Regressão logística com resposta contínua

Adrilayne dos Reis Araújo

Dissertação apresentada ao
Instituto de Matemática e Estatística
da Universidade de São Paulo
para obtenção do título de
Mestre em Estatística

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

*Durante parte da elaboração deste trabalho a autora
recebeu apoio financeiro da CAPES e do CNPq*

– São Paulo, Dezembro de 2002 –

Regressão logística com resposta contínua

Este exemplar corresponde à redação
final da dissertação devidamente corrigida e
defendida por Adrilayne dos Reis Araújo
e aprovada pela comissão julgadora.

São Paulo, 05 de dezembro de 2002.

Comissão Julgadora:

- Profa. Dra. Carmen Diva Saldiva de André (orientadora) - IME/USP
- Profa. Dra. Silvia Lopes de Paula Ferrari - IME/USP
- Prof. Dr. Paulo Hilário Nascimento Saldiva - FM/USP

*“Ao meu ver, era o açaí o maravilhoso
licor, o decantado néctar que as cunhantãs
serviam aos heróis quando, triunfadores,
mereciam, nas tabas, as
honrarias gloriosas”.*

Jaques Flores

*Aos meus pais
Célia e Pedro,
e ao meu irmão
Robson.*

Agradecimentos

- ★ A Deus por me dar forças necessárias nos momentos que mais precisei.
- ★ A minha orientadora, Professora Carmen Saldiva por acreditar em minha capacidade e pelo grande apoio dado em todos os momentos de dificuldade.
- ★ À Professora Lúcia Barroso pelo incentivo no início desta jornada.
- ★ Aos Professores Adilson Simonis, Antonio Carlos, Caio Dantas, Carlinhos, Cláudia Peixoto, Clélia Toloi, Denise Botter, Elisete Aubin, Gilberto Alvarenga, Heleno Bolfarine, Júlio Singer, Marcos Magalhães, Mariane Streibel, Mônica Sandoval, Nelson Tanaka, Pablo Ferrari, Pedro Morettin, Sérgio Wechsler, Sílvia Ferrari e Silvia Nagib que de alguma forma contribuíram para o crescimento de meu conhecimento.
- ★ Aos funcionários da Secretaria de Estatística que me socorriam nas horas de pane na sala 250, em especial a Simoni e Helena.
- ★ Aos amigos Domingo, Sandra, Jairo, Gladys Salcedo, Hernan e Paola pela boa amizade e, é claro, pelas boas festas ao ritmo de salsa e merengue.
- ★ As minhas amiguinhas do “Clube da Luluzinha” Luciana, Suzi, Sílvia, Susana, Gleice, Delhi e Carolina por participarem do clube comigo e pela rica amizade.
- ★ Aos amigos Liane Dorneles, Carlos Pinheiro, Valéria Pereira e Leandro Prudente pela boa convivência e alegre amizade.
- ★ À amiga Laila Nahum pela denominação de princesa e sempre alegre amizade.
- ★ Aos amigos Pedro Abreu, Alberto, Gisela, Edivaldo, Caio, Juvêncio, Marcos, Daniela, Cléber, Diana, Liliam Pereira e Regina Albanese pela companhia na 250 e amizade.
- ★ Aos Professores Héilton e Regina Tavares pela amizade e incentivo desde o início.
- ★ Ao querido amigo Carlos Paixão pela presença constante, mesmo que de tão longe.
- ★ A minha querida amiga Maria Ana pela amizade, apoio e compreensão em todos os momentos.
- ★ Ao meu querido Edson por sempre acreditar no meu potencial, incentivar o meu crescimento, e principalmente, conseguir as minhas melhores risadas.
- ★ Aos meus sobrinhos Henrique, Daniel e Jully que mesmo de longe sempre foram especiais.
- ★ Aos meus queridos pais e irmão por aceitarem a distância e sempre me demonstrarem confiança.
- ★ Finalmente, a todos que de alguma forma contribuíram para a realização deste trabalho.

Resumo

A regressão logística com resposta contínua é uma alternativa à regressão logística usual quando a variável resposta possui distribuição contínua e o objetivo do estudo é estimar a probabilidade de ocorrência de valores acima ou abaixo de um determinado valor de corte. O modelo assim construído pode ser escrito na forma de um modelo linear generalizado com função de ligação composta. Quando corretamente especificada, a incorporação da informação sobre a distribuição da variável resposta no modelo faz com que os estimadores de máxima verossimilhança sejam mais eficientes. A técnica é apresentada para os casos em que a variável resposta tem distribuição normal ou log-normal. Como aplicação, considerando dados referentes à cidade de São Paulo nos anos de 1998 e 1999, um modelo de regressão logística com resposta contínua foi considerado na previsão do risco da concentração do poluente NO_2 ser maior que um valor de corte estabelecido por legislação. Variáveis climáticas e temporais foram consideradas como preditoras. Mostraram-se importantes para prever o risco a temperatura, a umidade relativa do ar, os dias da semana, as estações do ano, precipitação pluviométrica e velocidade do vento.

Abstract

Binary regression with continuous outcomes constitutes an alternative to logistic regression when the outcome is continuous and the investigator's interest focuses to estimate the probability of subjects who fall above or below a cut-off value. The model is based on a generalized linear model with composite link that takes advantage of the continuous structure of the outcome, typically gaussian or lognormal. Under correct response modeling, binary regression with continuous outcomes is more efficient than logistic regression. A binary regression with continuous outcomes was considered to predict the risk that a NO_2 pollutant concentration is above the limits set by environmental legislation in São Paulo city during 1998 and 1999. Climatic and temporal variables were considered as predictors. Temperature, humidity, days of the week, station of the year, precipitation and speed of the wind revealed important to predict the risk.

Índice

Agradecimentos	v
Resumo	vi
1 Introdução	2
2 Regressão logística com resposta contínua	4
2.1 Introdução	4
2.2 Modelo logístico com resposta normal	6
2.3 Estimação para resposta normal	7
2.4 Teste de Hipóteses para resposta normal	12
2.5 Modelo logístico com resposta log-normal	13
2.6 Simulação	14
3 Aplicação	16
3.1 Introdução	16
3.2 Análise Descritiva	20
3.3 Ajuste do Modelo	24
3.4 Riscos Previstos	26
3.5 Previsão de uma nova observação	32
4 Considerações finais	34
A Procedimento de estimação dos coeficientes do modelo, dado σ.	35
B Tabelas e Gráficos	38
C Programa no S-Plus	46
Lista de Tabelas	51
Lista de Figuras	53
Bibliografia	54

Capítulo 1

Introdução

Associações entre morbidade ou mortalidade e poluição atmosférica na cidade de São Paulo foram detectadas em diversos estudos epidemiológicos [Böhm *et al.* (2002) e Singer *et al.* (2002), por exemplo], comprovando que o ar que respiramos é um problema de saúde pública. Em particular foram detectadas associações positivas entre o dióxido de nitrogênio (NO_2) e mortalidade fetal tardia em estudo realizado por Pereira *et al.* (1998) na cidade de São Paulo, entre NO_2 e admissões em crianças por causas respiratórias em estudo feito também em São Paulo [Braga *et al.*, 2001]. Também foram encontradas associações entre esse poluente e mortalidade por todas as causas e cardiovascular em sete cidades da Espanha [Saez *et al.*, 2002], ocorrência de derrame em Seul, em estudo descrito em Hong *et al.* (2002), e doenças cardiovasculares e respiratórias em idosos na cidade de Tóquio [Ye *et al.*, 2001], entre outros.

Seria interessante, então, a construção de modelos que possibilitassem prever a ocorrência de altas concentrações dos poluentes de forma a permitir que ações preventivas fossem tomadas no sentido de minimizar os efeitos nocivos da poluição atmosférica à saúde da população. Para o NO_2 , essas medidas constituiriam em diminuir o tráfego de veículos, uma vez que o NO_2 é um marcador de emissão automotiva recentemente liberada, pouco afetada pelo tipo de combustível utilizado.

Neste trabalho, procuramos modelar, para a cidade de São Paulo, o risco da concentração do NO_2 ser maior que um valor de corte estabelecido por legislação, dados os valores das variáveis climáticas: temperatura, umidade relativa do ar, precipitação pluviométrica e velocidade do vento, e das variáveis de controle temporal: estações do ano e dias da semana.

Um modelo adotado com essa finalidade foi o de regressão logística para respostas binárias [ver, por exemplo, Hosmer e Lemeshow, 1989]. Nesse modelo a variável resposta (concentração do poluente) assumiu valores 0 ou 1, sendo o valor 1 associado à ocorrência de concentração maior ou igual ao valor de corte estabelecido, e o 0 à ocorrência de concentração menor que esse valor. Adotando este modelo, a variável resposta originalmente contínua, foi transformada em binária, sendo perdida a informação sobre sua distribuição. Esse modelo será denotado neste trabalho por modelo logístico usual.

Um procedimento de modelagem alternativo, proposto por Suissa e Blais (1995), no qual a informação sobre a distribuição da variável resposta é incorporada ao modelo logístico usual foi também considerado. No Capítulo 2 mostramos que o modelo assim construído pode ser escrito na forma de um Modelo Linear Generalizado [McCullagh e Nelder, 1989] com função de ligação composta [Thompson e Baker, 1981] e descrevemos um procedimento para seu ajuste pelo método de máxima verossimilhança, considerando os casos em que a variável resposta contínua tem distribuição normal ou log-normal. Apresentamos também nesse capítulo parte do estudo de simulação realizado por Suissa e Blais (1995) cujos resultados mostraram que os estimadores de máxima verossimilhança dos parâmetros do modelo assim obtidos são de 25% a 85% mais eficientes que os estimadores de máxima verossimilhança dos parâmetros do modelo logístico usual.

No Capítulo 3 consideramos um conjunto de dados constituído por valores diários de concentrações dos poluentes NO_2 , SO_2 , CO , PM_{10} e O_3 e variáveis climáticas e de controle temporal observados na cidade de São Paulo durante os anos de 1998 e 1999. Uma análise descritiva sugeriu ser razoável supor que as concentrações desses poluentes seguem distribuições log-normais, exceto para o PM_{10} que apresentou um pequeno desvio dessa suposição. Para o NO_2 , adotando um valor de corte igual a $100 \mu\text{g}/\text{m}^3$, os modelos logísticos ajustados com e sem a informação sobre a sua distribuição apresentaram estimativas dos coeficientes próximas, observando-se porém menores erros padrão quando essa informação é incorporada.

Fazemos algumas considerações finais sobre o trabalho no Capítulo 4, onde também apresentamos sugestões para pesquisas futuras. Para o ajuste do modelo com resposta contínua foi construído um programa no aplicativo S-Plus, listado no Apêndice C.

Capítulo 2

Regressão logística com resposta contínua

2.1 Introdução

Seja Y uma variável aleatória com função densidade de probabilidade ou função de probabilidade pertencente à família exponencial, ou seja, da forma:

$$f(y; \theta, \phi) = \exp[\phi\{y\theta - b(\theta)\} + c(y, \phi)], \quad (2.1)$$

onde $b(\cdot)$ e $c(\cdot)$ são funções especificadas, e $\phi^{-1} > 0$ é o parâmetro de dispersão, suposto conhecido. Caso ϕ seja desconhecido $f(y; \theta, \phi)$ pode pertencer, ou não, à família exponencial com dois parâmetros.

McCullagh e Nelder (1989) mostraram que, para Y com função densidade de probabilidade ou função de probabilidade dada por (2.1),

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

e

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

onde $V = d\mu/d\theta$ é chamada função de variância que depende somente do parâmetro canônico θ , e $\theta = \int V^{-1} d\mu = q(\mu)$, onde $q(\cdot)$ é uma função conhecida da média de Y .

Suponha que $Y_1, \dots, Y_n$ sejam n variáveis aleatórias independentes provenientes de uma mesma função densidade de probabilidade ou função de probabilidade da forma (2.1) com parâmetros θ_i , $i = 1, \dots, n$. Suponha ainda que a média de Y_i , μ_i , está relacionada com

um conjunto de variáveis preditoras (não aleatórias) $X_1, \dots, X_p$ por meio de uma função monótona e diferenciável

$$g(\mu_i) = \eta_i \quad (2.2)$$

denominada função de ligação, onde

$$\eta_i = \mathbf{x}_i^t \boldsymbol{\beta} \quad (2.3)$$

é chamado de preditor linear, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^t$, $p < n$, é um vetor de parâmetros desconhecidos a serem estimados e $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^t$ representa os valores de p variáveis explicativas, $i = 1, \dots, n$. Os modelos lineares generalizados são então definidos por (2.1), (2.2) e (2.3), com (2.1) representando o componente aleatório do modelo e (2.2) e (2.3), o componente sistemático.

Se a função de ligação é escolhida de tal forma que o parâmetro canônico θ coincide com o preditor linear, ou seja, $g(\mu_i) = \theta_i = \eta_i$, então o preditor linear modela diretamente o parâmetro canônico e tal função é chamada ligação canônica. Por exemplo, as ligações canônicas para os modelos normal e binomial são, respectivamente, dadas por

$$\mu = \eta \quad \text{e} \quad \log \left\{ \frac{\mu}{1 - \mu} \right\} = \eta.$$

Uma aplicação bastante freqüente dos Modelos Lineares Generalizados na área médica ocorre quando a variável resposta é binária, podendo ser representada por uma variável indicadora, Y_i , que assume valores 1 (sucesso) ou 0 (fracasso), ou seja, Y_i é uma variável aleatória de Bernoulli com distribuição de probabilidade dada por:

$$Y_i = \begin{cases} 1, & P(Y_i = 1) = \pi_i \\ 0, & P(Y_i = 0) = 1 - \pi_i \end{cases}, \quad (2.4)$$

que pertence à família exponencial.

Um modelo utilizado com freqüência para relacionar uma ou mais variáveis preditoras a uma variável resposta binária como em (2.4) é o modelo de regressão logística cuja equação é dada por:

$$E(Y_i) = \pi_i = \frac{\exp(\mathbf{x}_i^t \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^t \boldsymbol{\beta})}, \quad i = 1, \dots, n.$$

O modelo de regressão logística para dados binários pertence à classe dos Modelos

Lineares Generalizados [ver, por exemplo, Demétrio (2001)]. Funções de ligação bastante utilizadas na área de Epidemiologia [ver, Suissa e Blais (1995)] são: identidade ($\pi_i = \eta_i$) – adequada para a estimação da diferença de riscos, logarítmica ($\log \pi_i = \eta_i$) – adequada para estimação da razão de riscos e a função de ligação canônica *logito* – adequada para a estimação da razão de chances,

$$g(\pi_i) = \log \left[\frac{\pi_i}{1 - \pi_i} \right] = \beta_1 + \beta_2 x_{i2} + \dots + \beta_p x_{ip}, \quad i = 1, \dots, n, \quad (2.5)$$

sendo esta última a mais utilizada.

Suponhamos que Y_i definida em (2.4) tenha sido construída a partir de uma variável aleatória contínua, T_i , de acordo com o objetivo do estudo. Por exemplo, considere a variável aleatória contínua “concentração diária de NO_2 na cidade de São Paulo”, e suponhamos que o interesse do estudo é estimar a probabilidade de em um dia se observar concentração acima de $100 \mu\text{g}/\text{m}^3$. Neste caso, o parâmetro de interesse no estudo é o risco, π_i , ou seja, a probabilidade de que a resposta seja superior ao valor de corte $C = 100 \mu\text{g}/\text{m}^3$.

O método a ser apresentado consiste em incorporar a informação sobre a distribuição de T_i no ajuste do modelo logístico usual, considerando que $\pi_i = P(T_i > C)$, onde C é o ponto de corte de interesse.

Vamos considerar dois casos: no primeiro a variável resposta tem distribuição normal e no segundo a variável resposta tem distribuição log-normal.

2.2 Modelo logístico com resposta normal

Vamos supor que T_i seja uma variável aleatória contínua com distribuição normal com média μ_i e variância σ^2 , $i = 1, \dots, n$. Portanto,

$$\pi_i = P(T_i > C) = P\left(Z_i > \frac{C - \mu_i}{\sigma}\right) = \Phi\left[\frac{\mu_i - C}{\sigma}\right],$$

onde Z_i é uma variável aleatória com distribuição normal padrão com função distribuição acumulada Φ . Supondo ainda que a relação entre o risco π_i e as variáveis preditoras $X_1, \dots, X_p$ possa ser descrita pelo modelo (2.5), temos:

$$\pi_i = \Phi\left[\frac{\mu_i - C}{\sigma}\right] = g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta}), \quad (2.6)$$

ou ainda,

$$g[\Phi(\gamma_i)] = \mathbf{x}_i^t \boldsymbol{\beta} = \eta_i, \quad (2.7)$$

onde $\gamma_i = (\mu_i - C)/\sigma$, $i = 1, \dots, n$.

Assumindo σ conhecido, esse modelo pertence à classe dos Modelos Lineares Generalizados com componente aleatório representado por um conjunto de variáveis aleatórias independentes com distribuição $N(\gamma_i, 1)$ e componente sistemático representado pela função de ligação composta dada por $g[\Phi(\cdot)]$ e preditor linear $\eta_i = \mathbf{x}_i^t \boldsymbol{\beta}$, $i = 1, \dots, n$. Tais modelos foram estudados por Thompson e Baker (1981) ampliando a classe de modelos que podem ser tratados utilizando a abordagem geral desenvolvida para os Modelos Lineares Generalizados.

2.3 Estimação para resposta normal

Nesta seção vamos apresentar o procedimento de estimação dos parâmetros do modelo (2.7) pelo método de máxima verossimilhança supondo que $T_1, \dots, T_n$ sejam n variáveis aleatórias independentes com distribuição $N(\mu_i, \sigma^2)$, $i = 1, \dots, n$, respectivamente. O logaritmo da função de verossimilhança é dado por

$$\log L(\boldsymbol{\beta}, \sigma; \mathbf{t}) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n \{t_i - \sigma\Phi^{-1}[g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})] - C\}^2. \quad (2.8)$$

As equações escore para σ e $\beta_1, \beta_2, \dots, \beta_p$ são

$$\frac{\partial \log L(\boldsymbol{\beta}, \sigma; \mathbf{t})}{\partial \sigma} = 0 \quad (2.9)$$

e

$$\frac{\partial \log L(\boldsymbol{\beta}, \sigma; \mathbf{t})}{\partial \beta_j} = 0, \quad j = 1, \dots, p. \quad (2.10)$$

Suissa e Blais (1995) propuseram que as soluções de (2.9) e (2.10) sejam obtidas da seguinte forma: estima-se σ mantendo $\boldsymbol{\beta}$ fixado, e então estima-se $\boldsymbol{\beta}$, fixando-se σ . Esses dois passos são alternados até a convergência, obtendo-se assim os estimadores de máxima verossimilhança para os parâmetros. Na estimação de $\boldsymbol{\beta}$, dado σ , pode ser utilizado o algoritmo de ponderação de Fisher [Rao, 1973], possibilitando, assim, que rotinas existentes em aplicativos estatísticos, como por exemplo, S-plus [Chambers e Hastie, 1993], GLIM ou SAS sejam utilizadas.

Fixando β , o estimador de máxima verossimilhança para σ é obtido resolvendo a equação (2.9). Temos então que

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n \{t_i - \sigma \Phi^{-1} [g^{-1}(\mathbf{x}_i^t \beta)] - C\}^2 + \frac{1}{\sigma^2} \sum_{i=1}^n \{t_i - \sigma \Phi^{-1} [g^{-1}(\mathbf{x}_i^t \beta)] - C\} \Phi^{-1} [g^{-1}(\mathbf{x}_i^t \beta)] = 0,$$

que também pode ser escrita como

$$-\frac{n}{\sigma} + \frac{\sum_{i=1}^n (t_i - C)^2}{\sigma^3} - \frac{\sum_{i=1}^n (t_i - C) \Phi^{-1} [g^{-1}(\mathbf{x}_i^t \beta)]}{\sigma^2} = 0,$$

ou ainda

$$n\sigma^2 + \sigma \sum_{i=1}^n (t_i - C) \Phi^{-1} [g^{-1}(\mathbf{x}_i^t \beta)] - \sum_{i=1}^n (t_i - C)^2 = 0. \quad (2.11)$$

Da equação (2.11) obtemos que o estimador de máxima verossimilhança de σ , dado β , é:

$$\hat{\sigma} = \frac{1}{2n} \left\{ - \sum_{i=1}^n (t_i - C) \Phi^{-1} [g^{-1}(\mathbf{x}_i^t \beta)] + \left[\left\{ \sum_{i=1}^n (t_i - C) \Phi^{-1} [g^{-1}(\mathbf{x}_i^t \beta)] \right\}^2 + 4n \sum_{i=1}^n (t_i - C)^2 \right]^{\frac{1}{2}} \right\}. \quad (2.12)$$

Considerando agora σ fixado, o estimador para β_j , $j = 1, \dots, p$ é obtido resolvendo a equação (2.10), ou seja,

$$\frac{1}{\sigma^2} \sum_{i=1}^n \{t_i - \sigma \Phi^{-1} [g^{-1}(\mathbf{x}_i^t \beta)] - C\} \sigma \frac{\partial \{\Phi^{-1} [g^{-1}(\mathbf{x}_i^t \beta)]\}}{\partial \beta_j} = 0. \quad (2.13)$$

Como Φ é uma função monótona e contínua em qualquer intervalo $[a, b]$, $a < b$, a derivada da inversa de Φ [Piskunov, 1983] é dada por

$$\frac{d\{\Phi^{-1}(v)\}}{dv} = \frac{1}{f(\Phi^{-1}(v))},$$

onde f é a função densidade de probabilidade de uma variável aleatória normal padrão. Então, (2.13) pode ser escrita como

$$\frac{1}{\sigma} \sum_{i=1}^n \left\{ \frac{t_i - C}{\sigma} - \Phi^{-1} [g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})] \right\} \sigma \frac{1}{f \{ \Phi^{-1} [g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})] \}} \frac{\partial [g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})]}{\partial \beta_j} = 0,$$

ou ainda como

$$\sum_{i=1}^n \left\{ \frac{t_i - C}{\sigma} - \Phi^{-1} [g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})] \right\} \frac{1}{f \{ \Phi^{-1} [g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})] \}} \frac{d [g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})]}{d(\mathbf{x}_i^t \boldsymbol{\beta})} \frac{\partial \mathbf{x}_i^t \boldsymbol{\beta}}{\partial \beta_j} = 0. \quad (2.14)$$

Como

$$g(\pi_i) = \log \left(\frac{\pi_i}{1 - \pi_i} \right) = \mathbf{x}_i^t \boldsymbol{\beta},$$

sua inversa é

$$g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta}) = \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}}},$$

e então

$$\begin{aligned} \frac{d [g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})]}{d(\mathbf{x}_i^t \boldsymbol{\beta})} &= \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}} (1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}}) - e^{\mathbf{x}_i^t \boldsymbol{\beta}} e^{\mathbf{x}_i^t \boldsymbol{\beta}}}{(1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}})^2} \\ &= \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}}}{(1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}})^2}, \end{aligned}$$

que é a função densidade de probabilidade logística. Assim, podemos escrever (2.14) como

$$\sum_{i=1}^n \left\{ \frac{t_i - C}{\sigma} - \Phi^{-1} [g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})] \right\} \frac{1}{f \{ \Phi^{-1} [g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})] \}} \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}}}{(1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}})^2} x_{ij} = 0. \quad (2.15)$$

Fazendo $t_i^* = (t_i - C)/\sigma$ e lembrando que $\gamma_i = \Phi^{-1} [g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})]$, reescrevemos (2.15) como

$$\sum_{i=1}^n (t_i^* - \gamma_i) \frac{d\gamma_i}{d\eta_i} x_{ij} = 0, \quad j = 1, \dots, p \quad (2.16)$$

com

$$\frac{d\gamma_i}{d\eta_i} = \frac{1}{f \{ \Phi^{-1} [g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})] \}} \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}}}{(1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}})^2}, \quad i = 1, \dots, n. \quad (2.17)$$

Sejam

$$U_j = \sum_{i=1}^n (t_i^* - \gamma_i) \frac{d\gamma_i}{d\eta_i} x_{ij}, \quad j = 1, \dots, p,$$

os elementos que definem o vetor escore $\mathbf{U} = (U_1, \dots, U_p)$ para $\boldsymbol{\beta}$. Temos que

$$\mathbf{U} = (\mathbf{L}\mathbf{X})^t (\mathbf{t}^* - \boldsymbol{\gamma}),$$

onde

$$\mathbf{L} = \begin{bmatrix} \frac{d\gamma_1}{d\eta_1} & 0 & \cdots & 0 \\ 0 & \frac{d\gamma_2}{d\eta_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{d\gamma_n}{d\eta_n} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1p} \\ 1 & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{np} \end{bmatrix},$$

$$\mathbf{t}^* = [t_1^* \ t_2^* \ \cdots \ t_n^*]^t \quad \text{e} \quad \boldsymbol{\gamma} = [\gamma_1 \ \gamma_2 \ \cdots \ \gamma_n]^t.$$

As equações escore $U_j = 0$, $j = 1, \dots, p$, não são lineares em β_j e têm que ser resolvidas por procedimentos iterativos. Para resolvê-las vamos adotar o método de ponderação de Fisher [Rao, 1973], geralmente utilizado no ajuste de Modelos Lineares Generalizados. Segundo Cordeiro (1986), para modelos com função de ligação composta, deve-se considerar no passo m do processo iterativo a matriz de planejamento dada por

$$\mathbf{X}^{*(m)} = \mathbf{L}^{(m)}\mathbf{X}$$

e variável dependente modificada definida por

$$\mathbf{z}^{(m)} = \mathbf{L}^{(m)}\boldsymbol{\eta}^{(m)} + (\mathbf{t}^* - \boldsymbol{\gamma}),$$

onde $\boldsymbol{\eta}^{(m)} = \mathbf{X}\boldsymbol{\beta}^{(m)}$. Este resultado está demonstrado no Apêndice A.

O procedimento de estimação de $\boldsymbol{\beta}$ e σ consiste, então, na combinação de dois procedimentos iterativos como descrito a seguir.

Passo 0: O processo iterativo é iniciado considerando a estimativa inicial $\boldsymbol{\beta}^{(0)}$ obtida por meio da regressão logística usual.

Passo 1: Obtido $\boldsymbol{\beta}^{(m)}$, $m = 0, 1, 2, \dots$, calcula-se $\sigma^{(m)}$ substituindo-se $\boldsymbol{\beta}^{(m)}$ em (2.12), isto

é, obtém-se

$$\sigma^{(m)} = \frac{1}{2n} \left\{ - \sum_{i=1}^n (t_i - C) \Phi^{-1} \left[g^{-1} \left(\eta_i^{(m)} \right) \right] + \left[\sum_{i=1}^n (t_i - C) \Phi^{-1} \left[g^{-1} \left(\eta_i^{(m)} \right) \right] \right]^2 + 4n \sum_{i=1}^n (t_i - C)^2 \right\}^{\frac{1}{2}},$$

onde $\eta_i^{(m)} = \sum_{j=1}^p x_{ij} \beta_j^{(m)}$, $i = 1, \dots, n$.

Passo 2: O procedimento de ponderação de Fisher é utilizado para obter a estimativa $\beta^{(m+1)}$, dado $\sigma^{(m)}$, $m = 0, 1, 2, \dots$. Temos então neste passo um outro processo iterativo que pode ser resumido como:

Passo 2.1: Seja $\beta_{(k)}^{(m)} = \beta^{(m)}$. Para $k = 1, 2, \dots$ calcula-se a variável dependente modificada $z_{i(k)}^{(m)}$ e a matriz de planejamento $(x_{ij}^*)_{(k)}^{(m)}$, $i = 1, \dots, n$ e $j = 1, \dots, p$, sendo

$$z_{i(k)}^{(m)} = \left(\frac{d\gamma_i}{d\eta_i} \right)_{(k)}^{(m)} \eta_{i(k)}^{(m)} + t_i^* - \gamma_{i(k)}^{(m)},$$

$$\text{onde } \left(\frac{d\gamma_i}{d\eta_i} \right)_{(k)}^{(m)} = \frac{1}{f \left\{ \Phi^{-1} \left[g^{-1} \left(\eta_{i(k)}^{(m)} \right) \right] \right\}} \frac{e^{\eta_{i(k)}^{(m)}}}{\left(1 + e^{\eta_{i(k)}^{(m)}} \right)^2}, \quad \eta_{i(k)}^{(m)} = \sum_{j=1}^p x_{ij} \beta_j^{(m)},$$

$$t_i^* = \frac{t_i - C}{\sigma^{(m)}}, \quad \gamma_{i(k)}^{(m)} = \Phi^{-1} \left[g^{-1} \left(\eta_{i(k)}^{(m)} \right) \right],$$

e

$$(x_{ij}^*)_{(k)}^{(m)} = \left(\frac{d\gamma_i}{d\eta_i} \right)_{(k)}^{(m)} x_{ij}.$$

Passo 2.2: O vetor $\beta_{(k+1)}^{(m)}$ é obtido da regressão de mínimos quadrados de $z_{i(k)}^{(m)}$ em $(x_{ij}^*)_{(k)}^{(m)}$, $i = 1, \dots, n$ e $j = 1, \dots, p$, ou seja,

$$\beta_{(k+1)}^{(m)} = \left(\mathbf{X}_{(k)}^{*(m)T} \mathbf{X}_{(k)}^{*(m)} \right)^{-1} \mathbf{X}_{(k)}^{*(m)T} \mathbf{z}_{(k)}^{(m)},$$

$$\text{onde } \mathbf{X}_{(k)}^{*(m)} = (x_{ij}^*)_{(k)}^{(m)} \text{ e } \mathbf{z}_{(k)}^{(m)} = \left(z_{1(k)}^{(m)}, \dots, z_{n(k)}^{(m)} \right)^T.$$

Passo 2.3: Os passos 2.1 e 2.2 são repetidos até que

$$\sum_{j=1}^p \left(\frac{\beta_{j(k)}^{(m)} - \beta_{j(k+1)}^{(m)}}{\beta_{j(k+1)}^{(m)}} \right)^2$$

seja suficientemente pequeno. Na convergência $\beta^{(m+1)} = \beta_{(k+1)}^{(m)}$.

Passo 3: Os passos 1 e 2 são repetidos atualizando-se o valor de $\beta^{(m)}$, até que o módulo da diferença $\sigma^{(m+1)} - \sigma^{(m)}$ seja suficientemente pequeno.

Se σ for conhecido, a matriz de variância-covariância assintótica do vetor $\hat{\beta}$ é dada pela inversa da matriz de informação esperada de Fisher:

$$\text{Var}(\hat{\beta}) = \mathfrak{F}^{-1} = [(\mathbf{L}\mathbf{X})^t \mathbf{L}\mathbf{X}]^{-1}$$

e é estimada por

$$\widehat{\text{Var}}(\hat{\beta}) = \widehat{\mathfrak{F}}^{-1} = [(\widehat{\mathbf{L}}\mathbf{X})^t \widehat{\mathbf{L}}\mathbf{X}]^{-1},$$

onde $\widehat{\mathbf{L}}$ é uma matriz diagonal, cujo i -ésimo elemento é dado por

$$\frac{\widehat{d\gamma_i}}{d\eta_i} = \frac{1}{f\{\Phi^{-1}[g^{-1}(\hat{\eta}_i)]\}} \frac{e^{\hat{\eta}_i}}{(1 + e^{\hat{\eta}_i})^2},$$

$$\text{e } \hat{\eta}_i = \sum_{j=1}^p x_{ij} \hat{\beta}_j, \quad i = 1, \dots, n.$$

2.4 Teste de Hipóteses para resposta normal

A hipótese para testar se um modelo mais simples (com menos parâmetros) é adequado, ou seja, verificar se algumas variáveis preditoras, $X_{q+1}, \dots, X_p$ digamos, podem ser retiradas do modelo, é dada por

$$H_0: \beta_{q+1} = \beta_{q+2} = \dots = \beta_p = 0$$

contra

$$H_a: \text{pelo menos um dos parâmetros } \beta_{q+1}, \beta_{q+2}, \dots, \beta_p \text{ não é nulo.}$$

A estatística do teste da razão de verossimilhanças para testar H_0 é

$$\xi_{\text{RV}} = 2\{\log L(\hat{\beta}, \hat{\sigma}; \mathbf{t}) - \log L(\hat{\beta}^0, \hat{\sigma}^0; \mathbf{t})\},$$

onde $\hat{\beta}^0$ e $\hat{\sigma}^0$ são, respectivamente, as estimativas de máxima verossimilhança de β e σ sob H_0 , isto é, as estimativas obtidas no ajuste do modelo reduzido, no qual não estão incluídas as variáveis preditoras $X_{q+1}, \dots, X_p$. Substituindo $\hat{\beta}^0$ e $\hat{\sigma}^0$ em (2.8) obtemos $\log L(\hat{\beta}^0, \hat{\sigma}^0; \mathbf{t})$, o logaritmo da função de verossimilhança sob a hipótese nula H_0 . O termo $\log L(\hat{\beta}, \hat{\sigma}; \mathbf{t})$ é obtido substituindo em (2.8) as estimativas de máxima verossimilhança de β e σ , $\hat{\beta}$ e $\hat{\sigma}$, resultantes do ajuste do modelo completo, ou seja, com todas as variáveis preditoras. Assintoticamente, sob H_0 , a estatística ξ_{RV} segue distribuição Qui-quadrado com $p - q$ graus de liberdade.

2.5 Modelo logístico com resposta log-normal

Segundo Suissa e Blais (1995) muitas variáveis aleatórias observadas em pesquisas clínicas, apresentam distribuição assimétrica e assim, a suposição de normalidade não é adequada. A distribuição log-normal pode ser considerada nestes casos. Em nossa aplicação de interesse foi observado ser razoável supor-se que as concentrações dos poluentes NO_2 , SO_2 , CO e O_3 seguem essa distribuição. Assumindo então que T_i seja uma variável aleatória positiva contínua com distribuição log-normal com parâmetros μ_i e σ^2 , e consequentemente $\log T_i$ com distribuição normal com média μ_i e variância σ^2 , $i = 1, \dots, n$, temos que o risco π_i é dado por

$$\begin{aligned}\pi_i &= P(T_i > C) = P(\log T_i > \log C) \\ &= P\left(Z_i > \frac{\log C - \mu_i}{\sigma}\right) \\ &= \Phi\left[\frac{\mu_i - \log C}{\sigma}\right],\end{aligned}$$

onde Z_i é uma variável aleatória com distribuição normal padrão com função distribuição acumulada Φ . Supondo ainda que a relação entre o risco π_i e as variáveis preditoras $X_1, \dots, X_p$ possa ser descrita pelo modelo (2.5), temos

$$g[\Phi(\gamma_i)] = \mathbf{x}_i^t \beta, \quad (2.18)$$

com $\gamma_i = (\mu_i - \log C)/\sigma$, $i = 1, \dots, n$.

A extensão dos resultados de estimação e teste de hipóteses apresentados nas Seções 2.3 e 2.4 para o modelo (2.18) é direta, substituindo T_i por $\log T_i$, $i = 1, \dots, n$, e a constante C por $\log C$.

2.6 Simulação

Para estudar a eficiência dos estimadores de $\beta_1, \dots, \beta_p$ obtidos pelo método proposto, em relação aos estimadores desses parâmetros em um modelo logístico usual, principalmente em pequenas amostras, Suissa e Blais (1995) realizaram um estudo de simulação.

No estudo foram utilizados três tipos de função de ligação: *logito*, logarítmica e identidade. Com o objetivo de apontar quais as situações mais favoráveis à aplicação desse método, reproduzimos aqui os resultados obtidos com a função de ligação *logito*.

Foi considerada apenas uma variável explicativa X que assumiu os valores 0, 1, 2, 3 e 4. Foram considerados tamanhos de amostra de $n = 20, 30, 50$ e 100 observações igualmente distribuídas nos cinco valores de X e geradas 100 amostras de n observações de uma variável aleatória Y com distribuição normal com média $\mu = \sigma\Phi^{-1}[g^{-1}(\beta_0 + \beta_1 X)] + C$ e variância $\sigma^2 = 1$. Foram utilizados como verdadeiros valores de $\beta_0 : -0,800; -1,891; -2,468$ ou $-2,937$ e $\beta_1 = 0,400$.

Diferentes valores para C foram fixados, de forma que, no ponto $\bar{X} = 2$, ocorresse $C = E(Y)$, $C = E(Y) + 0,67$, $C = E(Y) + 1$ ou $C = E(Y) + 1,25$.

Os resultados obtidos são apresentados na Tabela 2.1. Observamos que o modelo logístico usual não pode ser ajustado para todas as amostras geradas. Isto ocorreu ou por não terem sido observados nessas amostras eventos $[Y_i > C]$, ou por uma distribuição não apropriada desses eventos nos diferentes valores de X . Este problema não ocorreu nos ajustes dos modelos com resposta contínua. As médias e variâncias das 100 estimativas obtidas para cada combinação de n e $C - E(Y)$ são também apresentadas na Tabela 2.1. As estimativas pontuais obtidas pelos dois métodos são, em média, próximas. Entretanto, a variabilidade das estimativas obtidas no ajuste do modelo com resposta contínua é uniformemente menor do que a observada no ajuste do modelo logístico usual.

À medida que n aumenta e a constante C se distancia da média de Y , mais eficientes são as estimativas obtidas no modelo com resposta contínua em relação às obtidas no modelo logístico usual.

Resultados similares foram obtidos pelos autores para as funções de ligação logarítmica e identidade.

Tabela 2.1 *Resultados obtidos no estudo de simulação, para diferentes tamanhos de amostra (n), distâncias $C - E(Y)$ e valores dos parâmetros, adotando o modelo logístico usual e o modelo com resposta contínua (função de ligação logito)*

n	$C - E(Y)$	Parâmetros		Usual			Contínua		Eficiência ²
				N^1	Estimativas		Estimativas		
					Média	Variância	Média	Variância	
20	0	β_0	-0,800	100	-0,831	1,004	-0,950	0,698	0,717
		β_1	0,400	100	0,390	0,191	0,424	0,114	0,600
	0,67	β_0	-1,891	96	-2,204	2,253	-2,201	1,148	0,529
		β_1	0,400	96	0,435	0,337	0,438	0,145	0,433
	1	β_0	-2,468	90	2,892	3,197	-2,875	1,564	0,054
		β_1	0,400	90	0,459	0,454	0,445	0,175	0,387
30	0	β_0	-0,800	100	-0,824	0,642	-0,899	0,339	0,543
		β_1	0,400	100	0,386	0,100	0,417	0,061	0,612
	0,67	β_0	-1,891	100	-2,147	1,456	-2,080	0,504	0,355
		β_1	0,400	100	0,445	0,204	0,424	0,071	0,347
	1	β_0	-2,468	98	-2,893	2,234	-2,713	0,665	0,300
		β_1	0,400	98	0,469	0,282	0,427	0,082	0,288
50	0	β_0	-0,800	100	-0,834	0,336	-0,886	0,195	0,600
		β_1	0,400	100	0,394	0,060	0,406	0,037	0,617
	0,67	β_0	-1,891	100	-2,079	0,823	-2,041	0,301	0,377
		β_1	0,400	100	0,432	0,107	0,413	0,044	0,409
	1	β_0	-2,468	99	-2,840	1,603	-2,658	0,401	0,251
		β_1	0,400	99	0,458	0,176	0,417	0,051	0,286
	1,25	β_0	-2,937	98	-3,431	2,290	-3,161	0,506	0,219
		β_1	0,400	98	0,474	0,279	0,419	0,058	0,205
100	0	β_0	-0,800	100	-0,833	0,136	-0,847	0,091	0,680
		β_1	0,400	100	0,402	0,027	0,402	0,017	0,630
	0,67	β_0	-1,891	100	-1,995	0,293	-1,971	0,136	0,469
		β_1	0,400	100	0,413	0,043	0,406	0,019	0,441
	1	β_0	-2,468	100	-2,675	0,655	-2,568	0,179	0,271
		β_1	0,400	100	0,436	0,075	0,407	0,022	0,289
	1,25	β_0	-2,937	100	-3,169	1,071	-3,054	0,223	0,210
		β_1	0,400	100	0,433	0,127	0,408	0,025	0,196

¹ N = número total de amostras utilizadas no ajuste do modelo logístico usual, em um total de 100 amostras

² Eficiência = Erro quadrático médio das estimativas dos parâmetros no caso contínuo dividida pelo erro quadrático médio das estimativas dos parâmetros do modelo logístico usual

Capítulo 3

Aplicação

3.1 Introdução

Neste capítulo vamos adotar o modelo logístico usual e o modelo com resposta contínua com o objetivo de modelar o risco de um poluente atmosférico assumir valores acima de concentrações pré-especificadas (pontos de corte) na cidade de São Paulo. Antes de construirmos e ajustarmos os modelos, vamos fazer algumas considerações sobre o ar que respiramos.

De uma forma simplificada, poluentes atmosféricos são componentes resultantes de processos de combustão e de atividades industriais, que, quando lançados à atmosfera, são agressivos à vida de seres humanos, animais e vegetais. Os poluentes regulamentados na lei brasileira são (CETESB, 2001):

- ✓ Dióxido de Enxofre (SO_2): gás resultante da transformação do enxofre no processo de combustão do petróleo e carvão mineral;
- ✓ Monóxido de Carbono (CO): resultado da queima incompleta de qualquer combustível, principalmente por parte de veículos;
- ✓ Óxidos de Nitrogênio (NO_x): gases liberados para a atmosfera pelo processo de combustão, principalmente por veículos;
- ✓ Oxidantes Fotoquímicos: combustíveis não queimados e óxidos de nitrogênio reagem na atmosfera, principalmente quando ativados pela luz solar, formando um conjunto de gases agressivos chamados oxidantes fotoquímicos, sendo o mais importante dentre eles

o ozônio (O_3), que é utilizado como indicador da presença de oxidantes fotoquímicos na atmosfera e

✓ Material Particulado: grande classe de poluentes constituída por poeiras, fumaças e todo tipo de material sólido e líquido que, devido a seu pequeno tamanho, se mantém suspenso na atmosfera. Por seus efeitos danosos à saúde, existe legislação brasileira para partículas totais em suspensão, partículas com diâmetros menores que $10\ \mu m$, denominadas Partículas Inaláveis (PM_{10}), e as fumaças.

Os padrões nacionais de qualidade do ar foram estabelecidos pelo IBAMA – Instituto Brasileiro de Meio Ambiente e aprovados pelo CONAMA – Conselho Nacional de Meio Ambiente. Os padrões primários definem legalmente o nível máximo para a concentração de um poluente na atmosfera, que garanta a proteção da saúde da população. Os secundários estabelecem níveis de concentrações abaixo dos quais mínimo dano é causado à saúde, à flora, aos materiais e ao meio ambiente em geral, podendo ser interpretados como níveis máximos desejáveis. Esses padrões são apresentados na Tabela 3.1.

Tabela 3.1 *Padrões Primários e Secundários de Qualidade do Ar referentes às concentrações dos poluentes, estabelecidos pelo CONAMA*

Poluente	Tempo de Amostragem	Padrão Primário	Padrão Secundário
NO_2	1 hora ⁽¹⁾	$320\ \mu g/m^3$	$190\ \mu g/m^3$
	MAA ⁽²⁾	$100\ \mu g/m^3$	$100\ \mu g/m^3$
O_3	1 hora ⁽¹⁾	$160\ \mu g/m^3$	$160\ \mu g/m^3$
SO_2	24 horas ⁽¹⁾	$365\ \mu g/m^3$	$100\ \mu g/m^3$
	MAA ⁽²⁾	$80\ \mu g/m^3$	$40\ \mu g/m^3$
PM_{10}	24 horas ⁽¹⁾	$150\ \mu g/m^3$	$150\ \mu g/m^3$
	MAA ⁽²⁾	$50\ \mu g/m^3$	$50\ \mu g/m^3$
CO	1 hora ⁽¹⁾	35 ppm	35 ppm
	8 horas ⁽¹⁾	9 ppm	9 ppm

⁽¹⁾ Não deve ser excedido mais que uma vez ao ano

⁽²⁾ Média aritmética anual

O monitoramento do ar do estado de São Paulo é feito pela Companhia de Tecnologia de Saneamento Ambiental (CETESB), por meio de duas redes: uma automática e outra

manual. A rede automática é constituída por 29 estações fixas (ver Figura 3.1) e 2 estações móveis. Das fixas, 23 estão na região da Grande São Paulo, das quais 14 na cidade de São Paulo, 2 em Cubatão, 1 em Paulínea, 1 em Campinas, 1 em Sorocaba e 1 em São José dos Campos. A rede manual é composta por 19 estações na região da Grande São Paulo e Cubatão, e 19 no interior. Nem todas as estações medem todos os poluentes. Apresentamos na Tabela B.1 a configuração da rede automática com os parâmetros monitorados pelas estações. As medições são feitas de forma contínua, gerando registros diários, dentre outros, dos poluentes SO_2 , CO , PM_{10} , O_3 e NO_2 . Os registros são feitos da seguinte forma: média de 24 horas para SO_2 e PM_{10} , maior média móvel de 8 horas para CO e pico horário de 24 horas para NO_2 e O_3 .

Fonte: CETESB (2001)

Figura 3.1 Mapa de localização das 29 estações fixas de monitoramento do ar operadas pela CETESB no estado de São Paulo

Motivados pelos padrões primários e secundários apresentados na Tabela 3.1, vamos modelar o risco da concentração do NO_2 ser superior a $100 \mu\text{g}/\text{m}^3$.

Na escolha das variáveis preditoras a serem colocadas nos modelos, consideramos que a poluição atmosférica em São Paulo é predominantemente gerada pela frota de veículos que trafega pela cidade, e a dispersão desses poluentes é desfavorecida pelas condições climáticas, principalmente durante o outono e o inverno, quando frequentemente ocorrem inversões térmicas. Sabe-se também que a ocorrência de vento é importante para a dispersão dos poluentes. Assim, as seguintes variáveis climáticas e de controle temporal foram consideradas:

Variáveis Climáticas:

- *temperatura*: Temperatura mínima diária ($^{\circ}\text{C}$);
- *umidade*: Média diária da umidade relativa do ar (%);
- *precipitação*: Precipitação pluviométrica diária em *mm*. Dias em que ocorre precipitação pluviométrica menor que 0,1 *mm* não foram considerados;
- *chuva*: variável indicadora de *precipitação* definida como:

$$chuva = \begin{cases} 1, & \text{se } precipitação \geq 0,1 \text{ mm;} \\ 0, & \text{se } precipitação = 0 \text{ mm.} \end{cases}$$

- *velocidade do vento*: Velocidade diária máxima do vento em *km/h*.

A partir do primeiro e terceiro quartis da *velocidade do vento*, dados, respectivamente, por 18,520 *km/h* e 24,076 *km/h*, foi criada a variável *vento* com três categorias:

$$\begin{cases} vento \text{ fraco,} & \text{se } velocidade \text{ do vento} \leq 18,520 \text{ km/h;} \\ vento \text{ médio,} & \text{se } 18,520 \text{ km/h} < velocidade \text{ do vento} \leq 24,076 \text{ km/h e} \\ vento \text{ forte,} & \text{se } velocidade \text{ do vento} > 24,076 \text{ km/h.} \end{cases}$$

Foram criadas também duas variáveis indicadoras, uma para *vento médio* e outra para *vento forte* definidas como:

$$vento \text{ médio} = \begin{cases} 1, & \text{se } 18,520 \text{ km/h} < velocidade \text{ do vento} \leq 24,076 \text{ km/h;} \\ 0, & \text{caso contrário.} \end{cases}$$

$$vento \text{ forte} = \begin{cases} 1, & \text{se } velocidade \text{ do vento} > 24,076 \text{ km/h;} \\ 0, & \text{caso contrário.} \end{cases}$$

Variáveis de Controle Temporal:

- Foram criadas 6 variáveis indicadoras para os dias da semana, considerando *domingo* como a categoria de referência. Por exemplo, para segunda-feira temos a variável

$$segunda = \begin{cases} 1, & \text{se o dia da semana for segunda-feira;} \\ 0, & \text{caso contrário.} \end{cases}$$

- *outono-inverno*: variável indicadora de *estação do ano* definida como:

$$outono-inverno = \begin{cases} 1, & \text{se a estação do ano for Outono ou Inverno;} \\ 0, & \text{se a estação do ano for Primavera ou Verão.} \end{cases}$$

Nosso banco de dados é constituído por valores diários de NO₂ e das variáveis climáticas e de controle temporal observados na cidade de São Paulo durante os anos de 1998 e 1999.

As fontes dos dados utilizados foram:

✓ Laboratório de Poluição Atmosférica Experimental (LPAE) da Faculdade de Medicina da Universidade de São Paulo (FMUSP) que nos cedeu os dados referentes aos poluentes, calculados a partir de informações da Companhia de Tecnologia de Saneamento Ambiental (CETESB);

✓ Divisão Operacional - Serviço Regional de Proteção ao Vão (SRPV) do Estado de São Paulo, localizado no Aeroporto de Congonhas que nos cedeu os dados referentes às variáveis climáticas.

Uma análise descritiva dos dados é apresentada na Seção 3.2. Na Seção 3.3 ajustamos os modelos logísticos usual e com resposta contínua para o poluente NO₂.

3.2 Análise Descritiva

Na Tabela 3.2 observamos que maiores valores médios das concentrações ocorreram no ano de 1998 para os poluentes NO₂, PM₁₀ e CO, e no ano de 1999 para os poluentes O₃ e SO₂. Observamos que no ano de 1998 a média do NO₂ ultrapassou o padrão primário anual e a média do PM₁₀ foi próxima do valor máximo anual estabelecido.

Na Figura B.1 apresentamos séries cronológicas para as concentrações dos poluentes

Tabela 3.2 *Médias (Erros padrão) das concentrações dos poluentes nos anos de 1998 e 1999*

Ano	NO ₂	O ₃	SO ₂	PM ₁₀	CO
1998	100,06 (39,86)	66,46 (47,32)	13,31 (7,40)	49,38 (20,07)	3,21 (1,43)
1999	90,30 (38,54)	79,85 (42,08)	16,14 (8,60)	47,20 (22,43)	2,63 (1,29)
Total	95,18 (39,48)	73,15 (45,25)	14,73 (8,14)	48,29 (21,29)	2,92 (1,39)

NO₂, O₃, SO₂, PM₁₀ e CO. Observam-se picos de alta concentração dos poluentes, principalmente nas estações do ano Outono e Inverno, sendo os poluentes NO₂ e O₃ os que apresentaram o maior número de concentrações superiores ao padrão secundário.

Na Tabela 3.3 observamos que pelo menos 25% das concentrações diárias de NO₂ e PM₁₀ estão, respectivamente, acima de 114,96 $\mu\text{g}/\text{m}^3$ e acima de 55,79 $\mu\text{g}/\text{m}^3$, que correspondem a concentrações próximas dos padrões primários para médias anuais desses poluentes.

Tabela 3.3 *Percentis observados das concentrações dos poluentes*

Poluente	5%	10%	25%	50%	75%	90%	95%
NO ₂	46,44	52,35	66,00	88,42	114,96	147,13	172,84
O ₃	24,87	31,54	42,66	61,98	92,79	127,52	147,01
SO ₂	5,34	6,44	9,13	12,94	18,08	25,55	30,89
PM ₁₀	25,78	28,60	33,60	42,82	55,79	75,41	93,64
CO	1,30	1,51	1,95	2,63	3,47	4,75	5,92

Na Tabela 3.4 para todos os poluentes observamos, em geral, menores valores médios das concentrações em dias com *chuva*. Os valores médios das concentrações dos poluentes NO₂, SO₂, PM₁₀ e CO diminuem à medida que a *velocidade do vento* aumenta; já para o poluente O₃ acontece o inverso.

As variáveis climáticas *temperatura*, *umidade*, *precipitação* e *velocidade do vento* têm valores médios bastante próximos nos dois anos, apresentando maior variabilidade no

Tabela 3.4 *Medidas descritivas para as variáveis Temperatura e Umidade e concentrações dos poluentes NO₂, O₃, SO₂, PM₁₀ e CO nas diferentes combinações de Chuva e Vento*

	Medidas Descritivas	Vento	Sem chuva			Com chuva		
			Fraco	Médio	Forte	Fraco	Médio	Forte
Temperatura	Média		16,11	16,94	17,13	18,11	17,39	18,17
	Desvio Padrão		3,09	3,18	3,55	3,21	3,38	3,03
	Mínimo		9,80	9,60	9,80	6,70	7,20	8,10
	Máximo		24,20	24,80	24,40	22,60	23,00	24,60
Umidade	Média		69,73	71,91	70,03	79,36	79,54	77,70
	Desvio Padrão		9,74	8,50	8,38	6,96	6,95	6,47
	Mínimo		35,00	37,00	42,00	61,00	54,00	58,00
	Máximo		85,00	87,00	88,00	92,00	91,00	91,00
NO ₂	Média		115,95	98,00	81,34	87,59	85,19	83,69
	Desvio Padrão		41,76	40,95	33,06	34,34	30,75	32,10
	Mínimo		39,18	38,32	27,53	31,71	33,13	32,06
	Máximo		257,60	234,00	190,16	303,00	189,51	181,66
O ₃	Média		84,57	87,69	90,26	51,69	56,98	58,40
	Desvio Padrão		43,24	61,24	46,80	29,07	28,75	26,27
	Mínimo		21,87	17,31	23,27	10,18	14,52	15,65
	Máximo		376,87	389,45	280,48	180,56	133,15	152,56
SO ₂	Média		20,54	15,53	12,81	11,26	11,19	10,70
	Desvio Padrão		9,94	7,25	5,50	5,12	5,14	4,63
	Mínimo		5,30	3,06	4,63	2,90	3,10	3,69
	Máximo		71,47	43,25	29,55	32,54	25,57	26,07
PM ₁₀	Média		63,34	50,51	43,62	38,85	39,35	38,17
	Desvio Padrão		25,04	19,99	14,73	12,61	13,85	11,96
	Mínimo		22,84	26,28	22,63	18,68	19,29	20,23
	Máximo		157,27	140,14	110,56	99,86	102,19	97,66
CO	Média		3,43	2,89	2,14	2,88	2,74	2,89
	Desvio Padrão		1,55	1,55	0,97	0,88	0,94	1,25
	Mínimo		0,90	1,03	0,73	0,79	1,14	0,54
	Máximo		8,81	8,06	6,10	6,38	5,87	8,41
Total	675		200	129	74	107	82	83

ano de 1999, conforme é observado na Tabela 3.5. Na Figura B.2 apresentamos as séries cronológicas dessas variáveis.

Na Tabela 3.6 observamos que nos anos de estudo no *outono-inverno* ocorrem com maior frequência dias sem *chuva* e com *vento fraco*; já na *primavera-verão* ocorrem mais frequen-

Tabela 3.5 *Médias (Erros padrão) das variáveis climáticas*

Ano	Temperatura (°C)	Umidade ¹ (%)	Precipitação ² (mm)	Velocidade do vento (km/h)
1998	17,35 (3,07)	75,68 (8,73)	3,78 (7,90)	21,53 (6,08)
1999	16,75 (3,44)	72,88 (9,42)	3,38 (9,18)	21,40 (6,42)
Total	17,05 (3,27)	74,28 (9,18)	3,58 (8,56)	21,47 (6,25)

¹ $N = 728$ de um total de 730 observações² $N = 676$ de um total de 730 observações

temente os dias com *chuva* e com *vento forte*. Maiores valores médios de *temperatura* e *umidade* ocorrem em dias com *chuva* (Tabela 3.4).

Tabela 3.6 *Distribuição conjunta de frequências das variáveis climáticas categorizadas nas estações do ano Primavera-Verão e Outono-Inverno*

Estação do ano	Vento	Sem chuva			Com chuva			Total
		Fraco	Médio	Forte	Fraco	Médio	Forte	
Primavera-Verão		42	59	54	62	48	67	332
Outono-Inverno		158	70	20	45	34	16	343
Total		200	129	74	107	82	83	675

A Figura B.3 sugere que o comportamento do NO_2 em função da *temperatura* depende da *umidade*: parece haver uma tendência linear positiva entre essas variáveis em valores altos de *umidade* ($> 75\%$); para *umidade* menor que 75% a concentração do NO_2 cresce até uma *temperatura* aproximadamente igual a 17°C apresentando a partir desse valor, um leve decréscimo.

Observamos na Figura B.4 um crescimento da concentração diária do NO_2 de segunda até quarta ou quinta-feira, passando então a decrescer e atingir valor mínimo no *domingo*.

3.3 Ajuste do Modelo

Um estudo das distribuições das concentrações dos poluentes, por meio do teste Anderson-Darling [Stephens, 1974] (Tabela 3.7) e construção de gráficos de probabilidade (Figura B.5) mostrou ser razoável assumir-se para essas concentrações distribuições log-normal, exceto para o poluente PM₁₀.

Tabela 3.7 *Níveis descritivos do teste de Anderson-Darling para a normalidade do log(concentrações dos poluentes)*

NO ₂	O ₃	SO ₂	PM ₁₀	CO
0,820	0,385	0,832	0,000	0,112

Considerando como variável resposta a concentração de dióxido de nitrogênio (NO₂) dicotomizada da seguinte forma:

$$\begin{cases} 1, & \text{se a concentração de NO}_2 > 100 \mu\text{g}/\text{m}^3 \\ 0, & \text{caso contrário} \end{cases},$$

foram ajustados os modelos logístico usual e com resposta contínua incluindo as variáveis preditoras *temperatura*, *umidade*, as 6 indicadoras para dia da semana, *outono-inverno*, *chuva*, *vento médio* e *vento forte*. Também foram incluídos no modelo termos de interação para *temperatura* e *umidade* e as interações de 1ª e 2ª ordem entre as variáveis indicadoras de dia da semana, *outono-inverno*, *chuva*, *vento médio* e *vento forte*. Um resumo dos resultados obtidos na ajuste dos modelos logístico usual e com resposta contínua é apresentado na Tabela B.2.

Observamos que as estimativas dos coeficientes obtidas pelos dois métodos são, em geral, próximas, mas o modelo logístico com resposta contínua produz erros padrão uniformemente menores.

Na tentativa de reduzir o número de parâmetros do modelo logístico com resposta contínua foram realizados testes de hipóteses da razão de verossimilhanças, conforme apresentado na Seção 2.4, sempre retirando do modelo os efeitos cujo teste de nulidade produziu maior nível descritivo. As interações não significantes e respectivos níveis descritivos são apresentadas na Tabela B.3. Os demais efeitos foram significantes a um nível descritivo $p \leq 0,046$.

O modelo ajustado resultante após a seleção de variáveis é apresentado na Tabela 3.8. Nessa tabela também são apresentadas as estimativas dos coeficientes e respectivos erros padrão obtidos no ajuste de um modelo logístico usual com os mesmos efeitos. Notamos que, com exceção do intercepto as estimativas dos coeficientes obtidas no ajuste dos dois modelos são próximas, porém as estimativas dos erros padrão são uniformemente menores no modelo logístico com resposta contínua. Os modelos produziram riscos previstos próximos.

Tabela 3.8 *Modelos finais usual e com resposta contínua ajustados*

Efeitos	Logística Usual		Logística com resposta contínua		
	Estimativa	Erro padrão	Estimativa	Erro padrão	$\hat{\sigma}$
<i>intercepto</i>	18,269	5,322	8,914	3,405	0,304
<i>temperatura</i>	-0,691	0,299	-0,265	0,194	
<i>umidade</i>	-0,296	0,073	-0,176	0,046	
<i>segunda</i>	2,182	0,408	2,211	0,269	
<i>terça</i>	2,205	0,406	2,283	0,267	
<i>quarta</i>	2,241	0,410	2,425	0,269	
<i>quinta</i>	1,962	0,417	2,412	0,274	
<i>sexta</i>	1,731	0,415	2,071	0,272	
<i>sábado</i>	0,785	0,421	0,969	0,276	
<i>outono-inverno</i>	0,204	0,234	0,553	0,166	
<i>chuva</i>	-0,849	0,327	-0,630	0,232	
<i>vento médio</i>	-1,231	0,281	-0,812	0,198	
<i>vento forte</i>	-2,250	0,379	-1,856	0,256	
<i>temperatura*umidade</i>	0,011	0,004	0,005	0,003	
<i>chuva*vento médio</i>	0,981	0,466	0,673	0,327	
<i>chuva*vento forte</i>	2,081	0,509	1,312	0,359	

O modelo ajustado indica que o risco da concentração de NO_2 ser maior que $100 \mu\text{g}/\text{m}^3$ é influenciado pelas variáveis preditoras *temperatura*, *umidade*, dias da semana, *estação do ano*, ocorrência ou não de *chuva* e *velocidade do vento*. Além disso, o efeito do *vento* depende da ocorrência ou não de *chuva* e o efeito da *temperatura* depende da *umidade*.

Com o propósito de examinar o ajuste do modelo adotamos duas técnicas informais de diagnóstico descritas em Neter *et al.* (1996).

Na primeira agrupamos os riscos previstos pelo modelo ($\hat{\pi}_i$, $i = 1, \dots, n$) em 10 intervalos.

los de classe de tal forma que cada classe contenha valores de risco similares e aproximadamente o mesmo número de casos. O ponto médio de cada classe, $\bar{\pi}_i$, foi então calculado. Calculamos também a proporção de respostas com valor igual a “1” (p_j , $j = 1, \dots, 10$) em cada classe (Tabela B.4) e construímos o diagrama de dispersão de $\bar{\pi}_i \times p_j$, $j = 1, \dots, 10$, apresentado na Figura 3.2.

Figura 3.2 *Gráfico da proporção estimada versus ponto médio*

Esse gráfico sugere que o ajuste do modelo logístico com resposta contínua é adequada, pois os pontos são dispostos em torno de reta $y = x$.

Na segunda, considerando os mesmos intervalos de classe adotados para a construção da Figura 3.2, foram obtidos os valores observados de respostas “0” e “1” e os ajustados pelo modelo. Os resultados obtidos estão na Tabela B.5. O valor observado da estatística qui-quadrado do teste é $\chi^2 = 9,720$, com nível descritivo $p = 0,285$, indicando a não rejeição da hipótese de que o modelo logístico com resposta contínua seja apropriado.

3.4 Riscos Previstos

Nas Tabelas 3.9 a 3.12 apresentamos os riscos médios de ocorrência de concentrações de NO_2 maiores que $100 \mu\text{g}/\text{m}^3$ previstos pelo modelo logístico com resposta contínua, para todas as combinações dos dias da semana, *estação do ano*, *chuva* e *vento*. São também

apresentadas as estimativas dos erros padrão [Neter *et al.*, 1996]. Embora as variáveis *temperatura* e *umidade* tenham sido consideradas de forma contínua no modelo, para facilitar a apresentação dos resultados, consideramo-as como categorizadas, cada uma delas tendo duas categorias definidas pelas suas medianas: 17°C para *temperatura* e 75% para *umidade*.

Dentre os resultados apresentados destaca-se a ocorrência de riscos médios previstos bastante altos, acima de 0,70 de segunda a sexta-feira, no *outono-inverno* e *primavera-verão*, sem *chuva*, *vento fraco* e *umidade* baixa. Nestas condições e com *temperatura* alta, os riscos observados são também elevados nos finais de semana (0,40 no *domingo* e 0,64 no *sábado* para o *outono-inverno* e 0,51 no *sábado* para a *primavera-verão*).

Dias sem *chuva* e *vento fraco* são bastante desfavoráveis à dispersão do poluente NO₂. Os riscos observados nestas condições, a cada dia da semana, nas diferentes combinações de *temperatura*, *umidade*, *estação do ano* são representados na Figura 3.3. Esta figura nos permite visualizar facilmente que, na ausência de *chuva* e com *vento fraco*, os riscos médios são, em geral, maiores para temperaturas altas e umidades baixas, qualquer que seja o dia da semana.

Tabela 3.9 *Riscos médios previstos e erros padrão estimados ($\widehat{EP}$) pelo modelo logístico com resposta contínua, para dias com Temperatura $\leq 17^\circ\text{C}$ e Umidade $\leq 75\%$*

Estação	Chuva	Vento	Dia da semana	n	$\hat{\pi}$	$\widehat{EP}$
Primavera-Verão	Sem	Fraco	Domingo	3	0,15	0,17
			Terça	2	0,70	0,17
			Quarta	2	0,70	0,18
			Quinta	1	0,73	0,24
			Sexta	1	0,76	0,26
		Médio	Segunda	3	0,46	0,16
			Terça	1	0,57	0,27
			Quinta	3	0,38	0,15
			Sexta	1	0,69	0,29
			Sábado	2	0,16	0,19
		Forte	Domingo	3	0,02	0,18
			Segunda	1	0,20	0,27
			Terça	3	0,15	0,17
			Quarta	1	0,20	0,32
			Quinta	3	0,21	0,17
			Sexta	2	0,13	0,20
			Sábado	4	0,05	0,15
Primavera-Verão	Com	Fraco	Domingo	1	0,10	0,29
		Médio	Segunda	2	0,50	0,25
			Terça	1	0,54	0,30
			Quarta	1	0,38	0,27
		Forte	Quinta	1	0,29	0,26
			Sexta	1	0,33	0,28
Outono-Inverno	Sem	Fraco	Domingo	12	0,29	0,08
			Segunda	12	0,72	0,06
			Terça	11	0,77	0,07
			Quarta	10	0,83	0,08
			Quinta	9	0,83	0,09
			Sexta	9	0,73	0,08
			Sábado	13	0,40	0,06
		Médio	Domingo	5	0,10	0,12
			Segunda	6	0,51	0,10
			Terça	4	0,50	0,14
			Quarta	3	0,52	0,14
			Quinta	7	0,50	0,10
			Sexta	3	0,55	0,15
			Sábado	5	0,40	0,13
		Forte	Domingo	1	0,03	0,31
			Segunda	1	0,19	0,29
			Terça	4	0,27	0,15
			Quarta	2	0,39	0,22
			Sexta	1	0,24	0,29
Outono-Inverno	Com	Fraco	Segunda	1	0,42	0,26
			Terça	1	0,77	0,29
			Quarta	1	0,62	0,26
			Sábado	3	0,27	0,16
		Médio	Domingo	2	0,11	0,22
			Terça	1	0,69	0,29
			Quarta	1	0,57	0,28
			Sexta	1	0,40	0,28
		Forte	Sexta	1	0,41	0,31

Tabela 3.10 *Riscos médios previstos e erros padrão estimados ($\widehat{EP}$) pelo modelo logístico com resposta contínua, para dias com Temperatura $> 17^{\circ}\text{C}$ e Umidade $\leq 75\%$*

Estação	Chuva	Vento	Dia da semana	n	$\hat{\pi}$	$\widehat{EP}$
Primavera-Verão	Sem	Fraco	Domingo	1	0,26	0,26
			Segunda	3	0,71	0,12
			Terça	4	0,78	0,14
			Quarta	3	0,74	0,12
			Quinta	4	0,84	0,13
			Sexta	5	0,76	0,12
			Sábado	4	0,51	0,12
		Médio	Domingo	4	0,11	0,12
			Segunda	2	0,45	0,17
			Terça	5	0,49	0,10
			Quarta	10	0,57	0,08
			Quinta	4	0,55	0,11
			Sexta	2	0,55	0,18
			Sábado	2	0,35	0,21
		Forte	Domingo	4	0,06	0,14
			Segunda	4	0,33	0,13
			Terça	3	0,33	0,15
			Quarta	4	0,41	0,13
			Quinta	6	0,47	0,12
			Sexta	1	0,37	0,29
			Sábado	3	0,08	0,14
Primavera-Verão	Com	Fraco	Domingo	3	0,16	0,15
			Segunda	6	0,58	0,10
			Terça	1	0,58	0,25
			Quarta	2	0,64	0,19
			Quinta	1	0,50	0,25
			Sexta	4	0,51	0,11
			Sábado	2	0,29	0,18
		Médio	Segunda	2	0,50	0,19
			Terça	1	0,61	0,27
			Quinta	1	0,48	0,26
			Sexta	2	0,50	0,20
		Forte	Sábado	1	0,22	0,28
			Domingo	4	0,09	0,15
			Segunda	4	0,42	0,13
			Terça	4	0,46	0,14
			Quarta	4	0,44	0,13
			Quinta	3	0,48	0,15
			Sábado	4	0,16	0,14
Outono-Inverno	Sem	Fraco	Domingo	4	0,40	0,14
			Segunda	4	0,88	0,14
			Terça	3	0,92	0,18
			Quarta	8	0,84	0,08
			Quinta	3	0,92	0,17
			Sexta	3	0,76	0,13
			Sábado	3	0,64	0,16
		Médio	Segunda	1	0,60	0,26
			Terça	2	0,71	0,20
			Quarta	2	0,79	0,25
			Quinta	4	0,71	0,12
			Sexta	2	0,79	0,23
			Sábado	1	0,63	0,35
		Forte	Terça	1	0,61	0,30
			Quarta	2	0,73	0,29
			Sexta	2	0,35	0,24
			Sábado	1	0,25	0,31
Outono-Inverno	Com	Fraco	Terça	1	0,68	0,27
			Quinta	1	0,81	0,31
			Sexta	2	0,62	0,19
			Sábado	1	0,29	0,26
		Médio	Terça	2	0,73	0,21
			Quinta	1	0,75	0,31
			Sexta	1	0,83	0,36
		Forte	Terça	1	0,75	0,33
			Quinta	1	0,58	0,29

Tabela 3.11 *Riscos médios previstos e erros padrão estimados ($\widehat{EP}$) pelo modelo logístico com resposta contínua, para dias com Temperatura $\leq 17^\circ\text{C}$ e Umidade $> 75\%$*

Estação	Chuva	Vento	Dia da semana	n	$\hat{\pi}$	$\widehat{EP}$
Primavera-Verão	Sem	Fraco	Quarta	2	0,37	0,18
			Quinta	1	0,37	0,26
			Sexta	1	0,39	0,26
			Sábado	1	0,16	0,26
		Médio	Terça	1	0,33	0,24
			Quarta	1	0,31	0,24
			Quinta	1	0,21	0,25
			Sexta	3	0,20	0,15
			Sábado	1	0,10	0,25
		Forte	Domingo	1	0,01	0,31
			Segunda	2	0,06	0,22
			Terça	1	0,11	0,28
			Quarta	3	0,09	0,17
Primavera-Verão	Com	Fraco	Domingo	1	0,03	0,30
			Terça	1	0,15	0,27
			Quarta	2	0,16	0,20
			Sexta	3	0,17	0,15
			Sábado	1	0,04	0,28
		Médio	Domingo	1	0,02	0,31
			Segunda	1	0,07	0,36
			Terça	2	0,13	0,21
			Quarta	1	0,26	0,28
			Quinta	4	0,18	0,14
			Sexta	2	0,17	0,20
		Forte	Domingo	3	0,02	0,17
			Segunda	1	0,21	0,26
			Terça	1	0,06	0,32
			Quarta	3	0,15	0,17
			Quinta	2	0,12	0,21
Outono-Inverno	Sem	Fraco	Sexta	1	0,15	0,27
			Sábado	5	0,05	0,12
			Domingo	5	0,06	0,12
			Segunda	8	0,45	0,08
			Terça	9	0,47	0,07
			Quarta	4	0,49	0,12
		Médio	Quinta	7	0,51	0,09
			Sexta	7	0,41	0,09
			Sábado	6	0,22	0,09
			Domingo	3	0,04	0,16
			Segunda	3	0,34	0,14
			Terça	1	0,14	0,30
		Forte	Quarta	2	0,40	0,17
			Quinta	1	0,22	0,26
			Sexta	4	0,27	0,12
			Sábado	3	0,12	0,15
			Terça	1	0,22	0,29
			Sexta	2	0,11	0,21
Outono-Inverno	Com	Fraco	Sábado	1	0,03	0,32
			Domingo	7	0,03	0,12
			Segunda	3	0,22	0,16
			Terça	3	0,34	0,15
			Quarta	2	0,35	0,18
			Quinta	4	0,24	0,15
		Médio	Sexta	3	0,25	0,15
			Sábado	3	0,09	0,16
			Domingo	4	0,03	0,15
			Segunda	4	0,17	0,15
			Terça	2	0,22	0,19
			Quarta	5	0,29	0,13
		Forte	Quinta	1	0,19	0,30
			Sexta	1	0,10	0,37
			Sábado	5	0,08	0,15
			Segunda	2	0,11	0,24
			Quarta	3	0,22	0,16
			Quinta	2	0,21	0,21
			Sexta	2	0,18	0,21
			Sábado	1	0,04	0,34

Tabela 3.12 *Riscos médios previstos e erros padrão estimados ($\widehat{EP}$) pelo modelo logístico com resposta contínua, para dias com Temperatura > 17°C e Umidade > 75%*

Estação	Chuva	Vento	Dia da semana	n	$\hat{\pi}$	$\widehat{EP}$
Primavera-Verão	Sem	Fraco	Segunda	1	0,65	0,25
			Qua	1	0,59	0,25
			Quinta	1	0,70	0,26
			Sexta	1	0,30	0,27
		Médio	Domingo	4	0,05	0,14
			Segunda	3	0,39	0,15
			Terça	2	0,34	0,17
			Quinta	1	0,46	0,32
			Sexta	2	0,24	0,18
			Sábado	1	0,12	0,25
		Forte	Domingo	1	0,03	0,32
			Segunda	1	0,20	0,29
			Quinta	1	0,16	0,32
			Sexta	2	0,19	0,21
Primavera-Verão	Com	Fraco	Domingo	6	0,06	0,12
			Segunda	3	0,32	0,14
			Terça	3	0,44	0,15
			Qua	4	0,35	0,13
			Quinta	3	0,37	0,16
			Sexta	6	0,32	0,10
			Sábado	9	0,15	0,09
		Médio	Domingo	5	0,06	0,13
			Segunda	4	0,31	0,14
			Terça	5	0,32	0,12
			Qua	2	0,47	0,19
			Quinta	3	0,35	0,16
			Sexta	2	0,33	0,20
			Sábado	5	0,15	0,13
		Forte	Domingo	2	0,05	0,21
			Segunda	3	0,20	0,15
			Terça	6	0,27	0,11
			Qua	5	0,30	0,12
			Quinta	3	0,30	0,16
			Sexta	5	0,23	0,12
Outono-Inverno	Sem	Fraco	Sábado	2	0,07	0,19
			Segunda	1	0,61	0,22
			Terça	2	0,71	0,18
			Quinta	2	0,73	0,18
		Médio	Sábado	3	0,40	0,15
			Domingo	3	0,07	0,17
			Segunda	1	0,50	0,29
			Terça	1	0,57	0,26
			Qua	2	0,53	0,21
			Quinta	1	0,37	0,26
		Forte	Segunda	1	0,34	0,34
			Segunda	1	0,08	0,30
Outono-Inverno	Com	Fraco	Segunda	2	0,54	0,19
			Terça	1	0,49	0,29
			Qua	1	0,58	0,26
			Quinta	1	0,55	0,27
			Sexta	1	0,37	0,26
			Sábado	2	0,22	0,18
		Médio	Domingo	1	0,08	0,32
			Terça	1	0,38	0,28
			Quinta	1	0,49	0,30
		Forte	Qua	1	0,42	0,29
			Sexta	2	0,23	0,21

L1 : Primavera-Verão – Temperatura $\leq 17^{\circ}\text{C}$ – Umidade $\leq 75\%$ L5 : Outono-Inverno – Temperatura $\leq 17^{\circ}\text{C}$ – Umidade $\leq 75\%$
 L2 : Primavera-Verão – Temperatura $> 17^{\circ}\text{C}$ – Umidade $\leq 75\%$ L6 : Outono-Inverno – Temperatura $> 17^{\circ}\text{C}$ – Umidade $\leq 75\%$
 L3 : Primavera-Verão – Temperatura $\leq 17^{\circ}\text{C}$ – Umidade $> 75\%$ L7 : Outono-Inverno – Temperatura $\leq 17^{\circ}\text{C}$ – Umidade $> 75\%$
 L4 : Primavera-Verão – Temperatura $> 17^{\circ}\text{C}$ – Umidade $> 75\%$ L8 : Outono-Inverno – Temperatura $> 17^{\circ}\text{C}$ – Umidade $> 75\%$

Figura 3.3 Gráfico com riscos médios previstos pelo modelo logístico com resposta contínua em dias sem chuva e vento fraco

3.5 Previsão de uma nova observação

O modelo de regressão logística pode ser usado para fazer previsão de novas observações. A grande dificuldade é determinar o ponto crítico (p_c), tal que a previsão será 1, se $\hat{\pi}_i \geq p_c$, ou 0, caso contrário, $i = 1, \dots, n$, onde o ponto crítico ótimo será aquele que determinar a menor proporção de previsões incorretas.

Na determinação do ponto crítico ótimo para prever a ocorrência de concentração de NO_2 acima de $100 \mu\text{g}/\text{m}^3$, consideramos as estimativas dos riscos fornecidas pelo modelo logístico com resposta contínua e, adotando diferentes pontos críticos, calculamos, com base nas concentrações de NO_2 observadas, as proporções de classificações incorretas. Os resultados obtidos são apresentados na Tabela 3.13.

Observamos que as classificações incorretas ocorreram em menor número para pontos

críticos entre 0,40 e 0,50, sendo 0,45 o valor que produziu o menor número de classificações incorretas na amostra (25,5%).

Tabela 3.13 *Proporções de previsões incorretas para diferentes pontos críticos*

Ponto crítico	Proporção incorreta
0,30	0,302
0,40	0,264
0,41	0,261
0,42	0,258
0,43	0,264
0,44	0,262
0,45	0,255
0,46	0,258
0,47	0,264
0,48	0,264
0,49	0,264
0,50	0,273
0,60	0,292
0,70	0,304
0,80	0,345

Portando, para prever se um novo dia de observação terá concentração de NO_2 acima de $100 \mu\text{g}/\text{m}^3$ podemos adotar a regra:

$$\begin{cases} \text{NO}_2 > 100 \mu\text{g}/\text{m}^3, & \text{se } \hat{\pi}_i \geq 0,45; \\ \text{NO}_2 \leq 100 \mu\text{g}/\text{m}^3, & \text{caso contrário.} \end{cases}$$

A determinação do ponto crítico e correspondente estimativa da probabilidade de classificação errada ficariam melhor estabelecidas se dispuséssemos, para sua determinação, de um novo banco de dados. Assim, utilizaríamos um banco de dados para estimar o modelo e outro para determinar o ponto crítico e estimar a probabilidade de erro de classificação.

Capítulo 4

Considerações finais

A regressão logística com resposta contínua é uma alternativa à regressão logística usual quando a variável resposta possui distribuição contínua. A maior vantagem de se incorporar a informação sobre a distribuição da variável resposta no modelo logístico é o aumento da eficiência dos estimadores dos parâmetros do modelo. Entretanto, Ronchetti *et al.* (1997) consideraram que esse ganho de eficiência pode não ocorrer quando o modelo é mal especificado, porque os estimadores de máxima verossimilhança são extremamente sensíveis a desvios das suposições do modelo. Sugerimos que técnicas robustas de estimação do modelo logístico com resposta contínua sejam estudadas.

Neste trabalho o modelo logístico com resposta contínua foi adotado com o objetivo de modelar o risco da concentração do poluente NO_2 ser maior que $100 \mu\text{g}/\text{m}^3$, e também prever dias com altas concentrações do poluente em estudo. O modelo foi ajustado com base em dados obtidos nos anos de 1998 e 1999. Sugerimos que o modelo seja reajustado com base em um banco de dados no qual um período de tempo maior seja considerado.

O modelo proposto deve ser encarado apenas como uma primeira exploração do assunto. As variáveis explicativas incluídas nos modelos aqui construídos foram selecionadas por serem de fácil obtenção e previsão. A qualidade do ajuste e a capacidade preditiva do modelo, entretanto, podem ser melhoradas à medida que informações adicionais trazidas por pesquisadores de outras áreas (medicina, meio ambiente, pesquisa atmosférica) sejam incorporadas ao modelo.

Apêndice A

Procedimento de estimação dos coeficientes do modelo, dado σ .

No Capítulo 2, mostramos que, dado σ , a equação escore para estimar $\boldsymbol{\beta}$ é dada por

$$\sum_{i=1}^n \left\{ \frac{t_i - C}{\sigma} - \Phi^{-1} [g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})] \right\} \frac{1}{f\{\Phi^{-1}[g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})]\}} \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}}}{(1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}})^2} x_{ij} = 0, \quad (\text{A.1})$$

onde f é a função densidade de probabilidade de uma variável aleatória $N(0, 1)$.

Para resolver a equação (A.1) vamos utilizar o procedimento de ponderação de Fisher (Rao, 1973). Por este procedimento o vetor de estimativas de $\boldsymbol{\beta}$ no passo $(m + 1)$ do processo iterativo é dado por

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathfrak{S}^{-1})^{(m)} \mathbf{U}^{(m)}, \quad (\text{A.2})$$

onde

$$\mathfrak{S}_{jk} = E \left[-\frac{\partial^2 \log L(\boldsymbol{\beta}, \sigma; \mathbf{t})}{\partial \beta_j \partial \beta_k} \right],$$

ou ainda, multiplicando ambos os lados de (A.2) por $\mathfrak{S}^{(m)}$:

$$\mathfrak{S}^{(m)} \boldsymbol{\beta}^{(m+1)} = \mathfrak{S}^{(m)} \boldsymbol{\beta}^{(m)} + \mathbf{U}^{(m)}. \quad (\text{A.3})$$

Neste caso, a matriz de informação esperada de Fisher é

$$\begin{aligned}\mathfrak{S}_{jk} &= E \left[-\frac{\partial^2 \log L(\boldsymbol{\beta}, \sigma; \mathbf{t})}{\partial \beta_j \partial \beta_k} \right] \\ &= E \left[-\sum_{i=1}^n \left\{ -\frac{\partial \{\Phi^{-1}[g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})]\}}{\partial \beta_k} \frac{1}{f\{\Phi^{-1}[g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})]\}} \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}}}{(1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}})^2} x_{ij} + \right. \right. \\ &\quad \left. \left. + \left\{ \frac{t_i - C}{\sigma} - \Phi^{-1}[g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})] \right\} \frac{\partial}{\partial \beta_k} \left[\frac{1}{f\{\Phi^{-1}[g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})]\}} \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}}}{(1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}})^2} x_{ij} \right] \right\} \right]. \quad (\text{A.4})\end{aligned}$$

Como $E\{(t_i - C)/\sigma - \Phi^{-1}[g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})]\} = 0$, então o segundo termo de (A.4) é igual a zero e temos

$$\begin{aligned}\mathfrak{S}_{jk} &= E \left[\sum_{i=1}^n \frac{1}{f\{\Phi^{-1}[g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})]\}} \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}}}{(1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}})^2} \frac{1}{f\{\Phi^{-1}[g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})]\}} \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}}}{(1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}})^2} x_{ij} x_{ik} \right] \\ &= \sum_{i=1}^n \left[\frac{1}{f\{\Phi^{-1}[g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})]\}} \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}}}{(1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}})^2} \right]^2 x_{ij} x_{ik}.\end{aligned}$$

Por (2.17) temos então que

$$\mathfrak{S}_{jk} = \sum_{i=1}^n \left(\frac{d\gamma_i}{d\eta_i} \right)^2 x_{ij} x_{ik}.$$

Matricialmente,

$$\mathfrak{S} = (\mathbf{L}\mathbf{X})^t \mathbf{L}\mathbf{X},$$

onde $\mathbf{L}$ e $\mathbf{X}$ são dadas na Seção 2.3. Então, a equação (A.3) fica

$$\begin{aligned}(\mathbf{L}^{(m)}\mathbf{X})^t \mathbf{L}^{(m)}\mathbf{X}\boldsymbol{\beta}^{(m+1)} &= (\mathbf{L}^{(m)}\mathbf{X})^t \mathbf{L}^{(m)}\mathbf{X}\boldsymbol{\beta}^{(m)} + (\mathbf{L}^{(m)}\mathbf{X})^t (\mathbf{t}^* - \boldsymbol{\gamma}) \\ (\mathbf{L}^{(m)}\mathbf{X})^t \mathbf{L}^{(m)}\mathbf{X}\boldsymbol{\beta}^{(m+1)} &= (\mathbf{L}^{(m)}\mathbf{X})^t (\mathbf{L}^{(m)}\mathbf{X}\boldsymbol{\beta}^{(m)} + \mathbf{t}^* - \boldsymbol{\gamma}).\end{aligned} \quad (\text{A.5})$$

Seja $\mathbf{z}^{(m)} = \mathbf{L}^{(m)}\mathbf{X}\boldsymbol{\beta}^{(m)} + \mathbf{t}^* - \boldsymbol{\gamma}$ e $\mathbf{X}^{*(m)} = \mathbf{L}^{(m)}\mathbf{X}$. Então, por (A.5) temos que

$$\boldsymbol{\beta}^{(m+1)} = \left(\mathbf{X}^{*(m)t} \mathbf{X}^{*(m)} \right)^{-1} \mathbf{X}^{*(m)t} \mathbf{z}^{(m)}$$

o que comprova a afirmação feita na Seção 2.3.

Figura B.1 Representação gráfica das séries cronológicas para as concentrações dos poluentes NO_2 , O_3 , SO_2 , PM_{10} e CO nos anos de 1998 e 1999

Figura B.2 Representação gráfica das séries cronológicas de Temperatura, Umidade, Precipitação e Velocidade do vento nos anos de 1998 e 1999

Figura B.3 Diagramas de dispersão do poluente NO_2 pela Temperatura com suavizador “lowess”, considerando duas categorias de Umidade

Figura B.4 *Box-plot das concentrações do NO₂ por Dia da semana para as diferentes combinações de Chuva e Vento*

Figura B.5 Gráficos de Probabilidade Normal para as concentrações dos poluentes NO_2 , O_3 , SO_2 , PM_{10} e CO

Tabela B.2 *Modelos iniciais usual e com resposta contínua ajustados*

Efeitos	Logística Usual		Logística com resposta contínua		
	Estimativa	Erro padrão	Estimativa	Erro padrão	$\hat{\sigma}$
<i>intercepto</i>	20,654	6,049	8,569	3,700	0,288
<i>temperatura</i>	-0,822	0,336	-0,219	0,208	
<i>umidade</i>	-0,341	0,082	-0,175	0,049	
<i>segunda</i>	1,958	1,618	1,596	1,092	
<i>terça</i>	1,065	1,414	0,840	0,999	
<i>quarta</i>	4,657	1,871	3,076	1,026	
<i>quinta</i>	2,368	1,541	1,599	1,021	
<i>sexta</i>	2,588	1,565	2,295	1,027	
<i>sábado</i>	0,050	1,551	0,241	1,082	
<i>outono-inverno</i>	1,073	1,295	0,906	0,885	
<i>chuva</i>	-0,039	1,453	-0,438	0,965	
<i>vento médio</i>	-7,016	11,779	-1,378	1,031	
<i>vento forte</i>	-7,123	11,801	-2,967	1,213	
<i>temperatura*umidade</i>	0,013	0,005	0,005	0,003	
<i>segunda*outono-inverno</i>	-0,001	1,758	0,537	1,157	
<i>terça*outono-inverno</i>	0,693	1,571	1,258	1,078	
<i>quarta*outono-inverno</i>	-1,328	2,001	-0,488	1,113	
<i>quinta*outono-inverno</i>	-1,388	1,696	0,377	1,112	
<i>sexta*outono-inverno</i>	-1,873	1,705	-0,846	1,111	
<i>sábado*outono-inverno</i>	0,485	1,683	0,635	1,150	
<i>segunda*chuva</i>	-0,847	1,911	0,171	1,240	
<i>terça*chuva</i>	1,644	1,858	2,125	1,223	
<i>quarta*chuva</i>	-4,319	2,113	-1,541	1,213	
<i>quinta*chuva</i>	-2,344	2,099	0,328	1,290	
<i>sexta*chuva</i>	-1,091	1,831	-0,518	1,182	
<i>sábado*chuva</i>	-1,958	2,202	0,749	1,242	
<i>outono-inverno*chuva</i>	-5,701	10,309	-0,556	1,120	
<i>segunda*vento médio</i>	5,859	11,845	0,457	1,318	
<i>terça*vento médio</i>	7,596	11,813	2,344	1,237	
<i>quarta*vento médio</i>	4,085	11,864	0,042	1,249	
<i>quinta*vento médio</i>	5,151	11,833	1,304	1,262	
<i>sexta*vento médio</i>	6,773	11,834	1,768	1,281	
<i>sábado*vento médio</i>	7,832	11,840	1,180	1,351	
<i>segunda*vento forte</i>	6,737	11,877	2,463	1,506	
<i>terça*vento forte</i>	5,873	11,864	1,396	1,508	
<i>quarta*vento forte</i>	2,861	11,913	0,235	1,470	
<i>quinta*vento forte</i>	5,157	11,861	2,268	1,448	
<i>sexta*vento forte</i>	3,366	11,922	1,353	1,505	
<i>sábado*vento forte</i>	0,607	16,941	0,955	1,632	
<i>outono-inverno*vento médio</i>	5,620	11,805	-0,005	1,116	
<i>outono-inverno*vento forte</i>	-0,195	38,513	1,855	2,368	
<i>chuva*vento médio</i>	0,752	13,214	0,406	1,158	
<i>chuva*vento forte</i>	6,414	11,878	1,644	1,519	
<i>segunda*outono-inverno*chuva</i>	5,418	10,416	-0,398	1,377	
<i>terça*outono-inverno*chuva</i>	4,857	10,383	-0,670	1,353	
<i>quarta*outono-inverno*chuva</i>	8,806	10,407	1,910	1,353	
<i>quinta*outono-inverno*chuva</i>	9,775	10,432	1,226	1,441	
<i>sexta*outono-inverno*chuva</i>	6,344	10,388	-0,248	1,366	
<i>sábado*outono-inverno*chuva</i>	6,664	10,443	-1,199	1,411	
<i>segunda*outono-inverno*vento médio</i>	-6,346	11,891	-0,252	1,411	
<i>terça*outono-inverno*vento médio</i>	-6,695	11,859	-1,882	1,362	
<i>quarta*outono-inverno*vento médio</i>	-7,203	11,902	-0,768	1,387	
<i>quinta*outono-inverno*vento médio</i>	-4,942	11,877	-0,775	1,385	
<i>sexta*outono-inverno*vento médio</i>	-5,666	11,878	-0,372	1,411	
<i>sábado*outono-inverno*vento médio</i>	-7,271	11,888	-0,870	1,449	
<i>segunda*outono-inverno*vento forte</i>	-0,862	38,564	-3,062	2,674	
<i>terça*outono-inverno*vento forte</i>	-0,055	38,544	-2,032	2,604	
<i>quarta*outono-inverno*vento forte</i>	-1,310	38,565	-2,687	2,597	
<i>quinta*outono-inverno*vento forte</i>	-9,644	43,186	-4,643	2,998	
<i>sexta*outono-inverno*vento forte</i>	1,735	38,562	-1,521	2,594	
<i>sábado*outono-inverno*vento forte</i>	-2,654	42,371	-1,485	2,815	
<i>segunda*chuva*vento médio</i>	0,201	13,293	-0,065	1,433	
<i>terça*chuva*vento médio</i>	-2,192	13,255	-1,722	1,385	
<i>quarta*chuva*vento médio</i>	3,614	13,286	1,318	1,415	
<i>quinta*chuva*vento médio</i>	2,009	13,292	0,228	1,445	
<i>sexta*chuva*vento médio</i>	-0,815	13,269	-0,520	1,436	
<i>sábado*chuva*vento médio</i>	0,881	13,301	-0,358	1,473	
<i>segunda*chuva*vento forte</i>	-4,285	11,996	-0,019	1,850	
<i>terça*chuva*vento forte</i>	-6,923	11,983	-1,733	1,868	
<i>quarta*chuva*vento forte</i>	-1,042	12,012	1,268	1,812	
<i>quinta*chuva*vento forte</i>	-1,730	12,021	-1,206	1,907	
<i>sexta*chuva*vento forte</i>	-2,850	12,017	-0,332	1,834	
<i>sábado*chuva*vento forte</i>	2,399	17,049	-0,259	1,966	
<i>outono-inverno*chuva*vento médio</i>	-0,607	1,223	0,422	0,719	
<i>outono-inverno*chuva*vento forte</i>	-2,667	1,812	-0,378	0,990	

Tabela B.3 *Efeitos não significantes no ajuste do modelo logístico com resposta contínua*

Efeito	Nível descritivo
<i>segunda*outono-inverno*vento médio</i>	0,797
<i>terça*outono-inverno*vento médio</i>	
<i>quarta*outono-inverno*vento médio</i>	
<i>quinta*outono-inverno*vento médio</i>	
<i>sexta*outono-inverno*vento médio</i>	
<i>sábado*outono-inverno*vento médio</i>	
<i>segunda*outono-inverno*vento forte</i>	
<i>terça*outono-inverno*vento forte</i>	
<i>quarta*outono-inverno*vento forte</i>	
<i>quinta*outono-inverno*vento forte</i>	
<i>sexta*outono-inverno*vento forte</i>	
<i>sábado*outono-inverno*vento forte</i>	
<i>segunda*chuva*vento médio</i>	0,454
<i>terça*chuva*vento médio</i>	
<i>quarta*chuva*vento médio</i>	
<i>quinta*chuva*vento médio</i>	
<i>sexta*chuva*vento médio</i>	
<i>sábado*chuva*vento médio</i>	
<i>segunda*chuva*vento forte</i>	
<i>terça*chuva*vento forte</i>	
<i>quarta*chuva*vento forte</i>	
<i>quinta*chuva*vento forte</i>	
<i>sexta*chuva*vento forte</i>	
<i>sábado*chuva*vento forte</i>	
<i>outono-inverno*chuva*vento médio</i>	0,388
<i>outono-inverno*chuva*vento forte</i>	
<i>segunda*outono-inverno*chuva</i>	0,144
<i>terça*outono-inverno*chuva</i>	
<i>quarta*outono-inverno*chuva</i>	
<i>quinta*outono-inverno*chuva</i>	
<i>sexta*outono-inverno*chuva</i>	
<i>sábado*outono-inverno*chuva</i>	
<i>segunda*outono-inverno</i>	0,293
<i>terça*outono-inverno</i>	
<i>quarta*outono-inverno</i>	
<i>quinta*outono-inverno</i>	
<i>sexta*outono-inverno</i>	
<i>sábado*outono-inverno</i>	
<i>segunda*chuva</i>	0,423
<i>terça*chuva</i>	
<i>quarta*chuva</i>	
<i>quinta*chuva</i>	
<i>sexta*chuva</i>	
<i>sábado*chuva</i>	
<i>outono-inverno*chuva</i>	0,139
<i>outono-inverno*vento médio</i>	0,137
<i>outono-inverno*vento forte</i>	
<i>segunda*vento médio</i>	0,113
<i>terça*vento médio</i>	
<i>quarta*vento médio</i>	
<i>quinta*vento médio</i>	
<i>sexta*vento médio</i>	
<i>sábado*vento médio</i>	
<i>segunda*vento forte</i>	
<i>terça*vento forte</i>	
<i>quarta*vento forte</i>	
<i>quinta*vento forte</i>	
<i>sexta*vento forte</i>	
<i>sábado*vento forte</i>	

Tabela B.4 *Intervalos de classe para os riscos estimados ($\hat{\pi}_i$) considerados na construção da Figura 3.2*

Classe j	Intervalo $\hat{\pi}_i$	Ponto médio	n_j ¹	p_j ²
1	0,01 – 0,06	0,04	63	0,02
2	0,07 – 0,13	0,10	73	0,08
3	0,14 – 0,19	0,17	63	0,13
4	0,20 – 0,26	0,23	66	0,33
5	0,27 – 0,35	0,31	73	0,37
6	0,36 – 0,44	0,40	69	0,39
7	0,45 – 0,52	0,49	71	0,49
8	0,53 – 0,64	0,59	67	0,67
9	0,65 – 0,78	0,72	68	0,69
10	0,79 – 0,99	0,89	62	0,89

¹ Número de casos na j -ésima classe

² Proporção estimada de uns na j -ésima classe

Tabela B.5 *Valores observados e esperados para o construção do teste Qui-quadrado de qualidade de ajuste*

Classe j	Intervalo $\hat{\pi}_i$	n_j	O_{j0}	E_{j0}	O_{j1}	E_{j1}
1	0,01 – 0,06	63	62	60,5	1	2,5
2	0,07 – 0,13	73	67	65,9	6	7,1
3	0,14 – 0,19	63	55	52,5	8	10,5
4	0,20 – 0,26	66	44	51,1	22	14,9
5	0,27 – 0,35	73	46	50,1	27	22,9
6	0,36 – 0,44	69	42	41,4	27	27,6
7	0,45 – 0,52	71	36	36,5	35	34,5
8	0,53 – 0,64	67	22	28,1	45	38,9
9	0,65 – 0,78	68	21	19,7	47	48,3
10	0,79 – 0,99	62	7	8,0	55	54,0
Total		675	402	413,8	273	261,2

Apêndice C

Programa no S-Plus

```
## Organização do banco de dados ##

attach(dados.98.99NA)
dim(dados.98.99NA)

dados <- na.omit(dados.98.99NA) ## Omite as linhas com observações perdidas
dim(dados)

n.obs <- dim(dados)[1] # número de observações = 675
n.par <- 75 # número de parâmetros

## Construção da matriz.X : matriz de planejamento do modelo ##

matriz.X <- matrix(NA, nrow=n.obs, ncol=n.par)
matriz.X[,1] <- 1
matriz.X[,2] <- dados$temp
matriz.X[,3] <- dados$umidade

# Construção das variáveis indicadoras para dias da semana - SEMANA (domingo como referência)

for(i in 1:n.obs) if (dados$semana[i] == 2) matriz.X[i,4] <- 1 else matriz.X[i,4] <- 0
for(i in 1:n.obs) if (dados$semana[i] == 3) matriz.X[i,5] <- 1 else matriz.X[i,5] <- 0
for(i in 1:n.obs) if (dados$semana[i] == 4) matriz.X[i,6] <- 1 else matriz.X[i,6] <- 0
for(i in 1:n.obs) if (dados$semana[i] == 5) matriz.X[i,7] <- 1 else matriz.X[i,7] <- 0
for(i in 1:n.obs) if (dados$semana[i] == 6) matriz.X[i,8] <- 1 else matriz.X[i,8] <- 0
for(i in 1:n.obs) if (dados$semana[i] == 7) matriz.X[i,9] <- 1 else matriz.X[i,9] <- 0

# Construção da variável indicadora para estação do ano - ESTACAO (primavera-verão como referência)

for(i in 1:n.obs) if (dados$estacao[i] == 1) matriz.X[i,10] <- 1 else matriz.X[i,10] <- 0

# Construção da variável indicadora para precipitação (sem chuva como referência) - CHUVA

for(i in 1:n.obs) if (dados$chuva[i] == 1) matriz.X[i,11] <- 1 else matriz.X[i,11] <- 0

# Construção das variáveis indicadoras para vento (vento fraco como referência) - VENTO

for(i in 1:n.obs) if (dados$vento[i] == 1) matriz.X[i,12] <- 1 else matriz.X[i,12] <- 0
for(i in 1:n.obs) if (dados$vento[i] == 2) matriz.X[i,13] <- 1 else matriz.X[i,13] <- 0

# Construção da interação TEMPERATURA*UMIDADE

matriz.X[,14] <- matriz.X[,2]*matriz.X[,3]
```

```

# Construção das variáveis indicadoras para a interação SEMANA*ESTACAO

matriz.X[,15] <- matriz.X[,4]*matriz.X[,10]
matriz.X[,16] <- matriz.X[,5]*matriz.X[,10]
matriz.X[,17] <- matriz.X[,6]*matriz.X[,10]
matriz.X[,18] <- matriz.X[,7]*matriz.X[,10]
matriz.X[,19] <- matriz.X[,8]*matriz.X[,10]
matriz.X[,20] <- matriz.X[,9]*matriz.X[,10]

# Construção das variáveis indicadoras para a interação SEMANA*CHUVA

matriz.X[,21] <- matriz.X[,4]*matriz.X[,11]
matriz.X[,22] <- matriz.X[,5]*matriz.X[,11]
matriz.X[,23] <- matriz.X[,6]*matriz.X[,11]
matriz.X[,24] <- matriz.X[,7]*matriz.X[,11]
matriz.X[,25] <- matriz.X[,8]*matriz.X[,11]
matriz.X[,26] <- matriz.X[,9]*matriz.X[,11]

# Construção da variável indicadora para a interação ESTACAO*CHUVA

matriz.X[,27] <- matriz.X[,10]*matriz.X[,11]

# Construção das variáveis indicadoras para a interação SEMANA*VENTO

matriz.X[,28] <- matriz.X[,4]*matriz.X[,12]
matriz.X[,29] <- matriz.X[,5]*matriz.X[,12]
matriz.X[,30] <- matriz.X[,6]*matriz.X[,12]
matriz.X[,31] <- matriz.X[,7]*matriz.X[,12]
matriz.X[,32] <- matriz.X[,8]*matriz.X[,12]
matriz.X[,33] <- matriz.X[,9]*matriz.X[,12]
matriz.X[,34] <- matriz.X[,4]*matriz.X[,13]
matriz.X[,35] <- matriz.X[,5]*matriz.X[,13]
matriz.X[,36] <- matriz.X[,6]*matriz.X[,13]
matriz.X[,37] <- matriz.X[,7]*matriz.X[,13]
matriz.X[,38] <- matriz.X[,8]*matriz.X[,13]
matriz.X[,39] <- matriz.X[,9]*matriz.X[,13]

# Construção das variáveis indicadoras para a interação ESTACAO*VENTO

matriz.X[,40] <- matriz.X[,10]*matriz.X[,12]
matriz.X[,41] <- matriz.X[,10]*matriz.X[,13]

# Construção das variáveis indicadoras para a interação CHUVA*VENTO

matriz.X[,42] <- matriz.X[,11]*matriz.X[,12]
matriz.X[,43] <- matriz.X[,11]*matriz.X[,13]

# Construção das variáveis indicadoras para a interação SEMANA*ESTACAO*CHUVA

matriz.X[,44] <- matriz.X[,4]*matriz.X[,10]*matriz.X[,11]
matriz.X[,45] <- matriz.X[,5]*matriz.X[,10]*matriz.X[,11]
matriz.X[,46] <- matriz.X[,6]*matriz.X[,10]*matriz.X[,11]
matriz.X[,47] <- matriz.X[,7]*matriz.X[,10]*matriz.X[,11]
matriz.X[,48] <- matriz.X[,8]*matriz.X[,10]*matriz.X[,11]
matriz.X[,49] <- matriz.X[,9]*matriz.X[,10]*matriz.X[,11]

# Construção das variáveis indicadoras para a interação SEMANA*ESTACAO*VENTO

matriz.X[,50] <- matriz.X[,4]*matriz.X[,10]*matriz.X[,12]
matriz.X[,51] <- matriz.X[,5]*matriz.X[,10]*matriz.X[,12]
matriz.X[,52] <- matriz.X[,6]*matriz.X[,10]*matriz.X[,12]
matriz.X[,53] <- matriz.X[,7]*matriz.X[,10]*matriz.X[,12]

```

```

matriz.X[,54] <- matriz.X[,8]*matriz.X[,10]*matriz.X[,12]
matriz.X[,55] <- matriz.X[,9]*matriz.X[,10]*matriz.X[,12]
matriz.X[,56] <- matriz.X[,4]*matriz.X[,10]*matriz.X[,13]
matriz.X[,57] <- matriz.X[,5]*matriz.X[,10]*matriz.X[,13]
matriz.X[,58] <- matriz.X[,6]*matriz.X[,10]*matriz.X[,13]
matriz.X[,59] <- matriz.X[,7]*matriz.X[,10]*matriz.X[,13]
matriz.X[,60] <- matriz.X[,8]*matriz.X[,10]*matriz.X[,13]
matriz.X[,61] <- matriz.X[,9]*matriz.X[,10]*matriz.X[,13]

# Construção das variáveis indicadoras para a interação SEMANA*CHUVA*VENTO

matriz.X[,62] <- matriz.X[,4]*matriz.X[,11]*matriz.X[,12]
matriz.X[,63] <- matriz.X[,5]*matriz.X[,11]*matriz.X[,12]
matriz.X[,64] <- matriz.X[,6]*matriz.X[,11]*matriz.X[,12]
matriz.X[,65] <- matriz.X[,7]*matriz.X[,11]*matriz.X[,12]
matriz.X[,66] <- matriz.X[,8]*matriz.X[,11]*matriz.X[,12]
matriz.X[,67] <- matriz.X[,9]*matriz.X[,11]*matriz.X[,12]
matriz.X[,68] <- matriz.X[,4]*matriz.X[,11]*matriz.X[,13]
matriz.X[,69] <- matriz.X[,5]*matriz.X[,11]*matriz.X[,13]
matriz.X[,70] <- matriz.X[,6]*matriz.X[,11]*matriz.X[,13]
matriz.X[,71] <- matriz.X[,7]*matriz.X[,11]*matriz.X[,13]
matriz.X[,72] <- matriz.X[,8]*matriz.X[,11]*matriz.X[,13]
matriz.X[,73] <- matriz.X[,9]*matriz.X[,11]*matriz.X[,13]

# Construção das variáveis indicadoras para a interação ESTACAO*CHUVA*VENTO

matriz.X[,74] <- matriz.X[,10]*matriz.X[,11]*matriz.X[,12]
matriz.X[,75] <- matriz.X[,10]*matriz.X[,11]*matriz.X[,13]

# Nomes dos efeitos:

nome.matriz.X <- c("intercepto", "temperatura", "umidade", "segunda", "terca", "quarta", "quinta", "sexta",
"sabado", "outono-inverno", "chuva", "vento medio", "vento forte", "temperatura*umidade",
"segunda*outono-inverno", "terca*outono-inverno", "quarta*outono-inverno", "quinta*outono-inverno",
"sexta*outono-inverno", "sabado*outono-inverno", "segunda*chuva", "terca*chuva", "quarta*chuva",
"quinta*chuva", "sexta*chuva", "sabado*chuva", "outono-inverno*chuva", "segunda*vento medio",
"terca*vento medio", "quarta*vento medio", "quinta*vento medio", "sexta*vento medio",
"sabado*vento medio", "segunda*vento forte", "terca*vento forte", "quarta*vento forte",
"quinta*vento forte", "sexta*vento forte", "sabado*vento forte", "outono-inverno*vento medio",
"outono-inverno*vento forte", "chuva*vento medio", "chuva*vento forte", "segunda*outono-inverno*chuva",
"terca*outono-inverno*chuva", "quarta*outono-inverno*chuva", "quinta*outono-inverno*chuva",
"sexta*outono-inverno*chuva", "sabado*outono-inverno*chuva", "segunda*outono-inverno*vento medio",
"terca*outono-inverno*vento medio", "quarta*outono-inverno*vento medio",
"quinta*outono-inverno*vento medio", "sexta*outono-inverno*vento medio",
"sabado*outono-inverno*vento medio", "segunda*outono-inverno*vento forte",
"terca*outono-inverno*vento forte", "quarta*outono-inverno*vento forte",
"quinta*outono-inverno*vento forte", "sexta*outono-inverno*vento forte",
"sabado*outono-inverno*vento forte", "segunda*chuva*vento medio", "terca*chuva*vento medio",
"quarta*chuva*vento medio", "quinta*chuva*vento medio", "sexta*chuva*vento medio",
"sabado*chuva*vento medio", "segunda*chuva*vento forte", "terca*chuva*vento forte",
"quarta*chuva*vento forte", "quinta*chuva*vento forte", "sexta*chuva*vento forte",
"sabado*chuva*vento forte", "outono-inverno*chuva*vento medio", "outono-inverno*chuva*vento forte")
dimnames(matriz.X) <- list(NULL,nome.matriz.X)

## Modelo logístico usual para encontrar a estimativa inicial de beta ##

# Construção de variáveis indicadoras dos dias da semana, estação do ano, chuva e vento

semana <- factor(dados$semana)
semana <- C(semana, treatment)
estacao <- factor(dados$estacao)
estacao <- C(estacao, treatment)

```

```

chuva <- factor(dados$chuva)
chuva <- C(chuva, treatment)
vento <- factor(dados$vento)
vento <- C(vento, treatment)

# Ajuste do modelo logístico usual

fit.NO2D <- glm(dados$NO2D ~ dados$temp + dados$umidade + semana + estacao + chuva + vento
+ dados$temp*dados$umidade + semana*estacao + semana*chuva + estacao*chuva + semana*vento + estacao*vento
+ chuva*vento + semana*estacao*chuva + semana*estacao*vento + semana*chuva*vento + estacao*chuva*vento,
family = binomial(link = logit), na = na.omit)
summary(fit.NO2D)

#####
# Modelo Completo #
#####

beta <- matrix(c(fit.NO2D$coefficients),nrow=n.par,ncol=1) # beta inicial
corte <- 100
NO2 <- matrix(dados$NO2)
log.corte <- log(corte, base=exp(1))
log.NO2 <- log(NO2, base=exp(1))

n.it <- 15 # número de iterações

eta <- matrix(nrow=n.obs,ncol=1)
for (i in 1:n.obs){ eta[i,] <- matriz.X[i,]%*%beta }
exp.eta <- exp(eta)
g.inv.eta <- exp.eta/(1+exp.eta)
inv.fdist <- matrix(qnorm(g.inv.eta,mean=0,sd=1),nrow=n.obs,ncol=1)
densidade <- matrix(dnorm(inv.fdist,mean=0,sd=1),nrow=n.obs,ncol=1)

dp <- 1
it.dp <- 0
dif.dp <- 1

while((abs(dif.dp) > 1e-006) && ((it.dp <- it.dp + 1) < n.it)){

  # Cálculo de sigma estimado

  a.sigma <- n.obs
  b.sigma <- (log.NO2-log.corte)*inv.fdist
  b.sigma.soma <- sum(b.sigma)
  c.sigma <- (log.NO2-log.corte)^2
  c.sigma.soma <- sum(c.sigma)
  sigma <- ((-b.sigma.soma)+sqrt(b.sigma.soma^2+4*a.sigma*c.sigma.soma))/(2*a.sigma)

  dif.dp <- sigma - dp
  dp <- sigma

  # Algoritmo para estimar beta

  b <- beta
  it.b <- 0
  dif.b <- 1
  converge <- 1

  while((converge > 1e-006) && ((it.b <- it.b + 1) < n.it)){

    # Calcular a matriz média

```

```

matriz.media <- inv.fdist

# Cálculo da matriz L e matriz W

logistica <- exp.eta/((1+exp.eta)^2)
dgamadeta <- (1/densidade)*logistica
matriz.L <- diag(dgamadeta,nrow=n.obs,ncol=n.obs)
matriz.W <- t(matriz.L)%*%matriz.L

# Variável dependente: matriz z

matriz.z <- eta + solve(matriz.L)%*%((log.NO2 - log.corte)/sigma - matriz.media)

# Matriz beta estimada

beta <- solve(t(matriz.X)%*%matriz.W)%*%(t(matriz.X)%*%matriz.z)

dif.b <- b - beta
converge <- t(abs(dif.b)/b)%*%(abs(dif.b)/b)
b <- beta

# Matriz Informação Esperada de Fisher

matriz.inform.fisher <- (t(matriz.L)%*%matriz.X)%*%(matriz.L)%*%matriz.X)
matriz.inform.fisher.inv <- solve(matriz.inform.fisher)
matriz.variancia <- matrix(nrow=n.par,ncol=1)
for (i in 1:n.par) matriz.variancia[i,] <- matriz.inform.fisher.inv[i,i]
matriz.erropadrao <- sqrt(matriz.variancia)

for (i in 1:n.obs){ eta[i,] <- matriz.X[i,]%*%beta }
exp.eta <- exp(eta)
g.inv.eta <- exp.eta/(1+exp.eta)
inv.fdist <- matrix(qnorm(g.inv.eta,mean=0,sd=1),nrow=n.obs,ncol=1)
densidade <- matrix(dnorm(inv.fdist,mean=0,sd=1),nrow=n.obs,ncol=1)

  cat(it.b,beta,"\n")
}
cat(it.dp,sigma,"\n")
}
print(beta)
print(matriz.erropadrao)
print(sigma)
}

```

Lista de Tabelas

2.1	Resultados obtidos no estudo de simulação, para diferentes tamanhos de amostra (n), distâncias $C - E(Y)$ e valores dos parâmetros, adotando o modelo logístico usual e o modelo com resposta contínua (função de ligação <i>logito</i>)	15
3.1	Padrões Primários e Secundários de Qualidade do Ar referentes às concentrações dos poluentes, estabelecidos pelo CONAMA	17
3.2	Médias (Erros padrão) das concentrações dos poluentes nos anos de 1998 e 1999	21
3.3	Percentis observados das concentrações dos poluentes	21
3.4	Medidas descritivas para as variáveis Temperatura e Umidade e concentrações dos poluentes NO ₂ , O ₃ , SO ₂ , PM ₁₀ e CO nas diferentes combinações de Chuva e Vento	22
3.5	Médias (Erros padrão) das variáveis climáticas	23
3.6	Distribuição conjunta de frequências das variáveis climáticas categorizadas nas estações do ano Primavera-Verão e Outono-Inverno	23
3.7	Níveis descritivos do teste de Anderson-Darling para a normalidade do log(concentrações dos poluentes)	24
3.8	Modelos finais usual e com resposta contínua ajustados	25
3.9	Riscos médios previstos e erros padrão estimados ($\widehat{EP}$) pelo modelo logístico com resposta contínua, para dias com Temperatura $\leq 17^{\circ}\text{C}$ e Umidade $\leq 75\%$	28
3.10	Riscos médios previstos e erros padrão estimados ($\widehat{EP}$) pelo modelo logístico com resposta contínua, para dias com Temperatura $> 17^{\circ}\text{C}$ e Umidade $\leq 75\%$	29

3.11	Riscos médios previstos e erros padrão estimados ($\widehat{EP}$) pelo modelo logístico com resposta contínua, para dias com Temperatura $\leq 17^\circ\text{C}$ e Umidade $> 75\%$	30
3.12	Riscos médios previstos e erros padrão estimados ($\widehat{EP}$) pelo modelo logístico com resposta contínua, para dias com Temperatura $> 17^\circ\text{C}$ e Umidade $> 75\%$	31
3.13	Proporções de previsões incorretas para diferentes pontos críticos	33
B.1	Configuração da Rede Automática	38
B.2	Modelos iniciais usual e com resposta contínua ajustados	43
B.3	Efeitos não significantes no ajuste do modelo logístico com resposta contínua	44
B.4	Intervalos de classe para os riscos estimados ($\hat{\pi}_i$) considerados na construção da Figura 3.2	45
B.5	Valores observados e esperados para o construção do teste Qui-quadrado de qualidade de ajuste	45

Lista de Figuras

3.1	Mapa de localização das 29 estações fixas de monitoramento do ar operadas pela CETESB no estado de São Paulo	18
3.2	Gráfico da proporção estimada versus ponto médio	26
3.3	Gráfico com riscos médios previstos pelo modelo logístico com resposta contínua em dias sem chuva e vento fraco	32
B.1	Representação gráfica das séries cronológicas para as concentrações dos poluentes NO ₂ , O ₃ , SO ₂ , PM ₁₀ e CO nos anos de 1998 e 1999	39
B.2	Representação gráfica das séries cronológicas de Temperatura, Umidade, Precipitação e Velocidade do vento nos anos de 1998 e 1999	40
B.3	Diagramas de dispersão do poluente NO ₂ pela Temperatura com suavizador “lowess”, considerando duas categorias de Umidade	40
B.4	Box-plot das concentrações do NO ₂ por Dia da semana para as diferentes combinações de Chuva e Vento	41
B.5	Gráficos de Probabilidade Normal para as concentrações dos poluentes NO ₂ , O ₃ , SO ₂ , PM ₁₀ e CO	42

Bibliografia

- Braga, A. L. F., Saldiva, P. H. N., Pereira, L. A. A., Menezes, J. J. C., Conceição, G. M. S., Lin, C. A., Zanobetti, A., Schwartz, J. and Dockery, D. W. (2001). Health effects of air pollution exposure on children and adolescents in São Paulo, Brazil. *Pediatric Pulmonology*, **31**, 106-113.
- Böhm, G. M., Saldiva, P. H. N., Pereira, L. A. A., Conceição, G. M. S., Miraglia, S. G. K. and Braga, A. L. F. (2002). Mortality caused by São Paulo air pollution in the last decade of the 20th century: a critical review. (Submetido para publicação).
- CETESB (2001). *Relatório de qualidade do ar no Estado de São Paulo 2001*. São Paulo. Disponível em: http://www.cetesb.sp.gov.br/Ar/ar_geral.htm. Acesso em: 02 de setembro de 2002.
- Chambers, J. M. and Hastie, T. J. (1993). *Statistical Models in S*. Wadsworth & Brooks, California.
- Cordeiro, G. M. (1986). *Modelos Lineares Generalizados*. VII Simpósio Brasileiro de Probabilidade e Estatística, UNICAMP. Campinas, São Paulo.
- Demétrio, C. G. B. (2001). *Modelos Lineares Generalizados em Experimentação Agrônômica*. 46ª Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria (RBRAS) e 9º Simpósio de Estatística Aplicada à Experimentação Agrônômica (SEAGRO), ESALQ/USP. Piracicaba, São Paulo.
- Hong, Y., Lee, J., Kim, H., Ha, E., Schwartz, J. and Christiani, D. C. (2002). Effects of air pollutants on acute stroke mortality. *Environmental Health Perspectives*, **2**, **110**, 187-190.
- Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression*. John Wiley, New York.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, 2nd Edition*. Chapman and Hall, London.

-
- Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996). *Applied linear statistical models, 4th Edition*. McGraw-Hill, Boston.
- Paula, G. A. (2002). *Modelos de Regressão com apoio computacional*. São Paulo: IME/USP.
- 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. and Saldiva, P. H. N. (1998). Association between air pollution and intrauterine mortality in São Paulo, Brazil. *Environmental Health Perspectives*, **106**, 325-329.
- Piskunov, N. S. (1983). *Cálculo Diferencial e Integral, 6ª edição*. Mir, Moscou.
- Rao, C. R. (1973). *Linear statistical inference and its applications*. John Wiley, New York.
- Ronchetti, E., Heritier, S. and Morabia, A. (1997). *Robust Binary Regression with Continuous Outcomes*. Genève: Cahiers du Département d'Econométrie, Université de Genève, 21p. (RePEc:gen:geneem:97.03).
- Saez, M., Ballester, F., Barceló, M. A., Pérez-Hoyos, S., Bellido, J., Tenías, J. M., Ocaña, R., Figueiras, A. Arribas, F., Aragonés, N. Tobías, A., Cirera, L. and Cañada A. (2002). A combined analysis of the short-term effects of photochemical air pollutants on mortality with the EMECAM Project. *Environmental Health Perspectives*, **3**, **110**, 221-228.
- Singer, J. M., André, C. D. S., Lima, L. P. and Conceição, G. M. S. (2002). Atmospheric Pollution and Mortality en São Paulo, Brasil. In Yadolah Dodge ed: *Statistics in Industry and Technology: Statistical Data Analysis*, 439-450. Birhhäuser, Basel.
- Stephens, M. A. (1974). EDF Statistics for goodness of fit and some comparisons, *Journal of the American Statistical Association*, **69**, 730-737.
- Suissa, S. and Blais, L. (1995). Binary regression with continuous outcomes. *Statistics in Medicine*, **14**, 247-255.
- Thompson, R. and Baker, R. J. (1981). Composite link functions in generalized linear models. *Applied Statistics*, **30**, 125-131.
- Ye, F., Piver, W. T., Ando, M. and Portier, C. J. (2001). Effects of temperature and air pollutants on cardiovascular and respiratory diseases for males and females older than 65 years of age in Tokyo, July and August 1980-1995. *Environmental Health Perspectives*, **4**, **109**, 355-359.