Ambiente & Sociedade
ISSN: 1414-753X
revista@nepam.unicamp.br
Associação Nacional de Pós-Graduação e Pesquisa em Ambiente e Sociedade
Brasil

Souza Tadano, Yara de; Lie Ugaya, Cássia Maria; Teixeira Franco, Admilson
Método de regressão de Poisson: metodologia para avaliação do impacto da poluição atmosférica na saúde populacional
Ambiente & Sociedade, vol. XII, núm. 2, julio-diciembre, 2009, pp. 241-255
Associação Nacional de Pós-Graduação e Pesquisa em Ambiente e Sociedade
Campinas, Brasil

MÉTODO DE REGRESSÃO DE POISSON: METODOLOGIA PARA AVALIAÇÃO DO IMPACTO DA POLUIÇÃO ATMOSFÉRICA NA SAÚDE POPULACIONAL

YARA DE SOUZA TADANO[1]
CÁSSIA MARIA LIE UGAYA[1]
ADMILSON TEIXEIRA FRANCO[1]

1 *Introdução*

Em estudos epidemiológicos, os modelos freqüentemente utilizados são os modelos estatísticos e analíticos que, segundo Conceição et al. (2001, p. 207), "constituem ferramentas extremamente úteis para resumir e interpretar dados. Em particular, estes modelos podem facilitar a avaliação da forma e da intensidade de associações de interesse em estudos epidemiológicos".

O modelo estatístico utilizado na análise da relação entre a poluição atmosférica e o impacto na saúde é a análise de regressão, pois é uma ferramenta útil para avaliar a relação entre uma ou mais variáveis explicativas (variáveis independentes, preditoras ou covariáveis) (x_1, x_2, ..., x_n) e uma única variável resposta (variável dependente, prevista) (y) (MARTINS, 2000).

Neste artigo será apresentada, primeiramente, a definição da análise de regressão. Em seguida, será descrito o modelo de regressão de Poisson dos modelos lineares generalizados (MLG), com o qual é possível analisar os dados da poluição atmosférica na saúde populacional.

[1]Programa de Pós-graduação em Engenharia Mecânica e de Materiais, Universidade Tecnológica Federal do Paraná – UTFPR, Curitiba, Paraná, Brasil.

Autor para correspondência: Yara de Souza Tadano, Universidade Tecnológica Federal do Paraná – UTFPR, Rua Enoque Antônio de Aquino, 520, CEP 79950-000, Naviraí, MS, Brasil. E-mail: yaratadano@gmail.com

Recebido: 3/4/2009. Aceito: 24/6/2009

Devido à grande importância das decisões a serem tomadas desde a coleta e análise dos dados até a verificação do ajuste do modelo escolhido, serão apresentados os passos envolvidos em uma análise de regressão.

Finalmente, encontram-se alguns comentários finais sobre as vantagens e limitações do modelo proposto.

2 *Análise de regressão*

A análise de regressão que envolve apenas uma variável explicativa é chamada de regressão simples, enquanto a análise envolvendo duas ou mais variáveis explicativas é denominada regressão múltipla (HAIR jr. et al., 2005).

A regressão linear múltipla é dada por (Equação 1):

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \varepsilon \quad (1)$$

onde y é a variável resposta e x_i (i = 1, 2, ..., n) são as variáveis explicativas. β_0 representa o valor de y quando as variáveis explicativas são nulas, os termos β_i são chamados de coeficientes de regressão e o resíduo (ε) é o erro de previsão, ou seja, a diferença entre os valores reais e os previstos da variável resposta, que é assumido normalmente distribuído com média zero e variância σ^2 (HAIR Jr. et al., 2005).

O objetivo da análise de regressão linear múltipla, assim como de todos os tipos de regressão, é encontrar uma equação (chamada de equação de regressão, variável estatística de regressão ou modelo de regressão) que prevê de maneira melhor a variável resposta a partir de uma combinação das variáveis explicativas, ou seja, deseja-se encontrar os valores dos β's que melhor se ajustem aos dados do problema (HAIR Jr. et al., 2005).

Encontrados os β's, é necessário validar o modelo de regressão, que consiste em verificar se sinais e magnitude dos coeficientes fazem sentido no contexto do fenômeno estudado, que pode ser feito através do teste *t* de *Student* como será apresentado na análise dos resultados (WERKEMA; AGUIAR, 1996).

Nem sempre é possível aplicar um modelo de regressão linear em estudos epidemiológicos, como por exemplo, estudos sobre o impacto da poluição atmosférica na saúde populacional, devido ao caráter não linear da variável resposta. Nestes casos, geralmente utilizam-se as classes de modelos que oferecem uma poderosa alternativa para a transformação de dados, chamadas de modelos lineares generalizados (MLG) e modelos aditivos generalizados (MAG) (SCHMIDT, 2003).

O estudo realizado por Conceição et al. (2001) descreveu e comparou estas duas classes de modelos que podem ser utilizadas para avaliar a associação entre poluição atmosférica e morbidade ou mortalidade por causas específicas: os modelos lineares generalizados (MLG) e os modelos aditivos generalizados (MAG).

Conceição et al. (2001) concluíram que as duas classes de modelos apresentadas produziram resultados coerentes, sendo que a abordagem baseada em curvas suavizadas[1] utilizada nos MAG permite que o padrão sazonal seja definido pelos próprios dados, sem a imposição de uma estrutura rígida e, talvez, menos fidedigna encontrada nos MLG, devido

à inclusão de uma variável referente aos meses do ano. Existem, entretanto, desvantagens em utilizar curvas suavizadas, pois os coeficientes estimados correspondentes nos modelos de regressão não são interpretáveis (CHAPRA; CANALE, 1987).

Nos MLG, foram consideradas 18 variáveis explicativas, enquanto apenas 4 foram requeridas pelos MAG. Assim, aparentemente, os MAG são modelos mais parcimoniosos, ou seja, necessitam de um número menor de variáveis explicativas, o que justificaria o fato de detectarem um número maior de associações (CONCEIÇÃO et al., 2001).

O grande desafio existente nas avaliações do impacto da poluição atmosférica na saúde é encontrar um modelo estatístico capaz de considerar todos os fatores envolvidos, pois para avaliar o impacto da poluição atmosférica na saúde, é necessário levar em consideração que cada pessoa reage de forma diferente a uma determinada concentração de poluente.

No corrente trabalho será apresentada, então, uma metodologia para avaliar o impacto da poluição atmosférica na saúde, com a aplicação da família de Poisson para os MLG.

As curvas suavizadas, utilizadas nos modelos MAG por Conceição et al. (2001), foram incluídas nos modelos MLG, para suavizar a sazonalidade. No presente trabalho a curva suavizada utilizada foi a *spline* cúbica (*cubic splines*) (CHAPRA; CANALE, 1987).

3 *Modelos lineares generalizados (MLG)*

Os modelos lineares generalizados (MLG) representam a união de modelos lineares e não-lineares com uma distribuição da família exponencial, que é formada pela distribuição normal, Poisson, binomial, gama, normal inversa e incluem modelos lineares tradicionais (erros com distribuição normal), bem como modelos logísticos (SCHMIDT, 2003).

Desde 1972, inúmeros trabalhos relacionados com modelos lineares generalizados foram publicados, resultando em diversas ferramentas computacionais, como por exemplo, GLIM (*Generalized Linear Interactive Models*), S-Plus, R, SAS, STATA e SUDAAN, bem como extensões desses modelos (PAULA, 2004).

Os MLG são definidos por uma distribuição de probabilidade, membro da família exponencial de distribuições, e são formados pelas seguintes componentes (McCULLAGH; NELDER, 1989; TADANO et al., 2006a):

- Componente aleatória: *n* variáveis explicativas $y_1, ..., y_n$, de uma variável resposta que segue uma distribuição da família exponencial com valor esperado $E(y_i) = \mu$;
- Componente sistemática: compõe uma estrutura linear para o modelo de regressão $\eta = \beta x^T$, chamado de preditor linear, onde $x^T = (x_{i1}, x_{12}, ..., x_{ip})^T$, $i = 1, ..., n$ são as chamadas variáveis explicativas; e
- Função de ligação: Uma função monótona e diferenciável g, chamada de função de ligação, capaz de conectar as componentes aleatória e sistemática, ou seja, relaciona a média da variável resposta (μ) à estrutura linear, definida nos MLG por $g(\mu) = \eta$, onde (TADANO, 2007) (Equação 2):

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \quad (2)$$

ou em forma matricial (Equação 3):

$$\eta = \beta x^T \quad (3)$$

Com o coeficiente de regressão $\beta = (\beta_1, \beta_2, ..., \beta_n)$ representando o vetor de parâmetros a ser estimado (McCULLAGH; NELDER, 1989).

Na Tabela 1, cada distribuição tem uma função de ligação especial, chamada de função de ligação canônica que ocorre quando $\eta = \theta_i$, onde θ é o chamado parâmetro de localização ou parâmetro canônico (McCULLAGH; NELDER, 1989).

Tabela 1. Funções de ligação canônica de algumas distribuições da família exponencial (McCULLAGH; NELDER, 1989).

Distribuição	Função de ligação canônica (η)
Normal	μ
Poisson	$\ln(\mu)$
Binomial	$\ln\{\mu/(1-\mu)\}$
Gamma	μ^{-1}
Gaussiana Inversa	μ^{-2}

De acordo com Myers e Montgomery (2002), a utilização da função de ligação canônica implica algumas propriedades interessantes, porém não quer dizer que deva ser utilizada sempre. Essa escolha é conveniente porque, além de simplificar as estimativas dos parâmetros do modelo, também facilita o cálculo do intervalo de confiança[2] para a média da variável resposta. Contudo, a conveniência não implica necessariamente em qualidade de ajuste do modelo.

Conforme o exposto acima, se η é a função logarítmica e y_i possui distribuição de Poisson, o modelo resultante é o modelo de regressão de Poisson com função de ligação canônica, utilizado para avaliar dados não-negativos em forma de contagens, frequentemente encontrados em estudos epidemiológicos.

4 *Modelo de regressão de Poisson*

Dentre as famílias dos MLG, a mais utilizada em estudos sobre o impacto da poluição atmosférica na saúde é a de Poisson (CONCEIÇÃO et al., 2001; MARTINS, 2000; TADANO, 2007; TADANO et al., 2006b).

O modelo de regressão de Poisson tem por característica a análise de dados contados na forma de proporções ou razões de contagem, ou seja, leva em consideração o total de pessoas com uma determinada doença (McCULLAGH; NELDER, 1989).

O modelo de regressão de Poisson é um tipo específico dos MLG e MAG que teve origem por volta de 1970, quando Wedderburn (1974) desenvolveu a teoria da quasi-verossimilhança, analisada com mais detalhes por McCullagh (1983).

A variável resposta de uma regressão de Poisson deve seguir uma distribuição de Poisson e os dados devem possuir igual dispersão, ou seja, a média da variável resposta deve ser igual à variância. Entretanto, conforme Ribeiro (2006), quando se trabalha com dados experimentais, esta propriedade é frequentemente violada. Assim, pode-se ter uma

superdispersão quando a variância é maior que a média; ou uma subdispersão quando a variância é menor que a média (SCHMIDT, 2003). Nestes casos, ainda é possível aplicar o modelo de regressão de Poisson realizando-se alguns ajustes.

A representação das probabilidades da distribuição de Poisson para $\mu = 1, 2, 4, 6$ encontra-se na Figura 1 (SCHMIDT, 2003), onde y indica uma variável qualquer.

A Figura 1 mostra o gráfico da distribuição de Poisson (P(y)) para quatro valores de μ. Observa-se um achatamento da curva e o seu deslocamento para a direita quando μ aumenta. À medida que μ aumenta, a curva aproxima-se de uma distribuição normal (TADANO, 2007).

Antes de escolher um modelo estatístico, é necessário especificar o escopo do estudo para realizar a coleta dos dados necessários.

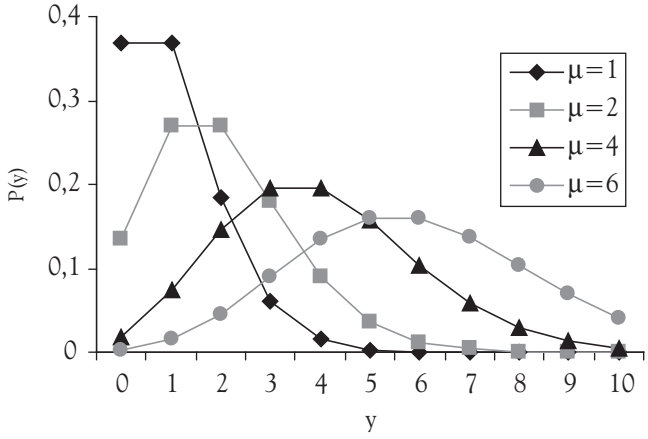

Figura 1. Gráfico da distribuição de Poisson para quatro valores de μ. Fonte: Adaptado de SCHMIDT (2003).

5 *Coleta e análise dos dados*

Em estudos epidemiológicos, é importante coletar dados separados por faixa etária devido à grande diferença existente na reação de cada uma a um determinado fator.

Na seleção do curso de ação para a melhoria do banco de dados, é essencial compreender os processos que conduzem a dados faltantes na amostra, pois são uma realidade em qualquer tipo de análise (HAIR Jr. et al., 2005).

Para alguns problemas em particular, existem na literatura aproximações padrões. Utilizar estas aproximações acarretará um aumento na confiabilidade dos dados. Por exemplo, os órgãos responsáveis pelo monitoramento da qualidade do ar seguem um padrão para o cálculo de médias diárias, mensais e anuais de variáveis meteorológicas e de concentração de poluentes.

O exame dos dados pode parecer uma tarefa comum e sem importância, porém é fundamental em qualquer análise de regressão, pois o banco de dados é a parte fundamental da análise, assim a confiabilidade do banco de dados é essencial (HAIR Jr. et al., 2005).

Com os devidos ajustes no banco de dados, parte-se para a escolha da modelagem estatística que melhor descreva os dados.

6 *Escolha do modelo estatístico*

Para confirmar que o modelo de regressão de Poisson dos MLG pode ser utilizado para avaliar o impacto da poluição atmosférica na saúde populacional, é preciso verificar primeiro se existe possibilidade de se aplicar a regressão estatística mais simples (regressão linear). Uma característica importante da regressão linear é a normalidade dos dados, ou seja, os dados devem possuir uma distribuição aproximadamente normal.

O teste de diagnóstico mais simples para verificar a normalidade é o histograma, que compara os valores dos dados observados com uma distribuição aproximadamente normal, como mostra a Figura 2. Este teste é atraente devido à simplicidade, porém é um método problemático para amostras com menos de 100 observações. Para estes casos, existem outros métodos mais confiáveis, como o teste de assimetria e curtose (HAIR Jr. et al., 2005; FONSECA; MARTINS, 1996).

Em estudos sobre o impacto da poluição atmosférica na saúde frequentemente tem-se mais de 100 observações, pois são necessários dados diários durante pelo menos dois anos para a realização de uma boa análise (MARTINS, 2000; LATORRE; CARDOSO, 2001). Portanto, a análise do histograma é suficiente para verificar a normalidade dos dados.

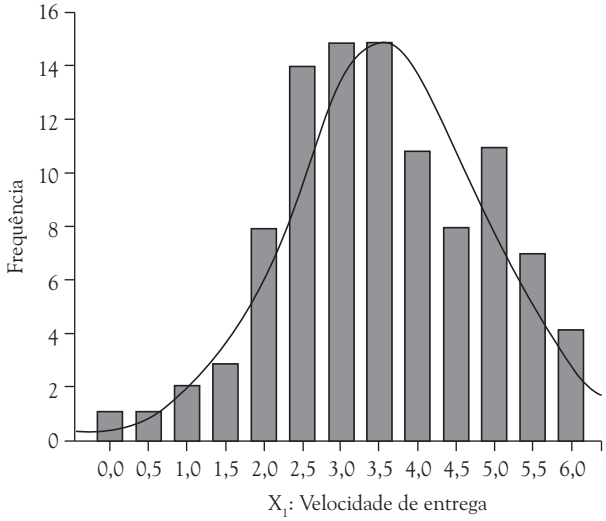

Figura 2. Representação gráfica de um histograma com distribuição aproximadamente normal. Fonte: HAIR Jr. et al. (2005).

Uma outra característica importante dos dados é a linearidade, ou seja, a magnitude da relação linear entre duas variáveis. A linearidade pode ser observada através de um diagrama de dispersão[3] (HAIR Jr. et al., 2005).

Em uma matriz de dispersão, como no exemplo da Figura 3, são apresentados o diagrama de dispersão abaixo da diagonal principal, o histograma na diagonal principal e os coeficientes de correlação entre as variáveis acima da diagonal principal (HAIR Jr. et al., 2005).

Dados que não seguem uma distribuição normal e são não negativos em forma de contagem indicam a possibilidade de aplicação da família de Poisson dos MLG. Para se

aplicar o modelo de Poisson, os dados devem possuir distribuição de Poisson, ou seja, o histograma deve apresentar uma tendência de acordo com as linhas da Figura 1, exceto para $\mu = 6$ que indica uma distribuição normal (SCHMIDT, 2003; TADANO, 2007).

Se os dados seguem uma distribuição de Poisson, inicia-se a aplicação do modelo de regressão de Poisson.

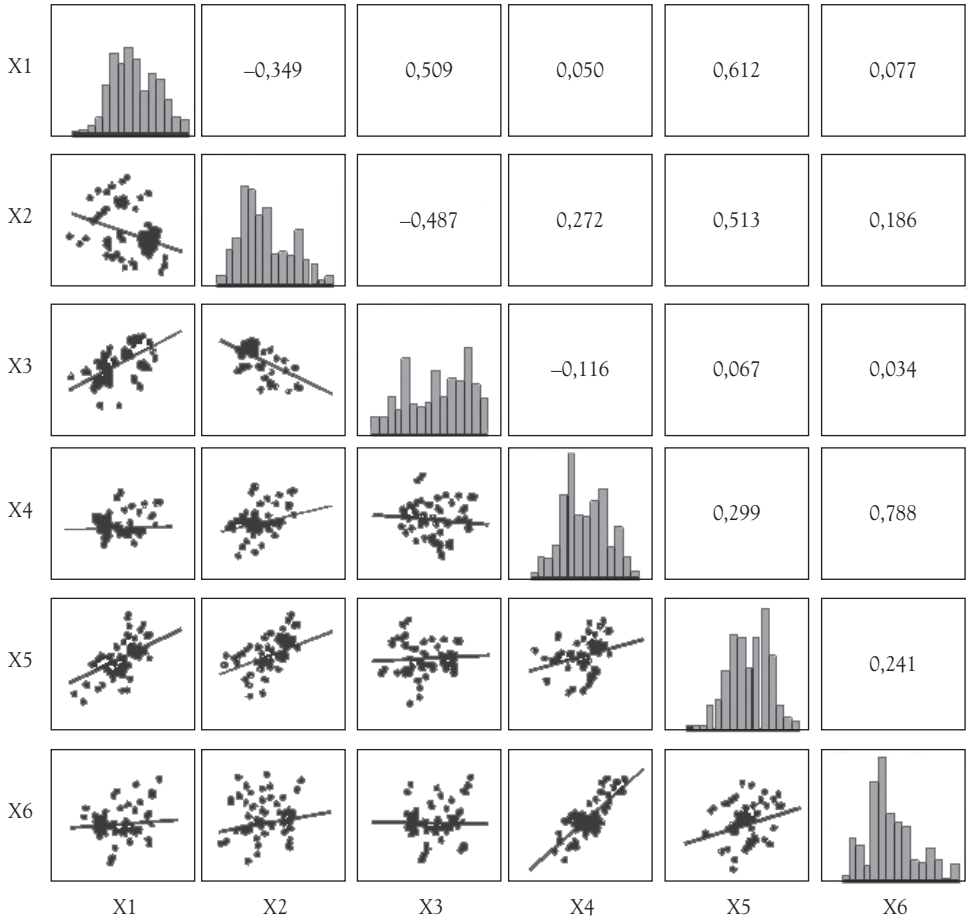

Figura 3. Matriz de dispersão. Fonte: Adaptado de HAIR Jr. et al. (2005).

6.1 Estimativa dos parâmetros

Com todas as suposições satisfeitas, inicia-se a estimativa do modelo que pode ser feita com o auxílio de programas estatísticos capazes de encontrar os coeficientes β_i da equação 2 (S-PLUS 7, 2005).

A estimativa dos parâmetros pode ser uma tarefa relativamente simples se houver somente um parâmetro a ser estimado, contudo não é o que ocorre normalmente. Nestes casos, processos iterativos devem ser empregados para resolver o sistema de equações.

Nos MLG, o método utilizado para estimar os valores dos parâmetros de regressão β_i é conhecido como método escore de Fisher para maximização da função de verossimilhança, que coincide com o método de Newton-Raphson quando a função de ligação é a canônica (PAULA, 2004; McCULLAGH; NELDER, 1989).

Desta forma, para o modelo de regressão de Poisson, a função densidade de probabilidade é (Equação 4):

$$f_y(y; \theta; \phi) = \exp\{y \ln(\mu) - \mu - \ln(y!)\} \quad (4)$$

sendo que y é a variável resposta, θ é o parâmetro de dispersão, ϕ é o parâmetro canônico e ! significa fatorial. Quando a função de ligação considerada é a ligação canônica ($\ln \mu = \eta$).

6.2 Ajuste de tendências temporais

a) Dias da semana

Em estudos sobre o impacto da poluição atmosférica na saúde, é necessário levar em consideração a tendência temporal. Por exemplo, nos finais de semana, o número de atendimentos hospitalares é menor do que nos dias de semana. Uma forma de ajustar esta tendência é acrescentar uma variável qualitativa para dia da semana, que varia de 1 a 7, começando a contagem no domingo (TADANO, 2007).

O número de atendimentos hospitalares nos feriados também é menor do que nos dias em que não é feriado. Esta tendência é ajustada através do acréscimo de uma variável binomial, ou seja, os dias de feriado recebem valor 1 e os dias em que não é feriado recebem valor 0.

b) Sazonalidade

Outra tendência temporal importante é a sazonalidade, pois as variáveis meteorológicas e concentração de poluentes variam no decorrer do ano. Para ajustar estas tendências frequentemente utiliza-se uma função chamada *spline*, já que essa função fornece uma aproximação melhor que as tendências polinomais (CHAPRA; CANALE, 1987).

Spline é um tipo de função que fornece uma aproximação do comportamento das funções que têm mudanças locais e abruptas. Nos *splines*, ao invés de utilizar apenas um polinômio para todo o conjunto de dados, definem-se alguns intervalos e estima-se uma função polinomial para cada um dos intervalos. O *spline* mais utilizado para suavização de curvas é o *spline* cúbico. O objetivo dos *splines* cúbicos é derivar um polinômio de terceira ordem para cada intervalo entre 2 nós (ponto onde dois *splines* se encontram) (CHAPRA; CANALE, 1987).

Em estudos sobre o impacto da poluição atmosférica na saúde, utilizam-se 4 a 5 nós por ano, pois a sazonalidade ocorre devido às diferenças existentes entre as estações do ano.

Apesar dos *splines* fornecerem uma boa aproximação para tendências temporais, não possuem interpretação física, portanto não podem ser utilizados para realizar previsões. Nestes casos, opta-se por utilizar uma variável qualitativa para os meses do ano que possui valores entre 1 (referente à janeiro) e 12 (referente à dezembro) (CONCEIÇÃO et al., 2001; CHAPRA; CANALE, 1987).

c) Correlação de dados com o tempo

As tendências temporais podem provocar autocorrelação entre os dados, ou seja, os dados de um dia podem estar correlacionados com os dados do dia anterior devido à diferença entre os atendimentos hospitalares nos finais de semana e feriados; mesmo após serem incluídas variáveis explicativas para dias da semana e feriado. Assim, para verificar a existência de correlação entre os dados e decidir quais as providências a serem tomadas, são construídos gráficos da função de autocorrelação parcial em relação ao tempo de defasagem (*lag*) (TADANO, 2007).

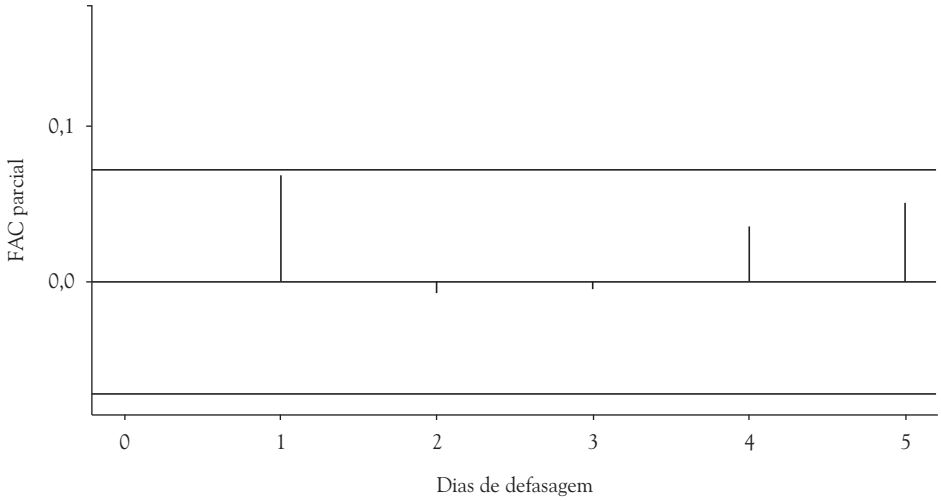

Figura 4. Exemplo de gráfico da função de autocorrelação (FAC) parcial em relação a dias de defasagem (*lag*) . Fonte: TADANO (2007).

A função de autocorrelação do resíduo é dada por (Equação 5):

$$FAC = \frac{c_k}{c_0} \tag{5}$$

onde $c_k = \frac{1}{n}\sum_{i=1}^{n-k}(y_i - \mu)(y_{i+k} - \mu)$ sendo que *n* é o número de observações e *k* representa os dias de defasagem (BOX et al., 1994).

No gráfico da função de autocorrelação, os resíduos devem ser os menores possíveis, encontrando-se numa faixa entre $[-2n^{-1/2}, 2n^{-1/2}]$, onde n é o número de observações presentes no problema, como mostra a Figura 4 (FERRAZ et al., 1999).

No caso de estudos epidemiológicos, as autocorrelações importantes e possíveis de serem interpretadas ocorrem nos primeiros 4 dias. Estas autocorrelações geralmente são devido ao reduzido número de atendimentos hospitalares nos finais de semana (TADANO, 2007).

Se os dados estão correlacionados, deve-se ajustar o modelo levando em consideração essas autocorrelações. Esta correção é feita através da inserção do resíduo no modelo.

Todas as considerações sobre as tendências temporais devem ser observadas quando se realiza um estudo, por exemplo, sobre o impacto de um determinado poluente na saúde populacional.

Outros fatores que geralmente são considerados nestes estudos são os efeitos da temperatura e da umidade.

Deseja-se então, após considerar os fatores mencionados, encontrar os valores dos coeficientes β's da seguinte Equação 6:

$$\ln(DS) = \beta_0 + \beta_1 temp + \beta_2 umid + \beta_3 conc + \beta_4 ns + \beta_5 dds + \beta_6 fer \quad (6)$$

onde DS - diagnóstico na saúde; temp - temperatura; umid - umidade; conc - concentração de um determinado poluente; ns - spline cúbico natural; dds - dia da semana e fer - feriado.

Uma vez escolhido o modelo que melhor se ajuste aos dados e determinados os coeficientes de regressão β's, é necessário avaliar este ajuste.

6.3 Avaliação do ajuste

Um teste interessante e fácil de ser aplicado para avaliar o ajuste de um modelo é chamado de estatística pseudo R^2 que é similar ao coeficiente de determinação obtido nos modelos lineares clássicos, e é definida por (Equação 7):

$$R_D^2 = 1 - \frac{D_c}{D_0} \quad (7)$$

onde D_c é o desvio do modelo ajustado e D_0 é o desvio do modelo nulo, ou seja, desvio antes da aplicação do modelo. Esta estatística mede a redução no desvio devido à inclusão de variáveis explicativas (RIBEIRO, 2006).

A principal ferramenta capaz de avaliar o ajuste do modelo, porém, é a análise de resíduos. Para os MLG, são definidos quatro tipos diferentes de resíduos capazes de avaliar o ajuste do modelo, o desvio residual, resíduo *working*, resíduo de Pearson e resíduo resposta. Dentre eles, será apresentado o desvio residual (*deviance residual*), que é capaz de detectar observações atípicas que influenciam o processo de ajuste do modelo e o resíduo de Pearson por ser uma versão aprimorada (*rescaled*) do resíduo *working* (S-PLUS 7, 2005).

O resíduo de Pearson está presente em uma estatística bastante utilizada para avaliar o ajuste do modelo, chamada estatística de Pearson ou Qui-quadrado χ^2 capaz de comparar a distribuição observada com a determinada pelo modelo através da seguinte expressão (SCHMIDT, 2003) (Equação 8):

$$\chi^2 = \sum (y - \hat{\mu})^2 / v(\hat{\mu}) \quad (8)$$

onde $v(\hat{\mu})$ é a função de variância estimada para a distribuição em questão.

Pode-se dizer que a estatística de Pearson é a soma dos resíduos de Pearson para cada observação.

Um modelo que se ajuste bem aos dados possui estatística de Pearson χ^2 aproximadamente igual ao seu grau de liberdade (gl), ou seja, $\chi^2/gl \sim 1$, caso contrário, pode se dizer que o modelo é inadequado, podendo tratar-se de um problema de superdispersão (RUSSO, 2002).

6.4 Análise gráfica

Entre os tipos de análises gráficas comumente utilizadas para avaliar o ajuste de um MLG tem-se (BOX et al., 1994):

1) Gráfico dos desvios residuais de cada observação em relação aos valores ajustados pelo modelo.

Um modelo bem ajustado possui o gráfico com os pontos o mais próximo possível de zero no intervalo entre –2 e 2, como mostra a Figura 5 (BAXTER et al., 1997).

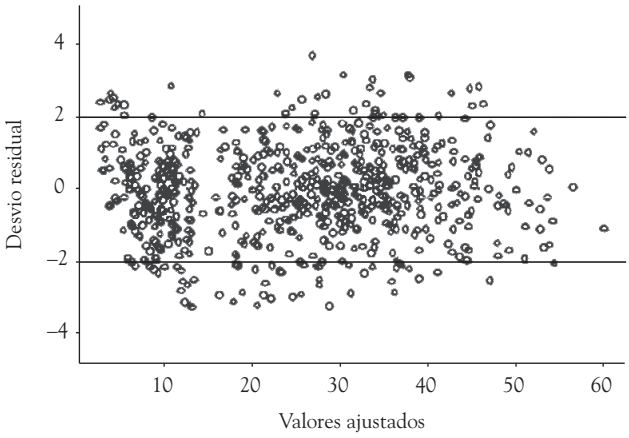

Figura 5. Exemplo de gráfico dos desvios residuais em função dos valores ajustados pelo modelo. Fonte: TADANO (2007).

O desvio residual, que em alguns trabalhos aparece como *deviance residual* ou resíduo *deviance*, é definido por (Equação 9):

$$r_i^D = |y_i - \hat{\mu}_i| \sqrt{d_i} \qquad (9)$$

onde d_i é a contribuição da i-ésima observação para o resíduo (S-PLUS 7, 2005).

2) Gráfico dos valores observados da variável resposta em relação aos valores ajustados pelo modelo.

Um exemplo deste gráfico é apresentado na Figura 6. Os pontos deste gráfico devem estar próximos da linha em que y = x, indicando que os valores ajustados estão próximos dos valores observados.

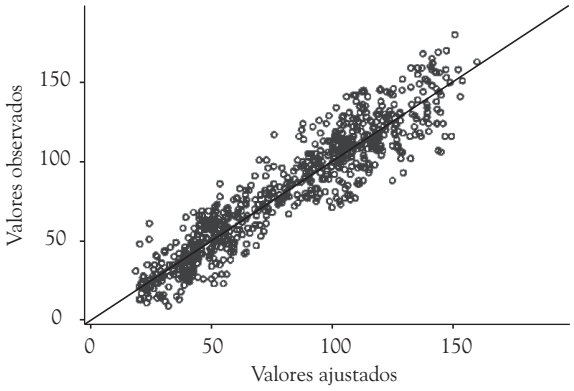

Figura 6. Exemplo de gráfico dos valores observados em função dos valores ajustados pelo modelo. Fonte: TADANO (2007).

7 *Análise dos resultados*

Para verificar a significância estatística dos coeficientes de regressão ajustados utiliza-se um teste de hipótese (TADANO et al., 2006b). A hipótese estatística a ser testada é designada por H_0, chamada hipótese nula, expressa por uma igualdade. A hipótese alternativa é dada por uma desigualdade (FONSECA; MARTINS, 1996).

O teste de hipótese possui várias finalidades, uma delas é verificar se o coeficiente de regressão estimado pode ser desprezado. Neste caso, consideram-se as hipóteses H_0: $\beta = 0$ e H_1: $\beta \neq 0$.

O teste estatístico utilizado para verificar esta hipótese é (Equação 10):

$$t_0 = \beta/\varepsilon \qquad (10)$$

onde ε é o erro padrão do coeficiente de regressão (β) estimado.

A rejeição da hipótese nula ocorre quando $|t_0| > t_{\alpha/2,n-k-1}$ (n = número de observações, k = número de variáveis explicativas), indicando que o valor encontrado para o coeficiente de regressão é estatisticamente significativo, ou seja, a variável explicativa considerada influencia nos resultados da variável resposta (FONSECA; MARTINS, 1996).

Os valores $t_{\alpha/2,n-k-1}$ são apresentados na tabela de distribuição t de *student*[4], onde gl é o grau de liberdade, dado por n – k – 1 e α é o nível de significância considerado (FONSECA; MARTINS, 1996).

Para a análise dos resultados em estudos epidemiológicos frequentemente utiliza-se uma medida chamada risco relativo, calculada com o uso dos parâmetros estimados no modelo (BAXTER et al., 1997).

8 *Risco relativo*

O risco relativo é uma medida da associação entre um fator particular (por exemplo, a concentração de poluentes atmosféricos) e o risco de um dado resultado (por exemplo, o número de pessoas com problemas respiratórias em uma região) (EVERITT, 2003).

De forma mais específica, a função risco relativo para um nível x de um poluente (Y) é definida por (BAXTER et al., 1997) (Equação 11):

$$RR(x) = \frac{E(Y|x)}{E(Y|x=0)} \tag{11}$$

Para o modelo de regressão de Poisson, o risco relativo é dado por (Equação 12):

$$RR(x) = e^{\beta x} \tag{12}$$

Isto indica, por exemplo, que o risco de uma pessoa exposta a uma concentração de poluente (x) adquirir uma doença específica é RR(x) vezes maior que uma pessoa que não foi exposta a esta concentração.

Um risco relativo igual a cinco para uma concentração de um determinado poluente de 100 $\mu g/m^3$, por exemplo, indica que uma pessoa exposta a uma concentração de 100 $\mu g/m^3$ possui cinco vezes mais chance de adquirir uma doença que uma pessoa que não foi exposta a essa concentração (TADANO, 2007).

9 *Comentários finais*

O processo de busca pelo modelo que melhor se ajusta a um conjunto de dados, apresentado neste trabalho, pode ser utilizado para qualquer problema que envolva uma variável resposta não negativa em forma de contagem.

Verificar o ajuste do modelo, dimensionando os erros envolvidos é essencial para confirmar se o modelo escolhido representa, adequadamente, os dados envolvidos.

Caso os erros sejam altos, é necessário encontrar um modelo que se ajuste melhor aos dados.

O risco relativo é uma medida muito importante para quantificar o impacto da poluição atmosférica na saúde populacional, bem como para alertar a população quanto aos riscos causados pela poluição.

De acordo com Conceição et al. (2001), para os MLG foram consideradas 18 variáveis explicativas, sendo 12 para os meses do ano, 4 para os anos de estudo, temperatura e umidade. Utilizando a função *spline* cúbica natural para o MLG, como apresentado neste trabalho, tem-se 6 variáveis explicativas. Assim, aparentemente, os MLG utilizando a função *spline* são mais parcimoniosos.

Existem, entretanto, desvantagens em utilizar a função *spline*, pois os coeficientes estimados nos modelos de regressão não são interpretáveis.

Substituindo a função *spline* por uma variável para os meses do ano, como apresentado por Conceição et al. (2001), tornam os coeficientes de regressão estimados possíveis de serem interpretados, porém, esta substituição pode provocar um aumento nos erros.

A fim de verificar se a aproximação utilizando a função *spline* se ajusta melhor que uma variável para os meses do ano, foi realizado um estudo de caso sobre o impacto do MP_{10} na saúde populacional do município de Araucária, PR (TADANO, 2007).

O estudo de caso será tema de um próximo artigo.

Agradecimentos

À Agência Nacional do Petróleo (ANP) e à Financiadora de Estudos e Projetos (FINEP) por meio do Programa de Recursos Humanos da ANP para o Setor de Petróleo e Gás – PRH – ANP/MCT (PRH10 – UTFPR) pelo apoio financeiro.

Referências bibliográficas

BAXTER, L. A. et al. Comparing estimates of the effects of air pollution on human mortality obtained using different regression methodologies. **Risk Analysis**, v. 17, n. 3, p. 273-278, 1997.

BERGAMASCHI, D. P.; SOUZA, J. M. P. **Bioestatística aplicada a Nutrição**. São Paulo: Universidade de São Paulo, 2005. (Aula 11). Disponível em: <http://www.fsp.usp.br/hep103/Aula11.pdf>. Acesso em: 27 Setembro 2006

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. **Time series analysis**: forecasting and control. EUA: Prentice-Hall, 1994.

CHAPRA, S. C.; CANALE, R. P. **Numerical methods for engineers with personal computer applications**. EUA: McGraw-Hill International Editions, 1987.

CONCEIÇÃO, G. M. S.; SALDIVA, P. H. N.; SINGER, J. M. Modelos MLG e MAG para análise da associação entre poluição atmosférica e marcadores de morbi-mortalidade: uma introdução baseada em dados da cidade de São Paulo. **Revista Brasileira de Epidemiologia**, v. 4, n. 3, p. 206-219, 2001.

EVERITT, B. S. **Modern medical statistics**. Nova Iorque: Oxford University Press Inc., 2003.

FERRAZ, M. I. F.; SÁFADI, T.; LAGE, G. Uso de modelos de séries temporais na previsão de séries de precipitação pluviais mensais no município de Lavras – MG. **Revista Brasileira de Agrometeorologia**, v. 7, n. 2, p. 259-267, 1999.

FONSECA, J. S.; MARTINS, G. A. **Curso de Estatística**. São Paulo: Atlas, 1996.

HAIR Jr., J. F. et al. **Análise multivariada de dados**. São Paulo: Bookman, 2005.

LATORRE, M. R. D.O.; CARDOSO, M. R. A. Time series analysis in epidemiology: an introduction to methodological aspects. **Revista Brasileira de Epidemiologia**, v. 4, n. 3, p. 145-152, 2001.

MARTINS, L. C. **Relação entre poluição atmosférica e algumas doenças respiratórias em idosos: avaliação do rodízio de veículos no município de São Paulo**. São Paulo, 2000. 97 f. Dissertação (Mestrado em Ciências) - Faculdade de Medicina, Universidade de São Paulo.

McCULLAGH, P. Quasi-likelihood functions. **Annals of Statistics**. v. 11, n. 1, p. 59-67, 1983.

McCULLAGH, P.; NELDER, J. A. **Generalized linear models**. 2 ed. Flórida, EUA: Chapman & Hall, 1989.

MYERS, R. H.; MONTGOMERY, D. C. **Response surface methodology:** process and product optimization using designed experiments. Nova Iorque: John Wiley & Sons, 2002.

PAULA, G. A. **Modelos de regressão com apoio computacional**. São Paulo: Instituto de Matemática e Estatística, Universidade de São Paulo, 2004. Disponível em: <http://www.ime.usp.br/~giapaula/livro.pdf>. Acesso em: 25 Janeiro 2006

RIBEIRO, A. J. F. **Um estudo sobre mortalidade dos aposentados por invalidez do regime geral da previdência social (RGPS)**. Belo Horizonte, 2006. 191 f. Tese (Doutorado em Demografia) - Centro de Desenvolvimento e Planejamento Regional, Faculdade de Ciências Econômicas, Universidade Federal de Minas Gerais.

RUSSO, S. L. **Gráficos de controle para variáveis não-conformes autocorrelacionadas**. Florianópilis, 2002. 120 f. Tese (Doutorado em Engenharia de Produção) - Universidade Federal de Santa Catarina.

SCHMIDT, C. M. C. **Modelo de regressão de Poisson aplicado à área da saúde**. Ijúi, 2003. 98 f. Dissertação (Mestrado em Modelagem Matemática) - Universidade Regional do Noroeste do Estado do Rio Grande do Sul.

S-PLUS 7. **Guide to Statistics**. Seattle, Washington: Insightful Corporation, 2005. (v. 1).

TADANO, Y. S. **Análise do impacto de MP10 na saúde populacional:** estudo de caso em Araucária, PR. Curitiba, 2007. 99 f. Dissertação (Mestrado em Engenharia Mecânica e de Materiais) - Universidade Tecnológica Federal do Paraná.

TADANO, Y. S.; UGAYA, C. M. L.; FRANCO, A. T. Análise estatística do impacto da poluição atmosférica na saúde populacional. In: RIO OIL & GAS 2006 EXPO AND CONFERENCE, 13, 2006, Rio de Janeiro. **Anais...** Rio de Janeiro: Editora IBP, 2006a.

TADANO, Y. S.; UGAYA, C. M. L.; FRANCO, A. T. Avaliação do impacto do ciclo de vida: efeitos dos poluentes e das condições meteorológicas na saúde da população de Araucária. In: RAA 2006 ENCONTRO DE PRHS REGIÃO SUL, 2006, Curitiba. **Anais...** Curitiba, 2006b.

WERKEMA, M. C. C.; AGUIAR, S. **Análise de regressão:** como entender o relacionamento entre as variáveis de um processo. Belo Horizonte: Fundação Christiano Ottoni da Escola de Engenharia da UFMG, 1996. (Série Ferramentas da qualidade, 7).

Notas

[1] Curvas suavizadas: São curvas ajustadas através de uma função que por definição deve ser mais "suave" do que os valores de y, ou seja, devem ter menor variabilidade do que os valores de y. Como exemplos de curvas suavizadas podem ser citados: média móvel, *splines*, *locally weighted running line smoother* (loess), entre outros (CONCEIÇÃO et al., 2001).

[2] O intervalo de confiança (IC) é um conjunto de valores calculados com base nos dados. Pressupõe-se que cubra o parâmetro de interesse com um 'certo' grau (nível) de confiança. O grau de confiança mais comumente utilizado é o de 95% (BERGAMASCHI; SOUZA, 2005).

[3] Diagrama de dispersão: "Gráfico de pontos baseado em duas variáveis, onde uma variável define o eixo horizontal e a outra define o eixo vertical. As variáveis podem ser observações, valores esperados ou mesmo resíduos. Uma forte organização dos pontos ao longo de uma linha reta caracteriza uma relação linear ou correlação" (HAIR Jr. et al., 2005).

[4] Distribuição t de *student* – É uma distribuição contínua semelhante à distribuição normal utilizada para realizar testes de hipóstese. A tabela da distribuição t de *student* pode ser encontrada em qualquer livro de estatística básica (FONSECA; MARTINS, 1996).

MÉTODO DE REGRESSÃO DE POISSON: METODOLOGIA PARA AVALIAÇÃO DO IMPACTO DA POLUIÇÃO ATMOSFÉRICA NA SAÚDE POPULACIONAL

YARA DE SOUZA TADANO
CÁSSIA MARIA LIE UGAYA
ADMILSON TEIXEIRA FRANCO

Resumo: Os modelos estatísticos mais utilizados para avaliar o impacto da poluição atmosférica na saúde populacional são os modelos de regressão, pois são capazes de relacionar uma ou mais variáveis explicativas com uma única variável resposta. O objetivo deste estudo foi apresentar o modelo estatístico de regressão de Poisson dos modelos lineares generalizados. Neste trabalho são apresentadas todas as etapas da avaliação, desde a coleta e a análise dos dados até a verificação do ajuste do modelo escolhido.

Palavras-chave: Análise de regressão. Modelo de regressão de Poisson. Poluição atmosférica. Saúde populacional.

Methodology to assess air pollution impact on the population's health using the poisson regression method

Abstract: *The most used statistical model to evaluate the relation between air pollution and population's health is regression analysis, as it is able to relate one or more explanatory variables with one response variable. This research aims to present the generalized linear model with Poisson regression. Every assessment step, from data collection and analysis to the verification of the chosen model adjustment, will be presented.*

Keywords: *Regression analysis. Poisson regression. Air pollution. Population's health.*