vizinho mais próximo, com l definido em (2.15), ou seja, colocando em ordem crescente as distâncias $|x_i - x_k|$, $i = 1,...,n$, $k = 1,...,n$, h_i é a distância que ocupa a l-ésima posição nesta seqüência. A função U deve ser especificada de forma a possuir as seguintes propriedades:

i. $U(g) > 0$ para $-1 < g < 1$;

ii $U(g)$ é uma função par, isto é, $U(g) = U(-g)$;

iii $U(g)$ é uma função decrescente para $g \geq 0$;

iii $U(g) = 0$ para $|g| = 1$

A função tricúbica dada por

$$U(g) = \begin{cases} \left(1 - |g|^3\right)^3 & para \quad |g| < 1 \\ 0 & caso \ \ contrário \end{cases} \tag{2.17}$$

apresenta as propriedades descritas acima e, de acordo com Cleveland (1979), fornece uma suavização adequada na maioria dos casos.

Com base na função (2.17) obtém-se a matriz de pesos referente ao ponto alvo (x_i, y_i)

$$\mathrm{U}_{x_i} = \text{diagonal } \{\mathrm{u}_{x_i,1}, ..., \mathrm{u}_{x_i,n}\} \tag{2.18}$$

com elementos dados por

$$u_{x_i,k} = \begin{cases} \left(1 - |h_i^{-1}(x_i - x_k)|^3\right)^3 & para \quad |h_i^{-1}(x_i - x_k)| < 1 \\ 0 & caso \ \ contrário \end{cases} . \tag{2.19}$$

Assim, por (2.18), em um ajuste local tendo como ponto alvo (x_i, y_i), este ponto fica associado a um peso 1; os pesos diminuem à medida que os pontos se afastam de (x_i, y_i) e pontos fora da vizinhança de x_i ficam associados a pesos nulos. A reta de regressão assim ajustada $\hat{y} = \hat{\alpha} + \hat{\beta}x$ fornece o valor previsto $\hat{f}(x_i) = \hat{\alpha} + \hat{\beta}x_i$.

Refazendo todos os passos considerando cada uma das n observações (x, y) como ponto alvo, obtêm-se os pontos $(x, \hat{f}(x))$ que formam a curva suavizada.

O *loess* é um suavizador linear; neste caso, os valores previstos de Y obtidos no procedimento de suavização podem ser escritos da forma dada em (2.2). Os elementos da matriz suavizadora S são denotados por

$$S = \begin{bmatrix} s_{x_1,1} & s_{x_1,2} & \cdots & s_{x_1,n} \\ s_{x_2,1} & s_{x_2,2} & \cdots & s_{x_2,n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{x_n,1} & s_{x_n,2} & \cdots & s_{x_n,n} \end{bmatrix} = \begin{bmatrix} s'_{x_1} \\ s'_{x_2} \\ \vdots \\ s'_{x_n} \end{bmatrix}$$

O vetor s'_{x_i}, referente à i-ésima linha da matriz S, corresponde também à i-ésima linha da matriz

$$S_{x_i} = X(X'U_{x_i}X)^{-1}X'U_{x_i}\ , \tag{2.20}$$

construída no ajuste da regressão ponderada local que tem (x_i, y_i) como ponto alvo e matriz de pesos U_{x_i} definida em (2.18), $i = 1,...,n$.

Dessa forma o valor previsto correspondente a x_i pode ser dado como em (2.3) e ser reescrito como

$$s'_{x_i}\, y\,,\ i = 1,...,n\,. \tag{2.21}$$

De forma geral, pode-se mostrar que o elemento ij da matriz suavizadora do *loess* é dado por

$$s_{x_i,j} = \frac{u_{x_i,j}\sum_{j=1}^{n} x_j^2 u_{x_i,j} - u_{x_i,j}(x_i + x_j)\sum_{j=1}^{n} x_j u_{x_i,j} + x_i x_j \sum_{j=1}^{n} x_j^2 u_{x_i,j} \sum_{j=1}^{n} u_{x_i,j}}{\sum_{j=1}^{n} u_{x_i,j} \sum_{j=1}^{n} x_j^2 u_{x_i,j} - \left(\sum_{j=1}^{n} x_j u_{x_i,j}\right)^2}, \tag{2.22}$$

onde $u_{x_i,j}$ é definido de acordo com (2.16).

A expressão 2.22 mostra claramente que os elementos de S dependem apenas das covariáveis e do parâmetro de suavização. Quanto menor o λ, maior o número de elementos iguais a zero na diagonal de U_{x_i} e maior o número de elementos nulos nas linhas de S.

3. MODELOS ADITIVOS GENERALIZADOS

Os suavizadores podem ser utilizados em modelos de regressão com o objetivo de descrever a relação entre a média da variável resposta e as variáveis preditoras. Neste capítulo é feita uma breve introdução dos Modelos Lineares Generalizados (MLG) e descritos os Modelos Aditivos Generalizados (MAG) – modelos de regressão que vêm ganhado destaque na literatura, principalmente por sua flexibilidade – mostrando o emprego dos suavizadores nesta classe.

3.1 Modelos lineares generalizados

Os Modelos Lineares Generalizados (MLG) são formados por um componente aleatório, um componente sistemático e uma função de ligação que "liga" os dois componentes. A resposta Y, componente aleatória do modelo, tem função de densidade de probabilidade (ou função de probabilidade) dada por

$$\rho_Y(y;\Theta;\Phi) = \exp\{\phi[y\Theta - b(\Theta)] + c(y,\Phi)\} \qquad (3.1)$$

onde Θ é chamado parâmetro natural e Φ parâmetro de dispersão, $b(\cdot)$ e $c(\cdot)$ são funções especificadas, $b(\cdot)$ é duas vezes diferenciável e $\Phi^{-1} > 0$. Assume-se também que a esperança de Y, denotada por μ, está relacionada às covariáveis $X_1,...,X_d$ por $g(\mu) = \eta$, onde $\eta = \alpha + X_1\beta_1 + ... + X_v\beta_d$; η é a componente sistemática do modelo linear generalizado chamada preditor linear e $g(\cdot)$ é a função de ligação. Um caso particular ocorre quando o preditor linear coincide com o parâmetro Θ, isto é, $\eta = \Theta$; neste caso, a função de ligação é chamada ligação canônica. Estas ligações desempenham papel muito importante na teoria dos MLG's e muitas vezes são escolhidas por possuírem propriedades estatísticas e matemáticas convenientes (ver Paula, 2000).

McCullagh e Nelder (1989) mostraram que se Y tem função densidade de probabilidade dada por (3.1) seu valor esperado e sua variância estão relacionados ao parâmetro natural da seguinte forma

$$\mu = \frac{\partial b(\Theta)}{\partial \Theta}$$

e (3.2)

$$Var(Y) = \Phi^{-1} \frac{\partial \mu}{\partial \Theta} = \Phi^{-1} V ,$$

onde V é chamada função de variância.

A estimação do vetor dos $d+1$ parâmetros dos MLG, $\boldsymbol{\beta} = (\alpha, \beta_1, ..., \beta_d)'$ é feita por máxima verossimilhança e a estimativa deste vetor é calculada resolvendo-se as seguintes equações escore

$$\sum_{i=1}^{n} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) V_i^{-1} (y_i - \mu_i) = 0, \ j = 1, ..., d \qquad (3.3)$$

onde $V_i = Var(Y_i)$ e $x_{i0} = 1$.

As equações (3.3) podem ser resolvidas pelo método *scoring* de Fisher (McCullagh e Nelder, 1989); um procedimento equivalente e também conveniente para a resolução destas equações é o procedimento de mínimos quadrados reponderados iterativamente (MQRI). Dado um vetor inicial $\boldsymbol{\beta}^{(0)}$, calcula-se a resposta modificada $z_i^{(m)}$ (3.4) e os pesos $w_i^{(m)}$ (3.5) no passo m.

$$z_i^m = \eta_i^{(m)} + (y_i - \mu_i^{(m)}) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)_{(m)} \qquad (3.4)$$

$$w_i^{(m)} = \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 (V_i^{(m)})^{-1}, \ i = 1, ..., n \qquad (3.5)$$

onde $\eta_i^{(m)} = \alpha^{(m-1)} + \sum_{j=1}^{d} \beta_j^{(m-1)} x_{ij}$, $\mu_i^{(m)} = g^{-1}(\eta_i^{(m)})$ e $V_i^{(m)} = \left(\frac{\partial \mu_i}{\partial \Theta_i} \right)_{(m)}$.

O vetor $\boldsymbol{\beta}^{(m)} = (X'W^{(m)}X)^{-1} X'W^{(m)} z^{(m)}$, $m = 1,2...,$ é obtido da regressão de $z_i^{(m)}$ em x_i com peso $w_i^{(m)}$, $i = 1,...,n$. A matriz X é a matriz de planejamento de dimensão $n \times (v+1)$, cuja i-ésima linha corresponde ao vetor $(1,\ x_i)$, $z^{(m)} = (z_1^{(m)},...,z_n^{(m)})$ e $W^{(m)} = \text{diagonal}\{w_1^{(m)},...,w_n^{(m)}\}$. Considerando agora o novo vetor $\boldsymbol{\beta}$, o critério de parada no *scoring* de Fisher se baseia numa medida de proximidade das estimativas; assim o processo é repetido até que

$$\frac{\sum_{j=1}^{d} \| \beta_j^{(m)} - \beta_j^{(m-1)} \|}{\sum_{j=1}^{d} \| \beta_j^{(m-1)} \|} \leq \delta, \qquad (3.6)$$

para um valor $\delta > 0$ pré-estabelecido.

3.2 Modelos aditivos generalizados

Na Seção 3.1 foi visto que o preditor linear é uma função linear de cada uma das variáveis preditoras $X_1,...,X_d$. No entanto, uma relação menos rígida pode ser adotada substituindo o termo linear correspondente a cada covariável por uma função não especificada dessa variável, obtendo-se o preditor aditivo

$$\eta = g(\mu) = \alpha + f(X_1) + ... + f(X_d). \qquad (3.7)$$

A classe de modelos assim obtida é denominada Modelos Aditivos Generalizados e pode ser vista como uma generalização dos MLG's.

O preditor (3.7) corresponde a um modelo totalmente não paramétrico. Porém, também fazem parte dos MAG's os modelos semiparamétricos, cujo preditor combina

formas paramétricas de algumas das r variáveis preditoras com termos não paramétricos das outras $(d-r)$ variáveis. Nestes casos o preditor pode ser escrito como

$$\eta = g(\mu) = \alpha + \beta_1 X_1 + ... + \beta_r X_r + f(X_{r+1}) + ... + f(X_d). \quad (3.8)$$

Considerando n observações $(X_{1i}, X_{2i}, ..., X_{di}, Y_i)$ do modelo Poisson semiparamétrico

$$Y_i \sim Poisson(\mu_i)$$

$i = 1, ..., n$, o preditor η assume a forma

$$\eta_i = \log(\mu_i) = \alpha + \beta_1 X_{1i} + ... + \beta_r X_{ri} + f(X_{(r+1)i}) + ... + f(X_{di}). \quad (3.9)$$

A estimação dos MAG's e testes de hipóteses sobre os componentes do modelo foram desenvolvidos em analogia a procedimentos utilizados com esses objetivos nos MLG's, modificando-os de forma que as funções f em (3.7) sejam estimadas por meio da utilização de suavizadores.

O processo de ajuste dos MAG's baseia-se na combinação de dois procedimentos iterativos: o procedimento de ponderação local – PPL – (*local scoring*) e o retroajuste (*backfitting*).

O PPL é um procedimento similar ao procedimento MQIR utilizado no ajuste dos MLG's e corresponde a um ciclo externo no processo e estimação necessário para o ajuste de um modelo com estrutura semelhante a um MLG; o retroajuste, um ciclo interno ao PPL, é o algoritmo responsável pela estimação de cada função f por meio da utilização de suavizadores ponderados.

3.2.1 Ajuste dos modelos não paramétricos

O ajuste de um MAG pode ser efetuado nas três etapas do algoritmo PPL esquematizado a seguir. O passo 2 é o retroajuste. Sejam $\alpha^{(m)}$ o valor estimado de α e $f_j^{(m)}$ a estimativa de f_j, $j = 1, ..., d$, no passo m do procedimento iterativo, com $m = 0$ denotando o passo inicial.

Dados valores iniciais para $\alpha^{(0)} = g\left(\sum_{i=1}^{n} \frac{y_i}{n}\right)$ e $f_1^{(0)} = \ldots = f_d^{(0)} = 0$, os valores da variável modificada $z_i^{(m)}$ e dos pesos $w_i^{(m)}$ são calculados da mesma forma que em (3.4) e (3.5), porém, neste caso, $\eta_i^{(m)} = \alpha^{(m-1)} + \sum_{j=1}^{d} f_j^{(m-1)} x_{ij}$. O algoritmo consiste em iterar as seguintes etapas para $m = 1,2,\ldots$

Passo 1 – Para $i = 1,\ldots,n$ calcular a variável resposta modificada e os pesos:

$$z_i^m = \eta_i^{(m)} + (y_i - \mu_i^{(m)})\left(\frac{\partial \eta_i}{\partial \mu_i}\right)_{(m)} \tag{3.10}$$

$$w_i^{(m)} = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 (V_i^{(m)})^{-1}, \tag{3.11}$$

onde $\eta_i^{(m)} = \alpha^{(m-1)} + \sum_{j=1}^{d} f_j^{(m-1)} x_{ij}$, $\mu_i^{(m)} = g^{-1}(\eta_i^{(m)})$ e $V_i^{(m)} = \left(\frac{\partial \mu_i}{\partial \Theta_i}\right)_{(m)}$.

Passo 2 – Os valores $f_1^{(m)},\ldots, f_d^{(m)}$ são obtidos com o uso do algoritmo de retroajuste:

Inicializa-se $\alpha^{(m)}$ com a média amostral da variável modificada no passo m, $\alpha^{(m)} = \bar{z}^{(m)} = \sum_{i=1}^{n} \frac{z_i^{(m)}}{n}$ e $f_j^{\,0} = f_j^{(m-1)}$, $j = 1,\ldots,d$ e calcula-se

$$f_{j(v)}^{(m)} = S_j^{(m)} r_{j(v)}^{(m)}, \tag{3.12}$$

até que $\| f_{j(v)}^{(m)} - f_{j(v-1)}^{(m)} \| \leq \varepsilon$, para um valor $\varepsilon > 0$ pré-estabelecido. Em (3.12), $r_{j(v)}^{m} = (r_{ij(v)}^{m},\ldots,r_{ij(v)}^{m})$ é o vetor de resíduos parciais com elementos dados por

$$r_{ij}^{(m)} = z_i^{(m)} - \bar{z}^{(m)} - \sum_{k=1}^{j-1} f_{k(v)}^{(m)}(x_{ik}) - \sum_{k=j+1}^{d} f_{k(v-1)}^{(m)}(x_{ik}),$$

para $v = 1,2,\ldots$, e $S_j^{(m)}$, de dimensão $n \times n$, é a matriz suavizadora ponderada relativa à j-ésima covariável. No caso do suavizador *loess*, a i-ésima linha de $S_j^{(m)}$ corresponde à i-ésima linha da matriz suavizadora ponderada $S_{x_i}^{(m)} = X(X'A_{x_i}X)^{-1}X'A_{x_i}$, onde $X = (\underline{1},\ X_j)$, $\underline{1}$ é um vetor $(n \times 1)$ de valores unitários, e A_{x_i} =diagonal$\{u_{x_i,1}w_1^{(m)},\ldots,u_{x_i,n}w_n^{(m)}\}$ com $u_{x_i,j}$ e $w_j^{(m)}$ definidos, respectivamente, em (2.19) e (3.11).

Passo 3 – Os passos 2 e 3 são repetidos até que

$$\frac{\sum_{j=1}^{d} \| f_j^{(m)} - f_j^{(m-1)} \|}{\sum_{j=1}^{d} \| f_j^{(m-1)} \|} \leq \delta,$$

para um valor $\delta > 0$ pré-estabelecido.

O algoritmo de retroajuste corresponde ao método de Gauss-Seidel para resolver o sistema de equações lineares:

$$\begin{bmatrix} I & S_1^{(m)} & S_1^{(m)} & \cdots & S_1^{(m)} \\ S_2^{(m)} & I & S_2^{(m)} & \cdots & S_2^{(m)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_d & S_d^{(m)} & S_d^{(m)} & \cdots & I \end{bmatrix} \begin{bmatrix} f_1^{(m)} \\ f_2^{(m)} \\ \vdots \\ f_d^{(m)} \end{bmatrix} = \begin{bmatrix} S_1^{(m)} z^{(m)} \\ S_2^{(m)} z^{(m)} \\ \vdots \\ S_d^{(m)} z^{(m)} \end{bmatrix}. \qquad (3.13)$$

O retroajuste é um método eficiente para resolvê-lo, principalmente quando o número de parâmetros é grande. Motivações para a utilização dessas equações na obtenção de $f_1^{(m)},\ldots, f_d^{(m)}$ podem ser encontradas em Hastie e Tibshirani (1990).

Se a variável resposta segue uma distribuição Normal, a função de ligação é a identidade, então $Z = Y$, $W = I$ e o procedimento MQIR é substituído por um método direto, ou seja, apenas o ciclo interno, correspondente ao retroajuste, é necessário. No caso de um MAG com apenas uma função não especificada, isto é, $d = 1$ em (3.7), o

algoritmo de retroajuste não é necessário, pois $f^{(m)}$ pode ser obtido diretamente com a utilização de um alisador ponderado aplicado aos resíduos aplicado aos resíduos $r_i^{(m)} = z_i^{(m)} - \bar{z}^{(m)}$ em função de x_i, $i = 1,...,n$ com matriz de pesos $W^{(m)}$.

Embora o retroajuste seja um algoritmo eficiente para resolver (3.13), pelo menos conceitualmente, a solução direta pode ser utilizada. Estimativas para $f_1,..., f_d$ podem ser obtidas pela relação

$$\hat{f} = M^{-1}Cz$$

com

$$\hat{f} = \begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \\ \vdots \\ \hat{f}_d \end{bmatrix}, \quad M = \begin{bmatrix} I & S_1 & S_1 & \cdots & S_1 \\ S_2 & I & S_2 & \cdots & S_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_d & S_d & S_d & \cdots & I \end{bmatrix} \text{ e } C = \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_d \end{bmatrix},$$

se a inversa de M existe.

Em particular, escrevendo

$$R_j = E_j M^{-1} C, \quad j = 1,...,d,$$

onde E_j denota uma matriz de dimensão $(n \times nd)$ composta por d "blocos" de dimensão $(n \times n)$, com todos os blocos nulos à exceção do j-ésimo, que é uma matriz identidade de tal maneira que

$$\hat{f} = R_j z.$$

Seja $f = f_1 + ... + f_d$, então

$$\hat{f} = R_1 z + ... + R_d z = R_{ND} z,$$

onde $R_{ND} = R_1 + ... + R_d$ é a matriz suavizadora ponderada que produz $\hat{f}$ a partir de z.

Para modelos que envolvem apenas duas matrizes suavizadoras em seu ajuste, $d = 2$, Hastie e Tibshirani (1990) fornecem expressões mais simples para R_1 e R_2:

$$R_1 = I - (I - S_1 S_2)^{-1}(I - S_1)$$
$$R_2 = I - (I - S_2 S_1)^{-1}(I - S_2). \tag{3.14}$$

Nesta caso,

$$R_{ND} = (R_1 + R_2) = I - (I - S_2)(I - S_1 S_2)^{-1}(I - S_1).$$

Expressões recursivas para modelos envolvendo mais de dois suavizadores foram deduzidas por Opsomer (2000). O custo computacional para obter R_j, $j = 1,...,d$, a partir destas expressões é, entretanto, elevado.

O algoritmo do retroajuste é mais eficiente do ponto de vista computacional, mas a solução direta pode ser utilizada como uma alternativa para obter expressões para $\hat{f}_j$ e $\hat{\eta}$ que tornem mais simples o estudo de suas propriedades estatísticas.

A convergência do procedimento de ajuste dos MAG's está condicionada à convergência do retroajuste, uma vez que o PPL não apresenta, em geral, problemas dessa ordem (Hastie e Tibshirani, 1990). Resultados sobre a convergência desse procedimento podem ser encontrados em Buja *et al.*. (1989) e Opsomer (2000).

3.2.2 Ajuste dos modelos semiparamétricos

Considere o modelo semiparamétrico

$$g(\mu) = \eta = \alpha + \beta_1 X_1 + ... + \beta_r X_r + f(x_{r+1}) + ... + f(x_d). \qquad (3.15)$$

Os parâmetros $\alpha, \beta_1, ..., \beta_r$ e as funções $f_{r+1}, ..., f_d$ podem ser estimados através do PPL e do retroajuste. Dado os valores iniciais $\boldsymbol{\beta}^{(0)} = (\alpha^{(0)}, \beta_1^{(0)}, ..., \beta_r^{(0)})'$ e $f_{r+1}^{(0)}, ..., f_d^{(0)}$ estimativas para $\boldsymbol{\beta}$ e $f_{r+1}, ..., f_d$ são obtidas resolvendo-se, iterativamente, as seguintes equações

$$\beta^{(m)} = (X'W^{(m)}X)^{-1} X'W^{(m)} \left(z^{(m)} - \sum_{j=r+1}^{d} f_j^{(m)} \right) \qquad (3.16)$$

e

$$f_j^{(m)} = S_j^{(m)} \left(z^{(m)} - X\beta^{(m)} - \sum_{\substack{i=r+1 \\ i \neq j}}^{d} f_i^{(m)} \right), \; j = r+1, ..., d \qquad (3.17)$$

onde $X = (\underline{1}, X_1, ..., X_r)$ é a matriz de especificação correspondente aos termos paramétricos do modelo com X_j, $j = 1, ..., r$, denotando o vetor dos valores observados da j-ésima covariável e $S_j^{(m)}$, $j = r+1, ..., d$, a matriz suavizadora ponderada relativa à j-ésima covariável no m-ésimo passo do procedimento iterativo PPL. Após obter as estimativas $\boldsymbol{\beta}^{(m)}$ e $f_j^{(m)}$, $j = r+1, ..., d$, pelo retroajuste, valores de $\eta^{(m+1)}$, $\mu^{(m+1)}$, $z^{(m+1)}$ e $w^{(m+1)}$ são calculados pelo PPL e o processo é repetido até a convergência.

Quando existe apenas uma função não especificada, isto é, $d = r+1$ em (3.8) as expressões (3.16) e (3.17) se reduzem a

$$\beta^{(m)} = (X'W^{(m)}X)^{-1} X'W^{(m)} \left(z^{(m)} - f^{(m)} \right) \qquad (3.18)$$

e

$$f^{(m)} = S^{(m)} \left(z^{(m)} - X\beta^{(m)} \right). \qquad (3.19)$$

Substituindo-se (3.19) em (3.18) obtém-se

$$\beta^{(m)} = [X'W^{(m)}(I - S^{(m)})X]^{-1} X'W^{(m)}\left(I - S^{(m)}\right)z^{(m)} \qquad (3.20)$$

e dessa forma o retroajuste pode ser evitado. Após a obtenção de uma estimativa $\beta^{(m)}$ segundo (3.20), uma estimativa $f^{(m)}$ é calculada segundo (3.19). Por intermédio do PPL estimam-se novos valores $\eta^{(m+1)}$, $\mu^{(m+1)}$, $z^{(m+1)}$ e $w^{(m+1)}$ e o processo é repetido até a convergência.

Um fato pouco evidenciado na literatura é que, em geral, os estimadores dos modelos semiparamétricos não são identificáveis quando incluem o intercepto (Opsomer e Ruppert, 1999). Neste caso, quando a soma dos elementos das linhas de S é igual a 1, o que acontece no caso *loess*, $X'W^{(m)}(I - S^{(m)})X$ e $X'W^{(m)}(I - R_{ND}^{(m)})X$ são singulares e uma solução simples para esse problema é substituir as matrizes $S_j^{(m)}$ por matrizes centradas, da forma $(I - \underline{1}\underline{1}'/n)S_j^{(m)}$; este procedimento faz com que a média de $\hat{f}$ seja igual a zero em cada passo e o modelo torne-se identificável.

3.2.3 Função desvio

Nos MLG's o desvio para o modelo ajustado $\hat{\mu}$, $D(y;\hat{\mu})$ e o desvio parcial para dois modelos ajustados $\hat{\mu}_1$ e $\hat{\mu}_2$, $D(\hat{\mu}_1;\hat{\mu}_2)$, são quantidades bem conhecidas, utilizadas, respectivamente, para avaliar a qualidade do ajuste e comparar dois modelos ajustados. Sem perda de generalidade, pode-se escrever $D(y;\hat{\eta})$ como sendo o desvio para o modelo ajustado, uma vez que $\hat{\mu}$ está relacionado com $\hat{\eta}$ por meio da função de ligação $g(\mu) = \eta$.

No caso dos MAG's, o desvio também pode ser usado como uma medida de ajuste e a comparação entre modelos. No modelo Poisson, o desvio é dado por $D(y;\hat{\mu}) = 2\sum_{i=1}^{n}(y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)$.

Embora as distribuições assintóticas dessas estatísticas não tenham sido determinadas, Hastie e Tibshirani (1990) mostram, por simulações, que a distribuição

χ^2 é uma boa aproximação. Uma medida aproximada para os graus de liberdade de $D(y;\hat{\eta})$ é dada por

$$gl = n - 1 - \sum_{j=1}^{d}[tr(S_j) - 1]. \tag{3.21}$$

3.2.4 Seleção do parâmetro de suavização

Um possível critério para selecionar parâmetros de suavização em um MAG com p termos não paramétricos é baseado na estatística VC

$$VC = \frac{1}{n}\sum_{i=1}^{n} D(y_i;\hat{\mu}_{\lambda}^{-1}). \tag{3.22}$$

A idéia é minimizar esta quantidade sob $\lambda_1,...,\lambda_d$, os parâmetros de suavização para cada curva ajustada; no entanto o custo computacional é muito grande já que são necessárias n aplicações completas do procedimento PPL para cada valor pré-fixado dos parâmetros de suavização.

Uma opção para a seleção destes parâmetros é baseado na estatística

$$AIC = \frac{1}{n}\sum_{i=1}^{n} D(y_i;\hat{\mu}_i) + \frac{2}{n}tr(R)\Phi \tag{3.23}$$

inspirada no critério de informação de Akaike (ver Hastie e Tibshirani, 1990 e Hastie,1992), valores pequenos desta estatística indicam um bom ajuste do modelo. O valor Φ, definido como parâmetro de dispersão na equação (3.1), está associado à distribuição de Y: se a variável resposta tem distribuição Normal, com variância σ^2, $\Phi = 1/\sigma^2$; se a variável resposta tem distribuição Poisson, $\Phi = 1$ (veja resultados para outras distribuições em Paula, 2000). Embora muito empregada na prática, não existem resultados teóricos sobre a adequação de sua utilização como um critério para seleção

do parâmetro de suavização. O ganho computacional é dado pelo fato de que a estatística AIC requer somente uma aplicação do PPL para cada valor $\lambda_1,...,\lambda_d$.

4. MODELOS AUTO-REGRESSIVOS GENERALIZADOS

Introduzido por Davis *et al.* (1999), o GLARMA – *autoregressive moving average generalized linear models* – é um modelo utilizado em estudos de séries temporais sendo capaz de capturar uma gama de estruturas de dependência nas observações. Neste capítulo o GLARMA é estendido para uma classe de modelos auto-regressivos aditivos generalizados para séries de contagem cuja distribuição condicional dada as observações passadas e variáveis explicativas segue uma distribuição de Poisson.

4.1 Modelos Poisson auto-regressivo média móvel linear generalizados (Poisson-GLARMA)

A classe GLARMA é uma classe de modelos que estende o processo ARMA (auto-regressivo médias móveis) Gausssiano de séries temporais para um modelo mais flexível para séries de contagem não-Gaussianas. A variável dependente é suposta ter uma distribuição condicional na família exponencial dado todo o passado do processo. Para introduzir o modelo Poisson-GLARMA, assuma que a observação Y_i dado o passado histórico F_{i-1} tem distribuição Poisson com média μ_i,

$$Y_i \mid F_{i-1} \sim Poisson(\mu_i),\ i = 1,...,n.$$

A função de ligação η segue a forma

$$\eta_i = \log(\mu_i) = \alpha + \beta_1 X_{1i} + ... + \beta_d X_{di} + T_i \qquad (4.1)$$

onde T_i, responsável pela estrutura de correlação de Y_i, é dado por $T_i = \sum_{j=1}^{\infty} \pi_j \varepsilon_{i-j}$, e ε_i assume a forma dos resíduos de Pearson

$$\varepsilon_i = \frac{Y_i - \mu_i}{\sqrt{\mu_i}}. \qquad (4.2)$$

A estrutura média móvel infinita de T_i pode ser especificada em termos de um número finito de parâmetros. Uma forma de parametrizar os pesos médias móveis π_j é e expressá-los como coeficientes de um filtro auto-regressivo médias móveis (Box e Jenkins, 1976)

$$\pi(B) = \sum_{i=1}^{\infty} \pi_i B^i = \frac{\theta(B)}{\phi(B)} - 1$$

onde

$$\phi(B) = 1 - \phi_1 B^1 - \phi_2 B^2 - \ldots - \phi_p B^p$$

$$\theta(B) = 1 + \theta_1 B^1 + \theta_2 B^2 + \ldots + \theta_q B^q$$

são os polinômio auto-regressivo e polinômio média móvel com todas raízes fora do círculo unitário. Dessa forma T_i pode ser expresso por

$$T_i = (\sum_{j=1}^{p} \phi_j T_{i-j}) + \varepsilon_i + \sum_{j=1}^{q} \theta_j \varepsilon_{i-j} \ . \qquad (4.3)$$

A estimação dos parâmetros do GLARMA, $\varphi = (\beta', \xi')'$, onde $\xi = (\phi', \theta')'$, é feita conjuntamente através da função de verossimilhança, maximizada pelo método numérico Newton-Raphson (Davis *et al.*, 2003).

Considere ℓ a densidade Poisson condicional de Y_i dado F_{i-1} e defina $L_i(\varphi) = \log \ell(y_i \mid F_{i-1})$. A log-verossimilhança pode ser escrita como

$$\sum_{i=1}^{n} L_i(\varphi)$$

que, ignorando termos que não envolvem os parâmetros, se torna

$$L(\varphi) = \sum_{i=1}^{n} (Y_i \eta_i\ (\varphi) - \varepsilon^{\eta_i\ (\varphi)}) \tag{4.4}$$

onde

$$\log(\mu_i) = \eta_i\ (\varphi) = \alpha + \beta_1\ X_{1i} + ... + \beta_d\ X_{di} + \sum_{j=1}^{\infty} \pi_i(\ \xi)\ \varepsilon_{i-j}\ (\varphi)$$

e

$$\varepsilon_i\ (\varphi) = (Y_i - \mu_i) \Big/ \sqrt{\mu_i}\ .$$

Para facilitar a compreensão dos cálculos, a dependência de ε_i em φ foi desconsiderada. A primeira e a segunda derivadas de L são dadas pelas expressões (4.5) e (4.6)

$$\frac{\partial L}{\partial \varphi} = \sum_{i=1}^{n} (Y_i - \mu_i) \frac{\partial \eta_i}{\partial \varphi} = \sum_{i=1}^{n} \varepsilon_i \sqrt{\mu_i}\ \frac{\partial \eta_i}{\partial \varphi} \tag{4.5}$$

$$\begin{aligned} \frac{\partial^2 L}{\partial \varphi \partial \varphi'} &= \sum_{i=1}^{n} \left[(Y_i - \mu_i) \frac{\partial^2 \eta_i}{\partial \varphi \partial \varphi'} - \mu_i \frac{\partial \eta_i}{\partial \varphi} \frac{\partial \eta_i}{\partial \varphi'} \right] \\ &= \sum_{i=1}^{n} \left[\varepsilon_i \sqrt{\mu_i}\ \frac{\partial^2 \eta_i}{\partial \varphi \partial \varphi'} - \mu_i \frac{\partial \eta_i}{\partial \varphi} \frac{\partial \eta_i}{\partial \varphi'} \right]. \end{aligned} \tag{4.6}$$

Maiores detalhes sobre expressões úteis no cálculo dessas derivadas e resultados assintóticos das estimativas dos parâmetros podem ser encontrados em Davis *et al.* (2003).

Para inicializar o método recursivo de Newton Raphson na maximização numérica da log-verossimilhança $L_i(\varphi)$, Davis *et al.* (2003) sugerem que os valores obtidos das estimativas do GLARMA sem os termos auto-regressivos média móveis sejam utilizados como valores iniciais. A convergência, na maioria dos casos, ocorre após 10 iterações. A matriz de covariância dos estimadores é estimada por

$$\hat{\Omega} = -\left(\frac{\partial^2 L(\hat{\theta})}{\partial \varphi \partial \varphi'} \right)^{-1} . \qquad 4.7$$

Maiores detalhes sobre as condições de estacionaridade, propriedades, estimação e inferência dos modelos GLARMA podem ser vistos em Benjamin *et al.* (2003) e Drescher (2005).

4.2 Modelos Poisson auto-regressivos aditivos generalizados (Poisson MAG-AR)

Considere que Y_i dado o passado histórico F_{i-1} tem distribuição Poisson com média μ_i como na Seção 4.1. Utilizando a construção MAG semiparamétrico e adicionando uma estrutura de correlação entre os dados proposta no modelo GLARMA, a equação (4.1) pode ser reescrita como

$$\eta_i = \log(\mu_i) = \alpha + \beta_1 X_{1i} + ... + \beta_r X_{ri} + f(X_{(r+1)i}) + ... + f(X_{di}) + T_i , \qquad (4.8)$$

com T_i sendo um processo auto-regressivo de ordem p, AR(p),

$$T_i = (\sum_{j=1}^{p} \phi_j T_{i-j}) + \varepsilon_i ,$$

ε_i definido em (4.2) e $i = 1,...,n$.

Muitos pesquisadores vêm utilizando o MAG na modelagem de séries temporais ignorando a dependência entre as observações. O modelo MAG-AR além de contar com a vantagem que o MAG oferece em eliminar a necessidade de especificar uma forma paramétrica para a associação de algumas covariáveis com a variável dependente ainda é capaz de capturar estruturas de correlação entre os dados.

A estimação do MAG-AR foi desenvolvida em analogia aos procedimentos de estimação do GLARMA, descritos na Seção 4.1, modificando-os de forma que as

funções f em (4.8) sejam estimadas por meio da utilização de suavizadores como nos procedimentos de estimação dos MAG's, descritos na Seção 3.2.1.

Para verificar se o procedimento proposto realmente funciona em processos que exibem uma estrutura auto-regressiva, um exercício de simulação simples foi implementado para T_i seguindo um processo AR(1). Assim, foram geradas uma variável resposta Y seguindo um processo de Poisson e duas séries temporais como variáveis explicativas, X_1 – relacionada linearmente com a variável resposta – e X_2 – que apresenta uma relação não-linear com Y. A Variável X_1 é um modelo auto-regressivo (AR) de ordem 1 da forma $X_{1i} = 0{,}4X_{1(i-1)} + e_i$, $e_i \sim N(0{,}1)$ e $X_2 = \cos(2\pi i/12) + v_i$, $v_i \sim N(0;0{,}01)$, $i = 1,...,100$. Considere Y_i gerado a partir de uma Poisson com média $\mu_i = \exp\{\alpha + \beta_1 x_{1i} + x_{2i} + T_i\}$, com α fixado em 0,01, β_1 fixado em 0,08 e T_i um AR de ordem 1 com $\phi = 0{,}6$. Os dados simulados foram ajustados através do MAG semiparamétrico, ignorando a correlação existente nas observações, e também através do MAG-AR(1), com o objetivo de se comparar as duas modelagens.

A Figura 4.1 apresenta a série Y_i gerada com a estrutura descrita acima, assim

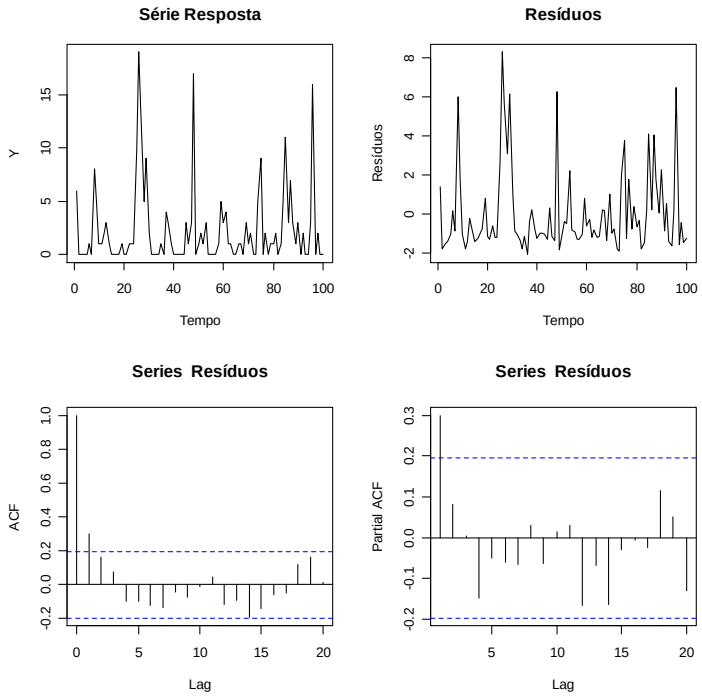

Figura 4.1 – Avaliação residual da série Y modelada via MAG

como gráficos de resíduos obtidos através do ajuste de um MAG a esta série. Dos gráficos de autocorrelação (ACF) e autocorrelação parcial (*partial* ACF) é clara a presença de um AR(1) nos resíduos, estrutura não capturada pelo modelo.

Para corrigir este problema, os dados são então ajustados através do MAG-AR(1) – os gráficos de resíduos são apresentados na Figura 4.2. Quando a nova modelagem é aplicada, a estrutura auto-regressiva é "captada" pelo modelo e os resíduos se tornam um ruído branco, logo parece que o modelo proposto é capaz de incorporar a estrutura auto-regressiva, fazendo com que se obtenha um melhor ajuste para as séries envolvidas.

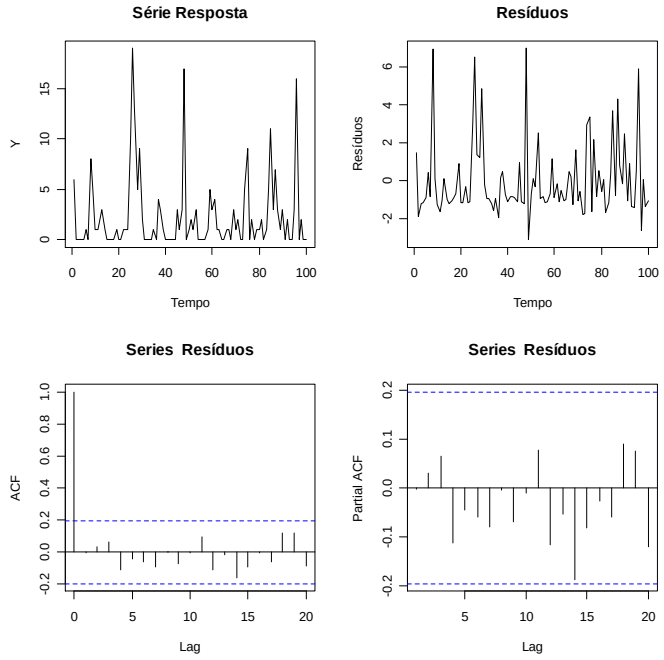

Figura 4.2 – Avaliação residual da série Y, modelada via MAG-AR

5. TÉCNICA BOOTSTRAP

Quando se deseja medir a precisão de estimadores para os parâmetros desconhecidos de uma dada distribuição, geralmente calcula-se uma medida que expresse a variabilidade dos mesmos. Mas se a distribuição exata do estimador é desconhecida ou se o pesquisador tem acesso apenas à sua distribuição assintótica, este cálculo pode ser complicado. Há mais de duas décadas surgiu um procedimento computacional – *Bootstrap* – uma técnica de reamostragem que pode ser utilizada para aproximar a distribuição teórica pela distribuição empírica de uma amostra finita de observações (Efron, 1979). Porém, sendo um método numérico, a sua operacionalidade somente se tornou viável com o advento dos computadores.

Em séries temporais, devido ao fato das observações serem correlacionadas, a aplicação desta técnica requer vários cuidados e a reamostragem direta das observações não pode ser feita. Nestes casos pode-se aplicar o *Bootstrap* amostrando os resíduos diretamente de sua distribuição (*Bootstrap* paramétrico) ou reamostrando os resíduos do modelo ajustado (*Bootstrap* não-paramétrico).

Em MAG's e GLARMA's o uso da técnica é ainda pouco aplicada e discutida. Em estudo recente, Härdle *et al.* (2004) mostram como o procedimento pode ser utilizado na correção do vício das estimativas paramétricas e não-paramétricas dos MAG's, em testes de hipótese e na construção de bandas de confiança.

5.1 *Bootstrap* nas observações

A proposta original da técnica *bootstrap* é a reamostragem direta, com reposição, das observações. Para ilustrar a técnica na modelagem aditiva generalizada, considere Y_i, $X_{1i},...X_{di}$, $i = 1,...,n$, vetores de dados independentes. Considere ainda que Y_i tenha distribuição Poisson e que sua média possa ser modelada por um MAG semiparamétrico através da relação

$$g[E(Y \mid X)] = g(\mu) = \log(\mu_i) = \alpha + \beta_1 X_{1i} + ... + \beta_r X_{ri} + f(X_{(r+1)i}) + ... + f(X_{di}),$$

$$i = 1,...,n. \qquad (5.1)$$

A técnica *bootstrap* consiste em reamostrar os pontos ($y_i, x_{1i},..., x_{di}$), com reposição, obtendo vetores bootstrap Y_i^*, $X_{1i}^*,..., X_{di}^*$.

5.2 *Bootstrap* não paramétrico nos resíduos

Com o objetivo de não reamostrar diretamente uma série temporal devido a não independência das observações, uma das alternativas é reamostrar os resíduos utilizando o método *bootstrap* (Efron e Tibshirani, 1993): inicialmente ajusta-se um modelo aos dados e reamostra-se os resíduos (que devem ser independentes e identicamente distribuídos).

A abordagem não-paramétrica é assim classificada por não utilizar nenhuma suposição quanto à distribuição dos resíduos ao reamostrá-los; neste caso usa-se uma distribuição empírica. Os procedimentos *bootstrap* não-paramétrico para o MAG e para o MAG-AR utilizando a distribuição de Poisson são descritas a seguir.

Seja Y_i, $i = 1,...,n$, um vetor de observações independentes com distribuição Poisson e esperança modelada por um MAG,

$$g[E(Y \mid X)] = g(\mu) = \log(\mu_i) = \alpha + \beta_1 X_{1i} + ... + \beta_r X_{ri} + f(X_{(r+1)i}) + ... + f(X_{di}),$$
$$i = 1,...,n \qquad (5.2)$$

onde $X_1,..., X_d$ são covariáveis também independentes.

Após estimar os parâmetros β's e as funções arbitrárias f's os resíduos de Pearson podem ser obtidos através da expressão

$$\varepsilon_i = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}},\ i = 1,...,n, \qquad (5.3)$$

onde $\hat{\mu}_i = \exp\{\hat{\alpha} + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_r x_{ri} + \hat{f}(x_{(r+1)i}) + ... + \hat{f}(x_{di})\}$. Em seguida reamostra-se, com reposição, ε_i atribuindo-se a cada um uma massa de probabilidade igual a $1/n$. Dessa forma obtêm-se os resíduos *bootstrap* ε_i^*. A partir daí, conforme expressão (5.4), constrói-se, recursivamente, a série *bootstrap* Y_i^*

$$Y_i^* = \varepsilon_i^* \sqrt{\hat{\mu}_i} + \hat{\mu}_i \ , \ i = 1,...,n, \tag{5.4}$$

Considerando Y_i, $X_{1i},...,X_{di}$, $i = 1,...,n$, séries temporais com distribuição Poisson e esperança modelada por um MAG-AR,

$$g[E(Y \mid X)] = g(\mu) = \log(\mu_i) = \alpha + \beta_1 X_{1i} + ... + \beta_r X_{ri} + f(X_{(r+1)i}) + ... + f(X_{di}) + T_i$$
$$i = 1,...,n \tag{5.5}$$

e os resíduos de Pearson definidos como em (5.3) com $\hat{\mu}_i = \exp\{\hat{\alpha} + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_r x_{ri} + \hat{f}(x_{(r+1)i}) + ... + \hat{f}(x_{di}) + T_i\}$ reamostra-se, com reposição, ε_i, cada um com a mesma massa de probabilidade, obtendo os resíduos *bootstrap* ε_i^* do MAG-AR. A série *bootstrap* Y_i^* é construída recursivamente da mesma forma que em (5.4).

5.3 *Bootstrap* condicional

O *bootstrap* condicional, sugerido por Figueiras *et al.* (2005), é um método que considera dados do tipo (x_i, y_i), $i = 1,...,n$ ou, de forma mais geral, $(y_i, x_{1i}, x_{2i},...)$, assumindo que a distribuição de Y_i é conhecida e que seus valores são condicionais aos valores $(x_{1i}, x_{2i},...)$. Os vetores das variáveis Y e X_j, $j = 1,...,d$, podem ser dados correlacionados ou não.

Para ilustrar a técnica na modelagem MAG, suponha que $Y_i \sim Poisson$ com esperança μ_i e considere $X_1,...,X_d$ relacionadas à Y_i pela expressão $\hat{\mu}_i = \exp\{\hat{\alpha} + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_r x_{ri} + \hat{f}(x_{(r+1)i}) + ... + \hat{f}(x_{di})\}$.

A técnica *bootstrap* condicional consiste em gerar um y_i^*, Y_i^* com distribuição Poisson de média $\hat{\mu}_i = \exp\{\hat{\alpha} + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_r x_{ri} + \hat{f}(x_{(r+1)i}) + ... + \hat{f}(x_{di})\}$, para cada ponto $(x_{1i},...,x_{di})$. O vetor Y_i^* é a série *bootstrap*.

Considerando $X_1,...,X_d$, séries temporais, relacionadas à série $Y_i \sim Poisson$ pela expressão $\hat{\mu}_i = \exp\{\hat{\alpha} + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_r x_{ri} + \hat{f}(x_{(r+1)i}) + ... + \hat{f}(x_{di}) + T_i\}$ a técnica segue a mesma forma.

5.4 Intervalos de confiança *bootstrap*

Existem vários métodos para calcular intervalos de confiança para um parâmetro, mas devido a algumas restrições tal como a imprecisão causada por aproximações feitas através da distribuição assintótica, Efron & Tibshirani (1986) propuseram métodos que fazem uso da técnica *bootstrap* na construção destes intervalos. Neste capítulo são apresentadas duas técnicas *bootstrap* de construção de intervalos – *bootstrap* percentílico e *bootstrap* com correção do vício – utilizadas na inferência da parte linear dos modelos MAG e MAG-AR.

5.4.1 Intervalos de confiança *bootstrap* percentílico

O intervalo percentílico $(1-\gamma)\%$ para o parâmetro β é definido pela expressão

$$\left(\hat{\beta}^*_{(\gamma/2)} \; ; \; \hat{\beta}^*_{(1-\gamma/2)}\right) \qquad (5.5)$$

onde, por definição, $\hat{\beta}^*_{(a)}$ é o $(100 \cdot a)$-ésimo percentil empírico da distribuição bootstrap (Efron e Tibshirani, 1993). Na prática, se são geradas B amostras *bootstrap* independentes, $x^{*1}, x^{*2},..., x^{*B}$, e estima-se $\hat{\beta}^*$ para cada uma delas tem-se que $\hat{\beta}^*_{(\gamma/2)}$

é o $B(\gamma/2)$º valor ordenado das replicações $\hat{\beta}^*$; a mesma interpretação é dada para $\hat{\beta}^*_{(1-\gamma/2)}$.

5.4.2 Intervalos de confiança *bootstrap* com correção do vício

Um dos principais objetivos da teoria *bootstrap* é produzir intervalos de confiança que realmente fornecem coberturas probabilísticas confiáveis para o parâmetro de interesse. O método *bootstrap* com correção do vício, apesar de ser um método mais complicado de se definir, é tão simples de ser usado quanto o método percentílico. Além disso, os intervalos de confiança correção do vício levam em consideração o vício do parâmetro estimado (Efron, 1982).

Os limites do intervalo de confiança correção do vício são encontrados através de percentis empíricos da distribuição *bootstrap*, mas não são necessariamente $\gamma/2$ e $1-\gamma/2$ tal como no método percentílico. Os percentis usados para o cálculo dos limites inferiores e superiores dos intervalos de confiança correção do vício dependem do número k_0 chamado *Bias-corrected*, ou corretor do vício, que é definido pela expressão

$$k_0 = \mathbf{\Phi}^{-1}\left(\frac{1}{B}\sum_{t=1}^{B} I(\hat{\beta}^*_t \leq \hat{\beta})\right),$$

em que I é uma função indicadora que recebe valor 1 se $(\hat{\beta}^*_t \leq \hat{\beta})$ e valor 0 caso contrário; B é o número de amostras *bootstrap* independentes; $\hat{\beta}$ é a estimativa do parâmetro para os dados observados; $\hat{\beta}^*$ é a estimativa do parâmetro para cada uma amostra *bootstrap* e $\Phi(\cdot)$ é a função de distribuição da Normal padronizada.

Os limites do intervalo *bootstrap* com correção do vício são dados por

$$\left(\hat{\beta}^*_{(p_1)}, \hat{\beta}^*_{(p_2)}\right)$$

sendo $p_1 = \Phi\left(2k_0 + z_{\frac{\gamma}{2}}\right)$ e $p_2 = \Phi\left(2k_0 + z_{1-\frac{\gamma}{2}}\right)$ – em que z_x é o $(100 \cdot x)$-ésimo ponto percentil da distribuição Normal padronizada – e $\hat{\beta}^*_{(p_1)}$ igual ao $B(p_1)$º valor ordenado das replicações $\hat{\beta}^*$.

Caso a distribuição do estimador $\hat{\beta}^*_a$ seja simétrica, tem-se que $k_0 = 0$, $p_1 = \gamma/2$ e $p_2 = 1 - \gamma/2$. Logo, nesta situação o intervalo de confiança obtido é o intervalo percentílico.

6. ANÁLISE DOS DADOS SIMULADOS

O desempenho da classe semiparamétrica dos MAG's no ajuste de dados independentes e no ajuste de séries temporais bem como a performance dos MAG-AR's no ajuste de séries temporais foi averiguada segundo a estimação do parâmetro linear β_1, via simulação de dados com variável resposta inteira e não negativa modelada pela distribuição Poisson

$$Y_i \sim poisson(\mu_i).$$

Para isto, foram consideradas duas covariáveis, sendo a primeira, X_1, relacionada linearmente com a variável resposta e a segunda, X_2, relacionada a Y de uma forma não linear através de uma função desconhecida. Dessa forma, a média da variável Y – vetor de dados independentes ou de séries temporais – quando modelada por um MAG é expressa pela relação

$$\eta_i = \log(\mu_i) = \alpha + \beta_1 X_{1i} + f(X_{2i}),\ i = 1,...,n. \qquad (6.1)$$

Considerando que a média da variável Y – agora uma série temporal – seja modelada por um MAG-AR o preditor aditivo assume a forma

$$\eta_i = g(\mu_i) = \log(\mu_i) = \alpha + \beta_1 X_{1i} + f(X_{2i}) + T_i,\ i = 1,...,n \qquad (6.2)$$

onde T_i é um AR de ordem 1 da forma, $T_i = \phi T_{i-1} + \varepsilon_i$.

As estatísticas que viabilizaram a comparação das performances do MAG, do MAG-AR e dos procedimentos *bootstrap* foram o vício e o erro quadrático médio (EQM) das estimativas. Os intervalos de confiança foram comparados via tamanho e probabilidade de cobertura – expressa pela razão entre o número de intervalos que contem o valor verdadeiro do parâmetro e o número total de intervalos construídos – com o nível nominal fixado em 95%, isto é, $\gamma = 0{,}05$. Para fins de comparação o

intervalo assintótico também foi construído, utilizando-se a distribuição assintótica normal para o estimador de β_1.

Dois valores foram definidos para n: 100 e 500. Para cada um deles, o número de simulações Monte Carlo (MC) e de replicações *bootstrap* foi fixado em 500. A técnica *loess* de suavização foi utilizada na estimação de f; o valor do parâmetro de suavização foi escolhido com base na estatística AIC dado os valores pré-fixados 0,5; 0,6; 0,7 e 0,8.

Os dados foram simulados e modelados através da linguagem de programação do *software* R; os algoritmos da modelagem MAG já estão implementados no software.

6.1 Resultados das simulações de dados independentes

O vetor resposta Y_i, $i = 1,...,n$, foi gerado a partir de uma Poisson com média

$$\mu_i = \exp\{\alpha + \beta_1 x_{1i} + x_{2i} + \tau_i\},$$

onde α e β_1 foram fixados, respectivamente, em 0,08 e 0,02. A covariável X_1 foi gerada a partir da distribuição Normal com média 3 e variância 4 e a variável X_2 gerada a partir da equação $X_2 = 0.001 \times (a + a^3)$, com a sendo uma variável aleatória de distribuição $N(10,4)$. O ruído $\tau, \tau \sim Normal(0,1)$, foi incluído no modelo para introduzir aleatoriedade no cálculo de Y.

As Tabelas 1 e 2 mostram os resultados sobre a estimação pontual e intervalar, respectivamente, do parâmetro β_1 obtidos através da modelagem MAG. Da Tabela 1, nota-se que a média das estimativas pontuais obtidas no Monte Carlo (MC) superestimaram o verdadeiro valor do parâmetro e, apesar do vício ser praticamente o mesmo tanto para as simulações de tamanho 100 quanto para as simulações de tamanho 500, os valores do EQM neste último cenário têm uma ordem de grandeza menor (da ordem de 10^{-4} para $n = 100$ e 10^{-5} para $n = 500$.

Os resultados do procedimento *bootstrap*, estão bem próximos à média das estimativas obtidas no MC, sendo que o melhor desempenho corresponde ao *bootstrap* nas observações. O *bootstrap* nos resíduos apresenta um EQM muito maior que os

outros dois tipos do *bootstrap*, do que se pode concluir que, se os dados são independentes, o melhor é utilizar o *bootstrap* direto nas observações, pois como o *bootstrap* nos resíduos é dependente do modelo ajustado, ele pode carregar o vício das estimativas calculadas para os parâmetros.

Tabela 1. Médias das estimativas, médias de vício e EQM na estimação de β_1 no MAG para dados independentes

n	MC	*Bootstrap* nas observações	*Bootstrap* condicional	*Bootstrap* nos resíduos
100	0,0217	0,0211	0,0220	0,0219
	0,0017	**0,0011**	**0,0020**	**0,0019**
	(5,81e-04)	(5,94e-04)	(5,87e-04)	(7,70e-04)
500	0,0216	0,0218	0,0219	0,0219
	0,0016	**0,0018**	**0,0019**	**0,0019**
	(3,45e-05)	(3,82e-05)	(3,95e-05)	(5,10e-05)

Nota: valores em negrito são as médias de vício, valores entre parênteses são os EQM's.

Da Tabela 2, pode-se concluir que os intervalos de confiança apresentam, em geral, probabilidade de cobertura bastante próxima ao valor nominal (0,95) e que os intervalos *bootstrap* com correção do vício têm melhor desempenho se comparados aos intervalos *bootstrap* percentílico, com a única exceção para o procedimento *bootstrap* nas observações. Em geral, os intervalos de confiança construídos através da técnica *bootstrap* apresentam-se mais próximos a 0,95 que o intervalo assintótico, tanto para $n = 100$ quanto para $n = 500$.

Tabela 2. Intervalos de confiança para β_1 no MAG para dados independentes

n	Assintótico	*Bootstrap* nas observações		*Bootstrap* condicional		*Bootstrap* nos resíduos	
		Percentílico	BC	Percentílico	BC	Percentílico	BC
100	**0,928**	**0,946**	**0,938**	**0,936**	**0,938**	**0,928**	**0,948**
	(0,090)	(0,095)	(0,094)	(0,091)	(0,091)	(0,099)	(0,097)
500	**0,932**	**0,944**	**0,942**	**0,940**	**0,942**	**0,934**	**0,950**
	(0,094)	(0,097)	(0,094)	(0,095)	(0,091)	(0,107)	(0,101)

Nota: valores em negrito são as probabilidades de cobertura, valores entre parênteses são as amplitudes dos intervalos.

6.2 Resultados das simulações de séries temporais

Para os dados que representam séries temporais, a variável resposta Y_i, $i = 1,...,n$, foi gerada a partir de uma Poisson com média

$$\mu_i = \exp\{\alpha + \beta_1 x_{1i} + x_{2i} + T_i\}, \quad (6.3)$$

onde α foi fixado em 0,08 e β_1 em 0,02. A covariável X_1 é um modelo auto-regressivo da forma $X_i = 0{,}4X_{i-1} + e_i$, $e_i \sim N(0{,}1)$, $X_2 = \cos(2\pi i/12) + v_i$, $v_i \sim N(0;0{,}01)$ e Z_i um modelo auto-regressivo de ordem 1, $T_i = \phi T_{i-1} + \varepsilon_i$, com $\phi = 0{,}4$ e $0{,}6$ e ε_i definido em (4.2). O procedimento *bootstrap* nas observações, conforme justificado no Capítulo 5, não é realizado para as séries temporais simuladas.

Nesta seção, o objetivo é avaliar as estimativas do parâmetro linear β_1 quando os dados são modelados através do MAG, ignorando a dependência entre as observações das séries temporais, e compará-las com as estimativas encontradas na modelagem MAG-AR, classe capaz de capturar estruturas de dependência entre os dados.

As Tabelas 3 e 4 apresentam as estimativas pontuais das séries temporais geradas, a primeira para o MAG e a segunda para o MAG-AR(1). Ao contrário das estimativas apresentadas na Seção 6.1 de dados independentes, todas as estimativas apresentadas nas Tabelas 3 e 4 subestimaram o verdadeiro valor do parâmetro linear, β_1.

Tabela 3. Médias das estimativas, médias de vício e EQM na estimação de β_1 no MAG para séries temporais

n	ϕ	MC	*Bootstrap* condicional	*Bootstrap* nos resíduos
100	0,4	-0,0043	-0,0043	-0,0028
		-0,0243	**-0,0243**	**-0,0228**
		(2,97e-02)	(2,97e-02)	(2,75e-02)
	0,6	0,0013	0,0015	0,0014
		-0,0187	**-0,0185**	**-0,0186**
		(3,93e-02)	(3,93e-02)	(3,47e-02)
500	0,4	-0,0031	-0,0033	-0,0063
		-0,0231	**-0,0233**	**-0,0263**
		(2,32e-02)	(2,32e-02)	(1,95e-02)
	0,6	0,0023	0,0024	0,0023
		-0,0177	**-0,0176**	**-0,0177**
		(3,44e-02)	(3,45e-02)	(2,77e-02)

Nota: valores em negrito são as médias de vício, valores entre parênteses são os EQM's.

Comparando o ajuste das séries de contagem utilizando o MAG e o MAG-AR(1) pode-se verificar que as estimativas apresentam maior vício e EQM para o MAG (Tabela

3) que para o MAG-AR(1) (Tabela 4), evidenciando a necessidade de se utilizar este último modelo para dados de séries temporais. Das Tabelas 3 e 4 verifica-se também que são mais viciadas e possuem EQM's maiores os resultados dos dados gerados com ϕ igual a $0{,}4$. Novamente, o desempenho dos procedimentos *bootstrap* é muito semelhante ao do MC. Neste caso, percebe-se que o EQM das estimativas utilizando o *bootstrap* nos resultados é ligeiramente menor que a do *bootstrap* condicional, com o vício das estimativas para o modelo MAG-AR(1) também menor para o *bootstrap* nos resíduos.

Tabela 4. Médias das estimativas, médias de vício e EQM na estimação de β_1 no MAG-AR(1) para séries temporais

n	ϕ	MC	*Bootstrap* condicional	*Bootstrap* nos resíduos
100	0,4	0,0169	0,0165	0,017
		-0,0031	**-0,0035**	**-0,0030**
		(7,80e-03)	(7,86e-03)	(7,78e-03)
	0,6	0,0179	0,0174	0,0176
		-0,0021	**-0,0026**	**-0,0024**
		(3,66e-03)	(3,91e-03)	(3,47e-03)
500	0,4	0,0178	0,0177	0,0181
		-0,0022	**-0,0023**	**-0,0019**
		(7,44e-03)	(7,54e-03)	(7,25e-03)
	0,6	0,0186	0,0186	0,0189
		-0,0014	**-0,0014**	**-0,0011**
		(2,14e-03)	(2,32e-03)	(1,93e-03)

Nota: valores em negrito são as médias de vício, valores entre parênteses são os EQM's.

Na Tabela 5 estão os resultados das estimações intervalares do parâmetro linear β_1 das séries temporais de contagem modeladas através do MAG e na Tabela 6 os resultados das séries com ajuste MAG-AR. Em todos os casos nota-se que a cobertura dos intervalos está mais próxima ao nível de 95% no modelo MAG-AR(1), com intervalos praticamente do mesmo tamanho.

Comparando os procedimentos *bootstrap*, o *bootstrap* nos resíduos apresenta melhor desempenho quanto à cobertura, apesar da amplitude dos mesmos ser um pouco maior. Os intervalos BC apresentam-se levemente superiores ao percentílico, mas ambos têm melhor desempenho se comparados ao intervalo assintótico. Nestes casos, os resultados foram melhores para $\phi = 0{,}4$.

Tabela 5. Estimação intervalar do parâmetro β_1 e comparação dos procedimentos MC e *bootstrap*'s – Séries temporais modeladas através do MAG

n	ϕ	Assintótico	*Bootstrap* condicional		*Bootstrap* nos resíduos	
			Percentílico	BC	Percentílico	BC
100	0,4	**0,870**	**0,884**	**0,886**	**0,922**	**0,924**
		(0,115)	(0,118)	(0,118)	(0,132)	(0,133)
	0,6	**0,866**	**0,880**	**0,882**	**0,914**	**0,916**
		(0,113)	(0,114)	(0,116)	(0,134)	(0,134)
500	0,4	**0,874**	**0,906**	**0,908**	**0,930**	**0,936**
		(0,117)	(0,121)	(0,122)	(0,138)	(0,138)
	0,6	**0,872**	**0,888**	**0,888**	**0,926**	**0,926**
		(0,117)	(0,119)	(0,119)	(0,136)	(0,137)

Nota: Valores em negrito são as probabilidades de cobertura, valores entre parênteses são as amplitudes dos intervalos

Tabela 6. Estimação intervalar do parâmetro β_1 e comparação dos procedimentos MC e *bootstrap*'s – Séries temporais modeladas através do MAG-AR

n	ϕ	Assintótico	*Bootstrap* condicional		*Bootstrap* nos resíduos	
			Percentílico	BC	Percentílico	BC
100	0,4	**0,898**	**0,902**	**0,914**	**0,938**	**0,938**
		(0,116)	(0,118)	(0,119)	(0,135)	(0,133)
	0,6	**0,884**	**0,888**	**0,902**	**0,938**	**0,940**
		(0,110)	(0,109)	(0,113)	(0,134)	(0,135)
500	0,4	**0,901**	**0,922**	**0,926**	**0,942**	**0,940**
		(0,121)	(0,119)	(0,119)	(0,137)	(0,138)
	0,6	**0,898**	**0,916**	**0,920**	**0,940**	**0,938**
		(0,116)	(0,118)	(0,118)	(0,134)	(0,134)

Nota: Valores em negrito são as probabilidades de cobertura, valores entre parênteses são as amplitudes dos intervalos

7. APLICAÇÃO A SÉRIES REAIS

Estudos epidemiológicos realizados em diferentes centros de pesquisa têm detectado associações significativas entre morbi-mortalidade por causas respiratórias e poluição atmosférica em populações urbanas – Schwartz (1994) e Saldiva *et al.* (1995), por exemplo. A maior parte desses estudos é do tipo ecológico, isto é, de base populacional e um grupo, ao invés de um indivíduo, constitui a unidade de observação seguida ao longo do tempo. Consistem, em geral, da observação de eventos tais como mortalidade, admissões hospitalares ou sintomas respiratórios. Esse tipo de planejamento é menos suscetível a variáveis de confusão individuais como fumo, pressão arterial e fatores sócio-econômicos (Rothman *et al.*, 1998), pois esses fatores não variam de dia para dia com a poluição atmosférica.

A cidade de São Paulo, o segundo maior centro urbano da América Latina, possui um cenário apropriado para o desenvolvimento de estudos dos efeitos da poluição atmosférica sobre a saúde de seres humanos; um dos principais motivos é a grande oferta de transporte coletivo e uma malha viária de uma frota de pouco mais de 6.100.000 veículos leves em toda região metropolitana (DETRAN-SP) que constituem a principal fonte da poluição do ar. A poluição atmosférica na cidade é predominantemente gerada por fontes poluidoras móveis, além disso, suas condições geográficas e meteorológicas desfavorecem a dispersão dos poluentes, principalmente durante os meses de inverno, quando, com freqüência, ocorrem inversões térmicas.

Neste Capítulo, os modelos MAG e MAG-AR(1) são ajustados a dados mensais do número de pacientes internados com afecção das vias aéreas superiores – AVAS – na cidade de São Paulo, nos anos de 1997 a 2000, para se verificar o desempenho da modelagem proposta, MAG-AR(1), em relação aos MAG's numa aplicação a dados reais, segundo a estimação do parâmetro linear β_1. As covariáveis consideradas são as séries dióxido de enxofre - $SO_2/(\mu g/m^3)$ - e monóxido de carbono - CO/ppm. A variável tempo i $(i = 1,...,n)$ e as variáveis harmônicas $sen(2\pi i/n)$ e $cos(2\pi i/n)$ foram inseridas no modelo com o objetivo de ajustar o efeito da tendência e da sazonalidade (Moretin *et al.*, 2004).

7.1 Análise descritiva

A seguir são apresentadas tabelas e gráficos construídos com o objetivo de resumir os dados mensais das variáveis de interesse: paciente com afecção das vias aéreas superiores – AVAS – e os poluentes SO_2 e CO, nos anos de 1997 a 2000.

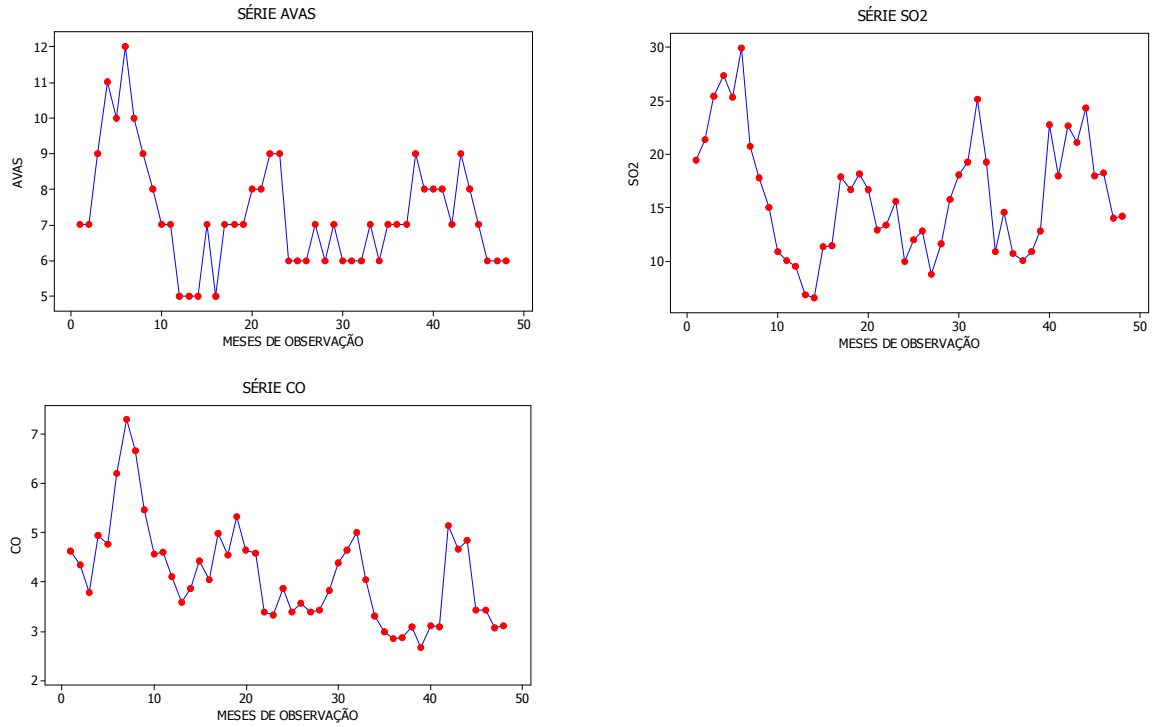

Figura 7.1 – Representação gráfica das séries AVAS, SO_2 e CO nos anos de 1997 a 2000.

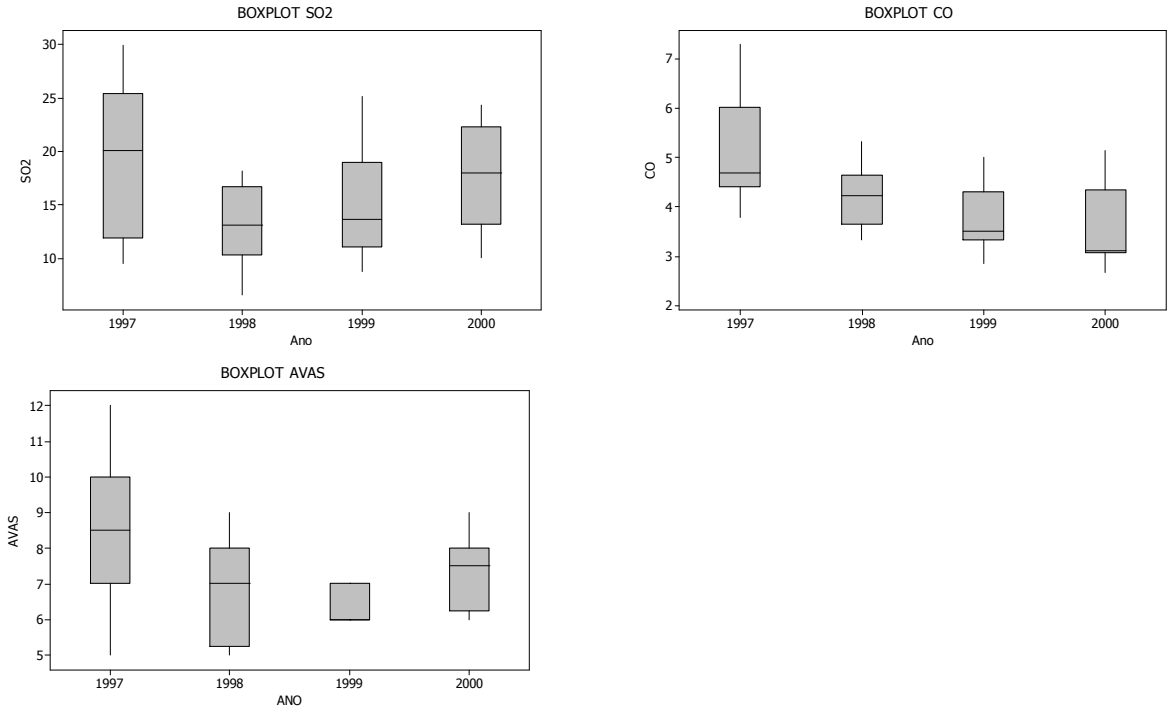

Figura 7.2 – Gráficos do tipo *Box-plot* para AVAS, SO_2 e CO nos anos de 1997 a 2000.

A Figura 7.1 apresenta as séries cronológicas e a Figura A.1, no Apêndice A, apresenta gráficos do tipo *box-plot*. Na Figura A.1 não se nota valores discrepantes da AVAS nos anos avaliados. Na Tabela 7, observa-se que o número mensal médio de pacientes com afecção das vias aéreas superiores foi maior no ano de 1997 (desvio padrão igual a 2,02). Nos anos de 1998, 1999 e 2000 essa média foi, respectivamente, igual a 6,92 (desvio padrão igual a 1,44), 6,42 (desvio padrão igual a 0,51) e 7,42 (desvio padrão igual a 1,08). Observando a representação gráfica da série AVAS (Figura 7.1) nota-se que em 1997, ano de maior ocorrência de pacientes com afecção das vias aéreas superiores os picos ocorrem nos meses mais frios do ano, em 1998 e 1999 os valores mais altos ocorrem nos últimos meses e em 2000 os picos ocorrem em fevereiro e julho.

As concentrações dos poluentes SO_2 e CO também apresentam valor médio mais alto em 1997: 19,39 (desvio padrão igual a 6,94) e 5,11 (desvio padrão igual a 1,08), respectivamente – veja Tabela 7.

Tabela 7 – Média ($\pm$ desvio padrão) para os dados AVAS e poluentes – SO_2 e CO.

Variáveis	N	1997	1998	1999	2000	Total
AVAS	48	8,50 ($\pm$2,02)	6,92 ($\pm$1,44)	6,42 ($\pm$0,51)	7,42 ($\pm$1,08)	7,31 ($\pm$1,55)
SO_2	48	19,39 ($\pm$6,94)	13,12($\pm$4,02)	14,88 ($\pm$4,75)	17,24 ($\pm$4,84)	16,16 ($\pm$5,62)
CO	48	5,11 ($\pm$1,08)	4,21 ($\pm$0,64)	3,73 ($\pm$0,67)	3,54 ($\pm$0,84)	4,15 ($\pm$1,01)

Os *box-plots* não apontam concentrações aberrantes de nenhum dos dois poluentes e evidenciam que os níveis de concentração de CO na cidade de São Paulo diminuíram, ano após ano, no período avaliado.

A relação entre a variável AVAS e os poluentes SO_2 e CO estão representadas na Figura 7.3. Tanto SO_2 quanto CO possuem correlação positiva com o número de pacientes com afecções das vias aéreas superiores – correlação de Person iguais a 0,614 (p-valor aproximadamente 0) e 0,417 (p-valor igual a 0,003) – porém, da avaliação da Figura 7.3 nota-se o fato do tipo de relação existente entre a resposta e o poluente CO poder não ser estritamente linear. Um polinômio envolvendo um termo de maior ordem ou uma função suavizadora podem ser, por exemplo, mais adequados para descrever a relação entre a resposta e esse poluente.

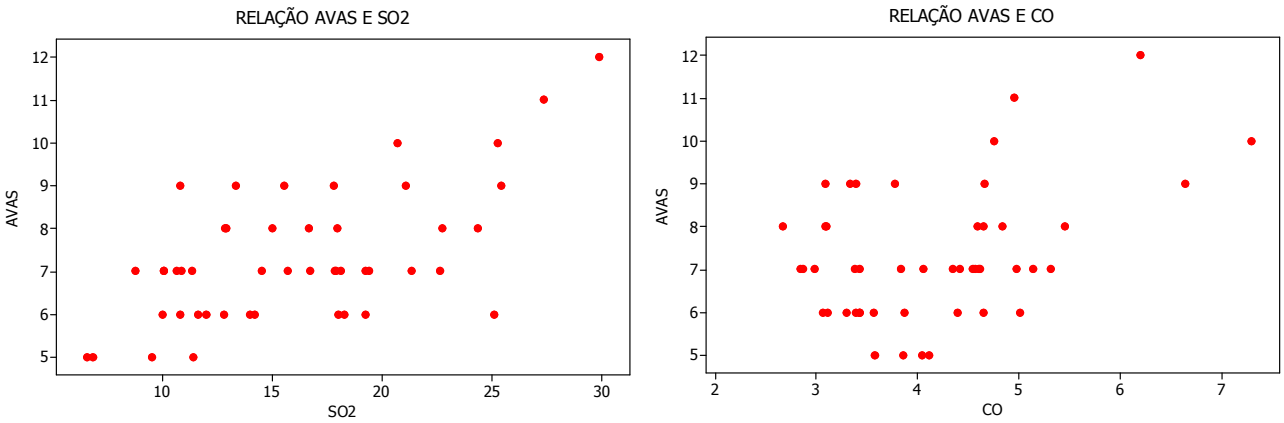

Figura 7.3 – Representação gráfica da relação entre as séries AVAS e CO, AVAS e SO_2.

A série de dióxido de enxofre apresenta uma estrutura auto-regressiva de ordem 1, conforme visto na Figura 7.4 e na Tabela 8 – note que o coeficiente AR(1) é significativo tal como a constante; o teste Ljung-Box, Tabela 9, não apresenta indícios que levem à rejeição de uma estrutura AR(1) na série SO_2 e a análise gráfica dos resíduos (Figura 7.5) conduzem à conclusão de normalidade e homocedasticidade dos mesmos.

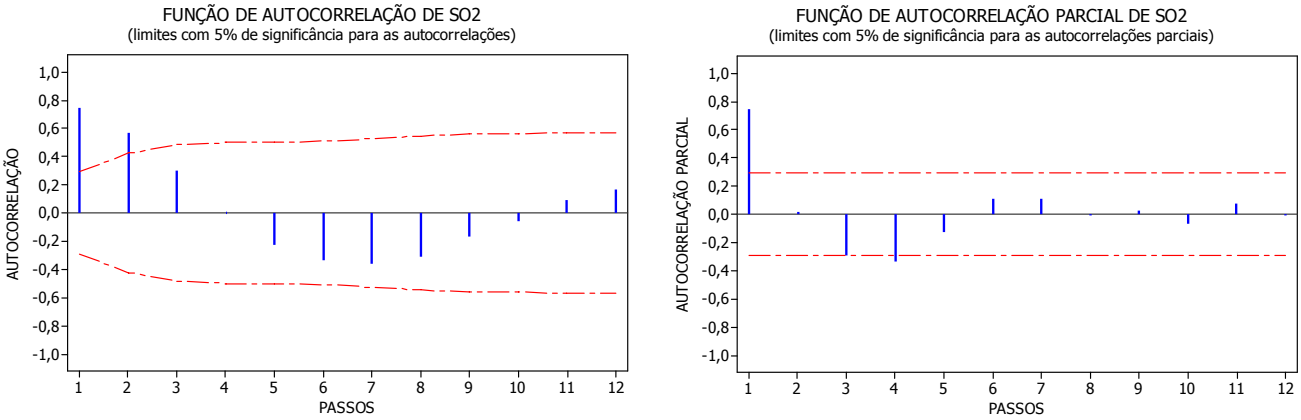

Figura 7.4 – Gráficos de autocorrelação e autocorrelação parcial de SO_2.

Tabela 8 – Modelagem AR(1) da série SO_2.

	Coeficiente	**Erro padrão**	**T**	**P**
AR(1)	0,75	0,09	7,77	~0
Constante	3,98	0,54	7,37	~0
Média	16,23	2,20		

Tabela 9 – Teste Ljung-Box para modelagem AR(1) da série SO_2.

	Passos		
	12	**24**	**36**
Qui-quadrado	16,10	28,9	52,8
Graus de liberdade	10	22	34
p-valor	0,10	0,15	0,02

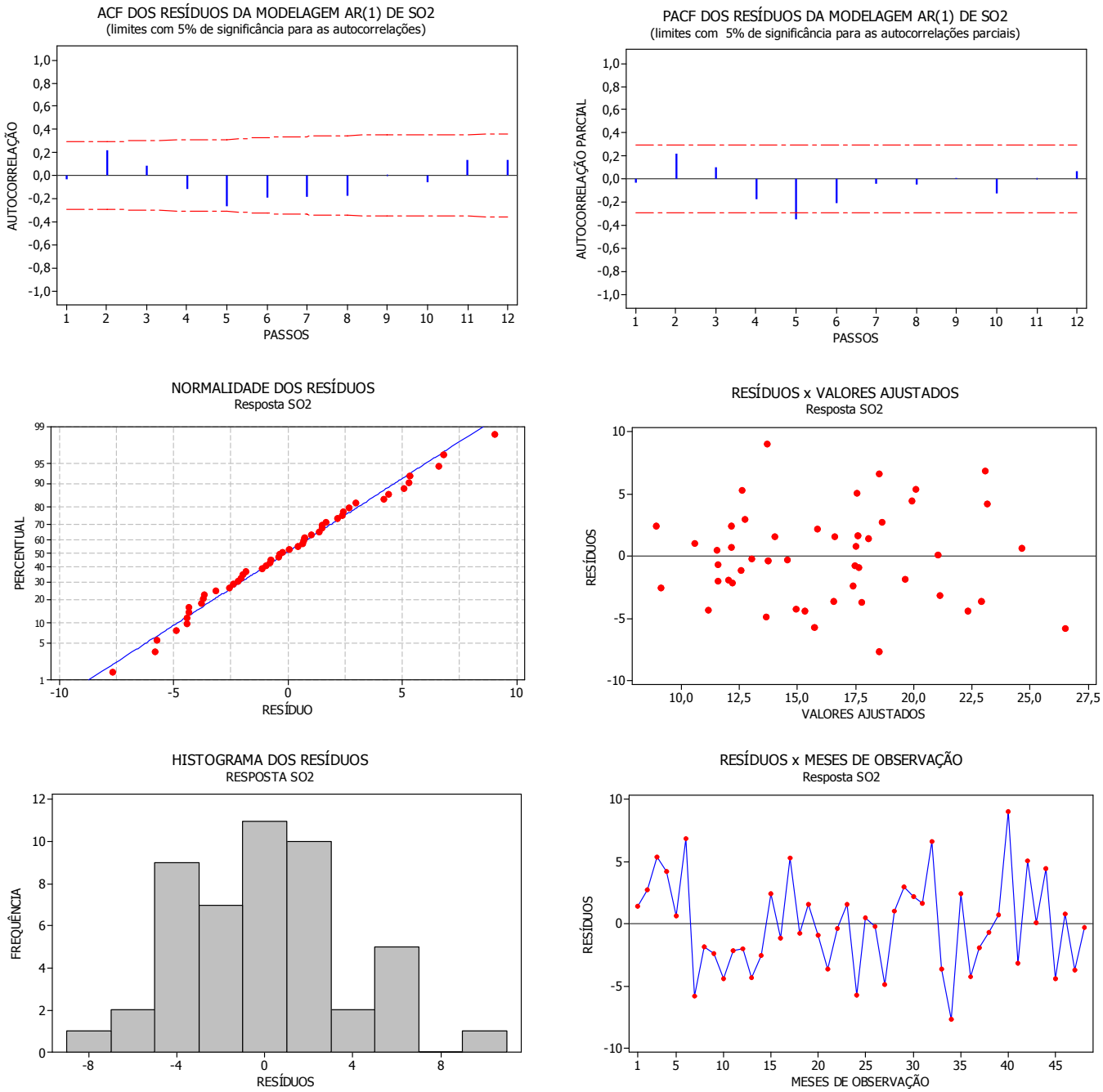

Figura 7.5 – Análise residual da modelagem AR(1) de SO_2.

7.2 Modelagem MAG

A seguir são apresentados resultados quando um MAG é adotado para relacionar a série AVAS aos poluentes SO_2 e CO, na forma

$$\text{AVAS} \sim poisson(\mu_i)$$
$$\log(\mu_i) = \alpha + \beta_1 \text{SO}_2 + f\ (\text{CO}) + \beta_2 i + \beta_3 sen(2\pi i/n) + \beta_4\ cos(\ 2\pi i/n\ ), \quad (7.1)$$

$i = 1,...,n$. O método de suavização utilizado para a estimação do termo não paramétrico foi o *loess*. A escolha do parâmetro de suavização, λ, foi feita a partir de valores pré-fixados em 0,3, 0,5, 0,7 e 0,8 de forma que o modelo final apresentasse valor mínimo para a estatística AIC. A Tabela 10 apresenta os resultados da estimação pontual e intervalar do parâmetro linear e os resultados dos procedimentos *bootstrap* condicional e *bootstrap* nos resíduos – médias das 500 replicações *bootstrap*. Os intervalos de confiança foram calculados com nível de 95% de confiança.

Tabela 10. Estimativas de β_1 na modelagem MAG do AVAS

	MAG	*Bootstrap* condicional	*Bootstrap* nos resíduos
$\hat{\beta}_1$	0,024 **(0,0139)**	0,024 **(0,0141)**	0,026 **(0,0146)**
AIC	208,10	-	-
IC assintótico	[-0,003;0,051]	-	-
IC *bootstrap* percentílico	-	[-0,006;0,054]	[0,016;0,037]
IC *bootstrap* com correção do vício	-	[-0,004;0,057]	[0,012;0,034]

Nota: Valores em negrito são o desvio padrão da estimativa de β_1.

A estimativa inicial obtida para o parâmetro linear β_1 é igual a 0,024. Quando a abordagem *bootstrap* é utilizada a média das estimativas de cada replicação apresenta valor bastante próximo ao valor encontrado na modelagem inicial – 0,024 no procedimento *bootstrap* condicional e 0,026 no procedimento *bootstrap* nos resíduos.
Quanto aos intervalos de confiança *bootstrap*, os percentílicos têm as menores amplitudes, porém, no estudo simulado apresentado no Capítulo 6, foi mostrado que,

apesar de menos acurado, o IC's *bootstrap*'s com correção do vício possuem probabilidade de cobertura mais próxima ao nível nominal e apresentam-se levemente superiores aos percentílicos, mas ambos com melhor desempenho que os intervalos assintóticos.

As estimativas para os demais parâmetros envolvidos no modelo, β_2, β_3 e β_4 são todas significativas a um nível de 10%: $\hat{\beta}_2 =$ -0,004 (desvio padrão igual a 0,007), $\hat{\beta}_3 =$ 0,029 (desvio padrão igual a 0,014) e $\hat{\beta}_4 =$ -0,025 (desvio padrão igual a 0,089).

A análise gráfica dos resíduos da modelagem da série AVAS através do MAG está representada na Figura 7.6; dos ACF's e PACF's é clara a presença de uma estrutura auto-regressiva de ordem 1 nos resíduos indicando que o MAG não foi capaz de capturar a estrutura de correlação presente no desfecho.

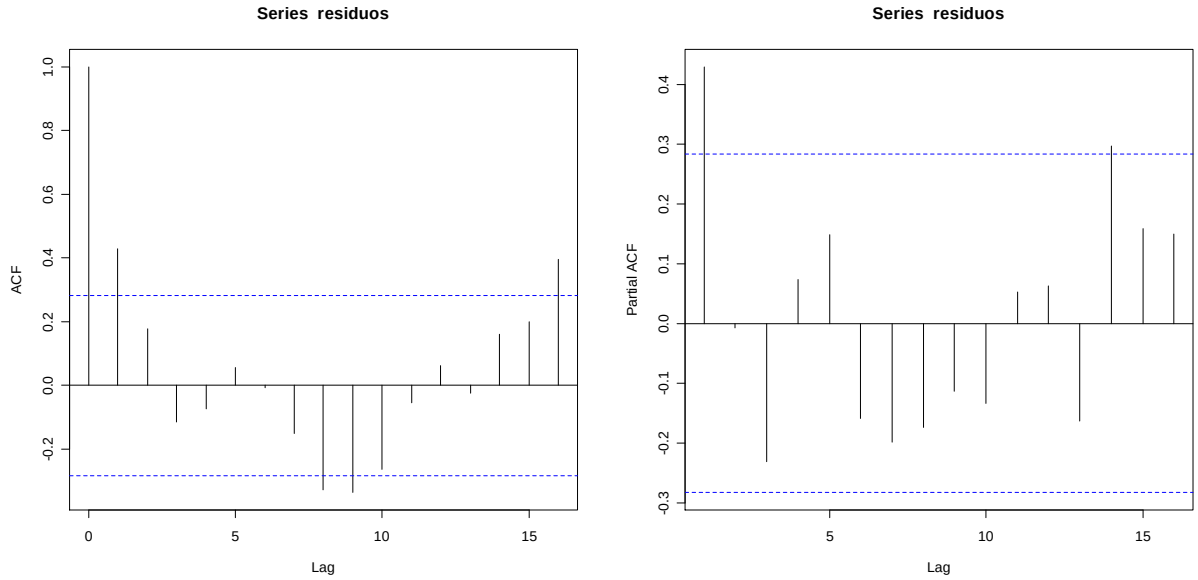

Figura 7.6 – Avaliação residual da série AVAS, modelada via MAG.

7.3 Modelagem MAG-AR(1)

Para corrigir a estrutura auto-regressiva presente nos resíduos da modelagem MAG, a relação entre a série AVAS e os poluentes SO_2 e CO foram ajustados através do MAG-AR(1) proposto neste trabalho; o modelo é da forma

$$\text{AVAS} \sim poisson(\mu_i)$$

$$\log(\mu_i) = \alpha + \beta_1 \text{SO}_2 + f\ (\text{CO}) + \beta_2 i + \beta_3 sen(2\pi i/n) + \beta_4\ cos(\ 2\pi i/n\ ) + T\ ,\ \ (7.2)$$

onde T representa a estrutura de correlação presente nas observações e $i = 1,...,n$.

O método de suavização utilizado na estimação do termo não paramétrico, os valores pré-fixados de λ e o nível de confiança utilizado no computo dos intervalos de confiança são os mesmo utilizados no ajuste do MAG na Seção 7.2 deste Capítulo. A Tabela 11 apresenta os resultados da estimação pontual e intervalar do parâmetro linear dos dados ajustados através do MAG-AR(1) conforme expressão (7.2).

As estimativas obtidas para o parâmetro linear β_1, tanto nos casos em que a abordagem *bootstrap* é utilizada quanto no caso em que não há qualquer reamostragem, apresentam valores muito próximos (iguais até pelo menos a terceira casa decimal) às estimativas encontradas no ajuste do MAG.

As estimativas intervalares do parâmetro linear quando AVAS é modelada através do MAG-AR(1) são bastantes semelhantes aos intervalos calculados na Seção 7.2, onde os dados são ajustados via MAG. Da mesma forma, as conclusões da comparação entre

Tabela 11. Estimativas de β_1 na modelagem MAG-AR(1) do AVAS

	MAG-AR	*Bootstrap* condicional	*Bootstrap* nos resíduos
$\hat{\beta}_1$	0,023 **(0,0139)**	0,024 **(0,0141)**	0,026 **(0,0147)**
AIC	209,62	-	-
IC assintótico	[-0,004;0,050]	-	-
IC *bootstrap* percentílico	-	[-0,005;0,054]	[0,015;0,037]
IC *bootstrap* com correção do vício	-	[-0,004;0,055]	[0,012;0,035]

Nota: Valores em negrito são o desvio padrão da estimativa de β_1.

as duas técnicas *bootstrap* de construção de intervalo – *bootstrap* percentílico e *bootstrap* com correção dos vícios – são as mesmas consideradas na seção anterior.

As estimativas para os demais parâmetros envolvidos no modelo, β_2, β_3 e β_4 são bastante parecidas às estimativas calculadas na modelagem MAG além de serem

significativas a um nível de 10%: $\hat{\beta}_2 = $-0,005 (desvio padrão igual a 0,007), $\hat{\beta}_3 = 0{,}019$ (desvio padrão igual a 0,014) e $\hat{\beta}_4 = $-0,022 (desvio padrão igual a 0,089).

A principal diferença entre a modelagem MAG e MAG-AR(1) está na capacidade deste último de capturar estruturas de correlação entre as observações. Percebe-se, da análise gráfica dos resíduos – Figura 7.7 –, que a estrutura auto-regressiva dos dados foi captada pelo modelo e que os resíduos se tornaram um ruído branco.

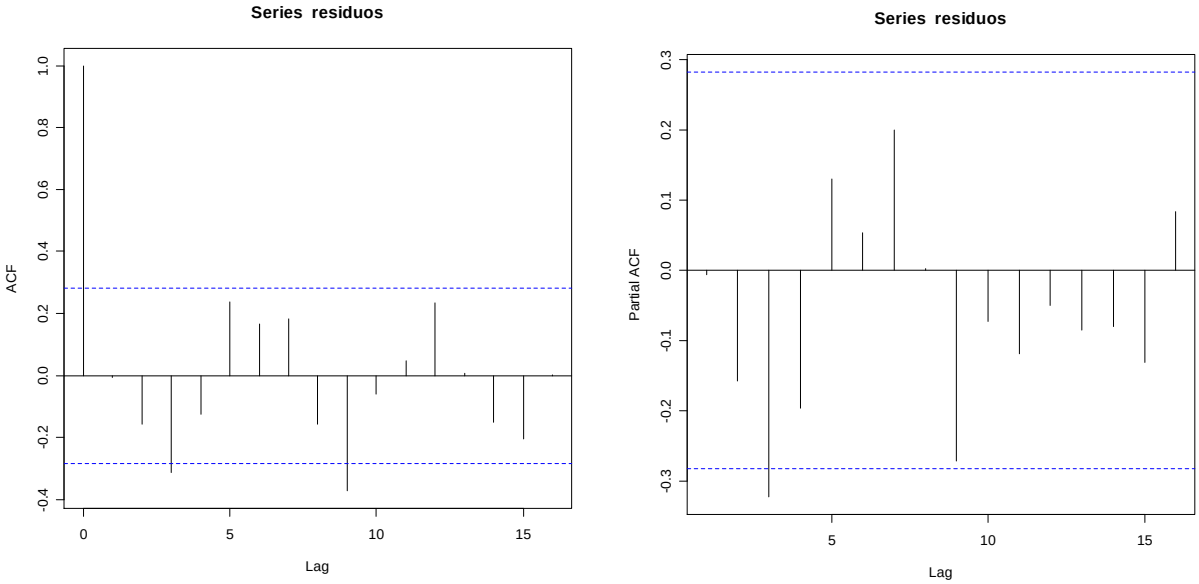

Figura 7.7 – Avaliação residual da série AVAS, modelada via MAG-AR(1).

8. CONCLUSÕES RELEVANTES

O Modelo Aditivo Generalizado e suas extensões constituem uma ampla classe de modelos de regressão, na qual o efeito das variáveis preditoras na variável resposta pode ser modelado de forma bastante flexível por meio de uma função não especificada. Apesar de bastante utilizados em estudos de séries temporais, principalmente em casos em que a variável resposta é uma contagem de eventos, os MAG's têm toda sua teoria de estimação e inferência construída sobre a hipótese de independência dos dados. Apesar de não apresentar a flexibilidade do MAG, o GLARMA, recentemente proposto na literatura estatística, tem a vantagem de ser um modelo capaz de capturar a estrutura de dependência existente entre as observações de séries temporais.

Neste trabalho é proposta uma nova classe de modelos, MAG-AR, que estende a estrutura linear do GLARMA para a estrutura semiparamétrica do MAG acomodando variáveis que tem relação não linear com a variável resposta em dados com estrutura AR de dependência. Além disto, a técnica *bootstrap* foi utilizada para se fazer inferências sobre o estimador dos parâmetros da parte linear do modelo.

O desempenho do MAG e do MAG-AR no ajuste de séries temporais foi, empiricamente, comparado através de alguns experimentos Monte Carlo com o cálculo do vício e do erro quadrático médio das estimativas do parâmetro linear dos modelos. Os resultados mostraram estimativas mais consistentes e menos viciadas para o MAG-AR(1) aplicados aos dados de séries temporais, quando comparados ao ajuste utilizando o MAG. A performance de algumas abordagens *bootstrap*'s – *bootstrap* condicional e *bootstrap* nos resíduos – nas séries temporais simuladas foi avaliada e os resultados indicaram que o *bootstrap* pode ser utilizado neste caso para se fazer inferências intervalares sobre os parâmetros lineares do modelo por apresentar resultados bastante semelhantes aos dos experimentos Monte Carlo.

Estimativas intervalares também foram calculadas e os intervalos de confiança *bootstrap* percentílico e *bootstrap* com correção do vício foram comparados ao intervalo assintótico quanto à probabilidade de cobertura e o tamanho dos mesmos. Em geral os intervalos *bootstrap* com correção do vício apresentaram resultados mais próximos à

cobertura nominal fixada (0,95) e os intervalos de confiança *bootstrap* tiveram melhor desempenho que o intervalo assintótico.

As metodologias MAG e MAG-AR foram utilizadas no ajuste da relação entre a série AVAS (número de pacientes com afecção das vias aéreas superiores) e os poluentes dióxido de enxofre (SO_2) e monóxido de carbono (CO). As análises mostraram as mesmas conclusões sobre o efeito dos poluentes na variável resposta, no entanto a estrutura de correlação presente entre as observações só foi capturada na modelagem MAG-AR(1).

Como continuidade do trabalho, podem ser sugeridas pesquisas futuras que incluam a extensão do modelo MAG-AR para estruturas de correlação mais complexas, como a adição de termos médias móveis, no caso um MAG-ARMA, ou modelos de longa dependência, MAG-ARFIMA.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Andrews, .F. (1974), A Robust Method for Multiple Linear Regression, *Technometrics*, 16, 523 – 531.

[2] Beaton, A.E.; Tukey, J.W. (1974), The fitting of power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data, *Technometrics*, 16, 147 – 185.

[3] Benjamin, R.A.; Rigby, M.A.; Stasinopoulos, M.D. (2003), Generalized autoregressive moving average models. *Journal of the American Statistical Association*, 98(461), 214-223.

[4] Bhattacharya, P.K.; Zhao, P.L. (1997), Semiparametric inference in a partial linear models. *The annals of statistics*, 25, 244-262.

[5] Box, G.E.P.; Jenkins, G.M. (1976), *Time series analysis: forecasting and control*. San Francisco: Holden-Day.

[6] Buja, A.; Hastie, T.J.; Tibshirani, R.J. (1989), Linear Smoothers and Additive Models. *The Annals of Statistics*, 16, 136-146.

[7] Cleveland, W.S. (1977), Locally Weighted Regression and Smoothing Scatterplots, Bell Laboratories memorandum.

[8] Cleveland, W.S. (1979), Robust locally-weighted regression and smoothing scatterplots. *J. Am. Statis. Assoc*., 74, 829 – 836.

[9] Conceição, G.M.S; Saldiva, P.H.N; Singer, J.M. (2001), Modelos MLG e MAG para análise da associação entre poluição atmosférica e marcadores de morbi-mortalidade: uma introdução baseada em dados da cidade de São Paulo. *Revista Brasileira de Epidemiologia*, 4.

[10] Cox D.D. (1983), Asymptotics for M-type smoothing splines. *Ann. Statist.* 11, 530-51.

[11] Craven, P.; Wahba, G. (1979), Smoothing noisy data with spline functions. *Nunmer. Math*., 31, 377-403.

[12] Davis, R.A.; Dunsmuir, W.T.M.; Wang,Y. (1999), Modelling time series of count data. *Asymptotics, Nonparametrics and Time Series*, 63-114, New York.

[13] Davis, R.A.; Dunsmuir, W.T.M.; Street, S.B. (2003), Observation-driven Models for Poisson Counts, *Biometrika*, 90, 4, 777-790.

[14] Deaton, A.; Muellbauer, J. (1980), Economics and Consumer Behavior. Cambridge University Press.

[15] Díez F.B.; Tenías, J.N.; Pérez-Hoyos, S. (1999), Efecto de la contaminación atmosférica sobre a salud: una introducción. *Rev Esp Salud Pública*, 73, 109-121.

[16] Dominici, F.; McDermott, A.; Zeger, S.L.; Samet, J.M. (2002), On the use of generalized additive models in time-series studies of air pollution and health. *Am. J. Epidemiol*. 156 (3), 193-203.

[17] Drescher, D. (2005). Alternative distributions for observation driven count series models. *Economics Working Paper*, 11, Christian-Albrechts-Universitat Kiel.

[18] Efron, B. (1979) Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics*, 7, 1-26.

[19] Efron, B.; Tibshirani, R. (1986), Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Satatistical Science* 1, 54-77.

[20] Efron, B.; Tibshirani, R. (1993), An introduction to the Bootstrap, New York: Chapman and Hall.

[21] Figueiras, A.; Roca-Pardiñas, J.; Suárez, C.C. (2005), A bootstrap method to avoid the effect of concurvity in generalised additive models in time series studies of air pollution. *J Epidemiol Community Health*, 59, 881-884.

[22] Goldberger, A.S. (1964), *Econometric Theory*. Wiley.

[23] Hall, P. (1988), Theoretical comparison of bootstrap confidence intervals (with discussion), *Annals Statistical*, 16, 927-953.

[24] Härdle, W. (1990), Applied Nonparametric Regression. Cambridge University Press. New York.

[25] Härdle, W.; Huet, S.; Mammen, E.; Sperlich, S. (2004), Bootstrap inference in semiparametric generalized additive models, *Econometric Theory*, 20, 265-300.

[26] Hastie, T.J.; Tibshirani, R.J. (1990) Generalized additive models, London:Chapman and Hall

[27] Hastie, T. J. (1992), Generalized additive models. Capítulo 7 de Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

[28] Hoaglin, D.C.; Welsch, R.E. (1978), The hat matrix in regression and ANOVA, *Amer. Statist*., 32, 17-22.

[29] Horowitz, J. (1998), Semiparametric methods in econometrics. *Lecture Notes in Statistics* 131, Springer, Heidelberg, Berlin, New York.

[30] Lee, D.K.C. (1990), Cross-Validation in Semiparametric Models: Some Monte Carlo Results. *Journal of Statistical Computation and Simulation*, 37, 171-187.

[31] Leontief, W. (1947), Introduction to a theory of the internal structure of functional relationships, *Econometrica*, 15, 361-373.

[32] Lima, L.P.; André, C.D.S.; Singer, J.M. (2001), Modelos Aditivos Generalizados: metodologia e prática, *R. Bras. Estat*., Rio de Janeiro, 62, 37-69

[33] McCullagh, P.; Nelder, J.A. (1989), Generalized linear models. 2.ed. London, Chapman and Hall.

[34] Moretin, P.A.; Toloi, C.M. (2004), Análise de séries temporais, São Paulo, Editora Blücher.

[35] Nelder, J.A.; Wedderburn, R.W.M.; (1972) Generalized linear models. *J.R. Statis. Soc. A* 135, 370-84.

[36] Opsomer, J.D.; Ruppert, D. (1999), A root-n consistent backfitting estimator for semiparametric additive modeling. *J. Comput. Graph. Statist*., 4, 715-732.
[37] Opsomer, J.D. (2000). *Asymptotic properties of backfitting estimators. J. Multivariate Anal*., 73, 166-179.
[38] Paula, G.A. (2004), Modelos de regressão com apoio computacional, Instituto de Matemática e Estatística, USP.
[39] Ramsay, T.O.; Burnett, R.T; Krewski, D. (2003), The Effect of Concurvity in Generalized Additive Models Linking Mortality to Ambient Particulate Matter, *Epidemiology*,14.
[40] Rice, J.A.; Rosenblatt, M. (1983), Smoothing splines, regression derivatives and convolution. *Ann. Statist*, 11, 141-56.
[41] Rothman, K.J.; Greenland, S. (1998). *Modern epidemiology*. 2.ed., Philadelphia, Lippincott-Raven, 737.
[42] Saldiva, P.H.N.; Pope III, C.A.; Schwartz, j.; Dockery, D.W.; Lichtenfels, A.J.F.C.; Salge, J.M.; Barone, I.A.; Böhm, G.M. (1995), Air pollution and mortality in elderly people: a time series study in São Paulo, Brazil. *Arch. Environmental Health*, 50, 150-163.
[43] Schick, A. (1986), On Asymptotically Efficient Estimation in Semiparametric Models. *The Annals of Statistics*, 14, 1139 – 1151.
[44] Schick, A. (1993), On efficient estimation in regression models. *The Annals of Statistics*, 21, 1486-1521.
[45] Schick, A. (1996), Root-n-consistent and Efficient estimation in semiparametric additive regression models. *Statistical & probability Letters*, 30, 45-51.
[46] Schwartz, J.; Slater, D.; Larson, T.V.; Pierson, W.E.; Koenig, J.Q. (1993), Particulate air pollution and hospital emergency room visits for asthma in Seattle, *Am Rev RespirDis*, 147, 826-831.
[47] Schwartz, J. (1994), Nonparametric smoothing in the analysis of air pollution and respiratory illness. *Canad. J. Statist.*, 22, 4, 471-487
[48] Silverman, B.W.(1985), Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser*, 47, 1-52.
[49] Stone, C.J. (1977), Consistent nonparametric regression, *Ann. Statist*. 5, 549-645.
[50] Stone, C.J. (1985), Additive Regression and Other Nonparametric Models. The *Annals of Statistics*, 13, 689-705.
[51] Whittaker, E. (1923), On a New Method of Graduation. *Proceedings of the Edinburgh Mathematical Society*, 41, 63-75.
[52] DETRAN-SP – Departamento Estadual de Trânsito de São Paulo. Frota de veículos. Disponível em: <http://www.detran.sp.gov.br>. Acesso em 21 abr. 2009.