Modelos Aditivos Generalizados: aplicação a um estudo epidemiológico ambiental

Liliam Pereira de Lima

DISSERTAÇÃO APRESENTADA AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA UNIVERSIDADE DE SÃO PAULO
PARA OBTENÇÃO DO GRAU DE
MESTRE EM ESTATÍSTICA

Área de Concentração: **Estatística**
Orientadora: **Profa. Dra. Carmen Diva Saldiva de André**

São Paulo, abril de 2001.

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Liliam Pereira de Lima e aprovada pela comissão julgadora.

São Paulo, 04 de abril de 2001.

Banca examinadora:

- Profa. Dra. Carmen Diva Saldiva de André (orientadora)
  Instituto de Matemática e Estatística / USP
- Prof. Dr. Julio da Motta Singer
  Instituto de Matemática e Estatística / USP
- Prof. Dr. Alfésio Luis Ferreira Braga
  Laboratório de Poluição Atmosférica Experimental - FM/USP

"Não desfrute somente o sol, aprecie também a lua. Não desfrute somente a calmaria, aproveite a tempestade. Tudo isso enriquece a existência."

(filosofia budista)

Aos meus queridos pais

Dora Cecilia e Laercio

Agradecimentos

Este trabalho de mestrado é resultado de uma jornada que começou na graduação quando vários professores do IME contribuíram para minha formação, dando exemplo de um trabalho sério e competente, fundamentado na dedicação e no estudo. Entre eles, não posso deixar de citar os professores Julio Singer e Lisbeth Cordani.

Nesta jornada, a presença dos amigos teve importância fundamental. Agradeço aqueles que me acompanharam nesta experiência única que é o IME, em especial, Lílian Natis e Adriana Sañudo. E a todos aqueles, incluindo meus irmãos Leonardo e Luis, que ficaram na torcida: saibam que eu pude sentir a energia de vocês e ela foi muito importante.

Ao meu querido Cesar Franco pelo incentivo e carinho que teve comigo durante todo o curso, e também por ter sido minha melhor fonte de força sempre que precisei.

Aos meus pais que tornaram meu caminho o mais confortável possível para que eu pudesse ir mais longe.

A todos aqueles que contribuíram para este trabalho. Entre eles, Alfésio Braga, Luiz Pereira, Paulo Saldiva e Simone Miraglia, do LPAE, Elisabete Oliveira, da biblioteca do IME, e a amiga Jacqueline David.

Um agradecimento especial a Carmen pela dedicada orientação, paciência e confiança. Este trabalho não teria sido possível sem sua contribuição e amizade.

Por fim, agradeço a Deus por ter, mais uma vez, colocado no meu caminho pessoas maravilhosas como as que conviveram comigo durante todo o curso.

Resumo

Os modelos aditivos generalizados (MAG's) constituem uma ampla classe de modelos de regressão, na qual o efeito de cada variável preditora na variável resposta é modelado de forma bastante flexível por meio de uma função f não especificada [Hastie e Tibshirani (1990). **Generalized Linear Models**. Chapman and Hall]. Os MAG's podem ser vistos como uma extensão dos modelos lineares generalizados (MLG's). Assim como estes modelos, os MAG's possuem uma metodologia unificada de análise que é apresentada neste trabalho. Inicialmente são apresentados métodos de suavização de diagramas de dispersão (especificamente, o *loess* e o *cubic spline*) que são ferramentas básicas para o ajuste dos MAG's. Procedimentos de estimação e testes para os parâmetros dos MAG's são, então, apresentados. A técnica é ilustrada com o uso de exemplos e foi aplicada a um estudo ambiental que buscou descrever a associação entre mortalidade fetal tardia e poluição atmosférica na cidade de São Paulo [Pereira et al. (1998). Association between Air Pollution and Intrauterine Mortality in São Paulo, Brazil. **Environmental Health Perspect.**, v. 106, n. 6, p. 325-329]. Estratégias de análise para este tipo de estudo também foram apresentadas. Os resultados obtidos foram comparados com os da análise do mesmo banco de dados via MLG, realizada por Pereira et al. (1998). A análise via MAG permitiu a construção de uma curva de risco relativo contínua que possibilitou uma melhor avaliação do impacto da poluição atmosférica na mortalidade fetal tardia.

Abstract

The generalized additive models (GAM's) consist of a large class of regression models. In which the effect for each of the covariates on the response variable is ajusted in a very flexible form using an unspecific function f [Hastie e Tibshirani. **Generalized Linear Models**. Chapman and Hall (1990)]. The GAM's could be seen like an extension of the generalized linear models (GLM's). Like these models, the GAM's possess a unified approach which is shown in this work. The methods shown first are scatterplot smoothers (specifically, loess and cubic spline) which are the basic tools to fit GAM's. The procedures of estimation and tests for the parameters of the GAM's are presented. The technique is illustrated with examples and were applied in an environmental study that described the association with still borns and atmospheric pollution in the city of São Paulo, Brazil [Pereira et al. (1998). Association between Air Pollution and Intrauterine Mortality in São Paulo, Brazil. **Environmental Health Perspect.**, v. 106, n. 6, p. 325-329]. The strategies for analysis for this study are also presented. The results were compared with the analysis on the same data by GLM, realized by Pereira et al. (1998). The analysis by GAM permits the construction of the continual risk relative curve, that possibility of a better evaluation of the impact of the atmospheric pollution an still borns.

Índice

Capítulo 1

Introdução

1.1. Introdução

Este trabalho apresenta técnicas de análise de modelos de regressão que pertencem à classe dos *modelos aditivos generalizados* (MAG's), a qual pode ser considerada uma generalização dos *modelos lineares generalizados* (MLG's) introduzidos por Nelder e Wedderburn (1972).

Os MLG's são da forma:

(1.1) $$g[\, E(Y|X_1,\ldots,X_p)\,] = \alpha + \beta_1 X_1 + \ldots + \beta_p X_p\,,$$

onde Y é uma variável resposta, $X_1,\ldots,X_p$ são variáveis preditoras, $g(\bullet)$ é a função de ligação que relaciona a média da resposta com as variáveis preditoras e α, $\beta_1,\ldots,\beta_p$ são parâmetros a serem estimados. Assume-se que a distribuição da variável aleatória Y pertence à família exponencial. Esses modelos constituem uma ampla e bastante difundida técnica de modelagem para relacionar a média da variável resposta com as preditoras $X_1,\ldots,X_p$, permitindo um leque de alternativas para a distribuição de Y. Modelos de regressão importantes, como o logístico e o de Poisson, pertencem a essa classe e podem ser ajustados utilizando-se uma metodologia estatística unificada.

Uma característica dos MLG's é que a forma da relação funcional entre a média da variável resposta e as variáveis preditoras é completamente especificada por termos paramétricos $\beta_1 X_1,\ldots,\beta_p X_p$, como mostra (1.1). Após o ajuste do modelo, técnicas de diagnóstico (Cook e Weisberg, 1982) podem ser utilizadas para identificar

possíveis falhas em sua especificação. Em particular, gráficos de resíduos e resíduos parciais podem apontar desvios da forma funcional adotada, indicando a adoção de um modelo mais adequado que inclua termos não lineares ou polinomiais nas variáveis preditoras. Entretanto, a simples inspeção dos resíduos pode não deixar claro quais funções das covariáveis são apropriadas.

Um procedimento alternativo é adotar um modelo não paramétrico no qual a relação entre a resposta e cada uma das variáveis preditoras é ditada pelos próprios dados. Modelos assim definidos constituem os *modelos aditivos generalizados* descritos por Hastie e Tibshirani (1990), cuja forma geral é

$$g[\, E(Y|X_1,\ldots,X_p)\,] = \alpha + f_1(X_1) + \ldots + f_p(X_p), \tag{1.2}$$

onde $g(\bullet)$ é definida como em (1.1), α é um parâmetro a ser estimado e $f_1,\ldots,f_p$ são funções arbitrárias não especificadas. A única restrição sobre essas p funções é que sejam *suaves*. O termo "aditivo" deve-se ao fato do modelo (1.2) ser constituído pela soma de funções das variáveis preditoras, o que permite avaliar o efeito de cada uma delas na variável resposta, condicionalmente à presença das outras covariáveis no modelo. A existência de interação entre duas covariáveis X_i e X_j, $i, j=1,\ldots,p$, $i\neq j$, pode ser avaliada definindo-se a variável $X_{p+1} = X_i X_j$, e adicionando-se o termo $f_{p+1}(X_{p+1})$ ao modelo (1.2). Alternativamente, a relação entre X_i e X_j com a variável resposta, controlando-se as demais variáveis preditoras, poderia ser descrita por uma função não especificada bivariada $f(X_i,X_j)$. Entretanto, funções multivariadas não serão consideradas neste trabalho. Os leitores interessados podem consultar Hastie e Tibshirani (1990) ou Hobert et al. (1997).

Na forma (1.2), o MAG é denominado totalmente não paramétrico. Entretanto, como será visto no **Capítulo 3**, modelos semi-paramétricos, constituídos pela soma de termos paramétricos de algumas variáveis preditoras e funções não especificadas de outras, também fazem parte dessa classe de modelos. Assim como os MLG's, os

MAG's podem ser analisados por meio de uma metodologia estatística unificada. O processo de ajuste do modelo requer a utilização de ferramentas denominadas *alisadores* ou *suavizadores*.

Sejam Y uma variável aleatória e X uma variável fixada. Uma técnica de suavização pode ser utilizada para descrever a dependência de Y em X. Por exemplo, considere o diagrama de dispersão apresentado na Figura 1.1 onde são representados 726 pontos (x_i,y_i) correspondentes a valores observados de X e Y. A relação entre essas variáveis é melhor visualizada com o auxílio da curva representada na Figura 1.2, obtida aplicando-se um procedimento de suavização aos pontos (x_i,y_i), i=1,...,726. Essa curva foi construída desenhando-se os pontos $(x_i, \hat{y}_i)$, onde $\hat{y}_i$ é o valor previsto de Y em $X = x_i$ obtido no processo de suavização, i=1,...,726. Essa curva também é considerada uma estimativa da função f no modelo

$$E(Y|X) = f(X),$$

para cada valor x_i, onde f é uma função não especificada a priori. A forma da curva estimada é determinada pelos próprios dados e, por esse motivo, os procedimentos de suavização são também denominados *técnicas de regressão não paramétrica*.

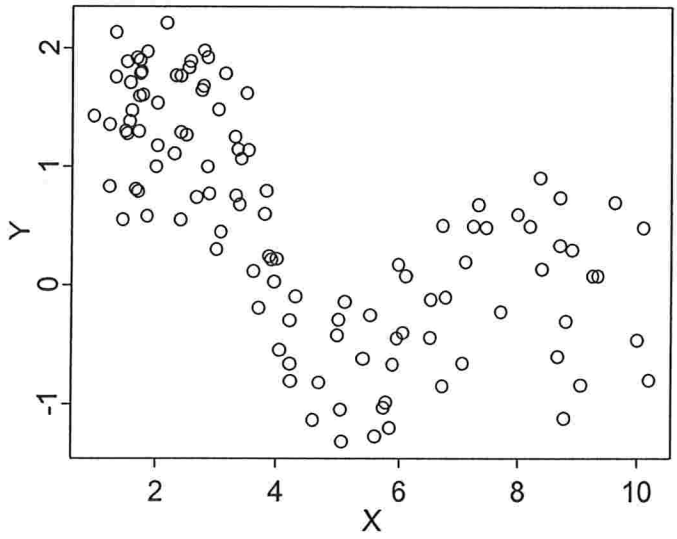


Figura 1.1: Diagrama de dispersão de X e Y.

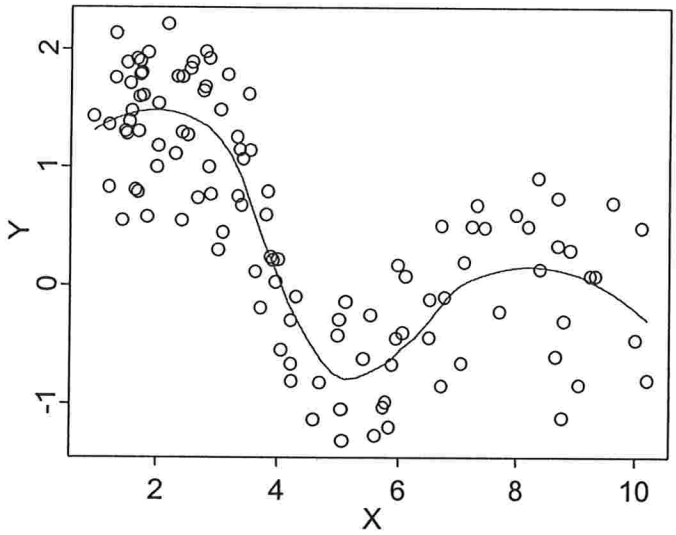


Figura 1.2: Diagrama de dispersão de X e Y com curva suavizada pelo método *loess*.

Existem várias técnicas de suavização (ver, por exemplo, Hastie e Tibshirani, 1990). A literatura sobre suas propriedades é bastante ampla, e não é objetivo deste trabalho esgotar o assunto. No **Capítulo 2** são apresentados alguns resultados gerais sobre alisadores e ilustradas com detalhes as técnicas *loess* e *cubic spline*. Estes alisadores foram escolhidos por apresentarem diferentes propriedades estatísticas e também por serem bastante utilizados em estudos sobre os efeitos da poluição atmosférica na saúde, que serão objeto da aplicação deste estudo.

O ajuste do modelo (1.2) é feito utilizando-se procedimentos similares aos adotados na estimação dos parâmetros dos MLG's pelo método da máxima verossimilhança. Esses procedimentos são descritos com detalhes no **Capítulo 3**, e basicamente consistem da combinação do algoritmo *scoring* de Fisher (McCullagh e Nelder, 1989) com procedimentos de suavização.

Como ilustração, no **Capítulo 4**, um MAG foi adotado para descrever a associação entre poluição atmosférica e óbitos em fetos com mais de 28 semanas (óbitos fetais tardios) na cidade de São Paulo, durante os anos de 1991 e 1992. O banco de dados utilizado foi, originalmente, analisado por Pereira et al. (1998) por meio de um modelo de regressão de Poisson. Como em geral é feito no estudo de possíveis associações entre poluição atmosférica e mortalidade ou morbidade, o modelo foi construído de forma a relacionar uma série cronológica de natimortalidade com séries cronológicas de concentrações de poluentes, controlando-se variáveis preditoras climáticas e temporais (variáveis confundidoras). Uma nova análise desses dados utilizando a mesma técnica foi feita e seus resultados foram comparados com os obtidos por MAG. Assumiu-se a distribuição de Poisson para a variável resposta, que corresponde a um processo de contagem de eventos raros.

No ajuste do modelo foi adotada uma estratégia de análise que busca controlar adequadamente as variáveis confundidoras e eliminar o efeito de autocorrelação dos resíduos.

Características do estudo, tais como informações sobre suas fontes de dados, bem como algumas características demográficas da cidade de São Paulo que possibilitam sua realização foram também abordadas.

As conclusões sobre a existência dos efeitos dos poluentes via MLG e via MAG foram as mesmas. O Dióxido de Nitrogênio foi o poluente que melhor se associou com a mortalidade fetal tardia, mostrando um comportamento dose-dependente.

Algumas considerações finais sobre o trabalho e propostas para estudos futuros são feitas no **Capítulo 5**.

Capítulo 2

Métodos de Suavização

2.1. Introdução

Um suavizador é uma ferramenta que descreve a tendência de uma variável Y como função de uma ou mais variáveis $X_1,\ldots,X_p$. A estimativa obtida de um procedimento de suavização tem variabilidade menor que a de Y, como pode ser observado no exemplo ilustado na Figura 1.2. Esta é a razão do nome suavizador ou alisador.

Quando a tendência de Y é descrita em função de apenas uma variável X, o alisador é denominado *unidimensional*. Quando p variáveis $X_1,\ldots,X_p$ são consideradas, diz-se que o alisador é *multidimensional*. Apenas alisadores unidimensionais serão apresentados neste trabalho.

Neste capítulo serão apresentados os procedimentos de suavização (ou suavizadores) que serão aplicados a conjuntos de dados consistindo de n pares de pontos (x_i,y_i), i=1,...,n, que se referem a valores x_i de uma variável X e correspondentes valores y_i de uma variável Y. Será assumido de início que os dados não apresentam réplicas nos valores de X. No final do capítulo serão feitas algumas considerações sobre o processo de suavização quando ocorrem empates.

O resultado de um procedimento de suavização consiste de valores ajustados de Y, $\hat{y}_i$, para cada x_i, i=1,...,n. Estes valores são obtidos sem a adoção de um modelo

paramétrico relacionando Y e X. Os pontos $(x_i, \hat{y}_i)$ podem ser representados graficamente resultando em uma *curva suavizada* ou *curva ajustada*.

Formalmente, um *suavizador* é definido como uma função de $\mathbf{x} = (x_1,\ldots,x_n)$ e $\mathbf{y} = (y_1,\ldots,y_n)$ cujo resultado é uma função $s = S(\mathbf{y}|\mathbf{x})$ com mesmo domínio de $\mathbf{x}$. Para alguns suavizadores, $s(x_o)$, isto é, a função $S(\mathbf{y}|\mathbf{x})$ calculada em x_o, é definida para todo x_o. Outras vezes ela é definida apenas para $x_1,\ldots,x_n$ que são os valores observados de X, e neste caso, algum tipo de interpolação é necessária para obter estimativas em outros valores de X.

As curvas suavizadas podem ser utilizadas com diferentes objetivos:

• *descrever a tendência dos dados em diagramas de dispersão* – as curvas suavizadas podem ser usadas para indicar a tendência em diagramas de dispersão. Uma aplicação interessante foi feita por Schwartz (1994), na qual um alisador foi utilizado em gráficos de resíduos a fim de detectar possíveis tendências dos resíduos em relação aos valores previstos da variável resposta, ou aos valores observados de uma variável preditora;

• *verificar a adequação de um modelo de regressão* – Neter et al. (1996) ilustraram a utilização de alisadores no diagnóstico de modelos de regressão com o objetivo de verificar se a escolha de um determinado modelo é adequada. Para isto, os autores descrevem o seguinte procedimento: considere o ajuste de um modelo de regressão linear simples tendo Y como variável resposta e X como preditora; num mesmo diagrama de dispersão, desenhe as bandas de confiança resultantes desse ajuste e a curva suavizada de Y em X. Se a curva suavizada estiver dentro das bandas de confiança para o modelo de regressão ajustado há indicação de que esse modelo é adequado aos dados;

• *auxiliar na escolha de valores iniciais de um processo iterativo* – os alisadores podem ser utilizados para auxiliar a escolha de valores iniciais dos parâmetros em

processos iterativos para estimação de modelos de regressão não linear, facilitando a convergência do processo. Um exemplo no qual tal procedimento foi adotado pode ser encontrado em André e Rancan (1999);

• *estudar a dependência da esperança da variável resposta Y em função de uma variável preditora X* – neste caso, o alisador é utilizado com o objetivo de ajustar o modelo

$$y_i = f(x_i) + \varepsilon_i, \quad i=1,\ldots,n, \tag{2.1}$$

onde f é uma função arbitrária desconhecida e os ε_i's são erros aleatórios distribuídos independentemente com média zero e variância σ^2. Este modelo pode ser considerado uma generalização do modelo de regressão linear simples que tem Y como variável resposta e X como variável preditora. Desde que $E(Y|X=x_i) = f(x_i)$, qualquer estimativa de f pode ser vista como uma estimativa desta esperança condicional.

Como neste trabalho a principal utilização dos estimadores é no ajuste de modelos de regressão, a notação utilizada na relação (2.1) será adotada no restante do capítulo.

Na maioria das técnicas de suavização, o valor suavizado $\hat{y}_i$ é obtido com base em uma *média* de *r* observações na vizinhança de um dado valor x_i. Diferentes formas de cálculo dessa média em uma vizinhança de x_i definem os diferentes métodos de suavização.

A escolha do *tamanho da vizinhança* é um problema importante no processo de suavização. Ele é associado a um parâmetro λ denominado *parâmetro de suavização* que deve ser fixado antes do início do processo. A escolha de valores para este parâmetro está relacionada à relação de ganho e perda entre o *viés* e a *variância* da curva estimada, que serão definidos na seção 2.4.

Os suavizadores são classificados como *lineares* ou *não lineares*. Um suavizador é dito linear quando o vetor de valores previstos $\hat{\mathbf{y}} = (\hat{y}_1,\ldots,\hat{y}_n)' = (\hat{f}(x_1),\ldots,\hat{f}(x_n))' = \hat{\mathbf{f}}$ pode ser escrito como

$$\hat{\mathbf{f}} = \mathbf{S}\mathbf{y}, \tag{2.2}$$

onde $\mathbf{y} = (y_1,\ldots,y_n)'$ é o vetor dos valores observados da variável resposta e $\mathbf{S} = \{s_{ij}\}$ é uma matriz de dimensão n x n chamada *matriz suavizadora*, que não depende de $\mathbf{y}$. $\mathbf{S}$ depende apenas de X e do parâmetro de suavização λ. O valor ajustado de Y em x_i é

$$\hat{y}_i = \hat{f}(x_i) = s_{x_i,1}\, y_1 + s_{x_i,2}\, y_2 + \ldots + s_{x_i,n}\, y_n, \quad i=1,\ldots,n,$$

onde $(s_{x_i,1},\ldots, s_{x_i,n})$ é a i-ésima linha da matriz $\mathbf{S}$.

Pelo fato de suas matrizes suavizadoras não dependerem da variável resposta Y, a análise por meio de suavizadores lineares torna-se relativamente simples.

A literatura sobre alisadores lineares é bastante ampla. Buja et al. (1989) e Hastie e Tibshirani (1990) descrevem vários desses alisadores e fornecem bibliografia adicional sobre o assunto.

Neste capítulo serão descritos dois procedimentos de suavização: o *loess* que apresenta uma versão robusta que é não linear e uma não robusta que é linear, e o *cubic spline* que é uma técnica de suavização linear.

Estes procedimentos serão, em geral, ilustrados com base no conjunto de dados apresentado na Tabela 2.1. Os valores x_i, i=1,...,10, foram gerados a partir de uma distribuição N(25,16). De forma independente, foram gerados erros e_i, i=1,...,10, segundo uma distribuição Normal padrão. As respostas y_i foram obtidas pela relação $y_i = 10 + 0{,}1x_i + e_i$. O diagrama de dispersão de (x_i,y_i) é mostrado na Figura 2.1.

Tabela 2.1: Dados gerados segundo o modelo $y_i = 10 + 0{,}1x_i + e_i$, i=1,...,10.

i	x_i	y_i
1	–2,58	10,18
2	3,69	11,55
3	7,29	11,25
4	11,31	12,57
5	12,32	11,33
6	14,49	10,50
7	17,22	12,20
8	24,76	12,81
9	39,07	15,82
10	62,04	16,30

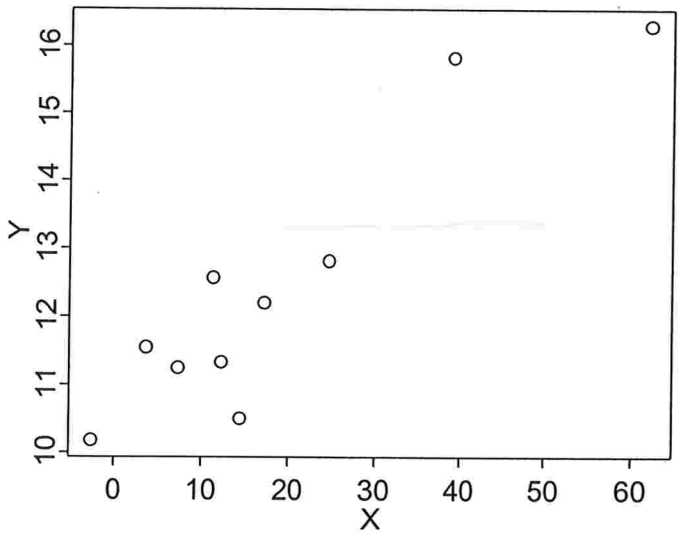


Figura 2.1: Diagrama de dispersão dos dados apresentados na Tabela 2.1.

2.2. Suavizador *loess* robusto

Suponha que se deseja suavizar um conjunto de pontos (x_i,y_i), i=1,...,n. O *loess* robusto (*robust locally-weighted scatterplot smoother)*, proposto por Cleveland (1979), é um método de suavização que se baseia no ajuste sucessivo de n modelos de

regressão pelo método de mínimos quadrados ponderados (MQP). Cada modelo é ajustado considerando observações cujo valor de X pertence a uma vizinhança da coordenada x_i de uma observação (x_i,y_i) fixada que é denominada *ponto alvo*, i=1,...,n. O valor ajustado é $\hat{y}_i = \hat{f}(x_i)$. Portanto, considerando sucessivamente as n observações (x_i,y_i) como ponto alvo, obtêm-se os pontos $(x_i, \hat{f}(x_i))$, i=1,...,n, que geram a curva suavizada.

Basicamente o método divide-se em duas etapas descritas a seguir.

Primeira etapa – Regressão ponderada local (*locally-weighted scatterplot smoother - loess*).

Para cada ponto alvo (x_i,y_i) define-se uma vizinhança, e aos pontos (x_j,y_j) nessa vizinhança é ajustado um polinômio de grau *d*: $y_j = \alpha + \beta_1 x_j + \ldots + \beta_d x_j^d + e_j$, j=1,...,n, por MQP com pesos dados por uma função U a ser definida.

A vizinhança de cada (x_i,y_i) é constituída pelos *r* pares de observações (x_j,y_j), que possuem as coordenadas x_j mais próximas a x_i. O número de pontos *r* a ser considerado é dado por

$$r = \lambda\, n,$$

onde λ $(0 < \lambda \leq 1)$ é o parâmetro de suavização, cujo valor deve ser fixado a priori, e corresponde à proporção do número total de observações a ser utilizado em cada ajuste local.

Não existe um critério rígido para a escolha do valor de λ. A literatura sugere que valores razoáveis para este parâmetro estão entre 1/3 e 2/3. No entanto, a escolha de um valor adequado, muitas vezes é obtida após o teste de vários valores de λ para um mesmo conjunto de dados. O parâmetro de suavização tem influência fundamental na variância e no viés da curva estimada: aumentar λ implica aumentar a suavização

da curva (diminuir a variância), porém, pode-se perder informação no ajuste (aumentar o viés). Este efeito é ilustrado na Figura 2.2, onde um conjunto de dados é suavizado pelo método *loess* com diferentes valores de λ.

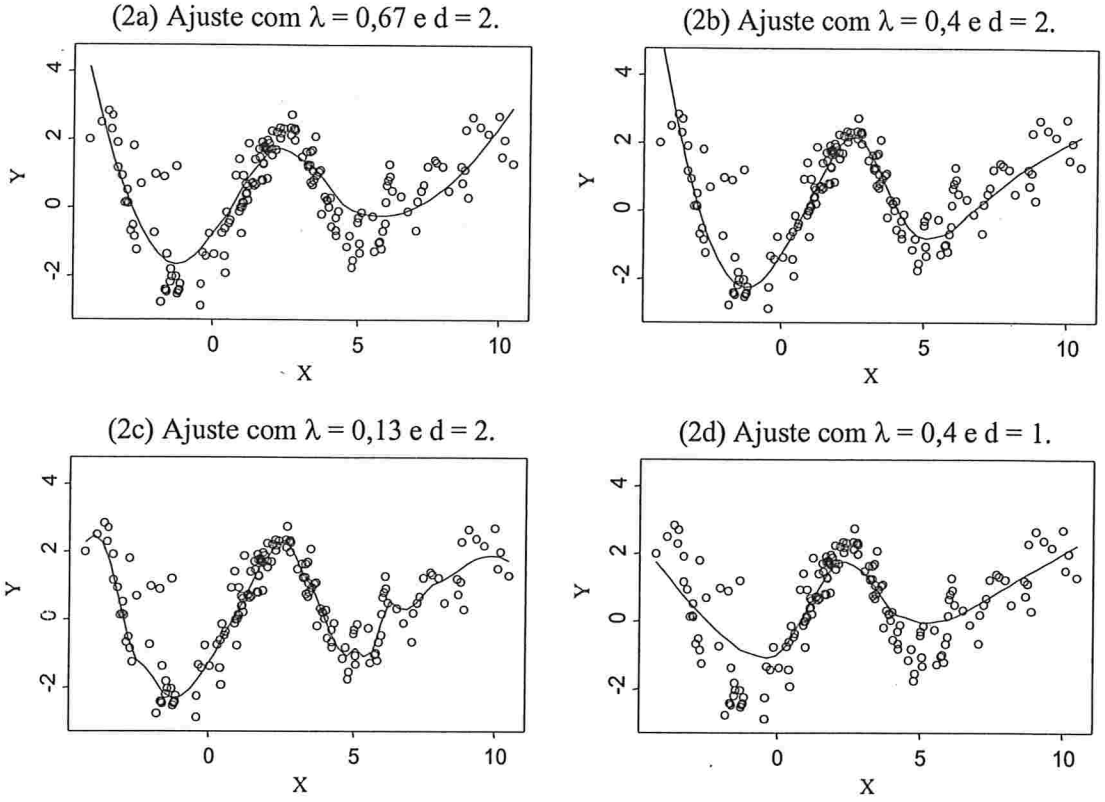


Figura 2.2: Curvas suavizadas pelo método *loess* com diferentes valores de λ e *d*.

A Figura 2.2a apresenta uma curva bastante suave, ou seja, pouco *ondulada* ($\lambda = 0,67$), porém, a curva não se adapta aos *picos* e aos *vales* adequadamente. Quando λ é reduzido para 0,4 (Figura 2.2b), a curva ajustada torna-se mais *ondulada* (menos suave), e consegue descrever melhor o comportamento dos dados (menor viés). Reduzindo ainda mais o valor de λ (Figura 2.2c), obtém-se uma curva muito *ondulada*, que não parece representar a verdadeira tendência dos dados. Um valor

adequado para λ neste exemplo seria 0,4, uma vez que este é o maior valor desse parâmetro (dentre os considerados) capaz de minimizar a variabilidade dos valores ajustados sem distorcer a tendência dos dados.

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 $y = \alpha + \beta x$, isto é $d = 1$, é adequado. Mas se existirem, regiões com máximos ou mínimos locais (como os dados da Figura 2.2), então um ajuste quadrático $y = \alpha + \beta_1 x + \beta_2 x^2$, isto é $d = 2$, normalmente produz uma curva que descreve melhor o padrão dos dados localmente.

As Figuras 2.2b e 2.2d foram suavizadas fixando-se $\lambda = 0{,}4$ e, respectivamente, $d = 2$ e $d = 1$. Observa-se que o ajuste linear (Figura 2.2d) não é capaz de acomodar os máximos e mínimos locais pois a curva permanece abaixo dos *picos* e acima dos *vales*. Para conseguir tal acomodação seria necessário diminuir muito o valor de λ. Já o ajuste local quadrático (Figura 2.2b), atinge os *picos* e *vales* com o mesmo valor de λ, sugerindo uma suavização mais adequada.

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

$$u_{x_i,j} = U(h_i^{-1}(x_j - x_i)), j=1,\ldots,n,$$

onde h_i é a distância entre x_i e o seu *r*-ésimo vizinho mais próximo, isto é, colocando em ordem crescente as distâncias $|x_i - x_k|$, $k = 1,\ldots,n$, h_i é a distância que ocupa a *r*-ésima posição nessa seqüência ordenada. Essa função deve ser especificada de forma a possuir as seguintes propriedades:

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

ii. $U(-g) = U(g)$;

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

iv. U(g) = 0 para $|g| \geq 1$.

A função tricúbica, dada por :

$$U(g) = \begin{cases} \left(1-|g|^3\right)^3 & \text{para } |g| < 1 \\ 0 & \text{para } |g| \geq 1, \end{cases} \tag{2.3}$$

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

Com base na função (2.3) obtém-se a matriz de pesos referente ao ponto alvo (x_i,y_i), denotada por

$$\mathbf{U}_{x_i} = \text{diagonal}\{u_{x_i,1},\ldots,u_{x_i,n}\}, \tag{2.4}$$

com elementos dados por:

$$u_{x_i,j} = \begin{cases} \left(1-\left|h_i^{-1}(x_j - x_i)\right|^3\right)^3 & \text{para } \left|h_i^{-1}(x_j - x_i)\right| < 1 \\ 0 & \text{caso contrário.} \end{cases} \tag{2.5}$$

Assim, por (2.4), 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.

O valor ajustado é, então, calculado a partir de um modelo de regressão ajustado por MQP com pesos dados por $\mathbf{U}_{x_i}$. A primeira etapa termina após a realização deste procedimento para cada x_i, i=1,...,n, obtendo-se valores suavizados $\hat{f}(x_1),\ldots,\hat{f}(x_n)$.

Segunda etapa – Ajuste robusto.

A segunda etapa visa tornar o ajuste robusto, introduzindo um novo conjunto de pesos δ_j definidos para cada (x_j,y_j), j=1,...,n, os quais serão usados no ajuste de MQP para obter os novos valores $\hat{f}(x_j)$. Estes pesos, chamados de pesos robustos, são baseados no tamanho dos resíduos $\hat{e}_j = y_j - \hat{f}(x_j)$ gerados pelos modelos de regressão ajustados na etapa anterior, sendo que, observações com pequenos resíduos ficam associadas a pesos maiores do que observações com grandes resíduos.

Os pesos desta etapa são definidos por:

$$\delta_j = B\left(\hat{e}_j / 6s\right), \quad j=1,\ldots,n, \tag{2.6}$$

onde $\hat{e}_j = y_j - \hat{f}(x_j)$, s = mediana de $|\hat{e}_j|$ e B = função biquadrada definida por:

$$B(x) = \begin{cases} (1-x^2)^2 \text{ para } |x| < 1 \\ 0 \qquad \text{para } |x| \geq 1. \end{cases}$$

Podemos dizer que s é uma medida da variabilidade dos resíduos. Se o resíduo correspondente ao valor observado y_j é pequeno em relação a $6s$, o peso robusto atribuído a (x_j,y_j) será próximo de 1; por outro lado, se for muito maior do que $6s$, esse peso será próximo de zero. Como valores atípicos costumam gerar resíduos grandes, observações espúrias recebem peso menor, e neste sentido, o *loess* torna-se um método de suavização robusto.

Após o cálculo dos pesos δ_j, novos n ajustes locais do polinômio de grau d são executados pelo método de mínimos quadrados ponderados. Quando o ponto alvo é a observação (x_i,y_i), os pesos são dados por $\delta_j u_{x_i,j}$, j=1,...,n, i=1,...,n.

Novamente, após a realização da segunda etapa para cada x_i, i=1,...,n, obtêm-se os correspondentes valores ajustados $\hat{f}(x_1),\ldots,\hat{f}(x_n)$.

Esta etapa do processo é repetida t vezes. Cada iteração da etapa robusta reduz o peso dado a valores aberrantes em Y. Segundo Cleveland (1979), estudos de simulação e análise de dados reais indicam que, na grande maioria dos casos, $t = 2$ iterações são suficientes para amenizar o efeito de observações aberrantes.

O processo total, incluindo as 2 etapas descritas anteriormente, é chamado de *loess* robusto. No entanto, quando não há valores aberrantes, pode-se usar apenas a primeira etapa do processo. Neste caso, o processo é conhecido apenas como *loess*.

Exemplo: ajuste da curva

Considerando os dados apresentados na Tabela 2.1, será ilustrado neste exemplo o método *loess* robusto de suavização. Os valores dos parâmetros escolhidos para a realização da primeira etapa foram:

- $d = 1$ (isto é, um ajuste linear local) pois o diagrama de dispersão (Figura 2.1) sugere uma tendência linear, e
- $\lambda = 0{,}5$ e portanto, $r = 0{,}5 \times 10 = 5$ será o número de pontos que efetivamente serão usados em cada regressão local.

Na primeira etapa do alisamento, para obter o valor $\hat{f}(x_i)$ correspondente a um dado valor alvo x_i, por exemplo, $x_3 = 7{,}29$, calcularam-se as distâncias entre x_3 e x_k, $k=1,\ldots,10$. Estas distâncias são apresentadas na Tabela 2.2, onde se observam os $r = 5$ valores mais próximos de x_3 colocados em destaque, sendo que a 5^{a} menor distância é $h_3 = 7{,}21$.

Os pesos $u_{x_3,j}$, $j=1,\ldots,10$, a serem usados na regressão local, foram obtidos a partir de (2.5) e são apresentados na Tabela 2.3. A função $u_{x_3,j}$ tem seu máximo em $x_j = x_3$ e decresce à medida que os valores x_j se distanciam deste valor, tornando-se

zero para os pontos que satisfazem $\left|h_3^{-1}(x_j - x_3)\right| \geq 1$. Assim, apenas os pontos representados na Figura 2.3b, são efetivamente considerados na obtenção de $\hat{f}(x_3)$.

Tabela 2.2: Distâncias relativas a x_3.

k	x_k	$\lvert x_3 - x_k \rvert$
1	–2,58	9,86
2	***3,69***	***3,59***
3	***7,29***	***0,00***
4	***11,31***	***4,03***
5	***12,32***	***5,04***
6	***14,49***	***7,21***
7	17,22	9,94
8	24,76	17,47
9	39,07	31,79
10	62,04	54,76

Tabela 2.3: Conjunto de pesos na primeira etapa do ajuste, com ponto alvo (x_3,y_3).

j	$u_{x_3,j}$
1	0
2	0,673
3	1,000
4	0,563
5	0,286
6	0,000
7	0
8	0
9	0
10	0

Calcularam-se então, as estimativas dos parâmetros da reta de regressão por MQP com pesos dados por $u_{x_3,j}$, j=1,...,10. A reta de regressão assim ajustada é

$\hat{y}_i = 11{,}03 + 0{,}08x_i$, a qual fornece o valor previsto $\hat{f}(x_3) = 11{,}59$, representado na Figura 2.3c.

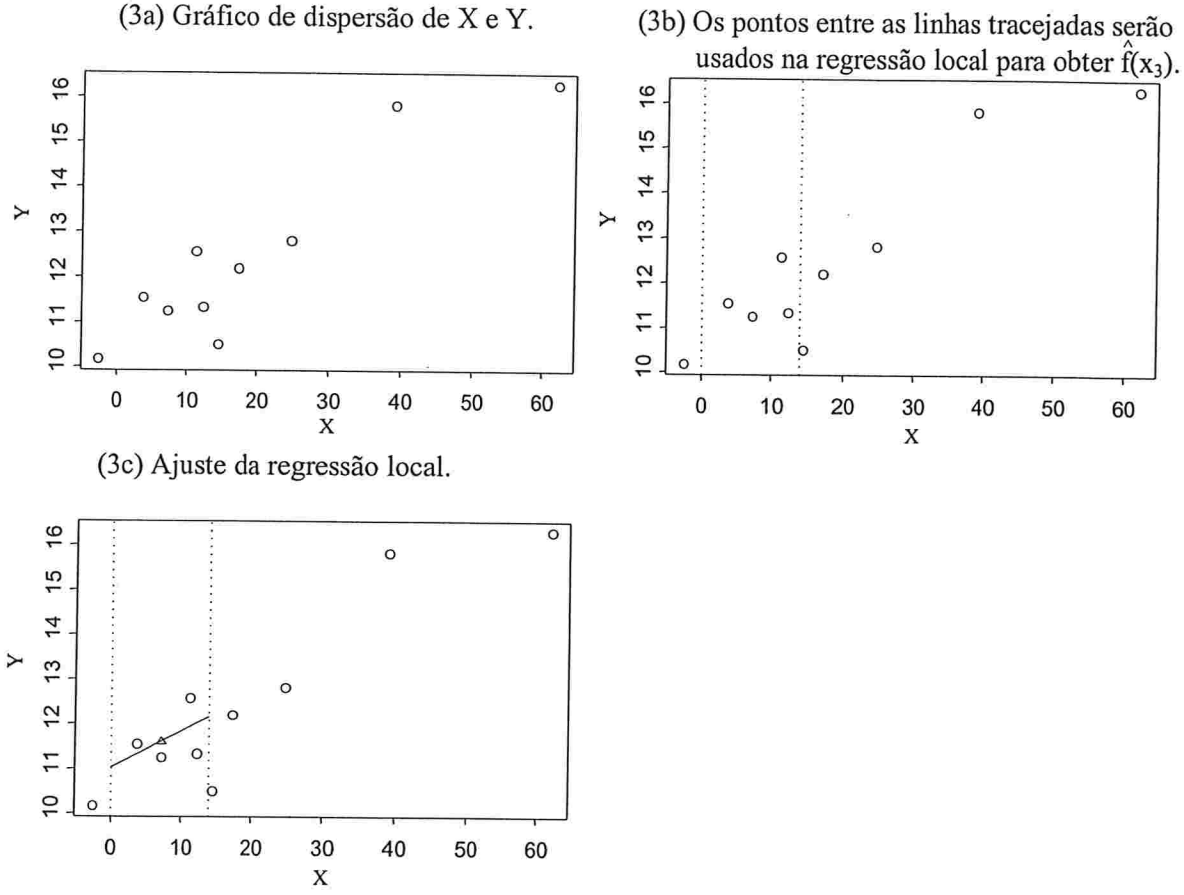


Figura 2.3: Etapas do ajuste *loess* no ponto alvo (x_3,y_3).

De forma semelhante, foram calculados os valores ajustados correspondentes às demais observações (x_i,y_i) quando cada uma delas foi considerada como ponto alvo, encerrando assim, a primeira etapa do ajuste *loess* robusto. Estes valores são mostrados na Tabela 2.4 e Figura 2.4.

Tabela 2.4: Estimativas dos parâmetros das retas de regressão e valores ajustados resultantes da primeira etapa do método *loess* robusto.

i	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{f}(x_i)$
1	10,66	0,15	10,27
2	10,89	0,09	11,22
3	11,03	0,08	11,59
4	13,53	–0,16	11,67
5	19,26	–0,61	11,69
6	11,97	–0,03	11,56
7	8,95	0,17	11,82
8	9,41	0,14	12,87
9	8,10	0,20	15,75
10	13,21	0,05	16,46

Para a realização da primeira iteração na segunda etapa, calcularam-se os resíduos $\hat{e}_j = y_j - \hat{f}(x_j)$ a partir dos valores ajustados na primeira etapa, e os pesos δ_j dados pela relação (2.6). Então, novos ajustes locais foram realizados, um para cada ponto alvo (x_i,y_i), i=1,...,10, por MQP com pesos dados por $\delta_j u_{x_i,j}$. Na Tabela 2.5 são apresentados os pesos $\delta_j u_{x_3,j}$, correspondentes ao ajuste local com ponto alvo (x_3,y_3).

A partir dos resultados desta iteração, a segunda etapa foi repetida completando um total de $t = 2$ iterações. Os valores previstos $\hat{f}(x_i)$, i=1,...,n, então obtidos, são apresentados na Tabela 2.6. Nota-se que os valores de $\hat{f}(x_i)$ apresentados nas Tabelas 2.4 e 2.6 são muito próximos, como seria esperado já que os y_i foram gerados a partir da distribuição Normal e, portanto, não é esperado que ocorram valores aberrantes. Na Figura 2.4 são apresentadas as curvas suavizadas obtidas ao final da primeira e segunda etapas.

Tabela 2.5: Conjunto de pesos utilizados no ajuste da regressão local para obter $(x_3, \hat{f}(x_3))$ na primeira iteração da segunda etapa do método *loess* robusto.

j	$\lvert\hat{e}_j\rvert$	δ_j	$\delta_j u_{x_3,j}$
1	0,10	0,9953	0,0000
2	0,33	0,9478	0,6380
3	0,34	0,9452	0,9425
4	0,90	0,6405	0,3607
5	0,36	0,9366	0,2681
6	1,05	0,5294	0,0000
7	0,09	0,9306	0,0000
8	0,08	0,9984	0,0000
9	0,07	0,9973	0,0000
10	0,15	0,9883	0,0000

Tabela 2.6: Valores previstos de Y obtidos na segunda etapa de ajuste pelo método *loess* robusto, após $t = 2$ iterações.

i	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{f}(x_i)$
1	10,65	0,15	10,27
2	10,88	0,09	11,21
3	11,23	0,04	11,50
4	12,06	–0,05	11,54
5	19,46	–0,64	11,56
6	10,73	0,07	11,75
7	9,19	0,17	12,07
8	10,28	0,10	12,83
9	8,09	0,20	15,75
10	13,21	0,05	16,46

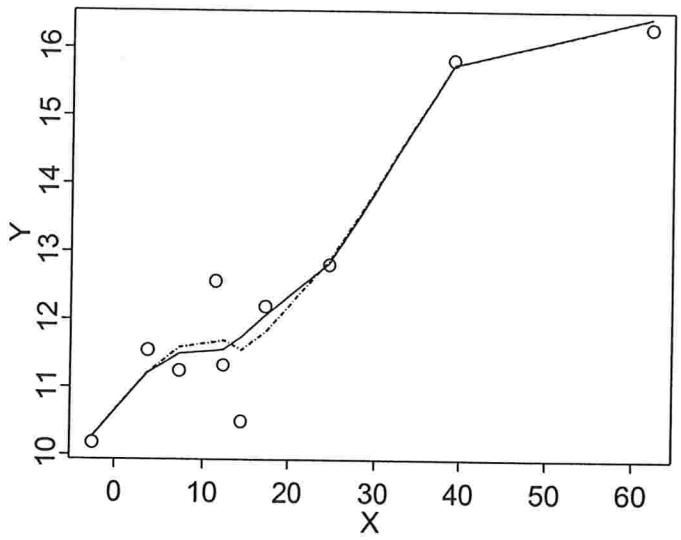


Figura 2.4: Curvas suavizadas obtidas pelo método *loess* robusto para os dados da Tabela 2.1. A linha pontilhada representa a curva estimada na primeira etapa do procedimento e a linha contínua representa a curva estimada na segunda etapa, após duas iterações.

Apenas como ilustração, suponha que o ponto (11,31;12,57) da Tabela 2.1 seja substituído por (11,31;14,00). Na Figura 2.5 são representados os dados após esta modificação e também as curvas suavizadas resultantes da primeira etapa (unidos por uma linha pontilhada) e da segunda etapa após 2 iterações (unidos por uma linha contínua). A comparação destas curvas mostra que o ajuste da segunda etapa fornece uma curva menos influenciada pelo valor modificado.

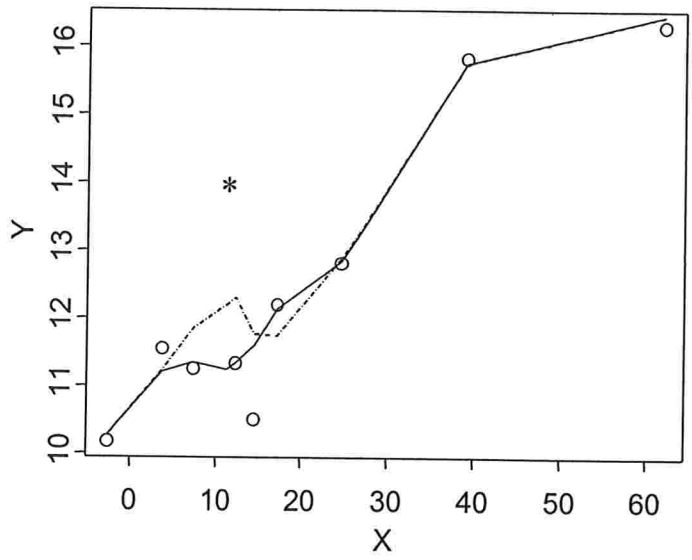


Figura 2.5: Curvas suavizadas pelo *loess* robusto para os dados da Tabela 2.1 após incluir um valor aberrante (*). A linha pontilhada representa a curva estimada na primeira etapa do procedimento e a linha contínua representa a curva estimada na segunda etapa, após duas iterações.

Matriz suavizadora

No início desta seção foi descrito o método de suavização *loess* robusto. Alguns autores, como Hastie e Tibshirani (1990), quando se referem ao *loess*, estão considerando apenas a primeira etapa deste método. Neste caso, os valores previstos de Y obtidos no procedimento de suavização, podem ser escritos na forma dada em (2.2). Os elementos da matriz suavizadora **S** são denotados por:

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

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

(2.7) $$\mathbf{S}_{x_i} = \mathbf{X}(\mathbf{X}'\mathbf{U}_{x_i}\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}_{x_i},$$

construída no ajuste da regressão ponderada local que tem (x_i,y_i) como ponto alvo e matriz de pesos $\mathbf{U}_{x_i}$ definida em (2.4), i=1,...,n.

O valor previsto correspondente a x_i é, então, dado por

$$\hat{f}(x_i) = s_{x_i,1}\, y_1 + s_{x_i,2}\, y_2 + \ldots + s_{x_i,n}\, y_n = \mathbf{s'}_{x_i}\, \mathbf{y},\ i=1,\ldots,n.$$

De uma forma geral, pode-se mostrar que o elemento ij da matriz suavizadora **S** é dado por:

$$(2.8) \qquad s_{x_i,j} = \frac{u_{x_i,j}\sum_{j=1}^{n} x_j^2 u_{x_i,j} - u_{x_i,j}(x_i + x_j)\sum_{j=1}^{n} x_j u_{x_i,j} + x_i x_j u_{x_i,j}\sum_{j=1}^{n} u_{x_i,j}}{\sum_{j=1}^{n} u_{x_i,j}\sum_{j=1}^{n} x_j^2 u_{x_i,j} - \left(\sum_{j=1}^{n} u_{x_i,j} x_j\right)^2},$$

onde $u_{x_i,j}$ é definido de acordo com (2.5).

As expressões (2.7) e (2.8) mostram claramente que os elementos da matriz **S** dependem apenas de $x_1,\ldots,x_n$ e do parâmetro de suavização λ. Portanto, o *loess* é um suavizador linear.

A matriz suavizadora **S** desempenha papel semelhante ao da matriz *hat* no método de estimação de mínimos quadrados e algumas de suas propriedades são demonstradas por Hoaglin e Welsch (1978), a saber

$$(2.9) \qquad \begin{aligned} &i.\ 0 \le s_{ii} \le 1, \\ &ii.\ -1 \le s_{ij} \le 1 \text{ para } i \ne j, \\ &iii.\ s_{ii} = 1 \text{ se e somente se } s_{ij} = 0 \text{ para todo } i \ne j, \text{ e} \\ &iv.\ \sum_{j=1}^{n} s_{ij} = 1. \end{aligned}$$

Porém, **S** não é simétrica nem idempotente, ou seja, não é um operador projeção como ocorre com a matriz *hat*. Como **S** não é simétrica, seus autovalores podem não ser reais.

Exemplo: cálculo da matriz S

Utilizando novamente os dados da Tabela 2.1, e fixando $\lambda = 0{,}5$ e $d = 1$, o cálculo da matriz **S** correspondente ao método *loess* é ilustrado a seguir.

Considere o ajuste da regressão ponderada local tendo (x_3,y_3) como ponto alvo. A matriz **X** e a matriz de pesos $\mathbf{U}_{x_3}$ são dadas por

$$\mathbf{X} = \begin{bmatrix} 1 & -2{,}58 \\ 1 & 3{,}69 \\ 1 & 7{,}29 \\ 1 & 11{,}31 \\ 1 & 12{,}32 \\ 1 & 14{,}49 \\ 1 & 17{,}22 \\ 1 & 24{,}76 \\ 1 & 39{,}07 \\ 1 & 62{,}04 \end{bmatrix} \text{ e } \mathbf{U}_{x_3} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0{,}673 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0{,}563 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0{,}286 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0{,}000 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Portanto,

$$\mathbf{S}_{x_3} = \mathbf{X}\,(\mathbf{X}'\mathbf{U}_{x_3}\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}_{x_3} = \begin{bmatrix} 0 & 1{,}44 & 0{,}61 & -0{,}62 & -0{,}44 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0{,}73 & 0{,}48 & -0{,}11 & -0{,}10 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0{,}32 & 0{,}41 & 0{,}18 & 0{,}09 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0{,}13 & 0{,}32 & 0{,}51 & 0{,}30 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0{,}25 & 0{,}30 & 0{,}59 & 0{,}35 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0{,}49 & 0{,}26 & 0{,}77 & 0{,}47 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0{,}80 & 0{,}20 & 0{,}99 & 0{,}61 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1{,}65 & 0{,}04 & 1{,}60 & 1{,}01 & 0 & 0 & 0 & 0 & 0 \\ 0 & -3{,}27 & -0{,}26 & 2{,}76 & 1{,}77 & 0 & 0 & 0 & 0 & 0 \\ 0 & -5{,}87 & -0{,}74 & 4{,}62 & 2{,}99 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

e a terceira linha da matriz **S** é:

$$[0 \;\; 0{,}32 \;\; 0{,}41 \;\; 0{,}18 \;\; 0{,}09 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0].$$

Repetindo o procedimento considerando outros pontos (x_i,y_i) como alvo, i=1,...,10, obtêm-se as demais linhas da matriz **S**, que é dada por

$$\mathbf{S} = \begin{bmatrix} 0,94 & 0,15 & -0,09 & -0,00 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0,18 & 0,52 & 0,30 & 0,01 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0,32 & 0,41 & 0,18 & 0,09 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0,18 & 0,37 & 0,32 & 0,13 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0,42 & 0,38 & 0,20 & 0,00 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0,15 & 0,21 & 0,30 & 0,34 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0,06 & -0,07 & 0,25 & 0,88 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0,00 & -0,04 & 0,06 & 0,98 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0,01 & 0,04 & 0,96 & 0,01 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0,00 & -0,05 & 0,09 & 0,96 \end{bmatrix}.$$

Para ilustrar a influência de λ nos elementos da matriz **S**, considere esta matriz para o mesmo conjunto de dados, com $\lambda = 0,4$. Neste caso, tem-se que

$$\mathbf{S} = \begin{bmatrix} 0,95 & 0,14 & -0,09 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0,17 & 0,52 & 0,31 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0,29 & 0,45 & 0,26 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0,72 & 0,41 & -0,13 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0,42 & 0,38 & 0,20 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0,36 & 0,36 & 0,28 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0,11 & 0,20 & 0,91 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0,04 & 0,05 & 0,99 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0,01 & 0,02 & 0,99 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0,06 & 0,09 & 0,97 \end{bmatrix}.$$

Nota-se que a matriz **S** calculada com $\lambda = 0,4$ apresenta maior número de elementos nulos e tem maior traço que a matriz **S** calculada com $\lambda = 0,5$. Quanto menor o λ, maior será o número de elementos iguais a zero na diagonal de $\mathbf{U}_{x_i}$ e maior o número de elementos nulos nas linhas da matriz **S**.

2.3. Suavizador *cubic spline*

Considere um diagrama de dispersão cujos pontos (x_i,y_i), i=1,...,n, correspondem aos valores observados de duas variáveis X e Y. Deseja-se determinar uma curva suave, $\hat{f}$, que resuma a dependência de Y em X. Se o critério para determinar $\hat{f}$ fosse simplesmente encontrar a função f tal que a soma de quadrados dos resíduos dada por

$$\sum_{i=1}^{n}(y_i - f(x_i))^2$$

seja mínima, o resultado seria qualquer curva interpolando os pontos do diagrama de dispersão. Assim a solução não é única e pode não ser uma curva suave. O suavizador *cubic spline* adiciona a este critério, a restrição da curva estimada ter primeira e segunda derivadas contínuas. Esse método procura, entre todas as funções f(x) com primeira e segunda derivadas contínuas no intervalo $[a,b] = [x_{(1)},x_{(n)}]$, aquela que minimiza

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_a^b (f''(u))^2 du, \text{ com } \lambda \geq 0, \tag{2.10}$$

onde $x_{(1)}$ e $x_{(n)}$ são, respectivamente, os valores mínimo e máximo de X na amostra. O termo $\int_a^b (f''(u))^2 du$ é uma maneira de medir a *ondulação* da função f: no caso de funções f lineares $\int_a^b (f''(u))^2 du = 0$, enquanto que funções não lineares estão associadas a valores maiores do que zero para esta expressão. O parâmetro $\lambda \geq 0$ regula a relação de ganho e perda entre o viés e a *ondulação* da curva estimada. Assim como no *loess*, λ é o parâmetro responsável por regular a suavização da curva: valores grandes de λ dão peso maior à segunda parcela da expressão (2.10), produzindo curvas mais suaves, isto é, menos *onduladas*. Num extremo, se $\lambda \to \infty$

tem-se f "(u) = 0, e então a solução é a reta de mínimos quadrados. Em outro extremo, se $\lambda \to 0$, a segunda parcela de (2.10) tem pouca influência, e a solução tende a ser uma função duas vezes diferenciável que interpola os n pontos.

Pode-se mostrar que (2.10) tem solução única e explícita no intervalo [a,b]; esta solução é denominada *cubic spline* (ver, por exemplo, Schoenberg, 1964).

Minimizar (2.10) é equivalente a deteminar $\mathbf{f} = (f(x_i),\ldots,f(x_n))'$ que minimiza o critério:

(2.11) $$MQP(\mathbf{f}) = (\mathbf{y} - \mathbf{f})'(\mathbf{y} - \mathbf{f}) + \lambda \mathbf{f}'\mathbf{K}\mathbf{f},$$

(ver, por exemplo, Fahrmeir e Tutz, 1994), onde $\mathbf{y} = (y_1,\ldots,y_n)'$ é o vetor de valores observados da variável resposta. A matriz de pesos **K** tem uma estrutura especial dada por

(2.12) $$\mathbf{K} = \mathbf{D}'\mathbf{C}^{-1}\mathbf{D},$$

onde $\mathbf{D} = \{d_{ij}\}$ é uma matriz superior tridiagonal (n–2) x n,

$$\mathbf{D} = \begin{bmatrix} \frac{1}{h_1} & -\left(\frac{1}{h_1}+\frac{1}{h_2}\right) & \frac{1}{h_2} & 0 & \cdots & 0 \\ 0 & \frac{1}{h_2} & -\left(\frac{1}{h_2}+\frac{1}{h_3}\right) & \frac{1}{h_3} & & \vdots \\ \vdots & & & \ddots & & 0 \\ 0 & \cdots & 0 & \frac{1}{h_{n-2}} & -\left(\frac{1}{h_{n-2}}+\frac{1}{h_{n-1}}\right) & \frac{1}{h_{n-1}} \end{bmatrix},$$

com $h_i = x_{i+1} - x_i$, e $\mathbf{C} = \{c_{ij}\}$ é uma matriz simétrica tridiagonal (n–2) x (n–2),

$$\mathbf{C} = \begin{bmatrix} 2(h_1+h_2) & h_2 & 0 & \cdots & 0 \\ h_2 & 2(h_2+h_3) & h_3 & & \vdots \\ 0 & & \ddots & & 0 \\ \vdots & & & & h_{n-2} \\ 0 & \cdots & 0 & h_{n-2} & 2(h_{n-2}+h_{n-1}) \end{bmatrix}.$$

Ou seja, $\hat{\mathbf{f}}$ é a estimativa de **f** pelo critério de mínimos quadrados penalizados.

Derivando a expressão (2.11) em relação à **f** e igualando o resultado a zero, obtém-se o estimador de **f**, que é dado por

(2.13) $$\hat{\mathbf{f}} = (\mathbf{I} + \lambda\mathbf{K})^{-1}\mathbf{y} = \mathbf{S}\mathbf{y},$$

onde $\mathbf{S} = (\mathbf{I} + \lambda\mathbf{K})^{-1}$ é a matriz suavizadora n x n e **I** denota a matriz identidade n x n.

Matriz suavizadora

A matriz suavizadora definida em (2.13) possui as propriedades (2.9) descritas para a matriz **S** do *loess*. No entanto, neste caso, **S** é simétrica e por este motivo seus autovalores são reais. Tem-se ainda que dois de seus autovalores são iguais a 1 e os demais assumem valor no intervalo (0,1). Os elementos dessa matriz dependem apenas de $x_1,\ldots,x_n$ e do parâmetro de suavização λ, e, portanto, o *cubic spline* também é um suavizador linear. Quanto menor o λ, menores serão os valores dos elementos da diagonal da matriz **S** e menores os valores dos elementos fora desta diagonal. Os valores absolutos destes elementos diminuem à medida que eles se afastam da diagonal.

Exemplo: ajuste da curva e cálculo da matriz S

Considerando os dados da Tabela 2.1 e fixando $\lambda = 0{,}5$, obteve-se a matriz suavizadora para o *cubic spline*,

$$\mathbf{S} = (\mathbf{I} + 0{,}5\mathbf{K})^{-1} =$$

$$\begin{bmatrix}
0{,}999 & 0{,}002 & -0{,}002 & 0{,}001 & -0{,}000 & -0{,}000 & 0{,}000 & -0{,}000 & 0{,}000 & -0{,}000 \\
0{,}002 & 0{,}991 & 0{,}012 & -0{,}008 & 0{,}004 & 0{,}000 & -0{,}000 & 0{,}000 & -0{,}000 & 0{,}000 \\
-0{,}002 & 0{,}012 & 0{,}979 & 0{,}028 & -0{,}017 & -0{,}001 & 0{,}000 & -0{,}000 & 0{,}000 & -0{,}000 \\
0{,}001 & -0{,}008 & 0{,}028 & 0{,}859 & 0{,}152 & -0{,}034 & 0{,}002 & -0{,}000 & 0{,}000 & -0{,}000 \\
-0{,}000 & 0{,}004 & -0{,}017 & 0{,}152 & 0{,}798 & 0{,}069 & -0{,}007 & 0{,}001 & -0{,}000 & 0{,}000 \\
-0{,}000 & 0{,}000 & -0{,}001 & -0{,}034 & 0{,}069 & 0{,}955 & 0{,}012 & -0{,}002 & 0{,}000 & -0{,}000 \\
0{,}000 & -0{,}000 & 0{,}000 & 0{,}002 & -0{,}007 & 0{,}012 & 0{,}991 & 0{,}003 & -0{,}001 & 0{,}000 \\
-0{,}000 & 0{,}000 & -0{,}000 & -0{,}000 & 0{,}001 & -0{,}002 & 0{,}003 & 0{,}999 & 0{,}001 & -0{,}000 \\
0{,}000 & -0{,}000 & 0{,}000 & 0{,}000 & -0{,}000 & 0{,}000 & -0.001 & 0{,}001 & 1{,}000 & 0{,}000 \\
-0{,}000 & 0{,}000 & -0{,}000 & -0{,}000 & 0{,}000 & -0{,}000 & 0.000 & -0{,}000 & 0{,}000 & 1{,}000
\end{bmatrix}.$$

Calculou-se então, a estimativa de **f**:

$\hat{\mathbf{f}} = \mathbf{Sy} = [10{,}18 \;\; 11{,}53 \;\; 11{,}29 \;\; 12{,}42 \;\; 11{,}45 \;\; 10{,}53 \;\; 12{,}17 \;\; 12{,}82 \;\; 15{,}82 \;\; 16{,}30]'$.

A Figura 2.6 apresenta os dados da Tabela 2.1 e os valores estimados pelo método *cubic spline* com $\lambda = 0{,}5$.

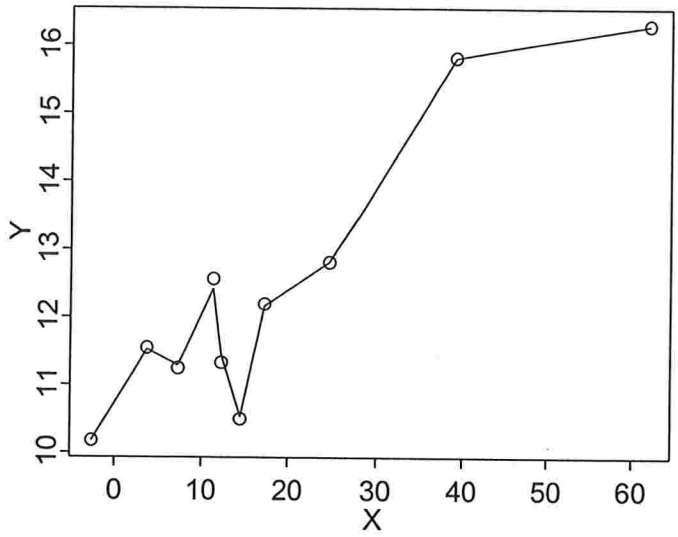


Figura 2.6: Valores ajustados pelo método *cubic spline*, com $\lambda = 0{,}5$, a partir dos dados da Tabela 2.1.

Aqui também pode-se ilustrar a influência de λ nos elementos da matriz **S**. O cálculo desta matriz para o mesmo conjunto de dados e com $\lambda = 0{,}9$ forneceu

$$\mathbf{S} = \begin{bmatrix} 0{,}999 & 0{,}004 & -0{,}003 & 0{,}001 & -0{,}000 & -0{,}000 & 0{,}000 & -0{,}000 & 0{,}000 & -0{,}000 \\ 0{,}004 & 0{,}983 & 0{,}020 & -0{,}012 & 0{,}004 & 0{,}001 & -0{,}000 & 0{,}000 & -0{,}000 & 0{,}000 \\ -0{,}003 & 0{,}020 & 0{,}965 & 0{,}043 & -0{,}021 & -0{,}004 & 0{,}001 & -0{,}000 & 0{,}000 & -0{,}000 \\ 0{,}001 & -0{,}012 & 0{,}043 & 0{,}797 & 0{,}211 & -0{,}043 & 0{,}002 & -0{,}000 & 0{,}000 & -0{,}000 \\ -0{,}000 & 0{,}004 & -0{,}021 & 0{,}212 & 0{,}718 & 0{,}109 & -0{,}022 & 0{,}001 & -0{,}000 & 0{,}000 \\ -0{,}000 & 0{,}001 & -0{,}004 & -0{,}043 & 0{,}109 & 0{,}897 & 0{,}043 & -0{,}004 & 0{,}000 & -0{,}000 \\ 0{,}000 & -0{,}000 & 0{,}001 & 0{,}002 & -0{,}022 & 0{,}043 & 0{,}971 & 0{,}006 & -0{,}001 & 0{,}000 \\ -0{,}000 & 0{,}000 & -0{,}000 & -0{,}000 & 0{,}001 & -0{,}004 & 0{,}006 & 0{,}998 & 0{,}001 & -0{,}000 \\ 0{,}000 & -0{,}000 & 0{,}000 & 0{,}000 & -0{,}000 & 0{,}000 & -0.001 & 0{,}001 & 1{,}000 & 0{,}000 \\ -0{,}000 & 0{,}000 & -0{,}000 & -0{,}000 & 0{,}000 & -0{,}000 & 0.000 & -0{,}000 & 0{,}000 & 1{,}000 \end{bmatrix}.$$

Comparando este resultado com o obtido para $\lambda = 0{,}5$, verifica-se que um valor de λ maior fornece uma matriz **S** com valores menores na diagonal e, conseqüentemente, maiores fora desta diagonal.

2.4. Alguns resultados para suavizadores lineares

Erro quadrático médio, variância e viés do estimador

Quando um suavizador é linear a *matriz de variâncias e covariâncias* de $\hat{\mathbf{f}}$ é dada por

$$\text{Var}(\hat{\mathbf{f}}) = \text{Var}(\mathbf{SY}) = \mathbf{S}\text{Var}(\mathbf{Y})\mathbf{S}'.$$

Sob a suposição de que os $Y_1,\ldots,Y_n$ são independentes com $\text{Var}(Y_i) = \sigma^2$, tem-se que

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

O vetor *viés* do estimador é definido como

$$\mathbf{b} = E(\mathbf{f} - \hat{\mathbf{f}}) = \mathbf{f} - E(\hat{\mathbf{f}}) = \mathbf{f} - E(\mathbf{SY}) = \mathbf{f} - \mathbf{Sf} = (\mathbf{I}-\mathbf{S})\mathbf{f}.$$

O *erro quadrático médio* global, dado por

$$\frac{1}{n}\sum_{i=1}^{n} E\left(f(x_i) - \hat{f}(x_i)\right)^2,$$

pode então, ser escrito como

(2.14)
$$\begin{aligned} EQM &= \frac{1}{n}\sum_{i=1}^{n}\text{Var}\left(\hat{f}(x_i)\right) + \frac{1}{n}\sum_{i=1}^{n} b_i^2 \\ &= \frac{\text{traço}(\mathbf{SS}')}{n}\sigma^2 + \frac{\mathbf{b}'\mathbf{b}}{n}. \end{aligned}$$

É interessante observar a influência de λ nos componentes do EQM. De forma geral, aumentando λ, o traço($\mathbf{SS}'$) tende a diminuir e os elementos de $\mathbf{b}$ tendem a aumentar, e vice-versa. De fato, isto pode ser observado no ajuste dos dados da Tabela 2.1 pelo *loess*: o ajuste com $\lambda = 0,5$ forneceu traço($\mathbf{SS}'$) = 6,21, enquanto o ajuste com $\lambda = 0,4$ forneceu traço($\mathbf{SS}'$) = 6,87.

O parâmetro σ^2 em (2.14) geralmente é desconhecido. Um estimador para este parâmetro, assumindo que $\hat{\mathbf{f}}$ é não viesado, é dado por:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2}{n - \text{traço}(2\mathbf{S} - \mathbf{SS}')} = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n - \text{traço}(2\mathbf{S} - \mathbf{SS}')}, \tag{2.15}$$

onde $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{f}} = (\mathbf{I} - \mathbf{S})\,\mathbf{y}$. Tem-se que $E[\hat{\mathbf{e}}] = 0$ e $\text{Var}[\hat{\mathbf{e}}] = \sigma^2(\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})'$.

Quando os erros são normais é fácil mostrar que $\hat{\sigma}^2$ é um estimador não viesado para σ^2. Assim, $\hat{\mathbf{e}}'\hat{\mathbf{e}}$ é uma forma quadrática em variáveis normais, e

$$E[\hat{\mathbf{e}}'\hat{\mathbf{e}}] = \text{traço}\left(\sigma^2(\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})'\right) = \sigma^2\left(n - \text{traço}(2\mathbf{S} - \mathbf{SS}')\right), \tag{2.16}$$

e portanto,

$$\sigma^2 = \frac{E[\hat{\mathbf{e}}'\hat{\mathbf{e}}]}{n - \text{traço}(2\mathbf{S} - \mathbf{SS}')}.$$

Bandas de confiança pontuais

Ainda sob o modelo (2.1), a matriz de variâncias e covariâncias do vetor $\hat{\mathbf{f}} = \mathbf{Sy}$ é simplesmente dada por:

$$\text{Var}(\hat{\mathbf{f}}) = \sigma^2\,\mathbf{SS}'.$$

Um estimador para σ^2 pode ser, por exemplo, $\hat{\sigma}^2$ dado em (2.15). Um estimador para o erro padrão de $\hat{\mathbf{f}}$ é, então, dado pela raiz quadrada dos elementos da diagonal da matriz $\hat{\sigma}^2\mathbf{SS}'$. Este resultado pode ser usado para construir bandas de confiança pontuais para $f(x_i)$, dadas por:

$$\hat{f}(x_i) \pm 2\,\hat{ep}(\hat{f}(x_i)), \tag{2.17}$$

onde $\hat{ep}(\hat{f}(x_i))$ é o valor estimado do erro padrão de $\hat{f}(x_i)$. Sob as suposições de erros normais e viés desprezível, estas bandas representam intervalos de confiança pontuais.

Se o viés não é desprezível (o que é difícil de ser verificado), as bandas fornecem intervalos de confiança pontuais com coeficiente de confiança γ para $\mathbf{g} = \mathbf{Sf}$, e não para $\mathbf{f}$.

As Figuras 2.7 e 2.8 apresentam as curvas suavizadas referentes aos dados da Tabela 2.1 obtidas, respectivamente, pelos métodos *loess* e *cubic spline*, ambos com $\lambda = 0{,}5$. As curvas pontilhadas representam as bandas de confiança pontuais para os valores de $\mathbf{f}$, com coeficiente de confiança aproximadamente igual a 0,95. Nota-se que, para um mesmo valor de λ, a curva suavizada pelo método *loess* apresenta-se mais suave do que a obtida pelo método *cubic spline*.

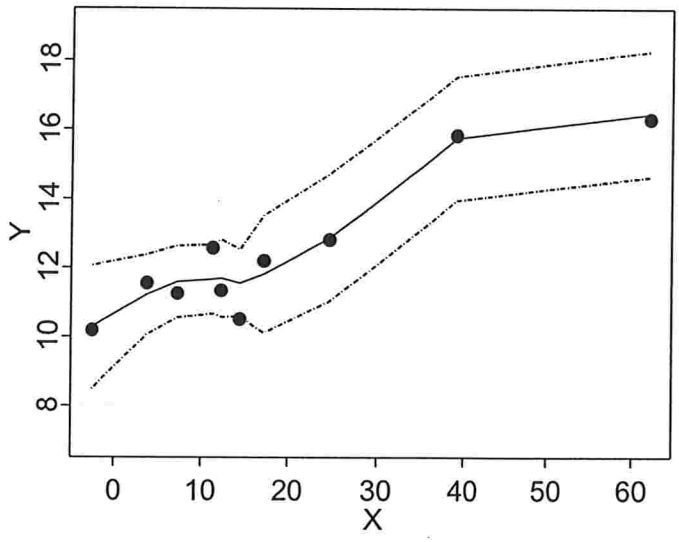


Figura 2.7: Curva suavizada obtida pelo método *loess* (com $\lambda = 0{,}5$) para os dados da Tabela 2.1 e bandas de confiança pontuais.

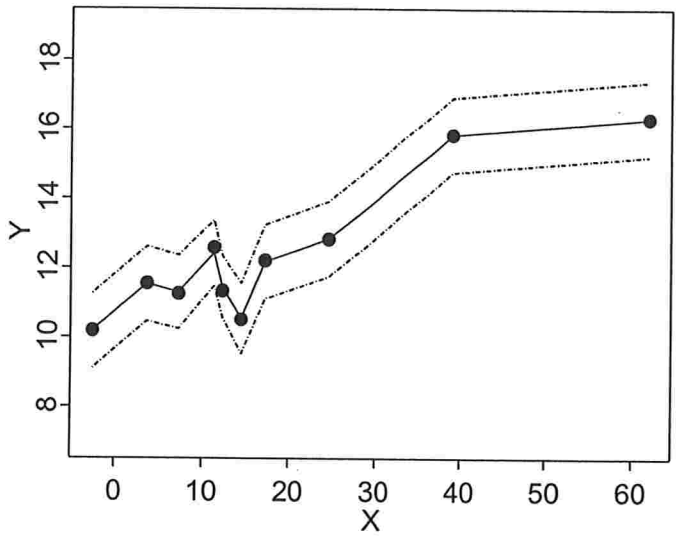


Figura 2.8: Curva suavizada obtida pelo método *cubic spline* (com $\lambda = 0{,}5$) para os dados da Tabela 2.1 e bandas de confiança pontuais.

Graus de liberdade de um suavizador

Uma maneira de tornar diferentes procedimentos de suavização comparáveis em relação à *quantidade de suavização* que apresentam é especificar os *graus de liberdade* de cada suavizador.

Dada uma matriz suavizadora **S** de um alisador linear com um λ fixado, o *número de graus de liberdade* do alisador, ou *número de parâmetros*, pode ser definido como

(2.18) $$gl = \text{traço}(\mathbf{SS}').$$

Quanto maior o número de graus de liberdade, menor a quantidade de suavização e, consequentemente, menor o valor de λ. Existem ainda, duas outras maneiras de definir os graus de liberdade de suavizadores lineares, são elas:

(2.19) $$gl = \text{traço}(\mathbf{S}),$$

e

(2.20) $$gl = \text{traço}(2\mathbf{S} - \mathbf{SS}').$$

As expressões para os graus de liberdade dadas em (2.18), (2.19) e (2.20) foram propostas por analogia à regressão de mínimos quadrados. Como já foi dito anteriormente, a matriz **S** desempenha papel semelhante à matriz *hat*, **H**. Na regressão de mínimos quadrados,

traço(**HH**') = traço(**H**) = posto(**H**) = número de parâmetros do modelo,

e daí decorrem as expressões (2.18) e (2.19). Em um caso extremo no qual a suavização é mínima, isto é, uma curva passando por todos os pontos, tem-se traço(**SS**') = n, que é igual ao número de parâmetros em um modelo de regressão saturado, para o qual traço(**H**) = n.

A motivação para a expressão (2.20) é dada pela expressão (2.16), considerando que, na regressão de mínimos quadrados, o valor esperado da soma de quadrados do resíduo é σ^2(n – número de parâmetros).

Quando a matriz suavizadora **S** é idempotente, as quantidades traço(**SS**'), traço(**S**) e traço(2**S** – **SS**') coincidem. Este não é o caso do *loess* nem do *cubic spline*. Para o *cubic spline*, como **S** é simétrica, tem-se

$$\text{traço}(\mathbf{S}) = \sum_{i=1}^{n} \theta_i ,$$

$$\text{traço}(\mathbf{SS}') = \sum_{i=1}^{n} \theta_i^2 \quad \text{e}$$

$$\text{traço}(2\mathbf{S} - \mathbf{SS}') = \sum_{i=1}^{n} \left(2\theta_i - \theta_i^2\right),$$

onde θ_i, i=1,...,n, são os autovalores da matriz **S**. Devido ao fato da matriz **S** possuir dois autovalores unitários e os demais no intervalo (0,1), tem-se ainda que

$$\text{traço}(\mathbf{SS}') \leq \text{traço}(\mathbf{S}) \leq \text{traço}(2\mathbf{S} - \mathbf{SS}').$$

Para os dados da Tabela 2.1, fixando $\lambda = 0{,}5$, o traço(**S**), o traço(**SS'**) e o traço(2**S** – **SS'**) são, respectivamente, 6,70, 6,21 e 7,19 para o *loess* e 9,57, 9,22 e 9,88 para o *cubic spline*, indicando que na obtenção dos valores ajustados pelo *cubic spline* foi utilizado um maior *número de parâmetros* do que no *loess*, obtendo-se uma curva menos suave.

Essas definições podem ser estendidas para os suavizadores não lineares, embora neste caso, passem a depender da distribuição de Y.

Suavizadores ponderados

A suposição de igualdade de variâncias dos erros assumida no modelo (2.1) pode ser verificada da mesma forma que em um modelo de regressão paramétrico. Embora os resíduos, $\hat{e}_i = y_i - \hat{f}(x_i)$, não tenham a propriedade de soma nula sob o enfoque da regressão de mínimos quadrados, eles podem ser examinados de forma usual para verificação da hipótese de homocedasticidade (ver, por exemplo, Neter et al., 1996).

Se a análise dos resíduos indicar que um modelo assumindo variâncias diferentes é mais adequado, isto é,

$$y_i = f(x_i) + \varepsilon_i, \quad i=1,\ldots,n,$$

onde os ε_i's são erros aleatórios distribuídos independentemente com média zero e variância σ_i^2, então, pode-se adotar um método de suavização ponderado no qual a i-ésima observação fica associada a um peso $w_i = 1/\sigma_i^2$, assumindo σ_i^2 conhecido.

Os métodos de suavização apresentados nas Seções 2.2 e 2.3, são facilmente adaptados ao caso ponderado, como descrito a seguir.

Quando o *loess* é utilizado, basta multiplicar $\mathbf{U}_{x_i}$, dada em (2.4), por $\mathbf{W} = \text{diagonal}\{w_1,\ldots,w_n\}$ e obter uma nova matriz de pesos $\mathbf{A}_{x_i} = \mathbf{U}_{x_i}\mathbf{W}$. Então, o valor previsto correspondente a x_i será dado por

$$\hat{f}(x_i) = s_{x_i,1}\, y_1 + s_{x_i,2}\, y_2 + \ldots + s_{x_i,n}\, y_n = \mathbf{s}'_{x_i}\mathbf{y},$$

onde $\mathbf{s}'_{x_i}$ é a i-ésima linha da matriz

$$\mathbf{S}_{x_i} = \mathbf{X}(\mathbf{X}'\mathbf{A}_{x_i}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}_{x_i}.$$

No caso do *cubic spline*, a função $\hat{\mathbf{f}}$ será obtida pelo o critério de mínimos quadrados penalizados ponderados, isto é, determinado-se $\mathbf{f}$ que minimiza a expressão

(2.21) $$\text{MQPP}(\mathbf{f}) = (\mathbf{y} - \mathbf{f})'\mathbf{W}(\mathbf{y} - \mathbf{f}) + \lambda\mathbf{f}'\mathbf{K}\mathbf{f},$$

onde $\mathbf{W} = \text{diagonal}\{w_1,\ldots,w_n\}$ é a matriz de pesos.

A solução de (2.21) é, novamente, um suavizador *cubic spline*, com vetor $\hat{\mathbf{f}}$ de valores ajustados dado por

$$\hat{\mathbf{f}} = (\mathbf{W} + \lambda\mathbf{K})^{-1}\mathbf{W}\mathbf{y} = \mathbf{S}\mathbf{y},$$

onde λ é o parâmetro de suavização, $\mathbf{K}$ é a matriz de penalização definida em (2.12) e $(\mathbf{W} + \lambda\mathbf{K})^{-1}\mathbf{W} = \mathbf{S}$ é a matriz suavizadora.

Como raramente as variâncias são conhecidas, torna-se necessário estimar σ_i^2. Existem várias formas de se obter estimadores para estes parâmetros na regressão de mínimos quadrados (ver, por exemplo, Neter et al., 1996, Capítulo 10), que podem ser aqui utilizadas. Uma possível forma seria construir o gráfico de dispersão dos resíduos versus x_i e estimar a relação existente entre a variância e a variável preditora. Um procedimento iterativo pode então, ser adotado para obtenção dos pesos finais.

As versões ponderadas do *loess* e do *cubic spline*, descritas anteriormente, consistem em ferramentas básicas para o uso destes suavizadores nos MAG's, como será mostrado no Capítulo 3.

Comparação entre suavizadores quanto às vizinhanças

Os suavizadores lineares podem ser comparados quanto à forma das suas vizinhanças em torno de um valor alvo x_i, e quanto aos pesos atribuídos às observações nessa vizinhança, dispondo em um gráfico os valores da i-ésima linha da matriz **S** em função dos valores observados de X. Para que esta comparação não dependa da quantidade de suavização fornecida pelos suavizadores a serem comparados, as matrizes de suavização devem ser construídas a partir de um mesmo número de graus de liberdade.

Considerando os dados da Tabela 2.1, na Figura 2.9 são representadas as linhas 1, 3, 5 e 10 da matriz **S** obtida no processo de suavização *loess*, com $\lambda = 0,5$, e *cubic spline*, com $\lambda = 40$. Em ambos os casos, os valores de λ forneceram traço(**SS'**) aproximadamente igual a 6,22. Esta figura deixa evidente o caráter local destes suavizadores e mostra que, quando os ajustes são feitos com equivalentes graus de liberdade, os pesos atribuídos a cada observação apresentam valores semelhantes.

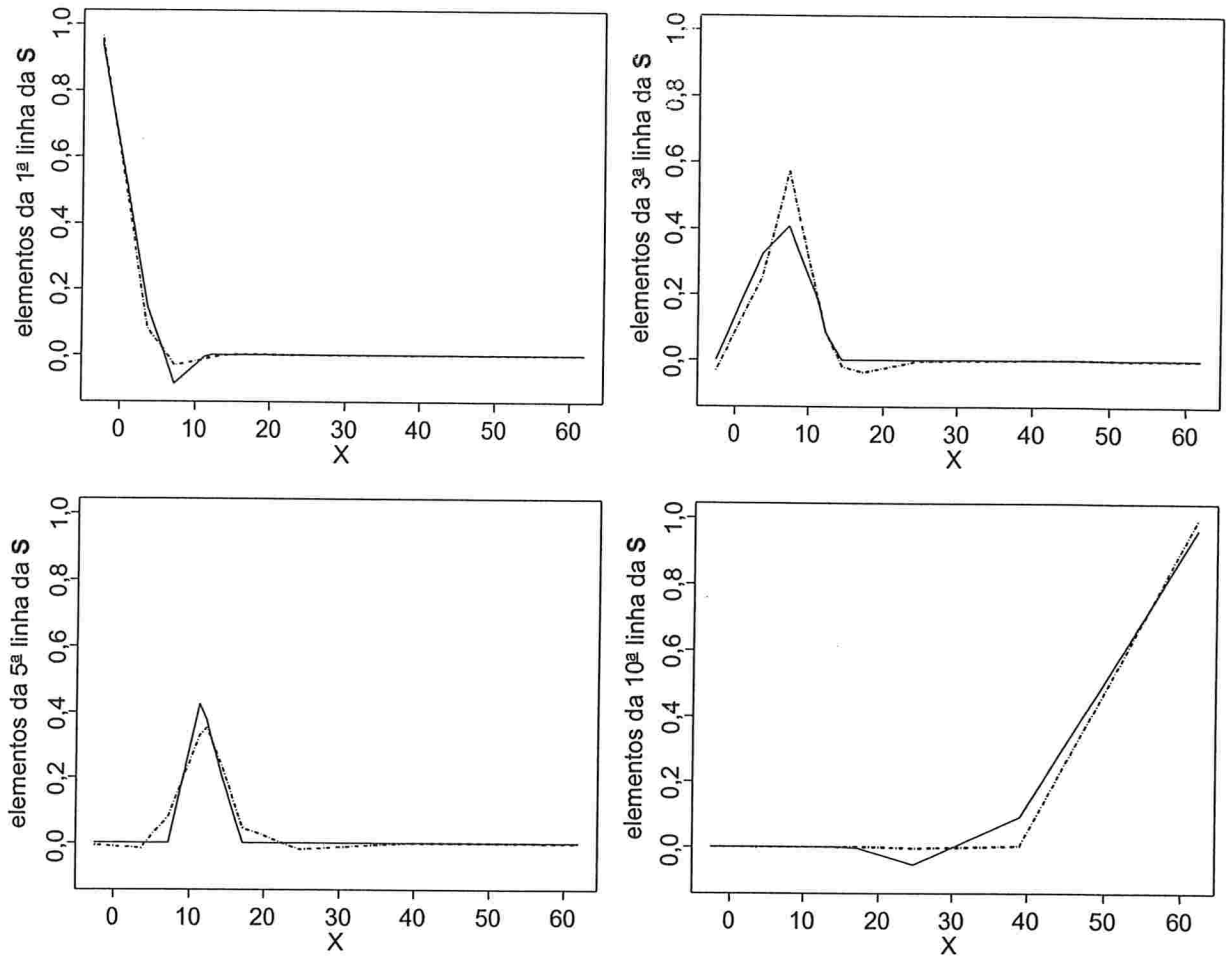


Figura 2.9: Representação dos elementos das linhas 1, 3, 5 e 10 da matriz **S** obtida pelo método *loess* (linha contínua) e *cubic spline* (linha pontilhada) no processo de suavização dos dados da Tabela 2.1, com aproximadamente 6 graus de liberdade.

2.5. Seleção do parâmetro de suavização

A utilização de alisadores requer a fixação do parâmetro de suavização. A escolha desse parâmetro é de extrema importância para o sucesso da técnica, sendo até mais importante do que a do método de suavização.

Uma forma automática de seleção do parâmetro de suavização é a validação cruzada (*cross-validation*), que é calculada para um conjunto pré-fixado de valores de

λ. Para cada λ considerado, o método consiste em retirar o ponto (x_i,y_i) e calcular $\hat{f}(x_i)$ com base nos n–1 pontos restantes, i=1,...,n. A estatística da validação cruzada é dada por

$$VC(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_\lambda^{-i}(x_i)\right)^2, \tag{2.22}$$

onde $\hat{f}_\lambda^{-i}(x_i)$ indica o valor de Y previsto quando o ponto (x_i,y_i) é eliminado, i=1,...,n. A expressão (2.22) é calculada considerando-se um conjunto de valores de λ fixados a priori, sendo selecionado o valor de λ que minimiza esta expressão.

Entretanto, Hastie e Tibshirani (1990) mostraram, em um estudo de simulação, 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 o uso de métodos gráficos, como foi feito para os dados apresentados na Figura 2.2, auxiliados por medidas dos graus de liberdade dos suavizadores.

2.6. Observações empatadas em X

No início deste capítulo foi assumido que os dados não apresentavam réplicas nos valores de X. No entanto, é comum os dados apresentarem mais de uma observação para cada x_i e neste caso, uma modificação deve ser feita para utilização dos suavizadores.

Uma solução simples que pode ser usada para todos os alisadores consiste em dividir o conjunto de dados em $m \leq n$ grupos determinados pelos m valores distintos de X. Assim, um novo conjunto de dados com m observações pode ser definido com cada observação representada por um valor distinto de X, x_i, uma média ponderada

dos valores de Y pertencentes ao i-ésimo grupo, denotada por y_i, e um peso w_i que é a soma dos pesos das observações nesse grupo, i=1,...,m. Nesse novo conjunto de dados é, então, aplicado um procedimento de suavização ponderada.

2.7. Propriedades assintóticas

Na seção 2.4 foram apresentados alguns resultados sob a suposição de ausência de viés de $\hat{\mathbf{f}}$. Porém, tanto o *loess* quanto o *cubic spline* fornecem funções $\hat{\mathbf{f}}$ viesadas para uma **f** arbitrária. O viés é nulo para uma classe restrita de funções. Por exemplo, o *cubic spline* é não viesado se **f** for linear. Uma solução para esse problema, que será adotada no Capítulo 3 na obtenção da matriz de variâncias e covariâncias de $\hat{\mathbf{f}}$, e na construção das bandas de confiança para **f**, é considerar o comportamento assintótico. Não é objetivo deste trabalho o estudo de propriedades assintóticas dos suavizadores. Sobre este tópico podem ser consultados, por exemplo, Stone (1977), Cox (1983) ou Rice e Rosenblatt (1983). Apenas o conceito de consistência será abordado informalmente.

Um suavizador é dito *consistente* se $\hat{f}(x)$ converge para f(x) em probabilidade a uma *taxa apropriada*. Neste caso, o viés assintótico do estimador é nulo. O termo "taxa apropriada" está associado ao fato de que, se o parâmetro de suavização é mantido fixo, quando n cresce, em geral, o viés assintótico é não nulo (Buja et al., 1989). Por outro lado, se a quantidade de suavização decresce segundo essa taxa à medida que n aumenta, então, sob determinadas condições de regularidade, os estimadores serão consistentes, ou seja,

$$\hat{f}_{\lambda_n}(x_i) \rightarrow f(x_i) \quad \text{em probabilidade, } n \rightarrow \infty,$$

onde λ_n é o parâmetro de suavização em uma amostra de tamanho n.

Capítulo 3

Modelos Aditivos Generalizados

3.1. Introdução

Os alisadores 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 mostra-se como os alisadores podem ser utilizados na classe dos MLG's.

Seja Y uma variável aleatória com função densidade de probabilidade ou função de probabilidade dada por

(3.1) $$f(y;\theta,\phi) = \exp\big(\phi(y\theta - b(\theta)) + c(y,\phi)\big)$$

onde b(•) e c(•) são funções especificadas, b(•) é duas vezes diferenciável, e $\phi^{-1} > 0$. Se ϕ é conhecido, tem-se um modelo da família exponencial unidimensional com parâmetro canônico (natural) θ. McCullagh e Nelder (1989) mostraram que, para Y com função densidade de probabilidade ou função de probabilidade dada por (3.1),

$$E(Y) = \mu = \frac{\partial b(\theta)}{\partial \theta}$$

e

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

onde $V = \partial\mu/\partial\theta$ é chamada função de variância, e depende somente do parâmetro canônico (e, portanto, da média de Y), e ϕ^{-1} é chamado parâmetro de dispersão, e não depende de θ. Será assumido neste trabalho que ϕ é conhecido.

Suponha agora que $Y_1,\ldots,Y_n$ sejam n variáveis aleatórias independentes, cada uma com função densidade de probabilidade ou função de probabilidade na forma (3.1) com $\theta = \theta_i$, i=1,...,n. Suponha ainda que a média de Y_i, μ_i, está relacionada com um conjunto de variáveis preditoras (não aleatórias) $X_1,\ldots,X_p$ por meio de uma função monótona e diferenciável

(3.2) $$g(\mu_i) = \eta_i$$

denominada função de ligação, onde

(3.3) $$\eta_i = \alpha + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

é chamado *preditor linear*, $\alpha, \beta_1,\ldots,\beta_p$ são os parâmetros a serem estimados e $x_{i1},\ldots,x_{ip}$ são os valores observados de $X_1,\ldots,X_p$ no i-ésimo elemento da amostra. As variáveis preditoras podem representar uma função de X_j, como X_j^2, ou ainda, o produto de duas covariáveis, como por exemplo $X_{j-1}X_j$.

As expressões (3.1), (3.2) e (3.3) definem os MLG's, com (3.1) representando o componente aleatório do modelo e (3.2) e (3.3) representando o componente sistemático.

Um caso particular importante ocorre quando o preditor linear coincide com o parâmetro canônico, isto é, quando $\eta_i = \theta_i$. 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, por exemplo, Paula, 2000).

A expressão (3.3) assume que o preditor linear η_i é uma função linear de cada uma das variáveis preditoras $X_1,\ldots,X_p$. No entanto, uma relação menos rígida pode ser adotada substituindo-se, em (3.3), o termo linear correspondente a cada covariável por uma função não especificada dessa variável, obtendo-se o *preditor aditivo*

(3.4) $$\eta_i = \alpha + f_1(x_{i1}) + \ldots + f_p(x_{ip}).$$

A classe dos modelos assim obtida é denominada *modelos aditivos generalizados* e pode ser vista como uma generalização da classe dos MLG's pelo fato da relação entre η_i e X_i não estar especificada.

O preditor (3.4) corresponde a um modelo totalmente não paramétrico. Porém, também fazem parte dos MAG's, os modelos cujo preditor combina formas paramétricas de algumas (r) variáveis preditoras com termos não paramétricos de outras (p – r) variáveis. Neste caso, o preditor pode ser escrito como:

(3.5) $$\eta_i = \alpha + \beta_1 x_{i1} + \ldots + \beta_r x_{ir} + f_1(x_{i,r+1}) + \ldots + f_{p-r}(x_{ip}).$$

Modelos nesta forma são denominados *modelos semi-paramétricos*.

Outras formas para o preditor podem ainda ser consideradas, são elas:

• $\eta_i = f(x_{i1}) + g_k(x_{i2})$, onde k indexa os níveis de um fator, e assim cria um termo de interação entre os efeitos do fator e a variável X_2, e

• $\eta_i = f(x_{i1}) + g(x_{i2},x_{i3})$, onde $g(\bullet)$ é uma função não especificada de duas variáveis. O ajuste de modelos envolvendo termos não paramétricos em mais de uma variável não será abordado neste trabalho, conforme mencionado no Capítulo 1.

A estimação dos MAG's e testes de hipótese 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 $\mathbf{f}_1,\ldots,\mathbf{f}_{p-r}$ em (3.5) sejam estimadas por meio da utilização de suavizadores.

3.2. Ajuste do modelo

Considere um MAG cujo preditor aditivo é dado por (3.4), isto é, um modelo não paramétrico nas p variáveis preditoras.

Seja $\mathbf{f}_j = (f_j(x_{1j}),\ldots,f_j(x_{nj}))'$ o vetor dos valores da função f_j, j=1,...,p, nos n valores observados de X_j.

O processo de estimação de um MAG equivale à estimação de α e $\mathbf{f}_j$, j=1,...,p. Para motivar a estimação dos MAG's serão apresentados, inicialmente, alguns resultados sobre estimação dos MLG's pelo método de máxima verossimilhança.

Ajuste dos modelos lineares generalizados

Considere um MLG onde $\beta = (\alpha, \beta_1,\ldots, \beta_p)'$ é o vetor de parâmetros a ser estimado. A estimação de β é usualmente feita por máxima verossimilhança e a estimativa deste vetor é calculada resolvendo-se as seguintes equações escore:

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

onde $V_i = \mathrm{Var}(Y_i)$ e $x_{i0} = 0$.

As equações acima podem ser resolvidas pelo método *scoring* de Fisher (McCullagh e Nelder, 1989) que é equivalente ao procedimento de mínimos quadrados reponderados iterativamente (MQRI). Em cada iteração desse processo, β é estimado pelo método de mínimos quadrados ponderados em uma regressão de uma variável resposta modificada Z em $\mathbf{X}$ com pesos W, como descrito a seguir.

Seja $\beta^{(m)}$ a estimativa de β no passo m do processo iterativo, com m = 0 denotando o passo inicial. O procedimento MQRI pode então, ser esquematizado da seguinte forma:

Algoritmo scoring de Fisher

Passo 1: O processo é iniciado considerando valores de $\beta^{(0)}$ obtidos da regressão de mínimos quadrados de $g(y_i)$ em $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})$, i=1,...,n.

Passo 2: A variável resposta modificada $z_i^{(m)}$ e os pesos $w_i^{(m)}$ no passo m, i=1,...,n, são calculados da forma

$$z_i^{(m)} = \eta_i^{(m)} + (y_i - \mu_i^{(m)})\left(\frac{\partial \eta_i}{\partial \mu_i}\right)_{(m)}, \tag{3.6}$$

onde $\eta_i^{(m)} = \alpha^{(m-1)} + \sum_{j=1}^{p} \beta_j^{(m-1)} x_{ij}$ e $\mu_i^{(m)} = g^{-1}(\eta_i^{(m)})$, e

$$w_i^{(m)} = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2_{(m)} (V_i^{(m)})^{-1}, \tag{3.7}$$

onde $V_i^{(m)} = \left(\frac{\partial \mu_i}{\partial \theta_i}\right)_{(m)}$.

Passo 3: O vetor $\beta^{(m)}$ é obtido da regressão de $z_i^{(m)}$ em $\mathbf{x}_i$ com peso $w_i^{(m)}$, i=1,...,n, isto é

$$\beta^{(m)} = (\mathbf{X}'\mathbf{W}^{(m)}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(m)}\mathbf{z}^{(m)}, \; m = 1,2,\ldots,$$

onde $\mathbf{X}$ é a matriz de planejamento de dimensão n x (p+1), cuja i-ésima linha corresponde ao vetor $(1,\mathbf{x}_i)$, $\mathbf{z}^{(m)} = (z_1^{(m)},\ldots,z_n^{(m)})'$ e $\mathbf{W}^{(m)} = \text{diagonal}\{w_1^{(m)},\ldots,w_n^{(m)}\}$.

Passo 4: Os passos 2 e 3 são repetidos até que a mudança no valor do desvio

$$D(\mathbf{y};\hat{\mu}) = 2\{\log(\mu_{max};\mathbf{y}) - \log(\hat{\mu};\mathbf{y})\},$$

onde $\log(\mu_{max};\mathbf{y})$ e $\log(\hat{\mu};\mathbf{y})$ representam, respectivamente, o logaritmo da função de verossimilhança calculado para o modelo saturado (com n parâmetros) e o logaritmo da função de verossimilhança calculado para o modelo ajustado (com p+1 parâmetros), seja suficientemente pequena.

Ajuste dos modelos aditivos generalizados

As estimativas de α e $\mathbf{f}_j$, j=1,...,p, são obtidas combinando dois processos iterativos:

• o de ponderação local (*local scoring*) – similar ao procedimento de MQRI utilizado na estimação dos MLG's, considerando o preditor aditivo dado em (3.4) no lugar do preditor linear apresentado em (3.3),

• e o de retroajuste (*backfitting*) – algoritmo de ajuste iterativo aninhado ao de ponderação local, responsável por estimar cada $\mathbf{f}_j$ individualmente, deixando os demais fixos, e então, repetindo o processo para j=1,...,p.

• Ponderação local

O procedimento de ponderação local (PPL) é um processo iterativo que pode ser considerado uma versão não paramétrica do procedimento de MQRI utilizado na estimação dos MLG's.

Sejam $\alpha^{(m)}$ o valor estimado de α e $\mathbf{f}_j^{(m)}$ a estimativa de $\mathbf{f}_j$, j=1,...,p, no passo m do processo iterativo, com m = 0 denotando o passo inicial.

Dados valores iniciais para $\alpha^{(0)}$ e $\mathbf{f}_1^{(0)},\ldots,\mathbf{f}_p^{(0)}$, os valores da variável resposta modificada $z_i^{(m)}$ e dos pesos $w_i^{(m)}$ são calculados da mesma forma que em (3.6) e (3.7), porém, neste caso, $\eta_i^{(m)} = \alpha^{(m-1)} + \sum_{j=1}^{p} f_j^{(m-1)}(x_{ij})$. O passo 3 do processo MQRI, no qual é ajustada uma regressão linear ponderada, é substituído pelo ajuste ponderado de um modelo aditivo não paramétrico com variável resposta modificada Z e matriz de pesos **W**. O ajuste desse modelo é feito por um novo procedimento iterativo, denominado retroajuste, a partir do qual novas estimativas $\mathbf{f}_1^{(m)},\ldots,\mathbf{f}_p^{(m)}$ são

obtidas. Então, são calculados novos valores $\eta_1^{(m)},\ldots,\eta_n^{(m)}$ e $\mu_1^{(m)},\ldots,\mu_n^{(m)}$. O processo é reiniciado com o cálculo de novos valores de $z_i^{(m)}$ e $w_i^{(m)}$, e pára quando valores estimados consecutivos das funções $\mathbf{f}_j$, j=1,...,p, são próximos.

O termo "local" deve-se à utilização de suavizadores ponderados na estimação das funções $\mathbf{f}_j$ dentro do retroajuste. Esquematicamente, o algoritmo de ponderação local é descrito da seguinte forma:

Algoritmo de ponderação local

Passo 1: O processo é iniciado com

$$\alpha^{(0)} = g\left(\sum_{i=1}^{n} \frac{y_i}{n}\right) \text{ e } \mathbf{f}_1^{(0)} = \ldots = \mathbf{f}_p^{(0)} = \mathbf{0}.$$

Passo 2: A variável resposta modificada $z_i^{(m)}$ e os pesos $w_i^{(m)}$ no passo m, i=1,...,n, são calculados da forma

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

onde $\eta_i^{(m)} = \alpha^{(m-1)} + \sum_{j=1}^{p} f_j^{(m-1)}(x_{ij})$ e $\mu_i^{(m)} = g^{-1}(\eta_i^{(m)})$, e

$$(3.8) \qquad w_i^{(m)} = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2_{(m)} (V_i^{(m)})^{-1},$$

onde $V_i^{(m)} = \left(\frac{\partial \mu_i}{\partial \theta_i}\right)_{(m)}$.

Passo 3: Os valores $\mathbf{f}_1^{(m)},\ldots,\mathbf{f}_p^{(m)}$ são obtidos com o uso do algoritmo de retroajuste.

Passo 4: Os passos 2 e 3 são repetidos até que

$$\frac{\sum_{j=1}^{p} \left\| \mathbf{f}_j^{(m)} - \mathbf{f}_j^{(m-1)} \right\|}{\sum_{j=1}^{p} \left\| \mathbf{f}_j^{(m-1)} \right\|} \leq \varepsilon,$$

para um valor $\varepsilon > 0$ preestabelecido.

Nota-se que, diferentemente do que acontece no *scoring* de Fisher, o critério de parada do PPL não se baseia no desvio do modelo, mas em uma medida da proximidade das estimativas dos $\mathbf{f}_j$'s em 2 passos consecutivos.

• Retroajuste

Sejam $\alpha^{(m)} = \bar{z}^{(m)}$, a média amostral da variável modificada no passo m, e $\mathbf{f}_1^0,\ldots,\mathbf{f}_p^0$, valores iniciais de $\mathbf{f}_1,\ldots,\mathbf{f}_p$ no início do retroajuste. Cada iteração do retroajuste atualiza o valor de $\mathbf{f}_j^0$ para $\mathbf{f}_j^1$. Os valores de $\mathbf{f}_j^1$ são obtidos por meio de um procedimento de suavização ponderado (como os descritos no Capítulo 2) dos resíduos parciais

$$r_{ij}^{(m)} = z_i^{(m)} - \bar{z}^{(m)} - \sum_{\substack{k=1 \\ k\neq j}}^{p} f_k^0(x_{ik}) \qquad (3.9)$$

em função de x_{ij} com pesos $w_i^{(m)}$, dados em (3.8), i=1,...,n, j=1,...,p. Assim, quando $\mathbf{f}_j^1$ é estimado, o efeito das demais preditoras é removido de $\mathbf{z}$ antes de suavizar $r_{ij}^{(m)}$ em função de x_{ij}, i=1,...,n, j=1,...,p. O processo é repetido seqüencialmente para j=1,...,p, e pára quando todos os $\mathbf{f}_j^1$ convergirem. Finalmente, o valor estimado de $\mathbf{f}_j$ no passo m do PPL é dado por $\mathbf{f}_j^{(m)} = \mathbf{f}_j^1$ na convergência. O retroajuste pode ser esquematizado da seguinte forma:

Algoritmo de retroajuste

(i) Inicializa-se $\alpha^{(m)} = \bar{z}^{(m)} = \sum_{i=1}^{n} \frac{z_i^{(m)}}{n}$ e $\mathbf{f}_j^0 = \mathbf{f}_j^{(m-1)}$, j=1,...,p.

(ii) Calculam-se seqüencialmente

$$\mathbf{f}_1^{\,1} = \mathbf{S}_1^{(m)}\mathbf{r}_1^{\,(m)} = \mathbf{S}_1^{(m)}\,(\,\mathbf{z}^{(m)} - \mathbf{1}\bar{z}^{\,(m)} - \mathbf{f}_2^{\,0} - \mathbf{f}_3^{\,0} - \ldots - \mathbf{f}_p^{\,0}\,)$$

$$\mathbf{f}_2^{\,1} = \mathbf{S}_2^{(m)}\mathbf{r}_2^{\,(m)} = \mathbf{S}_2^{(m)}\,(\,\mathbf{z}^{(m)} - \mathbf{1}\bar{z}^{\,(m)} - \mathbf{f}_1^{\,1} - \mathbf{f}_3^{\,0} - \ldots - \mathbf{f}_p^{\,0}\,)$$

$$\vdots$$

$$\mathbf{f}_p^{\,1} = \mathbf{S}_p^{(m)}\mathbf{r}_p^{\,(m)} = \mathbf{S}_p^{(m)}\,(\,\mathbf{z}^{(m)} - \mathbf{1}\bar{z}^{\,(m)} - \mathbf{f}_1^{\,1} - \mathbf{f}_2^{\,1} - \ldots - \mathbf{f}_{p-1}^{\,1}),$$

onde $\mathbf{S}_j^{(m)}$ de dimensão (n x n), j=1,...,p, é a matriz suavizadora ponderada relativa à j-ésima covariável, com matriz de pesos $\mathbf{W}^{(m)} = \text{diagonal}\{\,w_1^{(m)},\ldots,w_n^{(m)}\,\}$, $\mathbf{1}$ é um vetor (n x 1) de valores unitários, e $\mathbf{r}_j^{(m)}$ é o vetor dos resíduos parciais cujos elementos são dados por (3.9).

(iii) O passo *(ii)* é repetido até que

$$\left\| \mathbf{f}_j^1 - \mathbf{f}_j^0 \right\| \leq \varepsilon,\ j=1,\ldots,p.$$

para um valor $\varepsilon > 0$ preestabelecido.

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

$$\begin{bmatrix} \mathbf{I} & \mathbf{S}_1^{(m)} & \mathbf{S}_1^{(m)} & \cdots & \mathbf{S}_1^{(m)} \\ \mathbf{S}_2^{(m)} & \mathbf{I} & \mathbf{S}_2^{(m)} & \cdots & \mathbf{S}_2^{(m)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_p^{(m)} & \mathbf{S}_p^{(m)} & \mathbf{S}_p^{(m)} & \cdots & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{f}_1^{(m)} \\ \mathbf{f}_2^{(m)} \\ \vdots \\ \mathbf{f}_p^{(m)} \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1^{(m)}\mathbf{z}^{(m)} \\ \mathbf{S}_2^{(m)}\mathbf{z}^{(m)} \\ \vdots \\ \mathbf{S}_p^{(m)}\mathbf{z}^{(m)} \end{bmatrix}, \tag{3.10}$$

onde $\mathbf{S}_j^{(m)}$ é a j-ésima matriz suavizadora ponderada e $\mathbf{z}^{(m)}$ é o vetor dos valores da variável resposta modificada, calculados no passo m do PPL. Esse sistema possui np equações que correspondem aos np parâmetros a serem estimados, sendo o retroajuste

considerado um método eficiente para resolvê-lo, pois apresenta economia computacional em relação à solução direta, principalmente quando o número de parâmetros é grande.

As matrizes $\mathbf{S}_j^{(m)}$, j=1,...,n, em (3.10) podem corresponder a diferentes métodos de suavização (como o *loess* ou o *cubic spline*, descritos no Capítulo 2), mas, em geral, o mesmo suavizador é utilizado para estimar $\mathbf{f}_1,\ldots,\mathbf{f}_p$.

No caso do suavizador *loess* ser utilizado, a i-ésima linha de $\mathbf{S}_j^{(m)}$ corresponde à i-ésima linha da matriz suavizadora ponderada

$$\mathbf{S}_{x_i}^{(m)} = \mathbf{X}(\mathbf{X}'\mathbf{A}_{x_i}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}_{x_i},$$

onde $\mathbf{X} = (\mathbf{1}, \mathbf{X}_j)$ e $\mathbf{A}_{x_i} = \text{diagonal}\{u_{x_i,1}\, w_1^{(m)},\ldots,u_{x_i,n}\, w_n^{(m)}\}$ com $u_{x_i,j}$ e $w_j^{(m)}$ definidos, respectivamente, em (2.5) e (3.8).

No caso do *cubic spline*,

$$\mathbf{S}_j^{(m)} = (\mathbf{W}^{(m)} + \lambda_j\mathbf{K}_j)^{-1}\mathbf{W}^{(m)},$$

onde $\mathbf{W}^{(m)}$ é a matriz diagonal de pesos com elementos dados em (3.8), $\mathbf{K}_j$ é a matriz de penalização para o j-ésimo preditor, definida de maneira análoga a $\mathbf{K}$ em (2.12) e λ_j é o parâmetro de suavização fixado para $\mathbf{f}_j$.

• Casos particulares

Como descrito anteriormente, o processo de estimação dos MAG's baseia-se em dois procedimentos iterativos: o PPL que fornece um "ciclo externo" devido à estimação de um modelo com estrutura semelhante a um MLG, e o retroajuste que fornece um "ciclo interno" para estimar mais de uma função não especificada f_j.

Assim, se a variável resposta segue uma distribuição Normal, isto é, o modelo aditivo é normal, a função de ligação é a identidade, a variável modificada $Z = Y$, $\mathbf{W} = \mathbf{I}$ e o procedimento MQRI é substituído por um método direto, ou seja, o procedimento iterativo externo desaparece e apenas o ciclo interno, correspondente ao retroajuste, é necessário.

No caso de um MAG com apenas uma função não especificada, o algoritmo de retroajuste não é necessário pois $\mathbf{f}^{(m)}$ é obtido diretamente, utilizando-se um alisador ponderado aplicado aos resíduos $r_i^{(m)} = z_i^{(m)} - \bar{z}^{(m)}$ em função de x_i, i=1,...,n, com matriz de pesos $\mathbf{W}^{(m)}$.

Motivação para o procedimento de ajuste dos MAG's

O procedimento PPL não pode ser encarado apenas como uma modificação do *scoring* de Fisher. Existem duas motivações teóricas para sua utilização como um método de obtenção das estimativas em um MAG.

• Motivação 1

Denotando por $\ell(y_i;\eta_i)$ o logaritmo da função de verossimilhança escrita como uma função de $\eta_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij})$, o logaritmo da função de verossimilhança penalizada é dado por

$$\pi(\mathbf{f}_1,\ldots,\mathbf{f}_p) = \sum_{i=1}^{n} \ell(y_i;\eta_i) - \tfrac{1}{2}\sum_{j=1}^{p} \lambda_j \int (f_j''(u))^2 du,$$

onde $\lambda_j \geq 0$ é o parâmetro de suavização fixado para estimar f_j, j=1,...,p. Pode-se demonstrar que as soluções $\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_p$ obtidas pelo método de estimação descrito

quando o alisador *cubic spline* é utilizado, maximizam a função de verossimilhança penalizada.

• Motivação 2

O procedimento de ajuste dos MAG's pode ser considerado uma versão empírica de um método para determinação de esperanças condicionais.

Seja $\mathbf{X} = (X_1,\ldots,X_p)$ e $\ell(\mathbf{y};\eta)$ o logaritmo da função de verossimilhança, com $\eta = f_1(x_1) + \ldots + f_p(x_p)$.

Suponha que se deseja determinar a melhor aproximação aditiva para

$$\eta = g[\, E(Y|\mathbf{X}=\mathbf{x}) \,].$$

Isto é, deseja-se determinar a aproximação aditiva que maximiza

(3.11) $$E[\, \ell(\mathbf{y};\eta) \,],$$

entre todas as funções $\eta = f_1(x_1) + \ldots + f_p(x_p)$. Uma condição suficiente para η maximizar (3.11) é que

$$E\left(\frac{\partial \ell(\zeta)}{\partial \eta} \,\middle|\, X_j \right) = 0, \forall\, j.$$

Estas equações são não lineares em η e $\mathbf{f}_j$ e sua resolução necessita de um método iterativo para determinar os verdadeiros $\mathbf{f}_1,\ldots,\mathbf{f}_p$; esse método envolve o cálculo de esperanças condicionais.

O PPL é uma versão empírica desse método, sendo as esperanças condicionais substituídas por um suavizador ponderado, que fornece, na convergência, estimativas $\hat{\mathbf{f}}_1,\ldots,\hat{\mathbf{f}}_p$ para $\mathbf{f}_1,\ldots,\mathbf{f}_p$.

Solução direta: procedimento alternativo ao retroajuste

Embora o retroajuste seja um algoritmo eficiente para resolver (3.10), pelo menos conceitualmente, estimativas para $\mathbf{f}_1,\ldots,\mathbf{f}_p$ podem ser obtidas diretamente pela relação

$$\begin{bmatrix} \hat{\mathbf{f}}_1 \\ \hat{\mathbf{f}}_2 \\ \vdots \\ \hat{\mathbf{f}}_p \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{S}_1 & \mathbf{S}_1 & \cdots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \mathbf{S}_2 & \cdots & \mathbf{S}_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_p & \mathbf{S}_p & \mathbf{S}_p & \cdots & \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_p \end{bmatrix} \mathbf{z} = \mathbf{M}^{-1}\mathbf{C}\mathbf{z},$$

se a inversa de $\mathbf{M}$ existe.

A matriz que produz $\hat{\mathbf{f}}_j$ a partir de $\mathbf{z}$, é definida como

$$\mathbf{R}_j = \mathbf{E}_j\mathbf{M}^{-1}\mathbf{C}, \quad j=1,\ldots,p$$

onde $\mathbf{E}_j$ é uma matriz de dimensão (n x np) composta por p "blocos" de dimensão n x n, sendo o j-ésimo bloco a matriz identidade, e os demais elementos iguais a zero, de tal maneira que

$$\hat{\mathbf{f}}_j = \mathbf{R}_j\mathbf{z}, \quad j=1,\ldots,p.$$

Seja $\mathbf{f} = \mathbf{f}_1 + \ldots + \mathbf{f}_p$. Então,

$$\hat{\mathbf{f}} = \mathbf{R}_1\mathbf{z} + \ldots + \mathbf{R}_p\mathbf{z} = \mathbf{R}_{NP}\mathbf{z}.$$

onde $\mathbf{R}_{NP} = \mathbf{R}_1 + \ldots + \mathbf{R}_p$ é a matriz suavizadora ponderada que produz $\hat{\mathbf{f}}$ a partir de $\mathbf{z}$.

Para modelos que envolvem apenas duas matrizes suavizadoras em seu ajuste, Hastie e Tibshirani (1990) fornecem expressões mais simples para $\mathbf{R}_1$ e $\mathbf{R}_2$, dadas por:

(3.12)
$$\mathbf{R}_1 = \mathbf{I} - (\mathbf{I} - \mathbf{S}_1\mathbf{S}_2)^{-1}(\mathbf{I} - \mathbf{S}_1)$$
$$\mathbf{R}_2 = \mathbf{I} - (\mathbf{I} - \mathbf{S}_2\mathbf{S}_1)^{-1}(\mathbf{I} - \mathbf{S}_2).$$

Neste caso,

$$\mathbf{R}_{NP} = (\mathbf{R}_1 + \mathbf{R}_2) = \mathbf{I} - (\mathbf{I} - \mathbf{S}_2)\,(\mathbf{I} - \mathbf{S}_1\mathbf{S}_2)^{-1}\,(\mathbf{I} - \mathbf{S}_1).$$

Expressões recursivas para modelos envolvendo mais de dois suavizadores foram deduzidas por Opsomer (2000). O custo computacional para obter $\mathbf{R}_j$, j=1,...,n, a partir dessas expressões é, entretanto, elevado.

A apresentação da solução direta neste ponto do trabalho não foi feita com o objetivo de utilizá-la como uma alternativa ao retroajuste na estimação de $\mathbf{f}_1,\ldots,\mathbf{f}_p$, uma vez que este algoritmo é mais eficiente do ponto de vista computacional, mas sim para obter expressões para $\hat{\mathbf{f}}_j$ e $\hat{\eta}$ que tornem mais simples o estudo de suas propriedades, como será visto no decorrer do capítulo.

Convergência do procedimento de ajuste dos MAG's

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 de convergência (Hastie e Tibshirani, 1990). Resultados sobre a convergência do algoritmo de retroajuste necessitam da definição de *concurvidade*. A concurvidade pode ser pensada como um fenômeno análogo à multicolinearidade em modelos lineares: enquanto o termo "colinearidade" refere-se à dependência linear entre os preditores, o termo "concurvidade" tem sido usado (Buja et al., 1986) para referir-se a formas de relação não lineares entre as variáveis preditoras em modelos aditivos. Ocorre concurvidade se existe $\mathbf{g} = (g_1,\ldots,g_p)'$, $\mathbf{g} \neq 0$, tal que

$$\begin{bmatrix} \mathbf{I} & \mathbf{S}_1^{(m)} & \mathbf{S}_1^{(m)} & \cdots & \mathbf{S}_1^{(m)} \\ \mathbf{S}_2^{(m)} & \mathbf{I} & \mathbf{S}_2^{(m)} & \cdots & \mathbf{S}_2^{(m)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_p^{(m)} & \mathbf{S}_p^{(m)} & \mathbf{S}_p^{(m)} & \cdots & \mathbf{I} \end{bmatrix} \mathbf{g} = 0.$$

Isto implica que, se $\mathbf{f}$ é solução de (3.10), $\mathbf{f} + c\mathbf{g}$ é também uma solução para qualquer c.

Buja et al. (1989) obtiveram condições suficientes para a convergência do algoritmo de retroajuste. No caso do modelo envolvendo p funções não especificadas, os autores mostraram que quando todas as $\mathbf{S}_j$, j=1,...,p, são simétricas com autovalores em [0,1] e não existe concurvidade, o retroajuste converge para uma solução única de (3.10), independentemente dos valores iniciais dos $\mathbf{f}_j$'s. Nas mesmas condições, na presença de concurvidade, o retroajuste converge para uma solução de (3.10), que depende dos valores iniciais das funções.

Assim, quando não existe concurvidade e cada $\mathbf{S}_j$, j=1,...,p, é uma matriz suavizadora de um *cubic spline*, a convergência é garantida.

Condições menos restritivas para a convergência do retroajuste para uma única solução, válidas para qualquer suavizador linear, foram recentemente obtidas por Opsomer (2000). No referido trabalho demonstra-se que o retroajuste irá convergir para uma única solução se

$$\left\| \mathbf{S}_j \mathbf{R}_{NP}^{-j} \right\| < 1,$$

onde $\mathbf{R}_{NP}^{-j}$ é uma matriz tal que $\mathbf{R}_j = \mathbf{I} - (\mathbf{I} - \mathbf{S}_j \mathbf{R}_{NP}^{-j})^{-1} (\mathbf{I} - \mathbf{S}_j)$, $j \in (1,...,p)$ e qualquer norma matricial $\|\cdot\|$ (ver, por exemplo, Graybill, 1983).

Ajuste dos modelos semi-paramétricos

Considere o modelo semi-paramétrico:

$$g(\mu_i) = \alpha + \sum_{j=1}^{r} \beta_j x_{ij} + \sum_{j=r+1}^{p} f_j(x_{ij}).$$

Os parâmetros $\alpha, \beta_1,...,\beta_r$ e as funções $\mathbf{f}_{r+1},...,\mathbf{f}_p$ podem ser estimados pelo PPL. Dados os valores iniciais $\beta^{(0)} = (\alpha^{(0)}, \beta_1^{(0)},...,\beta_r^{(0)})'$ e $\mathbf{f}_{r+1}^{(0)},...,\mathbf{f}_p^{(0)}$, estimativas para β e $\mathbf{f}_{r+1},...,\mathbf{f}_p$ são obtidas resolvendo-se, iterativamente, as seguintes equações:

(3.13) $$\beta^{(m)} = (\mathbf{X}'\mathbf{W}^{(m)}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(m)}\left(\mathbf{z}^{(m)} - \sum_{j=r+1}^{p}\mathbf{f}_j^{(m)}\right)$$

e

$$\mathbf{f}_{r+1}^{(m)} = \mathbf{S}_{r+1}^{(m)}\left(\mathbf{z}^{(m)} - \mathbf{X}\beta^{(m)} - \sum_{j=r+2}^{p}\mathbf{f}_j^{(m)}\right)$$

(3.14) $$\vdots$$

$$\mathbf{f}_p^{(m)} = \mathbf{S}_p^{(m)}\left(\mathbf{z}^{(m)} - \mathbf{X}\beta^{(m)} - \sum_{j=r+1}^{p-1}\mathbf{f}_j^{(m)}\right),$$

onde $\mathbf{X} = (\mathbf{1}, \mathbf{X}_1,\ldots,\mathbf{X}_r)$ é a matriz de planejamento referente à parte paramétrica do modelo e $\mathbf{S}_j^{(m)}$ é a matriz suavizadora ponderada relativa à j-ésima covariável, j=r+1,...,p. Após obter as estimativas $\beta^{(m)}$ e $\mathbf{f}_j^{(m)}$, j=r+1,...,p, pelo retroajuste, o PPL estima novos valores $\eta^{(m)}$, $\mu^{(m)}$, $\mathbf{z}^{(m)}$ e $\mathbf{W}^{(m)}$, e o processo é repetido até a convergência.

Este procedimento é análogo ao de um modelo não paramétrico com (p – r)+1 suavizadores: um suavizador é um operador projeção, com $\mathbf{S}_1 = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$, que produz o valor ajustado $\mathbf{X}\hat{\beta}$, e os (p – r) suavizadores restantes produzirão $\hat{\mathbf{f}}_{r+1},\ldots,\hat{\mathbf{f}}_p$.

Quando existe apenas uma função não especificada, **f**, as expressões (3.13) e (3.14) se reduzem a:

(3.15) $$\beta^{(m)} = (\mathbf{X}'\mathbf{W}^{(m)}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(m)}(\mathbf{z}^{(m)} - \mathbf{f}^{(m)})$$

e

(3.16) $$\mathbf{f}^{(m)} = \mathbf{S}^{(m)}(\mathbf{z}^{(m)} - \mathbf{X}\beta^{(m)}).$$

Assim, o retroajuste pode ser evitado se $\mathbf{f}^{(m)}$ for eliminado de (3.15). Para isso, $\mathbf{f}^{(m)}$ é substituido em (3.15) e fornece

(3.17) $$\beta^{(m)} = (\mathbf{X}'\mathbf{W}^{(m)}(\mathbf{I} - \mathbf{S}^{(m)})\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(m)}(\mathbf{I} - \mathbf{S}^{(m)})\mathbf{z}^{(m)}.$$

Então, após obter uma estimativa $\beta^{(m)}$ segundo (3.17), uma nova estimativa $\mathbf{f}^{(m)}$ é calculada segundo (3.16). O PPL estima, então, novos valores para $\eta^{(m)}$, $\mu^{(m)}$, $\mathbf{z}^{(m)}$ e $\mathbf{W}^{(m)}$, e o processo é repetido até a convergência. As estimativas assim obtidas são idênticas às fornecidas quando o retroajuste é realizado.

Opsomer e Ruppert (1999) e Thurston et al. (2000) mostraram que no modelo envolvendo mais de um termo não paramétrico, o retroajuste pode ser evitado considerando os estimadores:

$$\beta^{(m)} = (\mathbf{X}'\mathbf{W}^{(m)}(\mathbf{I} - \mathbf{R}_{NP}^{(m)})\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(m)}(\mathbf{I}-\mathbf{R}_{NP}^{(m)})\mathbf{z}^{(m)}$$

e

$$\mathbf{f}_{NP}{}^{(m)} = \mathbf{R}_{NP}^{(m)}(\mathbf{z}^{(m)} - \mathbf{X}\beta^{(m)}),$$

onde $\mathbf{X} = (\mathbf{1}, \mathbf{X}_1,\ldots, \mathbf{X}_r)$, $\mathbf{f}_{NP}{}^{(m)} = \mathbf{f}_{r+1}{}^{(m)} + \ldots + \mathbf{f}_p{}^{(m)}$ e

$$\mathbf{R}_{NP}^{(m)} = \sum_{j=r+1}^{p} \mathbf{R}_j^{(m)}$$

denota a matriz suavizadora ponderada generalizada correspondente aos termos não paramétricos do modelo. As soluções obtidas por essas equações são equivalentes às soluções do retroajuste. Entretanto, no caso de mais do que um termo não paramétrico, o algoritmo de retroajuste é utilizado para efeito da estimação do modelo, pois é mais eficiente que o método direto devido ao custo computacional para obtenção de $\mathbf{R}_{NP}^{(m)}$.

Um fato pouco evidenciado na literatura é que, em geral, os estimadores dos modelos semi-paramétricos não são bem definidos quando incluem o intercepto (Opsomer e Ruppert, 1999). Neste caso, quando um suavizador possui a propriedade de que a soma dos elementos das linhas de $\mathbf{S}$ é igual a 1 (como é o caso do *loess* e do *cubic spline*), $\mathbf{X}'\mathbf{W}^{(m)}(\mathbf{I} - \mathbf{S}^{(m)})\mathbf{X}$ e $\mathbf{X}'\mathbf{W}^{(m)}(\mathbf{I} - \mathbf{R}_{NP}^{(m)})\mathbf{X}$ são singulares. Uma solução simples é substituir as matrizes $\mathbf{S}_j^{(m)}$ por matrizes $\mathbf{S}_j^{(m)}$ centradas, dadas por $(\mathbf{I}-$

$\mathbf{11}'/n)\mathbf{S}_j^{(m)}$. Este procedimento faz com que a média de $\hat{\mathbf{f}}$ seja igual a zero a cada passo, e o modelo torna-se identificável.

3.3. Testes de hipótese

Estatísticas para testes de hipótese sobre os parâmetros dos MLG's serão inicialmente apresentadas como motivação para os testes sobre os componentes dos MAG's.

Nos MLG's, o desvio (*deviance*) é uma medida usada para avaliar a qualidade do ajuste e para comparar dois modelos ajustados. O desvio para o modelo ajustado $\hat{\mu}$, é definido por

$$D(\mathbf{y};\hat{\mu}) = 2\{\log(\mu_{max};\mathbf{y}) - \log(\hat{\mu};\mathbf{y})\}, \tag{3.18}$$

onde $\log(\mu_{max};\mathbf{y})$ e $\log(\hat{\mu};\mathbf{y})$ denotam, respectivamente, o logaritmo da função de verossimilhança calculada para o modelo saturado (com n parâmetros) e logaritmo da função de verossimilhança calculada para o modelo ajustado (com p+1 parâmetros).

Esta estatística mede a proximidade do ajuste do modelo aos dados (quanto maior o desvio, pior é o ajuste do modelo) e desempenha papel análogo à soma de quadrados dos resíduos na regressão de mínimos quadrados. Quando a variável resposta tem distribuição normal, o desvio e a soma de quadrados dos resíduos coincidem.

Embora o desvio não tenha distribuição conhecida, nem mesmo assintoticamente, é comum utilizar a distribuição $\chi^2_{n-(p+1)}$ como uma distribuição de referência para (3.18), e comparar o valor observado do desvio com a média desta distribuição, ou seja, se o valor dessa estatística for muito maior do que n – (p+1), tem-se indicação de que o modelo não está bem ajustado.

Sem perda de generalidade, pode-se escrever $D(\mathbf{y};\hat{\eta})$ no lugar de $D(\mathbf{y};\hat{\mu})$, uma vez que $\hat{\mu}$ está relacionado com $\hat{\eta}$ por meio da função de ligação $g(\mu) = \eta$.

Suponha que η_1 e η_2 sejam preditores lineares de dois MLG's, dados por:

$$\eta_1 = \alpha + \beta_1 X_1 + \ldots + \beta_q X_q$$

e

$$\eta_2 = \eta_1 + \beta_{q+1} X_{q+1} + \ldots + \beta_p X_p$$

com $q < p$, $i=1,\ldots,n$. Suponha ainda que $D(\mathbf{y};\hat{\eta}_1)$ e $D(\mathbf{y};\hat{\eta}_2)$ sejam, respectivamente, os desvios correspondentes a $\hat{\eta}_1$ e $\hat{\eta}_2$.

Para verificar se o modelo mais simples (com menos parâmetros) é adequado, deve-se testar

H_0: $\beta_{q+1} = \ldots = \beta_p = 0$,

H_1: ao menos um dos parâmetros $\beta_{q+1},\ldots,\beta_p$ não é nulo.

Para isso, o *desvio parcial*, denotado por

$$D(\hat{\eta}_2;\hat{\eta}_1) = D(\mathbf{y};\hat{\eta}_1) - D(\mathbf{y};\hat{\eta}_2), \tag{3.19}$$

pode ser usado para avaliar a contribuição dos termos presentes apenas em η_2. Sob certas condições de regularidade e assumindo que η_1 é adequado, $\phi D(\hat{\eta}_2;\hat{\eta}_1)$ tem aproximadamente distribuição χ^2 com graus de liberdade dados pela diferença entre as dimensões dos dois modelos.

No caso dos MAG's, o desvio também pode ser usado como uma medida de ajuste e a comparação entre modelos pode ser feita utilizando-se o cálculo do desvio parcial. Porém, nem mesmo as distribuições assintóticas dessas estatísticas foram determinadas. Entretanto, Hastie e Tibshirani (1990) mostraram, por simulação, que

distribuições χ^2 com número de graus de liberdade determinado da maneira a seguir são boas aproximações para as distribuições dessas estatísticas.

Seja **R** o operador ponderado de ajuste aditivo obtido no último passo do PPL; então

$$\hat{\eta} = \mathbf{Rz},$$

e seja $D(\mathbf{y};\hat{\eta})$ o desvio correspondente a $\hat{\eta}$.

Assim como nos MLG's, a qualidade do ajuste de um MAG pode ser avaliada comparando o valor observado de $D(\mathbf{y};\hat{\eta})$ com a média da distribuição χ^2 com número de graus de liberdade dado por

(3.20) $$gl = n - \text{traço}(2\mathbf{R} - \mathbf{R}'\mathbf{W}\mathbf{R}\mathbf{W}^{-1}).$$

Se $D(\mathbf{y};\hat{\eta})$ for muito maior que o valor de gl, tem-se indicação de falta de ajuste do modelo.

Suponha agora que $\mathbf{R}_1$ e $\mathbf{R}_2$ sejam operadores de ajuste aditivo ponderado que produzem, respectivamente, $\hat{\eta}_1 = \mathbf{R}_1\mathbf{z}$ e $\hat{\eta}_2 = \mathbf{R}_2\mathbf{z}$, sendo

$$\eta_1 = \alpha + f_1(X_1) + \ldots + f_q(X_q)$$

e

$$\eta_2 = \eta_1 + f_{q+1}(X_{q+1}) + \ldots + f_p(X_p).$$

Neste caso, o valor de $\phi D(\hat{\eta}_1;\hat{\eta}_2)$ é usado para avaliar a contribuição dos termos presentes apenas em η_2, utilizando-se como referência a distribuição χ^2 com número de graus de liberdade dado por

$$gl(\hat{\eta}_1) - gl(\hat{\eta}_2) = \text{traço}(2\mathbf{R}_1 - \mathbf{R}_1'\mathbf{W}_1\mathbf{R}_1\mathbf{W}_1^{-1}) - \text{traço}(2\mathbf{R}_2 - \mathbf{R}_2'\mathbf{W}_2\mathbf{R}_2\mathbf{W}_2^{-1}).$$

Valores grandes de $\phi D(\hat{\eta}_1;\hat{\eta}_2)$ sugerem que a contribuição desses termos é significativa.

Os cálculos para obtenção dos graus de liberdade segundo as definições dadas acima apresentam elevado custo computacional. Por este motivo, a quantidade

$$gl = n - 1 - \sum_{j=1}^{p} \left(\text{traço}(\mathbf{S}_j) - 1\right)$$

é usada como uma medida aproximada para os graus de liberdade do desvio, dado em (3.18).

3.4. Medidas de precisão e bandas de confiança pontuais

Bandas de confiança pontuais para η podem ser obtidas usando a metodologia descrita em Hastie e Tibshirani (1990), que se baseia em procedimentos de linearização. A idéia é aproximar a variável resposta modificada $\mathbf{z}$ por uma quantidade assintoticamente equivalente $\mathbf{z}_0$, assumindo que o modelo é consistente. A matriz de variâncias e covariâncias assintótica de $\hat{\eta} = \mathbf{R}\mathbf{z}$ é, então, estimada com base na matriz de variâncias e covariâncias de $\mathbf{z}_0$ dada por $\mathbf{W}_0^{-1}\phi$ e considerando uma versão assintótica de $\mathbf{R}$, já que esta matriz não é um operador linear pois depende de y_i por meio dos pesos w_i, i=1,...,n. Obtém-se então,

$$\text{Var}(\hat{\eta}) \approx \mathbf{R}_0\mathbf{W}_0^{-1}\mathbf{R}_0'\phi$$

$$\approx \mathbf{R}\mathbf{W}^{-1}\mathbf{R}'\phi.$$

Similarmente,

$$\text{Var}(\hat{\mathbf{f}}_j) \approx \mathbf{R}_j\mathbf{W}^{-1}\mathbf{R}_j'\phi,$$

onde $\mathbf{R}_j$ é a matriz que produz $\hat{\mathbf{f}}_j$ a partir de $\mathbf{z}$. O símbolo $\approx$ significa "assintoticamente igual a".

Para o desenvolvimento dessa metodologia, foram supostas válidas as mesmas condições de regularidade requeridas para o desenvolvimento de resultados assintóticos para os MLG's (ver, por exemplo, Sen e Singer, 1993).

Os resultados obtidos permitem mostrar que $\hat{\eta}$ tem distribuição assintótica $N(\eta_0, \mathbf{R}_0\mathbf{W}_0^{-1}\mathbf{R}_0'\phi)$.

Bandas de confiança pontuais aproximadas para $f_j(x_{ij})$ são dadas por:

$$\hat{f}_j(x_{ij}) \pm 2 \text{ Diagonal}\left(\mathbf{R}_j\mathbf{W}^{-1}\mathbf{R}_j'\phi\right)^{1/2}. \tag{3.21}$$

Nos modelos semi-paramétricos a suposição de consistência não é válida devido ao fato do estimador de $\hat{\beta}$ considerado não ser consistente. Quando taxas de convergência "típicas" são utilizadas para os termos não paramétricos, a consistência de $\hat{\beta}$ não pode ser demonstrada (Opsomer e Ruppert, 1999 e Speckman, 1988). Uma expressão aproximada para a matriz de variâncias e covariâncias de $\hat{\beta}$ (Thurston et al., 2000) é dada por

$$\text{Var}(\hat{\beta}) = (\mathbf{X}'\mathbf{W}(\mathbf{I} - \mathbf{R}_{NP})\mathbf{X})^{-1},$$

onde $\mathbf{X} = (\mathbf{1}, \mathbf{X}_1, \ldots, \mathbf{X}_r)$ e $\mathbf{R}_{NP} = \sum_{j=r+1}^{p} \mathbf{R}_j$.

3.5. Seleção do parâmetro de suavização

Na Seção 2.5 foi apresentada a validação cruzada como um critério para seleção automática do parâmetro de suavização de um alisador. Esse critério pode ser adotado para selecionar parâmetros de suavização $\lambda_1, \ldots, \lambda_p$ em um MAG, no qual existem p termos não paramétricos $f_1, \ldots, f_p$.

Sejam $\mathbf{S}_1, \ldots, \mathbf{S}_p$ matrizes suavizadoras lineares com correspondentes parâmetros de suavização $\lambda_1, \ldots, \lambda_p$. Seja $\hat{\mu}_\lambda^{-i}$ o valor ajustado para a i-ésima observação, obtido após retirar o ponto (x_i, y_i) da amostra. A estatística da validação cruzada do desvio é definida por

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

A idéia é minimizar esta quantidade sob $\lambda_1,\ldots,\lambda_p$. No entanto, isto gera um custo computacional muito grande, necessitando n aplicações completas do procedimento PPL para cada valor pré selecionado de $\lambda_1,\ldots,\lambda_p$. Por isso, medidas aproximadas foram desenvolvidas para reduzir o trabalho computacional.

Uma opção muito usada na prática é a estatística

$$AIC = \frac{1}{n}\sum_{i=1}^{n} D(y_i; \hat{\mu}_i) + \frac{2}{n}\text{traço}(\mathbf{R})\,\phi, \tag{3.22}$$

inspirada no critério de informação de Akaike (Hastie e Tibshirani, 1990), que julga o desvio do modelo e o *número de parâmetros*. Assim, valores pequenos desta estatística indicam um bom ajuste do modelo.

A estatística AIC requer somente uma aplicação do PPL para cada valor $\lambda_1,\ldots,\lambda_p$ e por isso, é muito mais fácil de ser calculada do que a VC.

Entretanto, não existem resultados sobre a adequação de sua utilização como um critério para a seleção do parâmetro de suavização, e gráficos com a curva suavizada estimada sobreposta aos dados podem ser usados para visualizar a adequação da *quantidade de suavização* de cada alisador a um dado λ, como já foi comentado no Capítulo 2.

Além disso, a escolha do λ pode depender da natureza do problema estudado. Como ilustração, considere o método *loess*. Neste caso, o λ está relacionado ao tamanho da vizinhança, ou seja, ao número de pontos em cada ajuste local e, por este motivo, pode ser selecionado subjetivamente. Por exemplo, na análise de séries cronológicas para estudar os efeitos da poluição atmosférica sobre a morbi-mortalidade, recomenda-se que a escolha do parâmetro de suavização para o termo que controla a sazonalidade de longa duração seja tal, que as observações referentes a

períodos de aproximadamente 180 dias sejam incluídas em cada ajuste local. Por outro lado, Schwartz (1999) comenta que o valor de λ pode afetar a autocorrelação dos resíduos, e este fato deve também ser considerado na escolha de λ.

3.6. Exemplo: modelo de Poisson semi-paramétrico

O PPL para estimação de um modelo de Poisson semi-paramétrico é ilustrado a seguir.

Sejam Y_i, i=1,...,n, variáveis aleatórias independentes, cada uma com distribuição de Poisson com parâmetro μ_i, isto é:

$$f(y_i;\mu_i) = \exp\big(y_i \log(\mu_i) - \mu_i + (-\log(y_i!))\big).$$

Comparando a função de probabilidade acima com (3.1), é fácil identificar

$$\phi = 1,\ \theta_i = \log(\mu_i),\ b(\theta_i) = \exp(\theta_i) = \mu_i \text{ e } c(y_i,\phi) = -\log(y_i!).$$

Portanto,

$$E(Y_i) = \exp(\theta_i) = \mu_i,$$

e

$$Var(Y_i) = \exp(\theta_i) = \mu_i.$$

Considere o conjunto de dados apresentado na Tabela 3.1 obtido de acordo com o seguinte procedimento: de forma independente, foram geradas $n = 10$ observações x_{i1} de uma variável X_1 com distribuição N(15,3), e x_{i2} de uma variável X_2 com distribuição N(10,2). Calcularam-se então os valores x_{i3} da variável $X_3 = X_2 + (X_2)^3$. Os valores y_i da variável resposta Y foram gerados a partir de uma distribuição de Poisson com média $\exp(0{,}01x_{i2} + 0{,}001x_{i3})$, i=1,...,10. Para estes dados, foi ajustado o modelo semi-paramétrico

$$\eta = \log[\, E(Y)\,] = \alpha + \beta X_1 + f(X_3), \tag{3.23}$$

sendo α e β parâmetros desconhecidos e f uma função não especificada, que será estimada pelo PPL, considerando o suavizador *loess* com $\lambda = 0,7$.

Tabela 3.1: Dados gerados para exemplo.

i	x_{i1}	x_{i3}	y_i
1	15,42	545,01	1
2	16,90	687,36	2
3	17,08	736,07	3
4	12,79	902,40	2
5	14,59	904,53	3
6	17,21	992,43	3
7	18,00	1121,91	1
8	11,22	1730,46	9
9	13,12	1985,39	9
10	14,48	2075,14	8

Na iteração m do PPL são calculados:

$$\eta_i^{(m)} = \alpha^{(m-1)} + \beta^{(m-1)} x_{i1} + f^{(m-1)}(x_{i3}),$$

$$\mu_i^{(m)} = \exp(\eta_i^{(m)}),$$

$$z_i^{(m)} = \eta_i^{(m)} + \frac{y_i - \mu_i^{(m)}}{\mu_i^{(m)}},$$

$$w_i^{(m)} = \mu_i^{(m)}, \quad i=1,\ldots,10,$$

e as matrizes suavizadoras:

$$\mathbf{S}_1^{(m)} = \mathbf{X}(\mathbf{X}'\mathbf{W}^{(m)}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(m)},$$

onde $\mathbf{X} = (\mathbf{1}, \mathbf{X}_1)$ e $\mathbf{W}^{(m)} = \text{diagonal}\{w_1^{(m)},\ldots,w_{10}^{(m)}\}$, e $\mathbf{S}_2^{(m)}$ que tem a i-ésima linha correspondente à i-ésima linha da matriz suavizadora ponderada

$$\mathbf{S}_{x_i}^{(m)} = \mathbf{X}_{np}(\mathbf{X}_{np}'\mathbf{A}_{x_i}\mathbf{X}_{np})^{-1}\mathbf{X}_{np}'\mathbf{A}_{x_i},$$

onde $\mathbf{X}_{np} = (\mathbf{1},\ \mathbf{X}_3)$ e $\mathbf{A}_{x_i} = \text{diagonal}\{u_{x_i,1}\, w_1^{(m)},\ldots,u_{x_i,10}\, w_{10}^{(m)}\}$ com $u_{x_i,j}$ definido em (2.5).

Como tem-se apenas um termo não paramétrico no modelo, o algoritmo de retroajuste é eliminado do PPL, sendo as estimativas β e $\mathbf{f}$ na m-ésima iteração desse processo, dadas por:

$$\beta^{(m)} = (\mathbf{X}'\mathbf{W}^{(m)}(\mathbf{I} - \mathbf{S}_2^{(m)})\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(m)}(\mathbf{I} - \mathbf{S}_2^{(m)})\mathbf{z}^{(m)}),$$

e

$$\mathbf{f}^{(m)} = \mathbf{S}_2^{(m)}(\mathbf{z}^{(m)} - \mathbf{X}\beta^{(m)}).$$

A matriz $\mathbf{S}_2^{(m)}$ foi centrada de forma a evitar não identificabilidade do modelo.

Alguns resultados obtidos durante a execução do processo são apresentados no Quadro 3.1.

• Passo 1 - Inicialização: Os valores iniciais $\alpha^{(0)} = 4{,}59$ e $\beta^{(0)} = -0{,}23$ foram obtidos da regressão de mínimos quadrados de $\log(y_i)$ em x_{i1}, i=1,...,10, e $\mathbf{f}^{(0)} = \mathbf{0}$.

• Passo 2 - 1ª iteração (m = 1):

i	η_i	$\mu_i^{(1)}$	$= w_i^{(1)}$	$z_i^{(1)}$
1	1,04	2,82		0,39
2	0,70	2,00		0,69
3	0,66	1,93		1,21
4	1,64	5,17		1,03
5	1,23	3,41		1,11
6	0,63	1,87		1,23
7	0,44	1,56		0,09
8	2,00	7,42		2,22
9	1,57	4,79		2,44
10	1,25	3,50		2,54

$$\mathbf{S}_2^{(1)} = \begin{bmatrix} 0{,}61 & 0{,}21 & 0{,}12 & -0{,}25 & -0{,}16 & -0{,}12 & -0{,}10 & -0{,}09 & -0{,}11 & -0{,}10 \\ 0{,}25 & 0{,}15 & 0{,}12 & -0{,}02 & -0{,}02 & -0{,}08 & -0{,}10 & -0{,}09 & -0{,}11 & -0{,}10 \\ 0{,}12 & 0{,}14 & 0{,}12 & 0{,}05 & 0{,}03 & -0{,}07 & -0{,}10 & -0{,}09 & -0{,}11 & -0{,}10 \\ -0{,}14 & -0{,}01 & 0{,}04 & 0{,}23 & 0{,}15 & 0{,}06 & -0{,}04 & -0{,}09 & -0{,}11 & -0{,}10 \\ -0{,}14 & -0{,}01 & 0{,}03 & 0{,}23 & 0{,}15 & 0{,}06 & -0{,}04 & -0{,}09 & -0{,}11 & -0{,}10 \\ -0{,}14 & -0{,}10 & -0{,}08 & 0{,}19 & 0{,}13 & 0{,}13 & 0{,}17 & -0{,}09 & -0{,}11 & -0{,}10 \\ -0{,}14 & -0{,}18 & -0{,}21 & 0{,}02 & 0{,}03 & 0{,}26 & 0{,}52 & -0{,}09 & -0{,}11 & -0{,}10 \\ -0{,}14 & -0{,}07 & -0{,}05 & -0{,}15 & -0{,}10 & -0{,}07 & -0{,}03 & -0{,}61 & 0{,}05 & -0{,}05 \\ -0{,}14 & -0{,}07 & -0{,}05 & -0{,}15 & -0{,}10 & -0{,}09 & -0{,}12 & 0{,}12 & 0{,}31 & 0{,}29 \\ -0{,}14 & -0{,}07 & -0{,}05 & -0{,}15 & -0{,}10 & -0{,}09 & -0{,}15 & -0{,}10 & 0{,}42 & 0{,}43 \end{bmatrix}$$

$$\beta^{(1)} = \begin{bmatrix} 1{,}60 \\ -0{,}02 \end{bmatrix} \quad \text{e} \quad \mathbf{f}^{(1)} = \begin{bmatrix} -0{,}75 & -0{,}50 & -0{,}41 & -0{,}25 & -0{,}25 & -0{,}41 & -0{,}70 & 0{,}77 & 1{,}18 & 1{,}31 \end{bmatrix}'$$

⋮

• Passo 2 - 5ª iteração (m = 5):

i	η_i	$\mu_i^{(5)}$	$= w_i^{(5)}$	$z_i^{(5)}$
1	0,43	1,54		0,08
2	0,68	1,97		0,69
3	0,77	2,16		1,16
4	1,02	2,78		0,74
5	0,97	2,65		1,11
6	0,76	2,14		1,16
7	0,50	1,65		0,11
8	2,08	8,00		2,20
9	2,20	9,07		2,20
10	2,20	9,05		2,09

$$\mathbf{S}_2^{(5)} = \begin{bmatrix} 0{,}51 & 0{,}29 & 0{,}18 & -0{,}21 & -0{,}20 & -0{,}17 & -0{,}11 & -0{,}08 & -0{,}10 & -0{,}11 \\ 0{,}19 & 0{,}19 & 0{,}18 & -0{,}02 & -0{,}02 & -0{,}11 & -0{,}11 & -0{,}08 & -0{,}10 & -0{,}11 \\ 0{,}08 & 0{,}17 & 0{,}17 & 0{,}04 & 0{,}03 & -0{,}08 & -0{,}11 & -0{,}08 & -0{,}10 & -0{,}11 \\ -0{,}11 & -0{,}02 & 0{,}04 & 0{,}16 & 0{,}16 & 0{,}10 & -0{,}03 & -0{,}08 & -0{,}10 & -0{,}11 \\ -0{,}11 & -0{,}03 & 0{,}03 & 0{,}16 & 0{,}16 & 0{,}10 & -0{,}03 & -0{,}08 & -0{,}10 & -0{,}11 \\ -0{,}11 & -0{,}11 & -0{,}09 & 0{,}13 & 0{,}13 & 0{,}17 & 0{,}17 & -0{,}08 & -0{,}10 & -0{,}11 \\ -0{,}11 & -0{,}20 & -0{,}24 & 0{,}02 & 0{,}03 & 0{,}28 & 0{,}51 & -0{,}08 & -0{,}10 & -0{,}11 \\ -0{,}11 & -0{,}10 & -0{,}09 & -0{,}10 & -0{,}09 & -0{,}09 & -0{,}03 & -0{,}60 & 0{,}10 & -0{,}08 \\ -0{,}11 & -0{,}10 & -0{,}09 & -0{,}10 & -0{,}09 & -0{,}10 & -0{,}12 & 0{,}10 & 0{,}28 & 0{,}34 \\ -0{,}11 & -0{,}10 & -0{,}09 & -0{,}10 & -0{,}09 & -0{,}11 & -0{,}14 & -0{,}12 & 0{,}34 & 0{,}51 \end{bmatrix}$$

$$\beta^{(5)} = \begin{bmatrix} 1{,}55 \\ -0{,}03 \end{bmatrix} \quad \text{e} \quad \mathbf{f}^{(5)} = \begin{bmatrix} -0{,}72 & -0{,}44 & -0{,}34 & -0{,}20 & -0{,}20 & -0{,}35 & -0{,}59 & 0{,}82 & 0{,}99 & 1{,}03 \end{bmatrix}'$$

Quadro 3.1: Alguns resultados do PPL para os dados da Tabela 3.1.

Após m = 5 iterações, o critério de parada foi satisfeito (com $\varepsilon = 0{,}001$), e o procedimento foi encerrado.

Para o cálculo da matriz de variâncias e covariâncias aproximadas para $\hat{\beta}$ e $\hat{\mathbf{f}}$, as matrizes $\mathbf{R}_1$ e $\mathbf{R}_2$, dadas em (3.12), foram calculadas considerando $\mathbf{S}_1$ e $\mathbf{S}_2$ obtidas no último passo. Obtiveram-se então,

$$\mathbf{R}_1 = \begin{bmatrix} 0{,}09 & 0{,}11 & 0{,}12 & 0{,}06 & 0{,}09 & 0{,}12 & 0{,}11 & 0{,}07 & 0{,}08 & 0{,}13 \\ 0{,}04 & 0{,}17 & 0{,}21 & -0{,}08 & 0{,}06 & 0{,}19 & 0{,}17 & -0{,}06 & 0{,}03 & 0{,}26 \\ 0{,}03 & 0{,}17 & 0{,}23 & -0{,}10 & 0{,}06 & 0{,}20 & 0{,}18 & -0{,}07 & 0{,}02 & 0{,}28 \\ 0{,}18 & 0{,}02 & -0{,}04 & 0{,}32 & 0{,}15 & -0{,}01 & 0{,}01 & 0{,}31 & 0{,}18 & -0{,}12 \\ 0{,}11 & 0{,}08 & 0{,}07 & 0{,}14 & 0{,}11 & 0{,}08 & 0{,}08 & 0{,}15 & 0{,}12 & 0{,}05 \\ 0{,}02 & 0{,}18 & 0{,}24 & -0{,}12 & 0{,}06 & 0{,}21 & 0{,}18 & -0{,}08 & 0{,}02 & 0{,}29 \\ -0{,}00 & 0{,}21 & 0{,}29 & -0{,}19 & 0{,}04 & 0{,}25 & 0{,}21 & -0{,}15 & -0{,}01 & 0{,}36 \\ 0{,}23 & -0{,}04 & 0{,}14 & 0{,}48 & 0{,}18 & -0{,}09 & -0{,}06 & 0{,}45 & 0{,}24 & -0{,}25 \\ 0{,}16 & 0{,}03 & 0{,}02 & 0{,}29 & 0{,}14 & 0{,}01 & 0{,}02 & 0{,}28 & 0{,}17 & -0{,}08 \\ 0{,}12 & 0{,}08 & 0{,}06 & 0{,}16 & 0{,}11 & 0{,}07 & 0{,}07 & 0{,}16 & 0{,}12 & 0{,}04 \end{bmatrix},$$

$$\mathbf{R}_2 = \begin{bmatrix} 0{,}56 & 0{,}24 & 0{,}09 & -0{,}05 & -0{,}17 & -0{,}25 & -0{,}17 & 0{,}05 & -0{,}05 & -0{,}25 \\ 0{,}22 & 0{,}16 & 0{,}12 & 0{,}06 & -0{,}01 & -0{,}15 & -0{,}15 & -0{,}01 & -0{,}07 & -0{,}19 \\ 0{,}10 & 0{,}14 & 0{,}13 & 0{,}10 & 0{,}05 & -0{,}11 & -0{,}14 & -0{,}02 & -0{,}08 & -0{,}17 \\ -0{,}10 & -0{,}03 & 0{,}02 & 0{,}19 & 0{,}16 & 0{,}09 & -0{,}04 & -0{,}06 & -0{,}09 & -0{,}13 \\ -0{,}10 & -0{,}04 & 0{,}02 & 0{,}19 & 0{,}16 & 0{,}09 & -0{,}04 & -0{,}06 & -0{,}09 & -0{,}13 \\ -0{,}09 & -0{,}13 & -0{,}13 & 0{,}20 & 0{,}14 & 0{,}13 & 0{,}15 & -0{,}02 & -0{,}08 & -0{,}17 \\ -0{,}05 & -0{,}26 & -0{,}35 & 0{,}20 & 0{,}06 & 0{,}19 & 0{,}45 & 0{,}07 & -0{,}04 & -0{,}27 \\ -0{,}21 & 0{,}01 & 0{,}09 & -0{,}38 & -0{,}15 & 0{,}05 & 0{,}08 & 0{,}34 & -0{,}01 & 0{,}17 \\ -0{,}17 & -0{,}03 & 0{,}02 & -0{,}28 & -0{,}13 & -0{,}01 & -0{,}05 & -0{,}07 & 0{,}21 & 0{,}50 \\ -0{,}15 & -0{,}05 & -0{,}01 & -0{,}22 & -0{,}12 & -0{,}04 & -0{,}09 & -0{,}24 & 0{,}30 & 0{,}63 \end{bmatrix},$$

$$\text{Vâr}(\hat{\beta}) = (\mathbf{X'W(I-R}_2)\mathbf{X})^{-1} = \begin{bmatrix} 1{,}22 & -0{,}09 \\ -0{,}09 & 0{,}01 \end{bmatrix},$$

e

$$\text{Vâr}(\hat{\mathbf{f}}) = \mathbf{R}_2\mathbf{W}^{-1}\mathbf{R}_2' = \begin{bmatrix} 0{,}30 & 0{,}14 & 0{,}09 & -0{,}06 & -0{,}06 & -0{,}09 & -0{,}13 & -0{,}07 & -0{,}06 & -0{,}06 \\ 0{,}14 & 0{,}08 & 0{,}06 & -0{,}01 & -0{,}01 & -0{,}05 & -0{,}09 & -0{,}05 & -0{,}04 & -0{,}04 \\ 0{,}09 & 0{,}06 & 0{,}05 & 0{,}00 & 0{,}00 & -0{,}03 & -0{,}08 & -0{,}04 & -0{,}03 & -0{,}03 \\ -0{,}06 & -0{,}01 & 0{,}00 & 0{,}04 & 0{,}04 & 0{,}03 & 0{,}02 & -0{,}03 & -0{,}02 & -0{,}02 \\ -0{,}06 & -0{,}01 & 0{,}00 & 0{,}04 & 0{,}04 & 0{,}04 & 0{,}02 & -0{,}02 & -0{,}02 & -0{,}02 \\ -0{,}09 & -0{,}05 & -0{,}03 & 0{,}03 & 0{,}04 & 0{,}07 & 0{,}12 & -0{,}02 & -0{,}03 & -0{,}03 \\ -0{,}13 & -0{,}09 & -0{,}08 & 0{,}02 & 0{,}02 & 0{,}12 & 0{,}26 & -0{,}01 & -0{,}05 & -0{,}05 \\ -0{,}07 & -0{,}05 & -0{,}04 & -0{,}03 & -0{,}02 & -0{,}02 & -0{,}01 & 0{,}12 & 0{,}07 & 0{,}05 \\ -0{,}06 & -0{,}04 & -0{,}03 & -0{,}02 & -0{,}02 & -0{,}03 & -0{,}05 & 0{,}07 & 0{,}09 & 0{,}09 \\ -0{,}06 & -0{,}04 & -0{,}03 & -0{,}02 & -0{,}02 & -0{,}03 & -0{,}05 & 0{,}05 & 0{,}09 & 0{,}11 \end{bmatrix}$$

Os valores previstos de **f** e bandas de confiança pontuais calculadas de acordo com (3.21) são representadas na Figura 3.1. Por esta figura, pode-se visualizar a relação da variável X_3 na resposta, após controlar pela variável X_2.

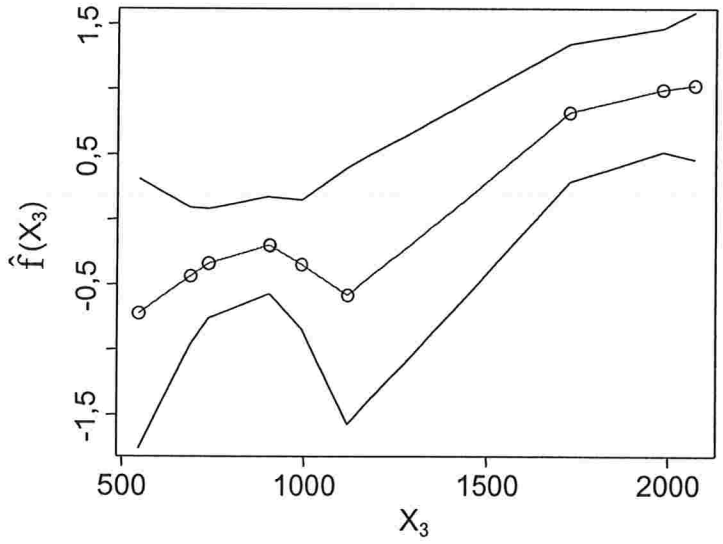


Figura 3.1: Valores previstos e bandas de confiança pontuais.

No modelo de Poisson, o desvio é dado por

$$D(\mathbf{y};\hat{\mu}) = 2\sum_{i=1}^{n}\left(y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\right).$$

Na convergência, obteve-se desvio igual a 1,65 com 4,61 graus de liberdade, calculados de acordo com (3.20).

A existência de efeito não linear de X_3 foi testada utilizando a estatística do desvio parcial, considerando que o modelo

$$\log[\ E(Y)\ ] = \alpha + \beta_1 X_1 + \beta_2 X_3$$

é um submodelo de (3.23). O valor da estatística (3.19) foi 1,31 com 2,39 graus de liberdade, e portanto, não houve evidências para considerar a existência de efeito não linear de X_3 ($p = 0{,}609$). Isto pode ter ocorrido por causa do pequeno número de pontos considerado, devido ao fato do objetivo do exemplo ser apenas ilustrar os procedimentos de estimação descritos no capítulo.

Capítulo 4

Aplicação

4.1. Apresentação do problema

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; Saldiva et al., 1995 e Braga et al., 1999, por exemplo). As populações mais vulneráveis são as constituídas por crianças, idosos e pessoas que apresentam doenças respiratórias prévias (Dockery e Pope, 1994; Saldiva et al., 1995; Gouveia e Fletcher, 2000, por exemplo).

A maior parte desses estudos são do tipo ecológico, isto é, são de base populacional, e um grupo, ao invés de um indivíduo, constitui a unidade de observação (Morgenstern, 1995) 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 socio-econômicos (Rothman et al., 1998), pois esses fatores não variam de dia para dia com a poluição atmosférica. Estas variáveis são, entretanto, possíveis confundidoras quando se comparam populações de localizações geográficas distintas sujeitas a diferentes níveis de poluição (Schwartz, 1994 e André et al., 2000). Em geral, variáveis temporais e climáticas são consideradas confundidoras nesses estudos.

A cidade de São Paulo possui um cenário apropriado para o desenvolvimento de estudos dos efeitos da poluição atmosférica sobre a saúde de sua população. Um dos motivos é o fato de constituir o maior centro urbano da América Latina, com insuficiente oferta de transporte coletivo, e uma malha viária aquém das necessidades de uma frota em torno de 5.500.000 veículos leves em toda região metropolitana, que constituem a principal fonte da poluição do ar (CETESB, 2000). A poluição atmosférica na cidade de São Paulo 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.

A CETESB, agência ambiental governamental, opera uma rede automática de monitoramento do ar composta por 25 estações fixas de amostragem (23 na região metropolitana e 2 na área de Cubatão) e 2 estações móveis. As estações móveis são deslocadas em função da necessidade de monitoramento em locais onde não existem estações de amostragem ou para estudos complementares à própria rede. Na cidade de São Paulo, estão situadas 13 estações de monitoramento da qualidade do ar, conforme mostra a Figura 4.1, responsáveis pela medição e registro contínuo das concentrações, dentre outros, dos seguintes poluentes: dióxido de enxofre (SO_2), em $\mu g/m^3$, monóxido de carbono (CO), em ppm, material particulado inalável (PM_{10}), em $\mu g/m^3$, ozônio (O_3), em $\mu g/m^3$, e dióxido de nitrogênio (NO_2), em $\mu g/m^3$. O registro diário destes poluentes é usualmente feito da seguinte forma: média de 24 horas para o SO_2, maior média móvel de 8 horas para o CO, média de 24 horas para o PM_{10}, pico horário de 24 horas para O_3 e pico horário de 24 horas para o NO_2.

Os dados de mortalidade na cidade de São Paulo são registrados e verificados por uma agência municipal centralizadora da emissão de certidões de óbitos, dentro do Programa de Aprimoramento das Informações de Mortalidade no Município de

São Paulo (PRO-AIM).

Já os dados de morbidade referentes aos atendimentos públicos, cobrindo aproximadamente 50% das internações hospitalares (Braga et al., 1999) são obtidos no Sistema Único de Saúde (SUS).

Os dados sobre as características climáticas (temperatura, umidade, etc.) podem ser obtidos junto ao Instituto de Astronomia e Geofísica da Universidade de São Paulo (IAG-USP).

Essas fontes de informação têm sido utilizadas em diversos estudos sobre os efeitos da poluição atmosférica na saúde dos habitantes do município de São Paulo realizados por pesquisadores do Laboratório de Poluição Atmosférica Experimental (LPAE) da Faculdade de Medicina da USP. Uma descrição de alguns desses trabalhos pode ser encontrada em André et al. (2000). Em um deles, Pereira et al. (1998), considerando as diferentes suscetibilidades das faixas etárias aos poluentes atmosféricos, buscou estabelecer um novo marcador dos efeitos nocivos da poluição sobre a saúde. Este estudo detectou a existência de associação entre poluição atmosférica e mortalidade fetal tardia (natimortalidade), com base em dados diários referentes aos anos de 1991 e 1992. Formalmente, os óbitos fetais tardios são definidos como os óbitos ocorridos intra-útero, em fetos com idade gestacional superior a 28 semanas, ou com peso superior a 1000 gramas, ou ainda com um comprimento crâneo-caudal superior a 35 cm. Quando um destes critérios for satisfeito, a legislação brasileira exige o preenchimento da Declaração de Óbito, e conseqüentemente, do Atestado de Óbito, para que seja efetuado o sepultamento.

A verificação da existência de associação entre mortalidade fetal tardia (variável resposta) e concentrações dos poluentes (variáveis independentes) foi feita utilizando-se um modelo de regressão de Poisson. Devido ao pioneirismo do estudo,

foram consideradas no modelo as mesmas variáveis de controle geralmente utilizadas em pesquisas de mortalidade e morbidade por causas respiratórias: meses do ano, dias da semana, temperatura mínima diária e umidade relativa do ar às 12 horas.

Vale ressaltar que, em estudos que avaliam os efeitos da poluição na morbi-mortalidade, é razoável considerar que os efeitos das concentrações dos poluentes e das variáveis climáticas, caso existam, não se dão necessariamente no mesmo dia em que ocorre o evento de interesse (óbito ou internação). Ou seja, o número de óbitos ou internações ocorridos no dia de hoje pode ser uma conseqüência das condições meteorológicas e da poluição não apenas de hoje, mas também de alguns dias anteriores. Por este motivo, é comum utilizarem-se modelos com defasagem ou médias móveis das variáveis meteorológicas e dos poluentes, como será visto mais adiante.

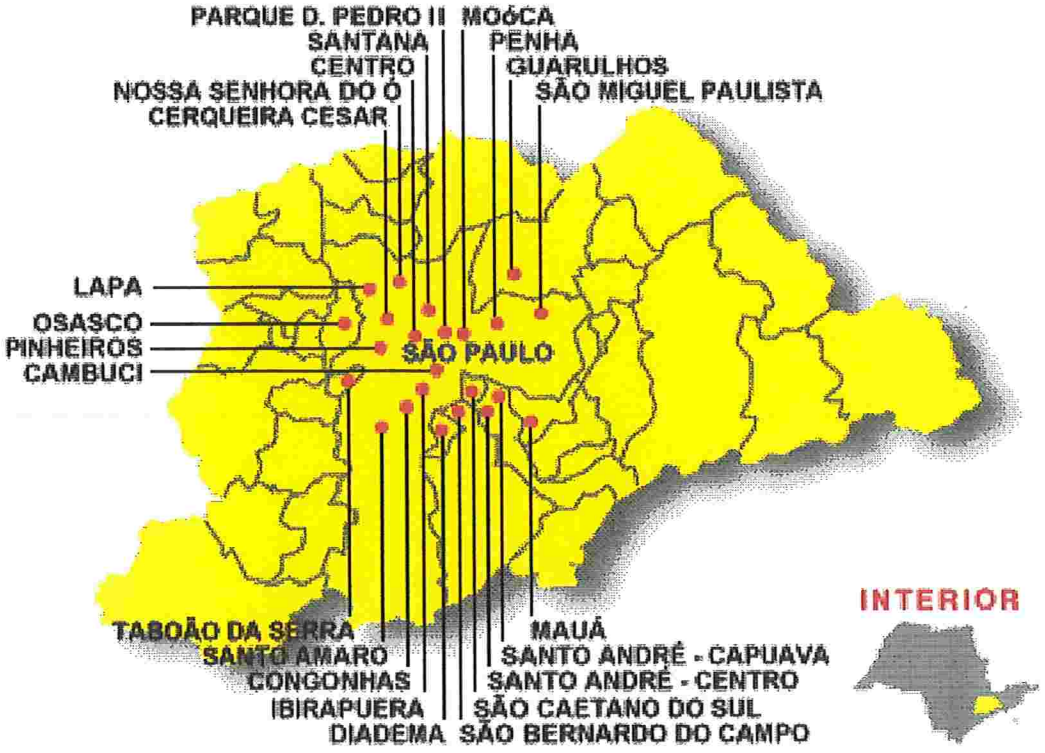
A análise feita por Pereira et al. (1998) foi repetida neste trabalho e o mesmo banco de dados foi analisado adotando-se um MAG. Os modelos ajustados foram comparados quanto aos seus desvios, comportamento dos resíduos e termos que mostraram associações significantes com a variável resposta.

Riscos relativos a diferentes concentrações dos poluentes foram estimados na análise via regressão de Poisson. Quando a técnica MAG é adotada, Schwartz (1994) mostrou que curvas dos riscos relativos em função das concentrações dos poluentes podem ser estimadas. Esses gráficos são úteis para verificar a existência de concentrações limites dos poluentes, acima das quais seus efeitos são nocivos à saúde da população. A partir deles, podem-se validar os limites não tóxicos à vida humana (denominados padrões primários) estabelecidos pelo Conselho Nacional do Meio Ambiente (CONAMA), listados na Tabela 4.1. A construção dessas curvas de risco também é ilustrada neste trabalho.

Na análise via MAG foi também estimada a variação relativa esperada na resposta correspondente a uma variação na concentração do poluente igual a um intervalo interquartil, que também é utilizada como uma medida do efeito da poluição sobre a saúde.

Tabela 4.1: Padrões primários referentes às concentrações dos poluentes, estabelecidos pelo CONAMA.

Poluente	Padrão primário
NO_2	média anual = 320 $\mu g/m^3$
O_3	média anual = 160 $\mu g/m^3$
SO_2	média anual = 365 $\mu g/m^3$
PM_{10}	máxima diária = 150 $\mu g/m^3$
CO	média anual = 9 ppm



(Fonte: http://www.cetesb.gov.br/, 2001)

Figura 4.1: Localização das 23 estações fixas de monitoramento do ar operadas pela CETESB na região metropolitana de São Paulo.

4.2. Análise descritiva

No Apêndice A são apresentadas tabelas e gráficos construídos com o objetivo de resumir os dados diários das variáveis de interesse, ou seja: mortalidade fetal tardia (NATMOR), temperatura mínima diária (TEMP), umidade relativa do ar medida às 12 horas (UMID) e poluentes SO_2, CO, PM_{10}, O_3 e NO_2 nos anos de 1991 e 1992.

A Figura A.1 apresenta gráficos do tipo *box-plot*, e a Figura A.2 apresenta séries cronológicas para NATMOR, TEMP, UMID e poluentes.

Na Figura A.1, notam-se valores discrepantes altos da NATMOR, maiores ou iguais a 15 no ano de 1991, e maiores ou iguais a 17 no ano de 1992. Na Tabela A.1, observa-se que o número diário médio de natimortos foi de 8,64 em 1991 (desvio padrão igual a 3,11) e 8,08 em 1992 (desvio padrão igual a 3,04) em 1992. Observando a representação gráfica da série de natimortalidade (Figura A.2), nota-se que valores discrepantes maiores do que 17 são observados principalmente na primavera e no verão.

Os valores médios de TEMP e UMID são bastante próximos nos dois anos. Os *box-plots* e gráfico da série de TEMP em 1991 e 1992 apontam a ocorrência de temperaturas bastante baixas nos invernos dos dois anos.

As concentrações dos poluentes SO_2, CO, PM_{10} e O_3 apresentaram maior valor médio em 1991, e o NO_2 apresentou comportamento oposto. Os *box-plots* apontam concentrações aberrantes altas dos poluentes nos dois anos, muitas delas superiores aos padrões primários. Nas séries cronológicas representadas na Figura A.2 notam-se picos de altas concentrações de SO_2, PM_{10} e NO_2, principalmente nos meses frios de 1991 e 1992. Vários picos com concentrações altas de CO foram observados desde aproximadamente o início do outono ao final do inverno de 1991. Em relação ao O_3,

picos de concentrações ocorreram durante todo o ano, exceto no período de abril a julho nos dois anos de estudo.

Da Tabela A.2 conclui-se que NO_2 e CO possuem correlação linear positiva com a mortalidade fetal tardia ($p < 0,05$ e $p < 0,01$, respectivamente). Os demais poluentes não apresentam correlação linear significante com a natimortalidade ($p > 0,05$).

4.3. Análise via MLG

Considerando que os valores da variável NATMOR resultam da contagem de eventos raros, Pereira et al. (1998) adotaram um modelo de regressão de Poisson com função de ligação logarítmica para o estudo da associação entre mortalidade fetal tardia e concentrações dos poluentes, controlando os efeitos de sazonalidade e das variáveis climáticas.

A estratégia de análise consistiu em, inicialmente, modelar a sazonalidade e as variáveis climáticas. O objetivo desta estratégia é construir um modelo básico incluindo somente as variáveis de controle sazonal e meteorológico, que explique ao máximo a variabilidade da resposta antes de adicionar a variável relativa às concentrações do poluente no modelo.

Foram consideradas como variáveis independentes:

• 23 variáveis indicadoras (*dummies*) dos meses no período do estudo para o controle da sazonalidade de longa duração (variáveis FEV91, MAR91,...,DEZ92, sendo o mês de janeiro de 1991 a categoria de referência para os meses);

• 6 variáveis indicadoras dos dias da semana para o controle da sazonalidade de curta duração (variáveis SEG,...,SÁB, sendo o domingo a categoria de referência para os dias da semana);

• médias móveis de dois dias da temperatura mínima e umidade relativa do ar (M2TEMP e M2UMID, respectivamente) definidas como a média dos valores dessas variáveis no dia e dia anterior ao registro da morte. Médias móveis de 2 até 7 dias foram pesquisadas e as de dois dias foram selecionadas para compor o modelo por apresentarem maior associação com a variável resposta;

• um controle adicional para TEMP e UMID foi feito categorizando estas variáveis em quatro classes delimitadas pelos seus quartis. Obtiveram-se as classes: "menor que 13,50°C", "de 13,50°C a 15,60°C", "de 15,60°C a 17,55°C" e "maior que 15,55°C" para temperatura mínima, e "menor que 59,0%", "de 59,0% a 67,5%", "de 67,5% a 76,5%" e "maior que 76,5%" para umidade relativa. Foram então adicionadas ao modelo variáveis indicadoras dessas categorias (variáveis TEMP1,...,TEMP3 para temperatura e UMID1,...,UMID3 para umidade). As classes "menor que 13,50 °C" e "menor que 59,0%" foram as categorias de referência para temperatura e umidade, respectivamente. O procedimento de categorizar as variáveis TEMP e UMID evita que uma relação paramétrica entre essas variáveis e a resposta seja introduzida no modelo.

Assim, o modelo adotado inicialmente apresentou a forma:

$$\log[E(NATMOR)] = \alpha + \beta_1 FEV91 + \ldots + \beta_{23} DEZ92 + \beta_{24} SEG + \ldots + \beta_{29} SÁB + \beta_{30} M2TEMP + \beta_{31} M2UMID + \beta_{32} TEMP1 + \ldots + \beta_{34} TEMP3 + \beta_{35} UMID1 + \ldots + \beta_{37} UMID3. \quad (4.1)$$

Schwartz (1994) salienta que a importância de verificar se o controle das

variáveis climáticas foi feito de forma adequada (isto é, se a relação entre a resposta e as variáveis climáticas foi estabelecida corretamente no modelo) deve-se ao fato de que um controle inadequado dessas variáveis pode confundir a associação entre a resposta e o poluente, quando este for incorporado ao modelo. Essa verificação pode ser feita através de um diagrama de dispersão suavizado dos resíduos (componentes do desvio) resultantes do ajuste de um modelo básico, como o modelo (4.1), e cada uma das variáveis climáticas. Tais gráficos são apresentados no Apêndice B, Figura B.1 para TEMP e Figura B.2 para UMID. Em ambos, as curvas suavizadas não apresentam tendência, indicando controle adequado das variáveis;

• finalmente, os poluentes foram considerados adotando-se, para cada um deles, o seguinte modelo:

$$\log[E(\text{NATMOR})] = \alpha + \beta_1 \text{FEV91} + \ldots + \beta_{23} \text{DEZ92} +$$

$$\beta_{24} \text{SEG} + \ldots + \beta_{29} \text{SÁB} + \beta_{30} \text{M2TEMP} + \beta_{31} \text{M2UMID} +$$

$$\beta_{32} \text{TEMP1} + \ldots + \beta_{34} \text{TEMP3} + \beta_{35} \text{UMID1} + \ldots + \beta_{37} \text{UMID3} + \beta_{38} \text{Poluente}.$$

Foram estudadas as associações entre NATMOR e concentrações dos poluentes em diferentes defasagens e entre NATMOR e médias móveis das concentrações dos poluentes nos períodos de 2 até 14 dias precedendo o registro da morte.

De todas as associações estudadas, a variável que apresentou maior associação com a resposta, avaliada pelo nível descritivo do teste de nulidade de seus coeficientes, foi a média móvel de cinco dias para NO_2 (M5NO2), com $p = 0{,}001$. Não foram observadas associações significantes entre a variável resposta e as concentrações de CO, SO_2, PM_{10} e O_3 ($p \geq 0{,}08$). O modelo ajustado com o poluente M5NO2 é apresentado na Tabela B.1, Apêndice B.

Assim, apenas o M5NO2 foi considerado na continuidade da análise.

O prosseguimento da análise deve-se ao fato do tipo de relação existente entre a resposta e o poluente poder não ser estritamente linear. Um polinômio envolvendo além do termo linear outros de maior ordem de M5NO2 pode ser, por exemplo, mais adequado para descrever a relação entre a resposta e esse poluente. Polinômios nos quais o termo linear não esteja presente descrevem um tipo de relação causa-efeito que não tem plausibilidade biológica, e esta foi a razão dos outros poluentes terem sido excluídos da análise.

Uma maneira de se evitar a modelagem da associação é dividir as concentrações do poluente em categorias, e introduzir no modelo uma variável categorizada no lugar da variável contínua. Esse procedimento foi adotado por Pereira et al. (1998) na continuidade da análise. Para isso, as concentrações de M5NO2 foram divididas em cinco intervalos de classes delimitados pelos seus quintis. Quatro variáveis indicadoras (M5NO2a,...,M5NO2d) foram criadas para identificar as cinco categorias, considerando a classe composta pelas menores concentrações como referência. O modelo adotado passou a ser:

$$\log[E(NATMOR)] = \alpha + \beta_1 FEV91 + \ldots + \beta_{23} DEZ92 + \beta_{24} SEG + \ldots + \beta_{29} SÁB + \beta_{30} M2TEMP + \beta_{31} M2UMID + \beta_{32} TEMP1 + \ldots + \beta_{34} TEMP3 + \beta_{35} UMID1 + \ldots + \beta_{37} UMID3 + \beta_{38} M5NO2a + \ldots + \beta_{41} M5NO2d. \quad (4.2)$$

O modelo ajustado é apresentado na Tabela B.2, onde se observa efeito significante do M5NO2 somente nas duas categorias formadas pelas concentrações mais altas do poluente, sugerindo um efeito não linear do NO_2 na média da NATMOR. Efeitos significantes também foram observados para as variáveis de controle da sazonalidade de longa e curta duração ($p < 0{,}001$). Não se observa efeito significante para as variáveis de controle de clima.

O ajuste do modelo (4.2) forneceu um desvio residual igual a 726,334 com 684 graus de liberdade, sugerindo um bom ajuste do modelo. A análise de resíduos também indicou um bom ajuste: o gráfico de envelope para a distribuição de Poisson (Figura B.3) mostrou pontos distribuídos dentro das bandas de confiança (ver Neter et al., 1996 para detalhes) e a avaliação da autocorrelação dos resíduos foi considerada adequada (Figura B.4). Além disso, o diagrama de dispersão suavizado dos valores observados e valores ajustados (Figura B.5) não apontou tendências.

Alternativamente, uma análise de resíduos poderia sugerir o tipo de relação existente entre M5NO2 e NATMOR. Na Figura B.6, os resíduos resultantes do modelo (4.1) foram suavizados em M5NO2 com o objetivo de investigar a relação desse poluente na resposta. A curva suavizada apresentada nessa figura mostra, inicialmente, um platô seguido de uma curva crescente a partir de uma concentração em torno de 300 $\mu g/m^3$, sugerindo uma possível relação não linear desse poluente. As curvas suavizadas considerando os demais poluentes (ver Figuras B.7 a B.10) não sugerem nenhum tipo de associação.

Um parâmetro de interesse em estudos epidemiológicos é o risco relativo. O risco de mortalidade fetal tardia em uma dada categoria (j) de concentração de um poluente relativa à categoria composta pelas concentrações mais baixas (categoria de referência), mantidos constantes os valores de todas as outras variáveis preditoras no modelo, é definido como:

$$(4.3) \qquad RR_j = \frac{E(\text{NATMOR} \mid \text{categoria j de concentração})}{E(\text{NATMOR} \mid \text{categoria de referência})} = \exp(\beta_j),$$

onde β_j é o coeficiente da variável indicadora da categoria de concentração j de interesse.

Um intervalo de confiança para RR com coeficiente de confiança $1-\alpha$ é dado por

$$\exp\left(\hat{\beta}_j \pm z_{\alpha/2}\hat{ep}(\hat{\beta}_j)\right),$$

onde $\hat{ep}(\hat{\beta}_j)$ é o valor estimado do erro padrão de $\hat{\beta}_j$ e $z_{\alpha/2}$ é o percentil de ordem $(1-\alpha/2)$ da distribuição Normal padrão.

A Figura 4.2 ilustra as estimativas por ponto e intervalo, com coeficiente de confiança de 0,95, do risco relativo de mortalidade fetal tardia em diferentes categorias de concentrações do M5NO2. Por esta figura, observa-se que o risco relativo do NO_2 apresenta comportamento dose-dependente, isto é, aumenta quando a concentração deste poluente cresce.

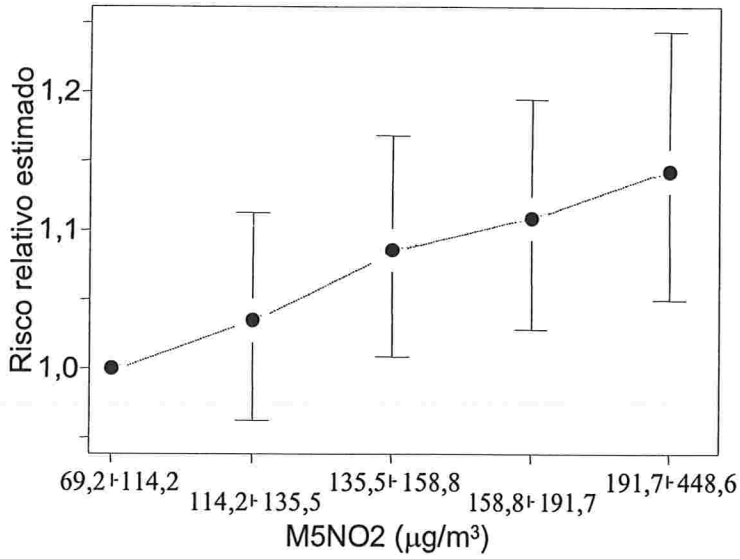


Figura 4.2: Estimativa por ponto e intervalo de confiança de 0,95 para o risco relativo obtido no ajuste do modelo (4.2).

Os resultados obtidos na análise indicam que o NO_2 é o poluente que mais se associou à mortalidade fetal tardia, apresentando um comportamento dose-dependente.

4.4. Análise via MAG

No ajuste do modelo (4.1) evitou-se a modelagem do efeito do tempo de observação e das variáveis climáticas criando-se categorias dessas variáveis. No modelo (4.2) este procedimento também foi adotado para o poluente. O motivo para se trabalhar com variáveis categorizadas no lugar de variáveis originalmente contínuas é a falta de conhecimento de uma função paramétrica adequada que relacione cada uma das variáveis preditoras com a resposta. Por exemplo, espera-se que a natimortalidade apresente comportamento sazonal, assim como a mortalidade por causas respiratórias em idosos (Miraglia et al., 1997) e em idosos e crianças (Conceição, 1999), e portanto, esperada-se uma relação não linear entre a variável resposta e a variável de controle no período de observação (dia ou mês de observação). A forma paramétrica da relação funcional existente entre essas variáveis é difícil de ser estabelecida. Uma alternativa seria incluir no modelo um termo não paramétrico para descrevê-la, isto é, adotar um *modelo aditivo generalizado*. O mesmo procedimento pode ser adotado em relação às variáveis TEMP, UMID e cada um dos poluentes.

A seguir é apresentada uma estratégia de análise quando um MAG é adotado para relacionar NATMOR com os poluentes, controlando-se as variáveis TEMP, UMID, Dias da semana e Dias de observação. A idéia aqui também é construir, inicialmente, um modelo básico envolvendo apenas as variáveis de controle sazonal e meteorológico. Essa estratégia foi inspirada na análise de outros estudos na área de epidemiologia ambiental, nos quais um MAG foi adotado para descrever a relação entre a morbi-mortalidade e poluentes (ver, por exemplo, Conceição et al., 2000 ou Schwartz, 1999). O método de suavização utilizado na estimação dos termos não paramétricos foi o *loess*.

O primeiro passo da análise foi modelar a sazonalidade de longa duração, isto é, a variação da natimortalidade em função dos dias de observação (726 dias), sendo, para isto, considerado o modelo:

$$\log[E(NATMOR)] = \alpha + f_1(DIAS). \quad (4.4)$$

Na estimação de f_1, o parâmetro de suavização foi fixado inicialmente em $\lambda_1 = 0{,}25$. Este valor foi determinado visando obter vizinhanças com 181 observações em cada ajuste local (isto é, cerca de 6 meses). Vale lembrar que observações feitas em dias consecutivos podem apresentar resíduos autocorrelacionados. Segundo Schwartz (1999), cada admissão hospitalar ou morte é um evento independente, e a existência de autocorrelação dos resíduos indica que uma covariável dependente do tempo pode ter sido omitida da análise, e a variabilidade dessa variável pode ter confundido o efeito do poluente. Se a autocorrelação é removida, a variação remanescente da covariável omitida não tem comportamento sistemático no tempo e o confundimento é menos provável. Por outro lado, parâmetros de suavização com valores muito pequenos podem induzir autocorrelação. Assim, o parâmetro de suavização deve ser fixado observando também o padrão de autocorrelação dos resíduos. Uma possível medida resumo da estrutura de autocorrelação dos resíduos é a soma dos quadrados dessas autocorrelações (sac). Para o ajuste do modelo (4.4) obteve-se sac = 0,06. Com o objetivo de diminuir esse valor, λ_1 foi aumentado procurando fazer com que o número de observações dentro da vizinhança não ultrapassasse um ano. A Figura C.1 apresenta o gráfico de autocorrelações dos resíduos resultantes do ajuste do modelo (4.4) considerando uma vizinhança contendo 363 dias de observação (isto é, $\lambda_1 = 0{,}5$). Neste caso, sac = 0,05.

Em seguida, 6 variáveis indicadoras dos dias da semana para o controle da sazonalidade de curta duração e termos não paramétricos para as variáveis climáticas

M2TEMP e M2UMID foram adicionados ao modelo. Assim, o modelo adotado nesta etapa foi:

$$\log[E(\text{NATMOR})] = \alpha + f_1(\text{DIAS}) + \beta_1 \text{SEG} + \ldots +$$

$$\beta_6 \text{SÁB} + f_2(\text{M2TEMP}) + f_3(\text{M2UMID}).$$

A escolha dos valores dos parâmetros de suavização para os termos f_2(M2TEMP) e f_3(M2UMID), respectivamente λ_2 e λ_3, foi feita a partir de valores pré-fixados de 0,65, 0,8 e 0,95. Segundo Schwartz (1994), embora a relação entre mortalidade e variáveis climáticas possa ser não linear, parâmetros de suavização muito pequenos para estas variáveis não têm plausibilidade biológica. Foram ajustados modelos considerando todas as possíveis combinações dos valores de λ_2 e λ_3 descritos acima e também modelos com termos lineares para M2TEMP e M2UMID. Entre eles, o modelo

(4.5)
$$\hat{\log}(\text{NATMOR}) = \hat{\alpha} + \hat{f}_1(\text{DIAS}) + \hat{\beta}_1 \text{SEG} + \ldots +$$

$$\hat{\beta}_6 \text{SÁB} + \hat{\beta}_7(\text{M2TEMP}) + \hat{\beta}_8(\text{M2UMID})$$

foi selecionado por ter sido aquele que apresentou o menor valor da estatística AIC, dada em (3.22).

Finalmente, cada poluente foi adicionado ao modelo por intermédio de um termo não paramétrico. O parâmetro de suavização de cada um deles também foi selecionado de forma que o modelo final apresentasse valor mínimo para a estatística AIC.

Para o poluente NO_2, o modelo final ajustado foi

$$(4.6) \qquad \text{lôg(NATMOR)} = \hat{\alpha} + \hat{f}_1(\text{DIAS}) + \hat{\beta}_1 \text{SEG} + \ldots + \hat{\beta}_6 \text{SÁB} +$$

$$\hat{\beta}_7(\text{M2TEMP}) + \hat{\beta}_8(\text{M2UMID}) + \hat{f}_4(\text{M5NO2}),$$

com $\lambda_4 = 0{,}8$ sendo o valor selecionado do parâmetro de suavização para $\hat{f}_4$. A escolha deste modelo foi feita considerando ajustes com valores de λ_4 pré-fixados em 0,6, 0,7 e 0,8, e também do ajuste de um modelo com termo linear para M5NO2.

No Apêndice C são apresentandos os resultados do ajuste do modelo com NO_2 e as representações gráficas dos efeitos estimados segundo este ajuste. Os poluentes PM_{10}, SO_2, O_3 e CO não são apresentados pois, assim como na análise via MLG, seus efeitos não foram significantes, mesmo com diferentes parâmetros de suavização.

A Tabela C.1 apresenta os resultados obtidos sob o modelo (4.6). De acordo com essa tabela, conclui-se que existem efeitos linear ($p = 0{,}033$) e não linear ($p = 0{,}043$) da variável M5NO2 na natimortalidade. Em relação às variáveis de controle da sazonalidade, foram encontrados efeitos linear ($p < 0{,}001$) e não linear ($p = 0{,}004$) da variável DIAS, e efeito significante para as variáveis indicadoras dos dias da semana ($p < 0{,}001$). Não foram encontrados efeitos significantes para M2TEMP e M2UMID.

O modelo (4.6) forneceu desvio residual de 747,727 com 710,896 graus de liberdade, sugerindo um bom ajuste. Comparado com o modelo (4.2) ajustado via MLG, o ajuste via MAG apresentou uma economia de quase 27 graus de liberdade, embora, neste tipo de estudo, onde o número de observações é grande, tal economia não representa uma vantagem da técnica MAG.

É possível representar graficamente o efeito estimado de cada variável preditora do modelo na variável resposta, mantidas constantes as outras variáveis. A curva representada na Figura C.2 é a contribuição estimada da M5NO2 no preditor

aditivo ajustado, após controlar por variáveis confundidoras. Esta figura indica uma relação não linear entre a resposta e o poluente. As linhas pontilhadas representam os valores ajustados ± 2 vezes os respectivos erros padrão estimados.

O maior interesse neste tipo de análise recai na estimação de medidas que avaliem o impacto do poluente na NATMOR. Uma medida com tal finalidade é o risco relativo. Neste sentido, a análise via MAG apresenta a vantagem de fornecer melhor interpretação dos resultados, pois permite estimar a curva do risco relativo sem a necessidade de categorizar os poluentes, o que acarreta perda de informação.

O risco de mortalidade fetal tardia em uma dada concentração M5NO2 do poluente, denotada por P_j, em relação à menor concentração observada desse poluente, denotada por P_0, mantidos constantes os valores de todas as outras variáveis preditoras no modelo, é dado por:

$$r_j = \frac{E(NATMOR \mid P_j)}{E(NATMOR \mid P_0)} = \exp[\, f_4(P_j) - f_4(P_0)\,]. \tag{4.7}$$

Denotando por $\hat{f}_4(P_j)$ o valor ajustado de f_4 em P_j e por $\hat{f}_4(P_0)$ o valor ajustado de f_4 para uma concentração de referência, um estimador para (4.7) é dado por

$$\hat{r}_j = \exp[\, \hat{f}_4(P_j) - \hat{f}_4(P_0)\,] = \exp(\hat{d}_j). \tag{4.8}$$

Considerando todas as concentrações observadas do poluente, pode-se calcular o valor de $\hat{r}_j$ a essas concentrações, e então construir o gráfico de $\hat{r}_j$ em P_j, obtendo-se a curva estimada do risco relativo.

Bandas de confiança aproximadas para essas curvas podem ser construídas representando no gráfico os pontos

$$\hat{r}_j \pm 2\, \hat{ep}(\hat{r}_j),$$

sendo $\hat{ep}(\hat{r}_j)$ o erro padrão estimado de $\hat{r}_j$.

Expressões para o erro padrão de $\hat{r}_j$ ($ep(\hat{r}_j)$) e $\hat{ep}(\hat{r}_j)$ não são fornecidas na literatura. Uma sugestão para obter $ep(\hat{r}_j)$ seria expandir $\exp(\hat{d}_j)$ em série de Taylor de em torno de $d_j = f_4(P_j) - f_4(P_0)$ até primeira ordem, obtendo:

$$\exp(\hat{d}_j) \approx \exp(d_j) + \exp(d_j)\,(\hat{d}_j - d_j).$$

Então,

$$\mathrm{Var}(\hat{r}_j) = \mathrm{Var}(\exp(\hat{d}_j)) \approx \exp^2(d_j)\,\mathrm{Var}(\hat{d}_j)$$

$$= \exp^2(d_j)\left(\mathrm{Var}[\hat{f}_4(P_j)] + \mathrm{Var}[\hat{f}_4(P_0)] - 2\mathrm{Cov}[\hat{f}_4(P_j), \hat{f}_4(P_0)]\right). \quad (4.9)$$

Um estimador da $\mathrm{Var}(\hat{r}_j)$ pode ser obtido substituindo-se d na expressão (4.9) por $\hat{d}_j = \hat{f}_4(P_j) - \hat{f}_4(P_0)$, assumindo consistência de $\hat{f}$. A estimação das variâncias e covariâncias nessa expressão requer, entretanto, o conhecimento de $\mathbf{R}_4$, que é a matriz que produz $\hat{\mathbf{f}}_4$ a partir de **z**.

A Figura 4.3 apresenta estimativas pontuais do risco relativo de mortalidade fetal tardia em diferentes concentrações da M5NO2. Por esta figura, observa-se um aumento acentuado do risco relativo estimado a partir da concentração mínima até concentrações em torno de 190 $\mu g/m^3$, seguindo-se um platô e um novo aumento constante a partir de uma concentração de aproximadamente 270 $\mu g/m^3$. Não se observa a presença de uma concentração limiar abaixo da qual não há relação entre NATMOR e M5NO2, indicando concentrações seguras.

A ocorrência do "platô" na curva de risco estimada foi observado em outros

estudos e uma explicação para esse comportamento está ainda sendo discutida por pesquisadores da área. A idéia mais aceita entre eles é que esse fenômeno ocorre devido à existência de grupos de pessoas dentro da população com diferentes suscetibilidades ao poluente.

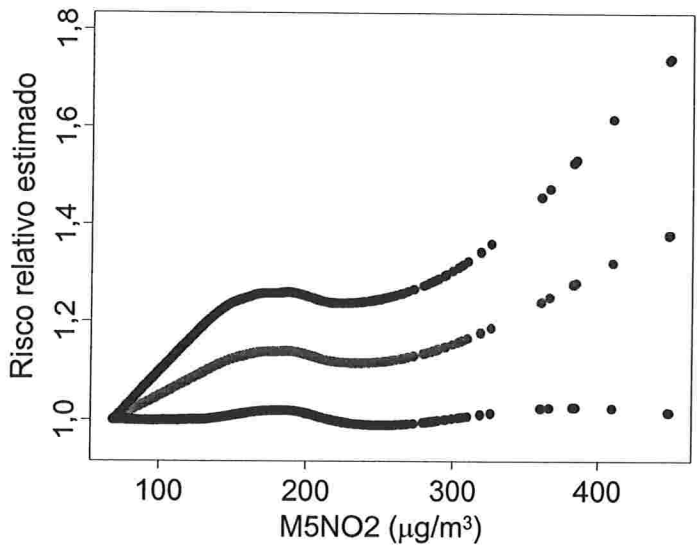


Figura 4.3: Estimativa por ponto para o risco relativo obtido do modelo (4.6) e bandas de confiança pontuais aproximadas.

Um outro parâmetro de avaliação do impacto do poluente na variável resposta em estudo (Schwartz, 1999) é o acréscimo relativo no valor esperado da resposta devido à variação de um intervalo interquartil na concentração do poluente, isto é,

$$\Delta = \left(\frac{\exp(f_4(P_{Q3}))}{\exp(f_4(P_{Q1}))} - 1 \right) 100\% = (\exp[f_4(P_{Q3}) - f_4(P_{Q1})] - 1) \, 100\%, \tag{4.10}$$

onde P_{Q1} e P_{Q3} são, respectivamente, os valores observados do 1º e 3º quartil do poluente.

No caso do modelo (4.6), uma estimativa para Δ pode ser obtida substituindo os valores $f_4(P_{Q3})$ e $f_4(P_{Q1})$, na expressão (4.10), pelos correspondentes valores estimados obtidos no ajuste desse modelo, ou seja:

(4.11) $$\hat{\Delta} = (\exp[\hat{f}_4(M5NO2_{Q3}) - \hat{f}_4(M5NO2_{Q1})] - 1) \, 100\% =$$

$$= (\exp[\hat{f}_4(180,7) - \hat{f}_4(119,9)] - 1) \, 100\% = 5,41 \, \%.$$

Como não se tem conhecimento de expressões para o erro padrão e para um estimador desse parâmetro, propõe-se aqui, que essas expressões sejam obtidas através dos mesmos argumentos utilizados para o risco relativo. Então,

$$\mathrm{Var}(\hat{\Delta}) = \mathrm{Var}(\exp[\hat{f}_4(M5NO2_{Q3}) - \hat{f}_4(M5NO2_{Q1})])$$

pode ser obtida a partir de (4.9), substituindo-se P_j por $M5NO2_{Q3}$ e P_0 por $M5NO2_{Q1}$. No exemplo, tem-se $\hat{\mathrm{Var}}(\hat{\Delta}) = 0,07 \, \%$ e $\hat{ep}(\hat{\Delta}) = 2,57 \, \%$.

Vale notar que $\hat{\Delta}$ pode assumir valor negativo. Por esta razão, a rigor, $\hat{\Delta}$ é melhor definido como sendo o máximo entre zero e o valor da expressão (4.11).

Para todos os modelos ajustados foram feitas análises de resíduos. No Apêndice C são mostrados apenas alguns gráficos relativos aos MAG, mas de uma forma geral, todos os modelos apresentaram um bom ajuste.

As Figuras C.3 a C.5 apresentam diagramas de dispersão suavizados dos resíduos do modelo (4.6) em função, respectivamente, dos dias de observação, M2TEMP e M2UMID. Nestas figuras, as curvas suavizadas não mostram tendência dos dados, indicando que o efeito de longa duração (Figura C.3) e o de clima (Figuras C.4 e C.5) estão bem controlados.

De forma semelhante, a Figura C.6 apresenta o diagrama de dispersão

suavizado dos resíduos do modelo (4.6) em função da M5NO2. De acordo com esta figura, pode-se dizer que o efeito do poluente não foi totalmente explicado pelo modelo, uma vez que se observa leve tendência de aumento da curva na presença de concentrações muito altas desse poluente. Uma solução para esse problema seria excluir da análise os dias que apresentam concentrações muito altas. Teria então que ser determinado um critério de corte adequado à realidade de São Paulo. Outra solução seria utilizar métodos não sensíveis a valores discrepantes da concentração do poluente (pontos de alavanca), mas esses métodos não são ainda descritos na literatura.

4.5. Conclusões

As análises via MLG e MAG mostraram as mesmas conclusões sobre o efeito dos poluentes na natimortalidade, indicando que o NO_2 é o poluente que mais se associou à mortalidade fetal tardia. Este resultado está de acordo com estudos prévios que mostram existir associação entre mortalidade por causas respiratórias e NO_X em crianças na cidade de São Paulo (Saldiva et al., 1994).

O fato das duas técnicas produzirem as mesmas conclusões é justificado pois os dois modelos, por elas adotados, estão controlando adequadamente os efeitos lineares e não lineares das mesmas variáveis preditoras. Porém, a categorização de variáveis contínuas, utilizada no MLG, acarreta perda de informação. A curva estimada do risco relativo obtida pelo MAG, apresentada na Figura 4.3, é mais informativa do que a obtida por MLG.

Capítulo 5

Considerações finais

Neste capítulo são feitas algumas considerações sobre a parte computacional dos MAG's e apresentadas algumas sugestões para a continuidade do trabalho.

Uma parte da análise apresentada no Capítulo 4 foi feita utilizando-se comandos do aplicativo S-Plus versão 4.5 (MathSoft, 1998), descritos em Chambers e Hastie (1993): *glm*(), *gam*(), *summary*(), *summary.glm*(), *step.gam*(), *acf*() e *loess.smooth*(). Uma rotina para o cálculo da estatística AIC (3.22) foi cedida por pesquisadores do Laboratório de Poluição Atmosférica Ambiental da FM/USP. O S-Plus não fornece, entretanto, as matrizes $\mathbf{R}_j$ necessárias ao cálculo das estimativas das matrizes de variâncias e covariâncias dos $\hat{\mathbf{f}}_j$, cujos elementos são necessários para a obtenção das estimativas dos erros padrão dos estimadores das medidas do impacto da poluição sobre a saúde (r_j e Δ). Para a obtenção dessas matrizes foram construídas rotinas de forma a extrair os elementos de $\mathbf{R}_j$ a partir de objetos gerados pelo S-Plus, de acordo com o que foi sugerido por Hastie e Tibshirani (1987). Nota-se, porém, que o custo computacional para a obtenção dessas matrizes é elevado. Uma listagem completa dos comandos utilizados no Capítulo 4 será incorporada na versão final deste trabalho.

Como continuidade do trabalho, pode ser sugerido o desenvolvimento de métodos *bootstrap* para a construção de bandas de confiança para o risco relativo. Técnicas robustas de estimação dos MAG's também merecem ser estudadas. Hastie e

Tibshirani (1990) extenderam a técnica de estimação M para a classe dos MAG's. Esse procedimento de estimação robusta está implementado no S-Plus. Entretanto, não são conhecidos métodos de estimação que produzam estimadores pouco sensíveis a valores extremos da variável preditora. Uma possível solução seria utilizar, dentro do *loess*, um critério de ponderação que atribua pesos pequenos às observações nas quais os valores observados de X são aberrantes, ou ainda, adotar como critério de ajuste das regressões locais o método de Mínima Mediana do Quadrado dos Resíduos (Rousseeuw, 1984).

Pode-se também sugerir um estudo das propriedades assintóticas dos estimadores em MAG's semi-paramétricos, seguindo resultados obtidos recentemente por Opsomer e Ruppert (1999) e Opsomer (2000) para modelos aditivos.

Apêndice A

Medidas descritivas

Tabela A.1: Média (± desvio padrão) para os dados de mortalidade, concentrações dos poluentes e variáveis climáticas.

variáveis	n	1991	1992	total
NATMOR	730	8,64 ± 3,11	8,08 ± 3,04	8,36 ± 3,08
Variáveis Climáticas				
UMID (%)	730	67,35 ± 13,34	69,42 ± 13,82	68,38 ± 13,61
TEMP (°C)	730	15,22 ± 2,87	15,53 ± 2,82	15,37 ± 2,85
Poluentes				
NO_2 ($\mu g/m^3$)	690	154,75 ± 78,43	159,05 ± 84,45	156,82 ± 81,36
CO (ppm)	726	6,59 ± 1,78	4,88 ± 1,58	5,73 ±1,89
SO_2 ($\mu g/m^3$)	730	19,46 ± 9,55	18,33 ± 7,34	18,90 ± 8,53
PM_{10} ($\mu g/m^3$)	730	69,37 ± 31,14	60,70 ± 21,97	65,04 ± 27,28
O_3 ($\mu g/m^3$)	727	74,01 ± 51,90	60,97 ± 35,80	67,50 ± 45,04

Tabela A.2: Coeficientes de Correlação de Pearson (r) para as principais variáveis do estudo.

		NATMOR	UMID	TEMP	NO2	SO2	CO	PM10
UMID	r	0,026						
	n	*730*						
TEMP	r	0,101**	–0,065					
	n	*730*	*730*					
NO2	r	0,078*	–0,248**	0,106**				
	n	*690*	*690*	*690*				
SO2	r	0,051	–0,274**	–0,141**	0,409**			
	n	*730*	*730*	*730*	*690*			
CO	r	0,151**	–0,042	0,113**	0,245**	0,337**		
	n	*726*	*726*	*726*	*686*	*726*		
PM10	r	0,019	–0,335**	–0,095*	0,451**	0,741**	0,415**	
	n	*730*	*730*	*730*	*690*	*730*	*726*	
O3	r	0,001	–0,332**	0,215**	0,166**	0,152**	0,038	0,248**
	n	*727*	*727*	*727*	*687*	*727*	*723*	*727*

*p < 0,05, **p < 0,01.

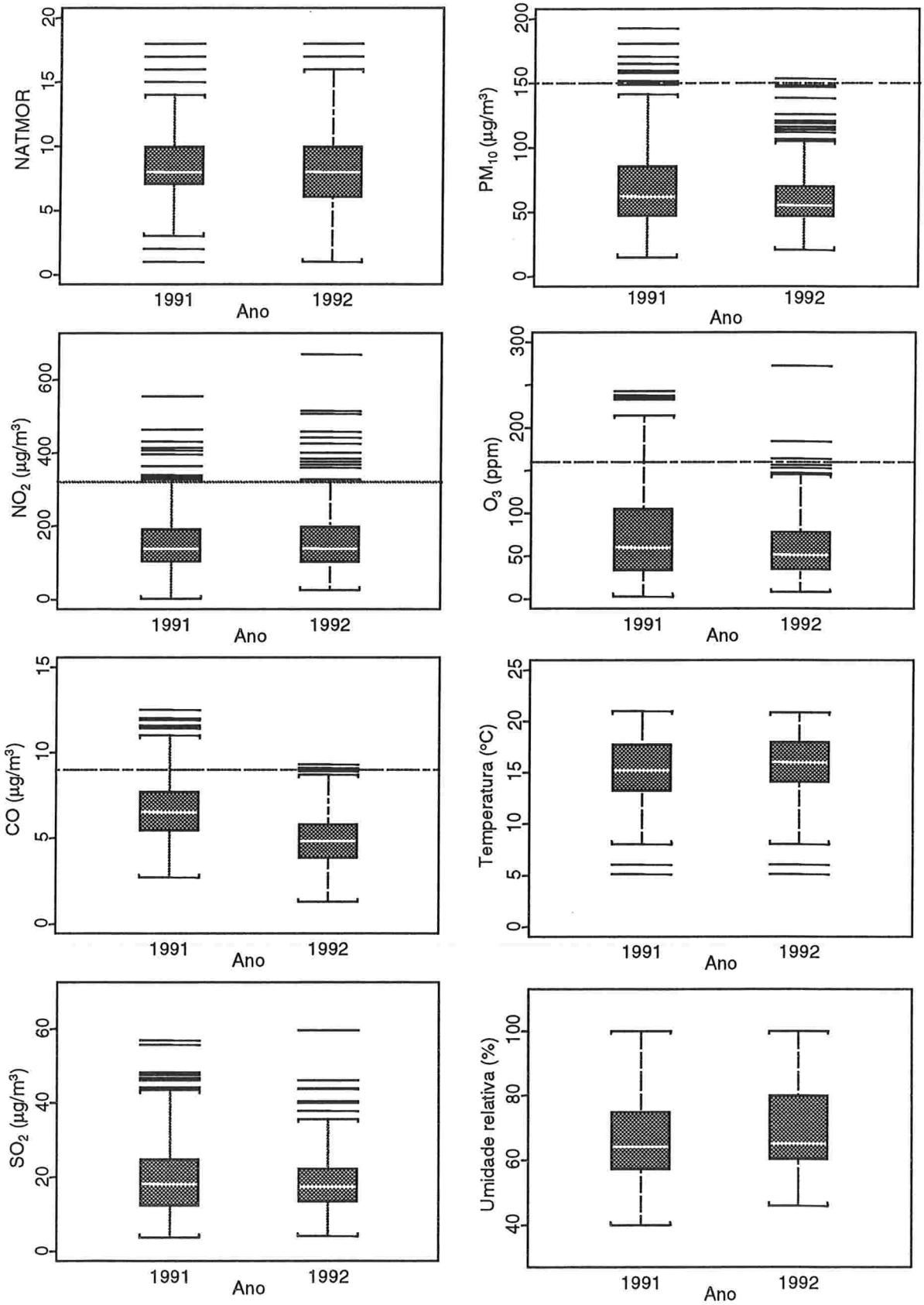


Figura A.1: Gráficos do tipo *box-plot* para a NATMOR, TEMP, UMID e poluentes nos anos de 1991 e 1992.

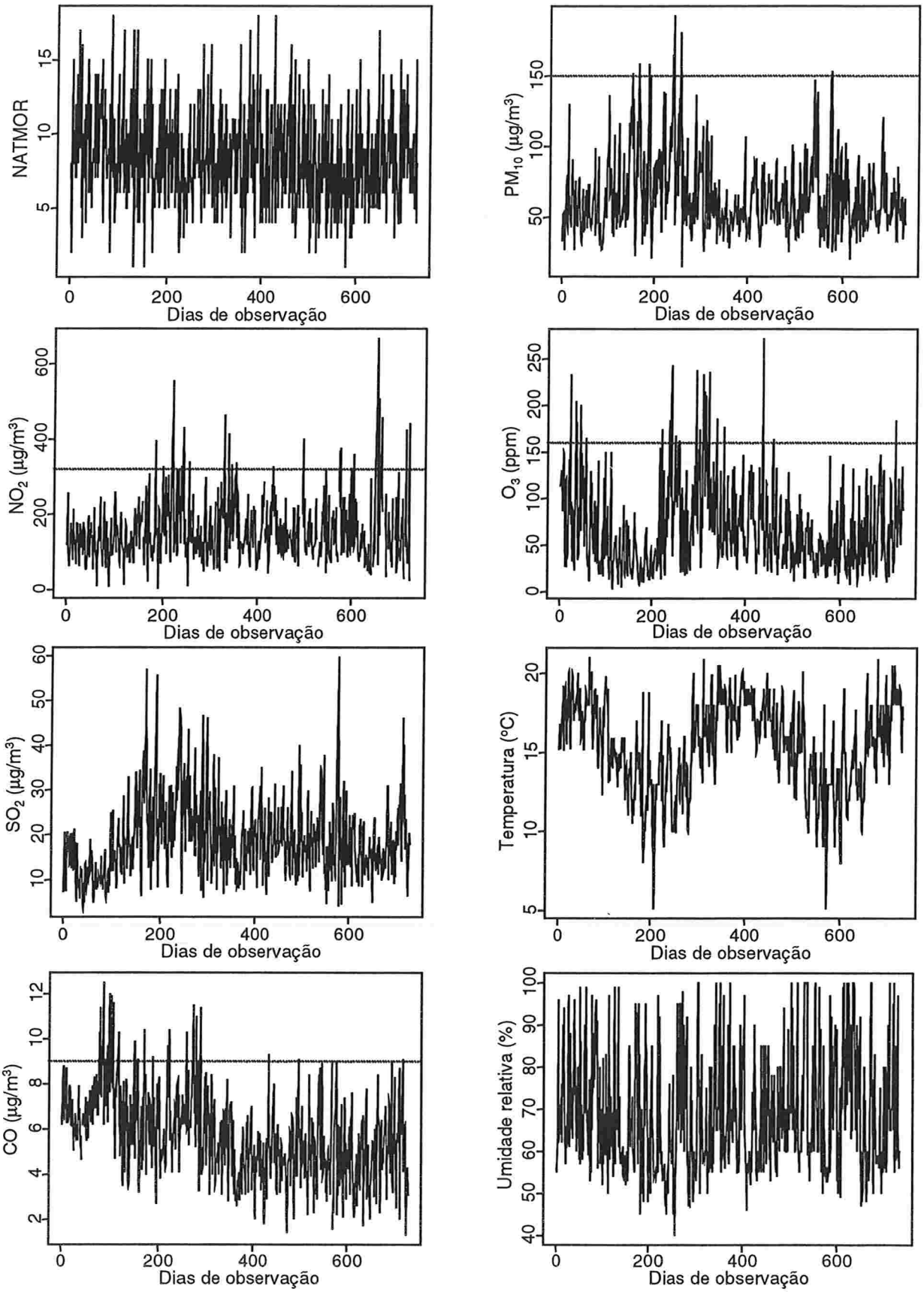


Figura A.2: Representação gráfica das séries cronológicas para NATMOR, TEMP, UMID e poluentes nos anos de 1991 e 1992.

Apêndice B

Resultados dos ajustes via MLG

Tabela B.1: Estimativas dos parâmetros do modelo (4.1) com o poluente M5NO2 e outras estatísticas.

Termo	Estimativa	Erro padrão	χ^2	Graus de liberdade	Valor de p
α	1,97610	0,23491			
M5NO2	0,00104	0,00032	10,701	1	0,001
FEV/91	–0,06887	0,08727			
MAR/91	–0,05714	0,08513			
ABR/91	–0,18400	0,08941			
MAI/91	–0,15676	0,09249			
JUN/91	–0,21359	0,09636			
JUL/91	–0,24343	0,10558			
AGO/91	–0,46408	0,10834			
SET/91	–0,27596	0,10182			
OUT/91	–0,21831	0,09212			
NOV/91	–0,27933	0,09256			
DEZ/91	–0,33256	0,09268			
JAN/92	–0,09137	0,08561			
FEV/92	–0,28278	0,09316			
MAR/92	–0,20986	0,08935			
ABR/92	–0,20051	0,08988			
MAI/92	–0,28456	0,09279			
JUN/92	–0,39607	0,09694			
JUL/92	–0,47786	0,10652			
AGO/92	–0,39324	0,10346			
SET/92	–0,29763	0,09611			
OUT/92	–0,32499	0,10012			
NOV/92	–0,21228	0,09099			
DEZ/92	–0,22926	0,08833	53,832	23	< 0,001
SEG	0,17112	0,05181			
TER	0,26061	0,05087			
QUA	0,27017	0,05061			
QUI	0,24307	0,05089			
SEX	0,22598	0,05102			
SÁB	0,15543	0,05208	42,544	6	< 0,001
M2UMID	0,00312	0,00204	2,403	1	0,121
M2TEMP	–0,01054	0,01233	0,722	1	0,386
UMID1	0,05112	0,04840			
UMID2	–0,03211	0,04572			
UMID3	–0,03536	0,06061	4,038	3	0,257
TEMP1	–0,01821	0,05045			
TEMP2	–0,02047	0,06732			
TEMP3	0,02761	0,08245	1,665	3	0,645

Desvio Residual: 724,695 com 687 graus de liberdade.
AIC aproximado: 805,274.

Tabela B.2: Estimativas dos parâmetros do modelo (4.2) e outras estatísticas.

Termo	Estimativa	Erro padrão	χ^2	Graus de liberdade	Valor de p
α	2,03347	0,23820			
M5NO2a	0,03421	0,04406	0,625	1	0,429
M5NO2b	0,08217	0,04474	3,499	1	0,061
M5NO2c	0,10311	0,04561	5,303	1	0,021
M5NO2d	0,13331	0,05170	6,892	1	0,009
FEV91	–0,06856	0,08761			
MAR91	–0,05484	0,08607			
ABR91	–0,18214	0,08997			
MAI91	–0,14677	0,09410			
JUN91	–0,20475	0,09824			
JUL91	–0,23155	0,10672			
AGO91	–0,42103	0,10717			
SET91	–0,26310	0,10202			
OUT91	–0,20761	0,09276			
NOV91	–0,26704	0,09311			
DEZ91	–0,31571	0,09475			
JAN92	–0,09220	0,08649			
FEV92	–0,27959	0,09528			
MAR92	–0,20297	0,08986			
ABR92	–0,20331	0,09032			
MAI92	–0,27656	0,09342			
JUN92	–0,38656	0,09852			
JUL92	–0,46362	0,10680			
AGO92	–0,38345	0,10425			
SET92	–0,28803	0,09743			
OUT92	–0,25721	0,09712			
NOV92	–0,21611	0,09168			
DEZ92	–0,22315	0,08929	49,825	23	< 0,001
SEG	0,16910	0,05198			
TER	0,25779	0,05108			
QUA	0,26334	0,05096			
QUI	0,23776	0,05127			
SEX	0,22213	0,05148			
SÁB	0,15241	0,05254	40,507	6	< 0,001
M2UMID	0,00320	0,00206	2,488	1	0,115
M2TEMP	–0,00929	0,01237	0,584	1	0,445
UMID1	0,05485	0,04874			
UMID2	–0,03019	0,04612			
UMID3	–0,03594	0,06114	4,229	3	0,238
TEMP1	–0,01689	0,05072			
TEMP2	–0,01641	0,06773			
TEMP3	0,02937	0,08296	1,517	3	0,678

Desvio Residual: 726,339 com 684 graus de liberdade.
AIC aproximado: 813,630.

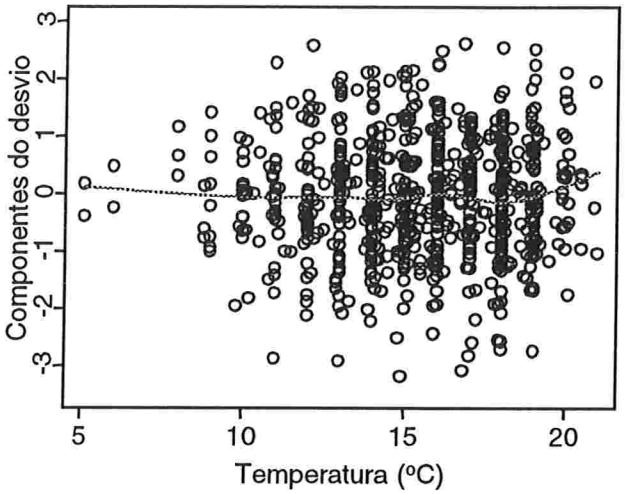


Figura B.1: Diagrama de dispersão suavizado dos componentes do desvio do ajuste do modelo (4.1) e TEMP.

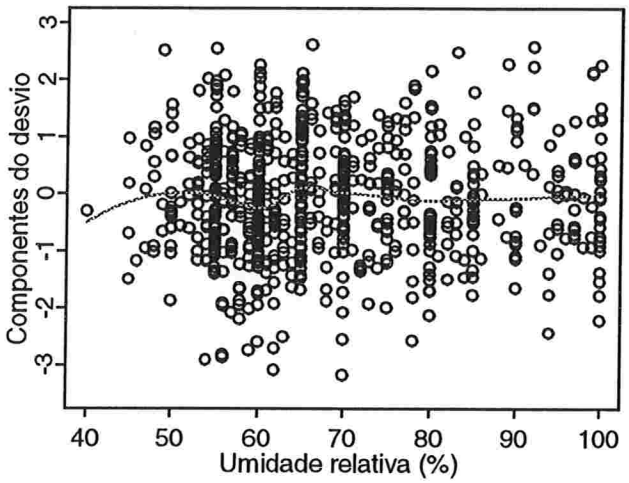


Figura B.2: Diagrama de dispersão suavizado dos componentes do desvio do ajuste do modelo (4.1) e UMID.

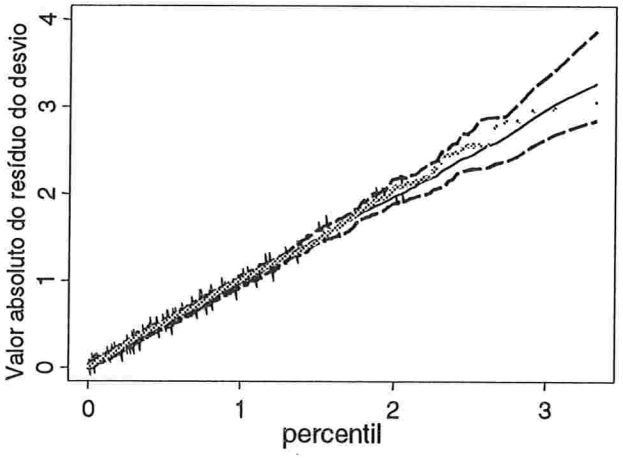


Figura B.3: Gráfico de envelope para a distribuição de Poisson para os resíduos do ajuste do modelo (4.2).

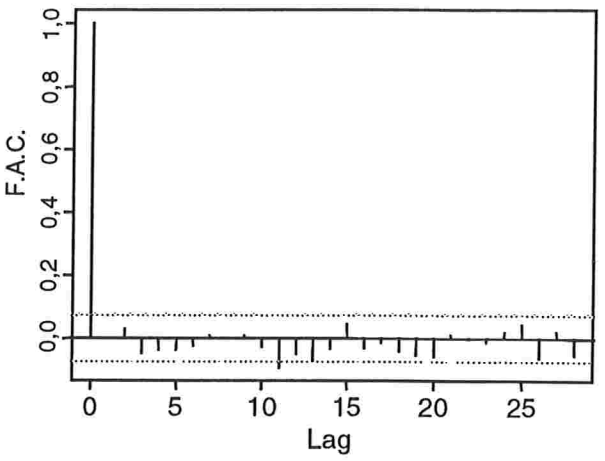


Figura B.4: Gráfico de autocorrelação para os resíduos do ajuste do modelo (4.2).

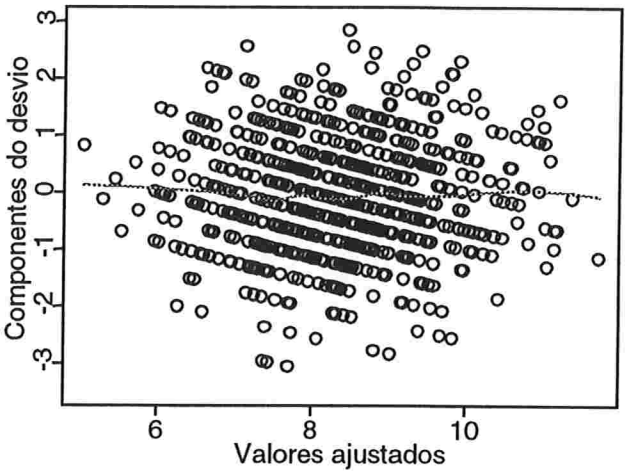


Figura B.5: Diagrama de dispersão suavizado dos componentes do desvio do ajuste do modelo (4.2) e valores ajustados.

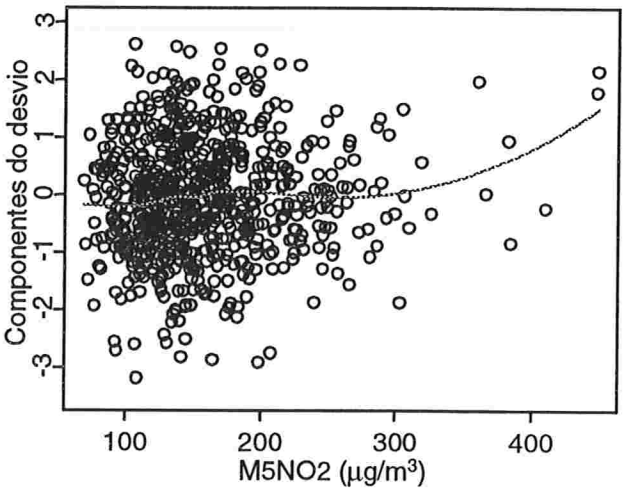


Figura B.6: Diagrama de dispersão suavizado dos componentes do desvio do ajuste do modelo (4.1) e M5NO2.

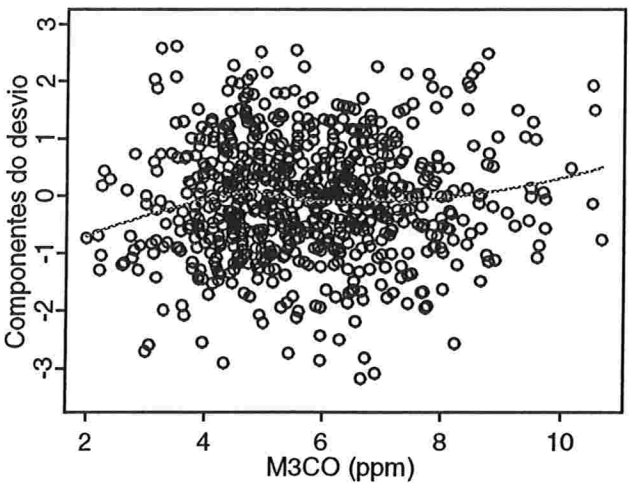


Figura B.7: Diagrama de dispersão suavizado dos componentes do desvio do ajuste do modelo (4.1) e M3CO.

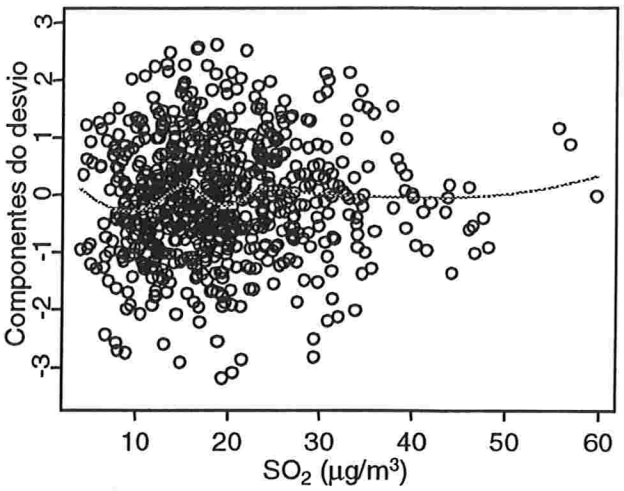


Figura B.9: Diagrama de dispersão suavizado dos componentes do desvio do ajuste do modelo (4.1) e SO_2.

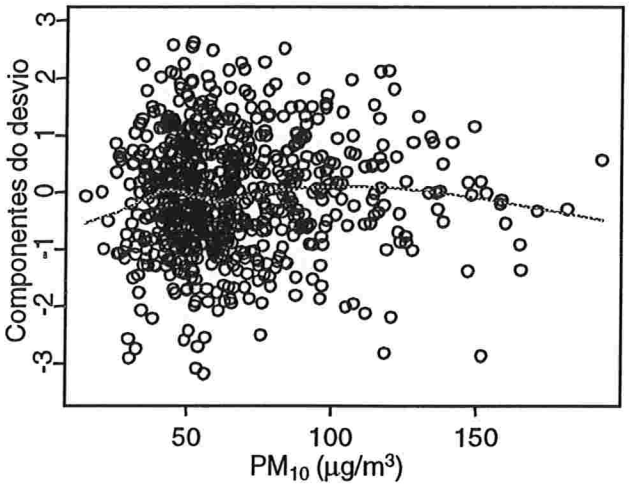


Figura B.8: Diagrama de dispersão suavizado dos componentes do desvio do ajuste do modelo (4.1) e PM_{10}.

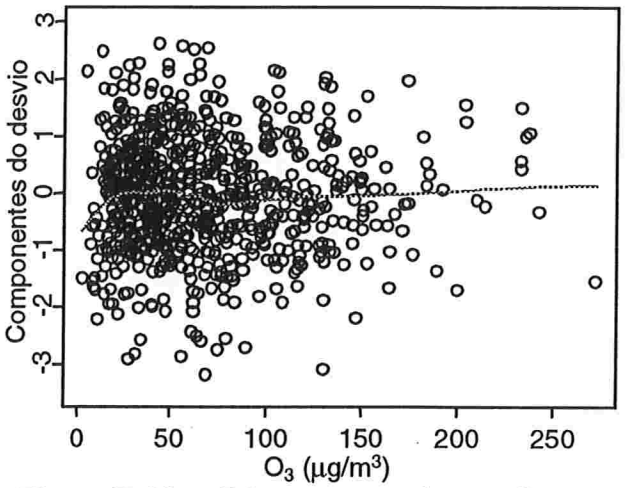


Figura B.10: Diagrama de dispersão suavizado dos componentes do desvio do ajuste do modelo (4.1) e O_3.

Apêndice C

Resultados dos ajustes via MAG

Tabela C.1: Estimativas dos parâmetros da parte linear do modelo (4.6) e outras estatísticas.

Parte não paramétrica						
	Efeito linear			Efeito não linear		
Termo	χ^2	Graus de liberdade	Valor de p	χ^2	Graus de liberdade	Valor de p
f_1(DIAS), $\lambda = 0,5$	19,280	1,0	< 0,001	11,603	2,3	0,004
f_4(M5NO2), $\lambda = 0,8$	4,547	1,0	0,033	5,938	1,8	0,043

Parte paramétrica					
Termo	Estimativa	Erro padrão	χ^2	Graus de liberdade	Valor de p
α	1,72850	0,11965			
SEG	0,16189	0,05145			
TER	0,25150	0,05045			
QUA	0,26625	0,05031			
QUI	0,24035	0,05047			
SEX	0,22662	0,05070			
SÁB	0,14541	0,05154	42,691	6,0	< 0,001
M2TEMP	0,00512	0,00484	0,000	1,0	> 0,999
M2UMID	0,00188	0,00119	2,197	1,0	0,138

Desvio residual: 747,727 com 710,896 graus de liberdade.
AIC aproximado: 778,797.

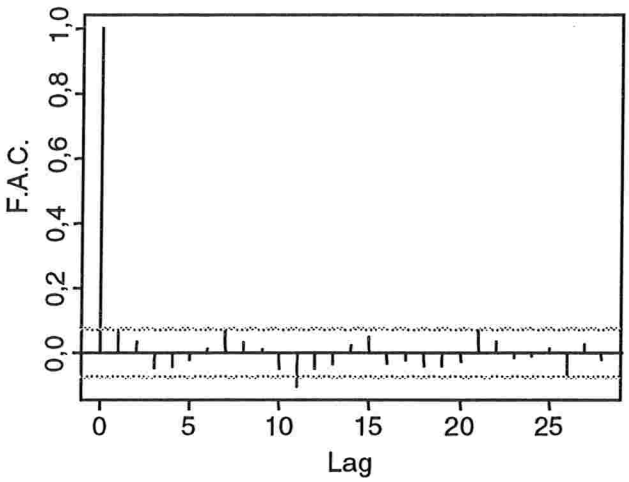


Figura C.1: Gráfico de autocorrelações dos resíduos resultantes do ajuste modelo (4.4), com $\lambda = 0{,}50$.

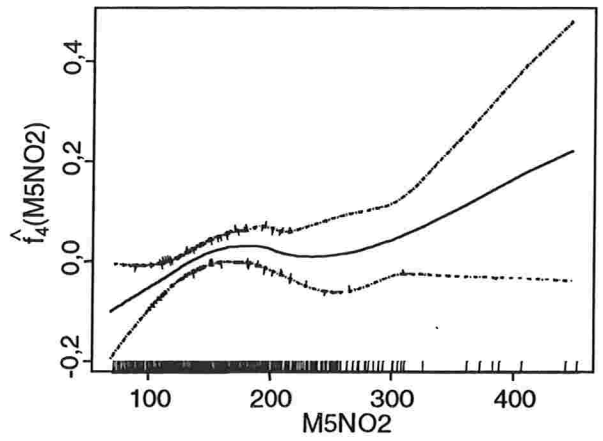


Figura C.2: Curva estimada $\hat{f}_4$(M5NO2) no modelo (4.6) e bandas de confiança.

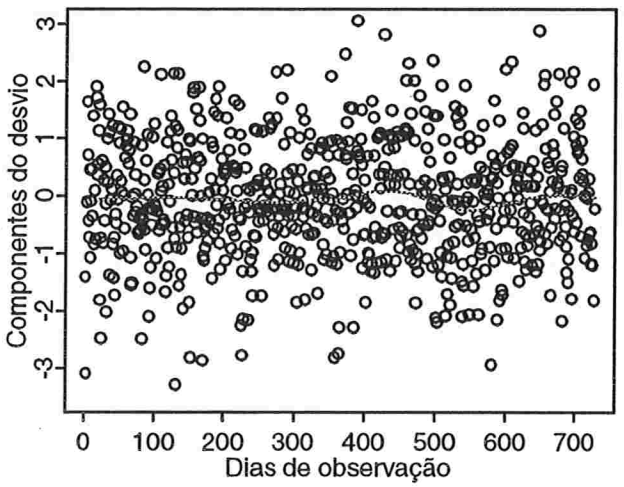


Figura C.3: Diagrama de dispersão suavizado dos componentes do desvio do modelo (4.6) e dias de observação.

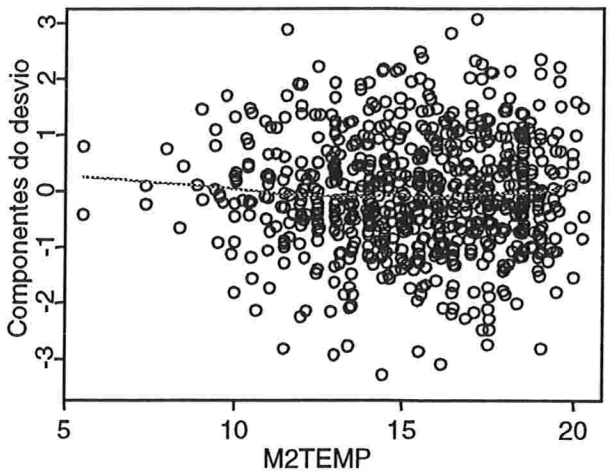


Figura C.4: Gráfico dos componentes do desvio para o modelo (4.6) e M2TEMP.

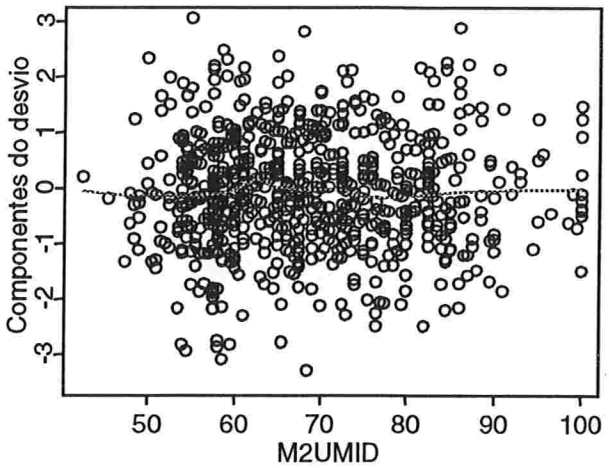


Figura C.5: Diagrama de dispersão suavizado dos componentes do desvio do modelo (4.6) e M2UMID.

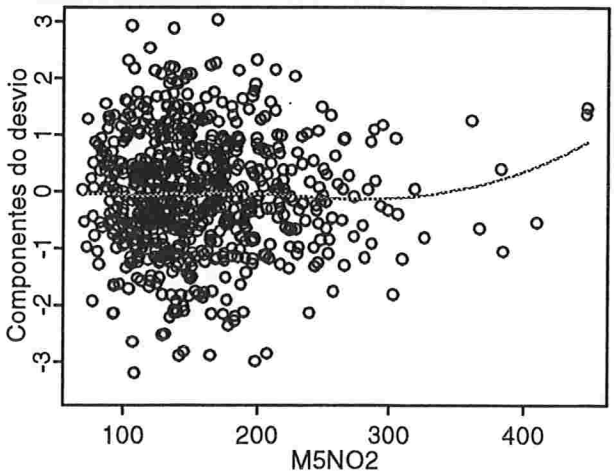


Figura C.6: Diagrama de dispersão suavizado dos componentes do desvio do modelo (4.6) e M5NO2.

Apêndice D

Listagem de alguns comandos utilizados no S-Plus para a análise dos dados apresentados no Capítulo 4

- **INICIALIZAÇÃO**

```
attach(fetal9192)
objects(2)
options(object.size=10E8)


## Função para cálculo da estatística AIC.

aic2<-function(x){
      n <- length(residuals(x))
      pres2_(residuals(x,type="pearson"))^2
      Dhat_sum(pres2)/x$df.residual
      aAIC_x$deviance+2*Dhat*(n-x$df.residual)
      cat("Approximate Akaike's Information Criterion is:", aAIC, "\n")
      }
```

- **PREPARAÇÃO DO BANCO DE DADOS**

```
## Criação de variáveis indicadoras dos dias da semana e dos meses do ano.

fetal9192$SEG<-ifelse(WEEK==2,1,0)
fetal9192$TER<-ifelse(WEEK==3,1,0)
fetal9192$QUA<-ifelse(WEEK==4,1,0)
fetal9192$QUI<-ifelse(WEEK==5,1,0)
fetal9192$SEX<-ifelse (WEEK==6,1,0)
fetal9192$SAB<-ifelse (WEEK==7,1,0)

fetal9192$FEV91<-ifelse(MOUNTH==2,1,0)
fetal9192$MAR91<-ifelse(MOUNTH==3,1,0)
fetal9192$ABR91<-ifelse(MOUNTH==4,1,0)
fetal9192$MAI91<-ifelse(MOUNTH==5,1,0)
fetal9192$JUN91<-ifelse(MOUNTH==6,1,0)
fetal9192$JUL91<-ifelse(MOUNTH==7,1,0)
fetal9192$AGO91<-ifelse(MOUNTH==8,1,0)
fetal9192$SET91<-ifelse(MOUNTH==9,1,0)
fetal9192$OUT91<-ifelse(MOUNTH==10,1,0)
fetal9192$NOV91<-ifelse(MOUNTH==11,1,0)
fetal9192$DEZ91<-ifelse(MOUNTH==12,1,0)
fetal9192$JAN92<-ifelse(MOUNTH==13,1,0)
fetal9192$FEV92<-ifelse(MOUNTH==14,1,0)
fetal9192$MAR92<-ifelse(MOUNTH==15,1,0)
fetal9192$ABR92<-ifelse(MOUNTH==16,1,0)
fetal9192$MAI92<-ifelse(MOUNTH==17,1,0)
fetal9192$JUN92<-ifelse(MOUNTH==18,1,0)
fetal9192$JUL92<-ifelse(MOUNTH==19,1,0)
fetal9192$AGO92<-ifelse(MOUNTH==20,1,0)
fetal9192$SET92<-ifelse(MOUNTH==21,1,0)
fetal9192$OUT92<-ifelse(MOUNTH==22,1,0)
fetal9192$NOV92<-ifelse(MOUNTH==23,1,0)
fetal9192$DEZ92<-ifelse(MOUNTH==24,1,0)
```

```
## Criação de variáveis indicadoras dos quartis da temperatura e da umidade.

fetal9192$TEMP1<-ifelse((TEMP>=13.5 & TEMP<15.6),1,0)
fetal9192$TEMP2<-ifelse((TEMP>=15.6 & TEMP<17.55),1,0)
fetal9192$TEMP3<-ifelse((TEMP>=17.55),1,0)

fetal9192$UMID1<-ifelse((UMID>=59.0 & UMID<67.5),1,0)
fetal9192$UMID2<-ifelse((UMID>=67.5 & UMID<76.5),1,0)
fetal9192$UMID3<-ifelse((UMID>=59.0),1,0)


## Criação de variáveis indicadoras dos quintis da variável M5NO2.

fetal9192$q1M5NO2<-ifelse((M5NO2>=114.2200 & M5NO2<135.5360),1,0)
fetal9192$q2M5NO2<-ifelse((M5NO2>=135.5360 & M5NO2<158.8320),1,0)
fetal9192$q3M5NO2<-ifelse((M5NO2>=158.8320 & M5NO2<191.6800),1,0)
fetal9192$q4M5NO2<-ifelse((M5NO2>=191.6800),1,0)
```

- **ANÁLISE DESCRITIVA**

```
summary(fetal9192)

boxplot(split(NATMOR,YEAR), ylim=c(0,20),xlab="Ano")
boxplot(split(NO2,YEAR), ylim=c(0,700),xlab="Ano")
abline(h=320,lty=2)
boxplot(split(TEMP,YEAR), ylim=c(0,25),xlab="Ano")
boxplot(split(UMID,YEAR), ylim=c(30,110),xlab="Ano")

tsplot(NATMOR,xlab="Dias de observação",type="l")
tsplot(NO2,xlab="Dias de observação",type="l")
abline(h=320,lty=2)
tsplot(TEMP,xlab="Dias de observação",type="l")
tsplot(UMID,xlab="Dias de observação",type="l")
```

- **ANÁLISE VIA MLG**

```
## Ajuste do modelo (4.2).

qNO2.glm<-glm(NATMOR~q1M5NO2+q2M5NO2+q3M5NO2+q4M5NO2+FEV91+MAR91+ABR91+MAI91+
JUN91+JUL91+AGO91+SET91+OUT91+NOV91+DEZ91+JAN92+FEV92+MAR92+ABR92+MAI92+JUN92+
JUL92+AGO92+SET92+OUT92+NOV92+DEZ92+SEG+TER+QUA+QUI+SEX+SAB+M2UMID+M2TEMP+UMID1+
UMID2+UMID3+TEMP1+TEMP2+TEMP3,family=poisson,na=na.omit)

summary(qNO2.glm,c=F,disp=0)
aic2(qNO2.glm)
acf(residuals(qNO2.glm,type="deviance"))
```

- **ANÁLISE VIA MAG**

```
## Ajuste do modelo (4.6).

# Função que seleciona o modelo com menor valor para a estatística AIC, a partir
de valores pré-fixados de λ:

NO2.inicial<-gam(NATMOR~lo(M5NO2,0.5)+lo(Dias,0.5)+SEG+TER+QUA+QUI+SEX+SAB+
M2TEMP+M2UMID,family=poisson,na=na.omit)
step.gam(NO2.inicial, scope=list(
      "Dias"    =~lo(Dias,0.5),
      "SEG"     =~SEG,
      "TER"     =~TER,
      "QUA"     =~QUA,
      "QUI"     =~QUI,
      "SEX"     =~SEX,
      "SAB"     =~SAB,
      "M2TEMP"  =~M2TEMP,
      "M2UMID"  =~M2UMID,
      "M5NO2"   =~lo(M5NO2,0.6)+lo(M5NO2,0.7)+lo(M5NO2,0.8)+M5NO2,
      ))

# O modelo selecionado foi:

NO2.smooth<-gam(NATMOR~lo(M5NO2,0.8)+lo(Dias,0.5)+SEG+TER+QUA+QUI+SEX+SAB+
M2TEMP+M2UMID,family=poisson,na=na.omit)
summary(NO2.smooth)
summary.glm(NO2.smooth,c=F,disp=0)
aic2(NO2.smooth)
acf(residuals(qNO2.glm,type="deviance"))


## Obtenção da matriz R4 = f̂4 z , modelo (4.6).

n<-726
I<-diag(n)
R.np<-array(dim=c(n,n))
Rli.np<-array(dim=c(n,n))
for (i in 1:n) {
      linhas<-gam(I[i,]~lo(M5NO2,0.8)+lo(Dias,0.5)+SEG+TER+QUA+QUI+SEX+SAB+
      M2TEMP+M2UMID,weight=NO2.smooth$weight,family=gaussian,na=na.omit)
      R.np[,i]<-linhas$smooth[,1]
      Rli.np[,i]<-(model.matrix(linhas)[,2]*linhas$coefficients[2])
      }
R4<-(Rli.np+R.np)


## Cálculo da matriz de variâncias e covariâncias de f̂4 , modelo (4.6).

W<-array(dim=c(n,n))
W<-diag(NO2.smooth$weight)
var.cov<-R4%*%solve(W)%*%t(R4)
```

```
## Construção do gráfico do risco relativo estimado.

f.total<-(model.matrix(NO2.smooth)[,2] * NO2.smooth$coefficients[2]) +
NO2.smooth$smooth[,1]
M4<-array(dim=c(n,1))
for (i in 1:n){
    M4[i,1]<-( var.cov[i,i] + var.cov[88,88] - (2*var.cov[i,88]) )
    }
rr<-array(dim=c(n,1))
dif.total<-array(dim=c(n,1))
for (i in 1:n){
      dif.total[i]<-exp(f.total[i]-f.total[88])
      rr[i]<-(dif.total[i])*(dif.total[i])* M4[i,1]
    }
rr.ep<-sqrt(rr)
super<-dif.total + (2*rr.ep)
infer<-dif.total - (2*rr.ep)

plot(M5NO2,dif.total, ylim=c(0.95,1.8),ylab="Risco relativo estimado")
points(M5NO2,super)
points(M5NO2,infer)
```

Bibliografia

[1] André, P.A.; Braga, A.L.F.; Lin, C.A.; Conceição, G.M.S.; Pereira, L.A.A.; Miraglia, S.G.E.K.; Böhn, G.M. (2000). Environmental epidemiology applied to urban atmospheric pollution: a contribution from the Experimental Air Pollution Laboratory (LPAE). **Rep. Public Heath**, v. 16, n. 3, p. 619-628.

[2] André, C.D.S.; Rancan, A. (1999). **Relatório de análise estatística sobre o projeto: Aplicação da microbiologia preditiva na avaliação da segurança de salsichas**. São Paulo, IME-USP, 62p. (RAE-CEA-99P02).

[3] Braga, A.L.F.; Conceição, G.M.S.; Pereira, L.A.A.; Kishi, H.S.; Pereira, J.C.R.; Andrade, M.F.; Gonçalves, F.L.T.; Saldiva, P.H.N.; Latorre, M.R.D.O. (1999). Air pollution and pediatric respiratory admissions in São Paulo, Brazil. **J. Environmental Med.**, v. 1, p. 95-102.

[4] Buja, A.; Donnell, D.; Stuetzle, W. (1986). **Additive principal components**. Technical Report, Department of Statistics, University of Washington, Seattle.

[5] Buja, A.; Hastie, T.J.; Tibishirani, R.J. (1989). Linear smoothers and additive models. **Ann. Statist.**, v. 17, n. 2, p. 453-555.

[6] CETESB (2000). **Relatório de qualidade do ar no Estado de São Paulo – 1999.** São Paulo, CETESB.

[7] Chambers, J.M.; Hastie, T.J. (1993). **Statistical models in S**. New York, Chapman and Hall, 608 p.

[8] Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scarterplots, **J. Amer. Statist. Assoc.**, v. 74, n. 368, p. 829-836.

[9] Conceição, G.M.S. (1999). **Métodos estatísticos para avaliação da associação entre poluição atmosférica e marcadores de morbi-mortalidade na cidade de São Paulo e aplicações**. São Paulo. 208p. Dissertação (Mestrado) – Faculdade de Medicina da Universidade de São Paulo.

[10] Conceição, G.M.S.; Miraglia, S.G.E.K.; Kishi, H.S.; Saldiva, P.H.N.; Singer, J.M. Air pollution and children mortality: a time series study in São Paulo, Brazil. **Environmental Health Perspect.** (Submetido para publicação).

[11] Cook, R.D.; Weisberg, S. (1982). **Residuals and influence in regression**. London, Chapman and Hall, 230 p.

[12] Cox, D.D. (1983). Asymptotics for M-type smoothing splines. **Ann. Statist.**, v. 11, p. 530-551.

[13] Dockery, D.W.; Pope III, C.A. (1994). Acute respiratory effects of particulate air pollution. **Annual Rev. Public Health**, v. 15, p. 107-132.

[14] Dockery, D.W.; Pope III, C.A.; Xu, X.; Spengler, J.D.; Ware, J.H.; Fay, M.E.; Ferris Jr., B.G.; Speizer, F.E. (1993). An association between air pollution and mortality in six US cities. **Engl. J. Med.**, v. 329, p. 1753-175.

[15] Fahrmeir, L.; Tutz, G. (1994). **Multivariate statistical modelling based on generalized linear models**. New York, Springer, 425 p.

[16] Graybill, F.A. (1983). **Matrices with applications in statistics**. 2.ed. Belmont, Wadsworth International Group, 461 p.

[17] Gouveia, N.; Fletcher, T. (2000). Time series analysis of air pollution and mortality: effects by cause, age and socioeconomic status. **J. Epidemiol. Community Health**, v. 54, p. 750-755.

[18] Hastie, T.J.; Tibshirani, R.J. (1987). Generalized additive models: some applications. **J. Amer. Statist. Assoc.**, v. 82 , n.398, p. 371-386.

[19] Hastie, T.J.; Tibshirani, R.J. (1990). **Generalized additive models**. London, Chapman & Hall, 335 p.

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

[21] Hobert, J.P.; Altman, N.S.; Schofield, C.L. (1997). Analyses of fish species richness with spatial covariate. **J. Amer. Statist. Assoc.**, v. 92, n. 439, p. 846-854.

[22] MathSoft (1998). **S-Plus User's Guide**. Version 4.5. Seattle, Data Analysis Products Division.

[23] McCullagh, P.; Nelder, J.A. (1989). **Generalized linear models**. 2.ed. London, Chapman and Hall, 511 p.

[24] Miraglia, S.G.E.K.; Conceição, G.M.S.; Saldiva, P.H.N.; Strambi, O. (1997). Analysis of the impact of fuel consumption on mortality rates in São Paulo. In: THIRD INTERNATIONAL CONFERENCE ON URBAN TRANSPORT AND THE ENVIRONMENT FOR THE 21ST CENTURY, Acquasparta, Boston. Computational Mechanics Publications, p. 434-444.

[25] Morgenstern, H. (1995). Ecologic studies in epidemiology: concepts, principles and methods. **Annual Rev. Public Health**, v. 16, p. 61-81.

[26] Nelder, J.A.; Wedderburn, R.W.M. (1972). Generalized linear models. **J. Roy. Statist. Soc. Ser. A**, v. 135, p. 370-384.

[27] Neter, J.; Kutner, M.H.; Nachtsheim, C.J.; Wasserman, W. (1996). **Applied linear statistical models**. 4.ed. Boston, McGraw-Hill, 720 p.

[28] Opsomer, J.D. (2000). Asymptotic properties of backfitting estimators. **J. Multivariate Anal.**, v. 73, n. 2, p. 166-179.

[29] Opsomer, J.D.; Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. **J. Comput. Graph. Statist.**, v. 4, n. 8, p. 715-732.

[30] Paula, G.A. **Modelos de regressão com apoio computacional**. IME-USP. (Não publicado).

[31] Pereira, L.A.A.; Loomis, D.; Conceição, G.M.S.; Braga, A.L.F.; Arcas, R.M.; Kishi, H.S.; Singer, J.M.; Böhm, G. M.; Saldiva, P.H.N. (1998). Association between air pollution and intrauterine mortality in São Paulo, Brazil. **Environmental Health Perspect.**, v. 106, n. 6, p. 325-329.

[32] Rice, J.; Rosenblatt, M. (1983). Smoothing splines, regression derivatives and convolution. **Ann. Statist.**, v. 11, p. 141-156.

[33] Rothman, K.J.; Greenland, S. (1998). **Modern epidemiology**. 2.ed. Philadelphia, Lippincott-Raven, 737 p.

[34] Rousseeuw, P. (1984). Least median of squares regression. **J. Am. Statist. Assoc.**, v. 79, p. 871-880.

[35] Saldiva, P.H.N.; Lichtenfels, A.J.F.C.; Paiva, P.S.O.; Barone, I.A.; Martins, M.A.; Massad, E.; Pereira, J.C.R.; Xavier, V.P.; Singer, J.M.; Böhm, G.M. (1994). Association between air pollution and mortality due to respiratory diseases in children in São Paulo, Brazil: a preliminar report. **Environmental Res.**, v. 65, p. 218-225.

[36] 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**, v. 50, p. 159-163.

[37] Schoenberg, I.J. (1964). Spline functions and the problem of graduation. **Proc. Nat. Acad. Sci., USA**, v. 52, p. 974-50.

[38] Schwartz, J. (1994). Nonparametric smoothing in the analysis of air pollution and respiratory illness. **Canad. J. Statist.**, v. 22, n. 4, p. 471-487.

[39] Schwartz, J. (1999). Air pollution and hospital admissions for heart disease in eight U.S. countries. **Epidemiology**, v. 10, p. 17-22.

[40] Sen, P.K.; Singer, J.M. (1993). **Large sample methods in statistics: an introduction with applications**. London, Chapman e Hall, 382 p.

[41] Speckman, P.E. (1988). Regression analysis for partially linear models. **J. Roy. Statist. Assoc. Ser. B**, v. 50, p. 413-436.

[42] Stone, C.J. (1977). Consistent nonparametric regression. **Ann. Statist.**, v. 5, p. 595-620.

[43] Thurston, S.W.; Wand, M.P.; Wiencke, J.K. (2000). Negative binomial additive models. **Biometrics**, v. 59, p. 139-144.