RT-MAE 2002-10

USE OF OVERDISPERSED REGRESSION MODELS IN ANALYSING THE ASSOCIATION BETWEEN AIR POLLUTION AND HUMAN HEALTH

by

Silvia L.P. Ferrari, Jacqueline S.E. David,
Paulo A. André
and
Luiz A.A. Pereira

Palavras-Chave: Air pollution, diagnostics, generalized linear models, human health, negative binomial distribution, overdispersion, Poisson distribution, quasi-likelihood.
Classificação AMS: 62J12, 92C60.
(AMS Classification)

- Junho de 2002 -

Use of overdispersed regression models in analysing the association between air pollution and human health

Silvia L.P. Ferrari, Jacqueline S. E. David,
Paulo A. André, and Luiz A.A. Pereira

Abstract. Poisson regression models are frequently used in analysing the association between count data and a set of covariates. However, the assumption that, given the values of the covariates, the mean and the variance of the response variable are equal is oftentimes violated. This is the case of a study of the association between daily emergency hospital visits for respiratory diseases and the levels of air pollution in Vitória, Espírito Santo, Brazil. In this paper we examine two overdispersed regression models for count data and apply them to this study. The main aspects addressed are inference on the regression and dispersion parameters and model adequacy checking.

1. Introduction

The Poisson distribution provides the basis for the standard model to the analysis of the association between count data and a set of covariates. A basic assumption of Poisson regression models is that, given the values of the covariates, the mean and the variance of the response variable are equal. The dispersion may, however, be greater than that predicted by the model. Possible causes for overdispersion include variability of experimental material, correlation between individual responses, omitted unobserved variables, among others. One important consequence of ignoring overdispersion is that the standard errors obtained from the Poisson regression model are incorrect and underestimate the variability of the regression parameter estimators.

In this paper we examine two regression models for overdispersed count data, namely the constant overdispersion and the negative binomial regression models.

1991 *Mathematics Subject Classification.* Primary 62J12; Secondary 92C60.

Key words and phrases. air pollution, diagnostics, generalised linear models, human health, negative binomial distribution, overdispersion, Poisson distribution, quasi-likelihood.

Correspondence to: Silvia L.P. Ferrari. Departamento de Estatística, Universidade de São Paulo, Caixa Postal 66281, São Paulo/SP, 05311-970, Brasil

Both models are applied to a study of the association between daily emergency hospital visits for respiratory diseases and the levels of air pollution in Vitória, Espírito Santo, Brazil. The present work is related to McNeney and Petkau's (1994) paper, although the emphasis in their paper is a simulation study. Studies of the association between air pollution and mortality or hospital admissions have been considered by various researchers; see Schwartz et. al (1996), Morgan et al. (1998), Saldiva et al. (1994), Saldiva et al. (1995), Braga et al., (1999, 2001), Lin et al. (1999), Conceição et al., (2001), Wong et al. (2001) and references therein. Although the constant overdispersion model is frequently used, the negative binomial model has not been considered in environmental studies. A comprehensive text on overdispersed regression models may be found in Hinde and Demétrio (1998).

In Section 2 we review the usual Poisson regression model and two alternative models which are suitable for overdispersed counts: the constant overdispersion and the negative binomial regression models. Inference on the regression and dispersion parameters is discussed. Section 3 is dedicated to model adequacy checking including residual plots and influence diagnostics. We have noticed that diagnostic techniques are rarely used in environmental studies. A study of the association between air pollution and human health is presented in Section 4. We show that the Poisson regression model is inappropriate for our data. A diagnostic analysis suggests that both the alternative regression models provide much better fits.

2. Models for overdispersed counts

Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables having Poisson distribution with means $\mu_1, \ldots, \mu_n$ respectively, and let $x_1, x_2, \ldots, x_n$ be $p \times 1$ vectors of known constants. The *Poisson regression model* assumes that

$$(2.1) \qquad g(\mu_i) = \eta_i = x_i^\top \beta,$$

for $i = 1, \ldots, n$, where $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a vector of p unknown parameters and $g(.)$ is a continuous monotone twice diferentiable link function. The usual log-linear Poisson regression model assumes that $g(\mu_i) = \log(\mu_i)$. Here,

$$(2.2) \qquad \mathrm{Var}(Y_i) = \mu_i,$$

i.e., given the value of the covariates, the mean and the variance of the response are assumed to be equal. This model belongs to the class of generalized linear models (McCullagh and Nelder, 1989). Maximum likelihood estimates for the regression parameters can be obtained using the standard iteratively re-weighted least squares (IRLS) algorithm which is implemented in statistical software such as S-PLUS (Venables and Ripley, 1999), GLIM 4 (Aitkin, Anderson, Francis and Hinde, 1989), STATA (Hardin and Hilbe, 2001) and SAS (Pedan, 2001). Model reductions may be based on changes in the Poisson deviance. It is defined as the difference between the log-likelihood function of the saturated model and of the model under investigation, i.e., $D_p = \sum_{i=1}^{n} d_i^2$ where

$$(2.3) \qquad d_i = 2\ \text{sign}(y_i - \widehat{\mu}_i) \left\{ y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i - \widehat{\mu}_i) \right\}^{1/2},$$

$\widehat{\mu}_i$ being the maximum likelihood estimate of μ_i. If the model has an intercept, we have $\sum_{i=1}^{n}(y_i - \widehat{\mu}_i) = 0$ and the deviance reduces to $D_p = 2\sum_{i=1}^{n} y_i \log(y_i/\widehat{\mu}_i)$. Under the null hypothesis $\text{H}_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$, the partial deviance, i.e., the difference between the deviances of model (2.1) and of the restricted model, has a chi-squared distribution with $p - q$ degrees of freedom asymptotically.

The mean of the response variable and the regression parameters are related by the link function $g(.)$. There may be more than one link function apparently appropriate for a particular application. The log link function $g(\mu_i) = \log \mu_i$ has an advantage over other usual link functions since it allows a simple interpretation for the regression parameters. Let $x_h = (x_{h1}, \ldots, x_{hp})^{\top}$ be a vector of covariate values. Now let $x_{h^*} = (x_{h1}, \ldots, x_{hj} + c, \ldots, x_{hp})^{\top}$. Under a log-linear model, the means of the response variable given x_h and x_{h^*} are $\mu_h = \exp(x_h^{\top}\beta)$ and $\mu_{h^*} = \exp(x_{h^*}^{\top}\beta)$, respectively. Notice that the relative risk is $\mu_{h^*}/\mu_h = \exp(c\beta_j)$ and hence $c\beta_j$ is the logarithm of the relative risk when the value of the j-th covariate is increased by an amount of c units and the others remain unchanged.

A simple extention of the Poisson regression model that allows the response variance to be greater than the respective mean is the *constant overdispersion model* which replaces (2.2) by

$$\text{Var}(Y_i) = \phi\mu_i.$$

This model belongs to the class of quasi-likelihood models (Wedderburn, 1974). The idea is to relax the assumption of a particular distribution for the response but instead consider the following relation between its mean and variance: $\text{Var}(Y_i) = \phi V(\mu_i)$. For our purposes, $V(\mu_i) = \mu_i$. Inferences may be based on the quasi-likelihood function

$$Q(\mu; y) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} dt,$$

which has properties similar to those of a genuine log-likelihood function. Here, $y = (y_1, \ldots, y_n)^{\top}$ and $\mu = (\mu_1, \ldots, \mu_n)^{\top}$. The quasi-score function $U(\beta) = \partial Q/\partial\beta$ can be written as $U(\beta) = D^{\top}V^{-1}(y - \mu)/\phi$, where $D = \partial\mu/\partial\beta^{\top} = W^{1/2}V^{1/2}X$, $V = \text{diag}\{V(\mu_1), \ldots, V(\mu_n)\}$, $W = \text{diag}\{w_1, \ldots, w_n\}$, $w_i = (d\mu_i/d\eta_i)^2/V(\mu_i)$ and X is an $n \times p$ matrix with rows $x_1^{\top}, \ldots, x_n^{\top}$. The quasi-score function is proportional to the score function for the Poisson regression model and hence the regression parameter estimates obtained from the quasi-likelihood equations $U(\beta) = 0$ coincide with the maximum likelihood estimates for the Poisson regression model.

The information matrix for β is $I(\beta) = -E(\partial U(\beta)/\partial\beta^{\top}) = \phi^{-1}D^{\top}V^{-1}D$ and the asymptotic covariance matrix for $\widehat{\beta}$ is given by

$$I(\beta)^{-1} = \phi(D^{\top}V^{-1}D)^{-1}$$

(McCullagh and Nelder, 1989, Chapter 9). It equals the asymptotic covariance matrix for the Poisson regression model multiplied by the dispersion parameter ϕ. It is then clear that if overdispersion is ignored, the variances of the regression parameter estimators will be underestimated. The usual estimate for ϕ is based on the Pearson residuals (see Section 3) and is given by

$$\widehat{\phi} = \frac{1}{(n-p)} \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}.$$

The constant overdispersion regression model with log link function may be fitted in the S-Plus software using the function `glm` with the option `family = quasi(link=log, variance=''mu'')`; other choices for the link function are allowed. See Hardin and Hilbe (2001, Chapter 16) and Pedan (2001) for details of the model fitting using STATA and SAS, respectively.

A quasi-deviance function (without scale factor) may be defined for the constant overdispersion regression model as

$$D(y;\hat{\mu}) = 2\phi\{Q(y;y) - Q(\hat{\mu};y)\} = 2\sum_{i=1}^{n} \int_{\mu_i}^{y_i} \frac{y_i - t}{V(t)} dt.$$

Notice that $D(y;\hat{\mu})$ does not depend on ϕ. It is easy to show that the unscaled quasi-deviance function above equals the deviance function for the Poisson regression model.

Now let $\mathrm{H}_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$ be the null hypothesis of interest to be tested against a two-sided alternative. A natural statistic for this test is $F = (D_p - D_q)/\{\widehat{\phi}(p-q)\}$ which is compared to the quantiles of a $F_{p-q,n-p}$ distribution. Here D_q and D_p denote the unscaled quasi-deviance functions of the restricted and unrestricted models respectively, and $\widehat{\phi}$ is obtained under the unrestricted model.

The negative binomial distribution is useful for defining another regression model for overdispersed counts. If Y has a negative binomial distribution with parameters $\mu > 0$ and $k > 0$ and probability function

$$f_Y(y;\mu,k) = \frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)} \left(\frac{k}{k+\mu}\right)^k \left(\frac{\mu}{k+\mu}\right)^y, \quad y = 0,1,\ldots,$$

where $\Gamma(.)$ is the gamma function, then

$$\mathrm{E}(Y) = \mu \quad \text{and} \quad \mathrm{Var}(Y) = \mu + \frac{\mu^2}{k}.$$

Here, a possible motivation for assuming that the observed counts come from a negative binomial distribution is the following fact: if, conditionally on θ, $Y \sim \mathrm{Poisson}(\theta)$ and $\theta \sim \mathrm{Gamma}(k,\lambda)$, then the marginal distribution of Y is a negative binomial distribution with $\mathrm{E}(Y) = \mu = k/\lambda$ and $\mathrm{Var}(Y) = \mu + \mu^2/k$.

The *negative binomial regression model* assumes that independent random variables $Y_1,\ldots,Y_n$ have negative binomial distribution with means $\mu_1,\ldots,\mu_n$ respectively, and a common parameter k, and that the means are related to the covariates $x_1, \ldots, x_n$ through (2.1). For known k, this model belongs to the class of generalized linear models but not otherwise.

The log-likelihood function for β and k is given by

$$l(\mu, k; y) = \sum_{i=1}^{n} \left\{ y_i \log \mu_i + k \log k - (k + y_i) \log(k + \mu_i) + \log(y_i + k) - \log\left(\frac{k}{y_i!}\right) \right\}.$$

The score function $U(\beta, k) = (\partial l/\partial\beta_1, \ldots, \partial l/\partial\beta_p, \partial l/\partial k)^\top$ has elements

$$\text{(2.4)} \qquad \frac{\partial l}{\partial \beta_r} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i) g'(\mu_i)} x_{ir},$$

and

$$\text{(2.5)} \quad \frac{\partial l}{\partial k} = \sum_{i=1}^{n} \left\{ \psi(y_i + k) - \psi(k) - \log(\mu_i + k) - \frac{k + y_i}{k + \mu_i} + \log k + 1 \right\},$$

where $V(\mu_i) = \mu_i + \mu_i^2/k$, x_{ir} is the (i, r)-th element of x_i and $\psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ represents the digamma function.

It is easy to show that the information matrix for (β, k) is given by

$$I(\beta, k) = \begin{bmatrix} X^\top W X & 0 \\ 0^\top & i_{k,k} \end{bmatrix}$$

where W is defined above and

$$i_{k,k}(\beta, k) = \sum_{i=1}^{n} \left\{ -\mathrm{E}(\psi'(Y_i + k)) + \psi'(k) + \frac{1}{\mu_i + k} - \frac{1}{k} \right\}.$$

Notice that $I(\beta, k)$ is a block-diagonal matrix and hence β and k are orthogonal parameters. The asymptotic covariance matrix for the maximum likelihood estimator of β is $(X^\top W X)^{-1}$ either if k is estimated from the data or assumed to be known.

The block-diagonal form of the information matrix allows the maximum likelihood estimates for β and k to be obtained using the Gauss-Seidel approximation by iterating between the two steps below:

(i) for a fixed k, solve (2.4) via an IRLS algorithm (see McCullagh and Nelder, 1989);

(ii) for a fixed β, solve (2.5) applying the Newton-Raphson algorithm (see Rustagi, 1994, for details).

An initial value for k is

$$k^{(0)} = \frac{\sum_{i=1}^{n} \widehat{\mu}_i (1 - c_i \widehat{\mu}_i)}{\sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i} - (n - p)},$$

where here $\widehat{\mu}_i$ is the maximum likelihood estimate for μ_i obtained from the Poisson regression model and $c_i = x_i^\top (X^\top \widehat{W} X)^{-1} x_i$, with $\widehat{W} = W(\widehat{\mu})$, is the estimated asymptotic variance of $\widehat{\eta}_i = x_i^\top \widehat{\beta}$. The idea comes from the comparison of the Pearson statistic of the Poisson fit, i.e., $X^2 = \sum_{i=1}^{n} (y_i - \widehat{\mu}_i)^2/\widehat{\mu}_i$ with its expected value under the negative binomial model (see Breslow, 1984, for details). An alternative approach for the estimation of k based on the method of moments is discussed in Breslow (1984). Lawless (1987) show that the method of moments estimates are more robust than the maximum likelihood estimates although the latter are more efficient than the former if the negative binomial regression model is correct. The S-PLUS functions `neg.bin(k)` and `glm.nb` found in the library *MASS* (Venables and Ripley, 1999) give the maximum likelihood estimates for the parameters of the negative binomial regression model under known and unknown k, respectively. See Hardin and Hilbe (2001, Chapter 13) and Pedan (2001) for details of the model fitting using STATA and SAS.

Model reductions may be based on the partial deviance. Let $\mathrm{H}_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$ be the null hypothesis to be tested against a two-sided alternative. For known k, the deviance residuals are given by

$$(2.6) \quad d_i = \mathrm{sign}(y_i - \widehat{\mu}_i)\left\{ y_i \left[\log\left(\frac{y_i}{y_i + k}\right) - \log\left(\frac{\widehat{\mu}_i}{\widehat{\mu}_i + k}\right)\right] + k\log\left(\frac{\widehat{\mu}_i + k}{y_i + k}\right)\right\}^2$$

and the partial deviance equals the likelihood ratio statistic. Under the null hypothesis it has a chi-squared distribution with $p - q$ degrees of freedom asymptoticaly. If k is unknown, an approximation for the partial deviance is obtained by estimating k under the unrestricted model.

3. Diagnostics

Diagnostic techniques are of great relevance for detecting regression problems such as lack of fit and the presence of outliers and influential observations. Plots of the Pearson or the deviance residuals against some function of the data, such as the estimated linear predictors $\widehat{\eta}_i = x_i^\top \widehat{\beta}$, may be helpful for detecting lack of fit or highlighting outliers. The Pearson residuals are defined as

$$(3.1) \qquad r_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{v}_i}},$$

where $\widehat{v}_i$ is the estimated variance of Y_i. For the Poisson, the constant overdispersion and the negative binomial regression models, the Pearson residuals are given by (3.1) with $v_i = \mu_i$, $v_i = \phi\mu_i$ and $v_i = \mu_i + \mu_i^2/k$, respectively. The deviance residuals are given by (2.3) and (2.6) for the Poisson and the negative binomial regression models, respectively. The deviance residuals for the constant overdispersion regression model coincide with the corresponding residuals for the Poisson case multiplied by $\phi^{-1/2}$. For diagnostics in regression models for count data, see Cameron and Trivedi (1998, Chapter 5). The use of bivariate smoothing for examining residual plots in the analysis of the association between air pollution and respiratory illness is discussed in Schwartz (1994).

Since the distribution of the residuals is not known, half-normal plots with simulated envelopes are helpful tools for diagnostic purposes (Atkinson, 1985, Neter et al., 1996, Section 14.6). The idea is to enhance the usual half-normal plots by adding a simulated envelope which can be used to decide whether the observed responses are consistent with the fitted model.

Half-normal plots with simulated envelope are constructed as follows:

1. fit the model and generate a simulated sample of n independent observations using the fitted model as if it were the true model;
2. fit the model to the new sample, and calculate the ordered absolute values of the diagnostic quantity of interest;
3. repeat the steps above 18 times;
4. consider the n sets of the 19 order statistics; for each set calculate its mean, minimum and maximum values;
5. plot these values and the ordered diagnostic quantities of the original sample against the half-normal scores $\Phi^{-1}((i+n-1/8)/(2n+1/2))$, where $\Phi(.)$ is the cumulative distribution function of the standard normal distribution.

The minimum and maximum values of the 19 order statistics provide the envelope. An informal check for overdispersion may be based on a half-normal plot with simulated envelope of the Pearson or the deviance residuals for the Poisson fit. A large portion of points failing over the envelope indicates that the variability of the residuals is greater than expected and hence an overdispersed regression model may be more appropriate for the data set being analysed.

For the constant overdispersion regression model, no particular distribution is assumed for the response. This leads to difficulties in step 1 above. Demétrio and Hinde (1997) suggest to generate the Y_i's as $Y_i = \widehat{\phi} Y_i^*$, where $Y_1^*, \ldots, Y_n^*$ come from independent Poisson distributions with means $\widehat{\mu}_1, \ldots, \widehat{\mu}_n$. Another approach is to generate $\theta_1, \ldots, \theta_n$ as n independent observations, θ_i having a gamma distribution with parameters $\zeta_i = \lambda\widehat{\mu}_i$ and $\lambda = 1/(\widehat{\phi}-1)$, and then $Y_1, \ldots, Y_n$ are generated as independent Poisson variables with means $\theta_1, \ldots, \theta_n$. In both cases, the desired relation between the mean and the variance of the response variable is satisfied.

A graphical technique for detecting overdispersion is discussed by Lambert and Roeder (1995) and a formal check is described by Lawless (1987). Assume that the negative binomial regression model is correct and let $\delta = 1/k$. If $\delta = 0$, no overdispersion is present and the data come from a Poisson distribution. The likelihood ratio statistic of $\mathrm{H}_0 : \delta = 0$ against $\mathrm{H}_1 : \delta \neq 0$ is $\omega = -2(l_P - l_{NB})$, where l_P and l_{NB} are the log-likelihood functions of the Poisson and the negative binomial regression models, respectively. Notice that the value for δ in H_0 is in the boundary of the parameter space and hence the usual properties of the likelihood ratio test are not valid here. Lawless (1987) shows that under the null hypothesis the cumulative distribution function of ω is $F_\omega(z) = 1/2 + P(\chi_1^2 \leq z)$, for $z \geq 0$, where χ_1^2 represents a random variable having a chi-squared distribution with one degree of freedom. For other tests for overdispersion, see Dean (1992).

For the negative binomial regression model, the ith diagonal element h_i of the matrix $H = W^{1/2}X(X^\top WX)^{-1}X^\top W^{-1/2}$ can be viewed as a leverage measure of the corresponding observation. A plot of $h_1, \ldots, h_n$ against the fitted values $\widehat{\mu}_1, \ldots, \widehat{\mu}_n$ may be helpful for detecting high leverage observations. Such observations are potentially influential in the model fit. For log linear models, $h_i = k\mu_i/(k+\mu_i)x_i^\top(X^\top WX)^{-1}x_i$. The Cook distance is a widely used influence measure (Cook, 1986). In our setting, it is not possible to write it in closed form but it may be approximated by $r_i^2 h_i/(1-h_i)^2$ with r_i given in (3.1). Local influence diagnostics and deviance residual analysis in log-linear negative binomial models are discussed in Svetliza and Paula (2001).

4. The Vitória study

Large Brazilian cities, particularly São Paulo as the biggest city in South America, have been subject of several epidemiological studies to evaluate the association between air pollution and human health; see Saldiva et al. (1994), Saldiva et al. (1995), Braga et al., (1999, 2001), Conceição et al., (2001) and references therein. In this paper our focus is Vitória city in Espírito Santo State, Brazil, which has a population of about one million and four hundred thousand people. Its industrial activities include a steel plant and the main Brazilian harbour for iron ore and mining products exportation in urban area.

Following the HEADLAMP Project recommendations (WHO, 1996a, 1996b), we performed an epidemiological ecologic study. Our aim is to evaluate the association between air pollution and human health in Vitória. The particulate material concentration (PM_{10}) is recommended as the pollution indicator. However, we replaced it by the daily average concentration of sulfur dioxide (SO_2) because of the availability of such information and the strong correlation with existing PM_{10} data. The total daily number of visits for respiratory causes in the emergency room of the unique local public child hospital represents the health indicator and is the response variable. The epidemiological design also requires the control of confounding variables, i.e., known or possible risk factors which may confound the relationship under analysis. To represent the environment the following variables were included: the temperature, represented by its minimum daily value, and the relative humidity, by its nearest value at midday. To represent seasonal factors we included year, month and day of the week. The occurrence of respiratory disease epidemic and the number of emergency hospital visits for non-respiratory diseases were also considered, the latter aiming to control external factors such as strike of hospital employees.

The study includes daily data from 1993 until 1997. Table 4.1 shows summary measures of some variables included in our study by year. It is clear that the average number of visits for respiratory causes and the average concentration of SO_2 decreased during the period under investigation while the average number of visits for other causes did not change much.

It is well known that the effect of temperature, humidity and pollution on health indicators does not necessarily occur on the same day of the event, in this case, the emergency room visit. A previous analysis of our data indicated that models with three day moving average for temperature and humidity and seven day moving average for SO_2 generally fit best. The moving average for a specific variable is the average of this variable in the referred day with the values of the precedent days. For example, a two day moving average is the variable average of the present and the previous day.

Table 4.1. Yearly number of observations (n), means (standard deviations) of the number of visits for respiratory (RESP) and non-respiratory causes (NRESP), temperature (T), humidity (H) and SO_2 concentration.

Year	RESP	NRESP	T (°C)	H (%)	SO_2 ($\mu g/m^3$)
1993 (n=249)	53,29 (33,01)	132,76 (31,23)	20,90 (3,55)	64,84 (20,65)	15,60 (14,78)
1994 (n=192)	54,54 (24,31)	126,78 (40,76)	19,38 (2,73)	69,36 (7,47)	9,67 (12,31)
1995 (n=237)	44,34 (22,11)	152,89 (51,33)	21,51 (3,95)	72,96 (17,33)	6,97 (7,74)
1996 (n=207)	39,71 (15,50)	136,96 (33,72)	19,35 (2,42)	72,51 (12,38)	4,81 (7,30)
1997 (n=335)	37,39 (24,09)	144,00 (50,02)	20,53 (2,54)	92,32 (6,58)	3,36 (5,44)
Total (n=1220)	45,08 (25,65)	139,53 (43,85)	20,41 (3,18)	75,98 (17,35)	7,80 (10,82)

Figures in the first column of Table 4.1 show that some variables have missing values and this leads to a much higher number of missing observations after computing the required moving averages. The final data set consists of 599 observations.

The model building strategy for the statistical analysis involves the construction of a basic model in which variations due to the confounding variables are removed. Once the variables most strongly associated with the response are determined (significance level equal to 5%), the pollutant is added to the model.

For similar model building strategies, see Conceição et al. (2001) and Schwartz et al. (1996). The statistical analysis that follows was carried out using S-Plus 4.5.

First, let us consider the log-linear Poisson regression model. The basic model was fitted and significant associations for all the confounding variables were observed ($p < 0.001$). The inclusion of the pollutant in the basic Poisson model is significant ($p < 0.001$). Nevertheless, the residual deviance equals 3,107.0 with 572 degrees of freedom (d.f.) indicating that the Poisson regression model does not provide a good fit. Moreover, all the points in the half-normal plot of the deviance residuals (Figure 4.1a) fall over the simulated envelope, leading to the conclusion that there is a strong evidence of overdispersion.

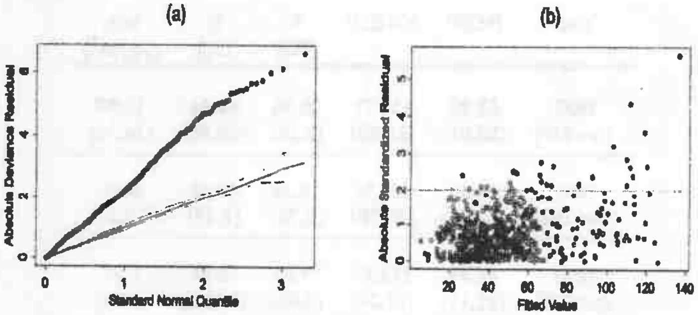

Figure 4.1. Half-normal plot of the deviance residuals for the Poisson model (a) and plot of absolute standardized residuals against the fitted values for the normal model (b).

The 25th and 75th percentiles, the average and the standard deviation for the observed response variable are 30, 59, 47.7 and 28.1 respectively. These figures suggest that emergency visits for respiratory diseases are not rare events and that a normal model may be more suitable for our data than the Poisson model. A normal model with independent identically distributed errors including all the confounding variables and the pollutant was fitted. The plot of the absolute standardized residuals against the fitted values (Figure 4.1b) clearly shows that the variability of the residuals is not constant. In fact, the Breusch-Pagan test (Breusch and Pagan, 1979) rejects the null hypothesis of homoscedasticity ($p < 0.001$) and leads to the conclusion that this model is not appropriate for our data.

Now, we consider the constant overdispersion regression model with log link function. The basic model was fitted and we found a non significant effect only for the relative humidity ($p = 0.064$). The inclusion of the pollutant in the basic model is significant ($p < 0.001$), the residual deviance equals 609.2 with 572 d.f. and we obtained $\widehat{\phi} = 5.11$. Figure 4.2a shows a half-normal plot of the deviance residuals. It indicates that the constant overdispersion model is much more appropriate than the Poisson model for our data. The plot of the Pearson residuals against the fitted values (Figure 4.2b) do not show any outliers or evidence of lack of fit.

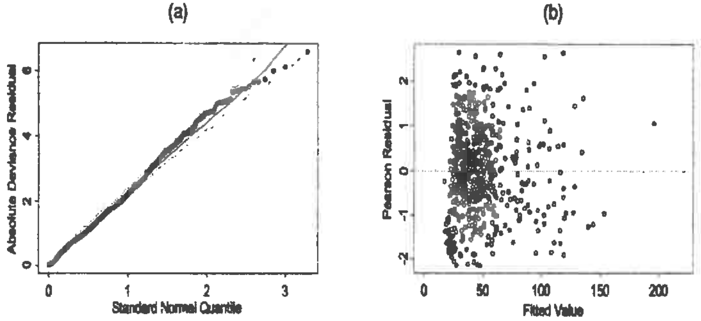

Figure 4.2. Half-normal plot of the deviance residuals (a) and plot of the Pearson residuals against the fitted values (b) for the constant overdispersion model.

We now move to the negative binomial regression model with log link function and unknown dispersion parameter k. As in the constant overdispersion model, only the relative humidity was eliminated from the basic model ($p = 0.315$). The inclusion of the pollutant is significant ($p < 0.001$), the residual deviance equals 663.7 with 572 d.f. and we obtained $\widehat{k} = 9.14$ with standard error equal to 0.69. The half-normal plot of the deviance residuals (Figure 4.3a) indicates that the negative binomial model is much more suitable than the Poisson model. However, a comparison between Figures 4.2a and 4.3a indicates that the constant overdispersion model seems to provide a better fit.

Figure 4.3c shows an index plot of Cook's distance for the negative binomial fit. We noticed outstanding Cook's distance for the 372nd and the 595th observations. For these cases, the Pearson residuals are -2.08 and 3.34 respectively, indicating that the observed count of visits is much smaller than expected for the first case and much higher than expected for the second case. Figure 4.3d shows a plot of h_i against the fitted values. We noticed that all the observations with high leverage ($h_i > 2.5p/n$) correspond to December, 1996. These are the only cases observed in December in our data set and this is the reason why they are potentially influential. However, the elimination of all the discrepant observations does not change the inferential conclusions. Plots of local influence, constructed as suggested by Svetliza and Paula (2001), were omitted here because they do not show any influential point.

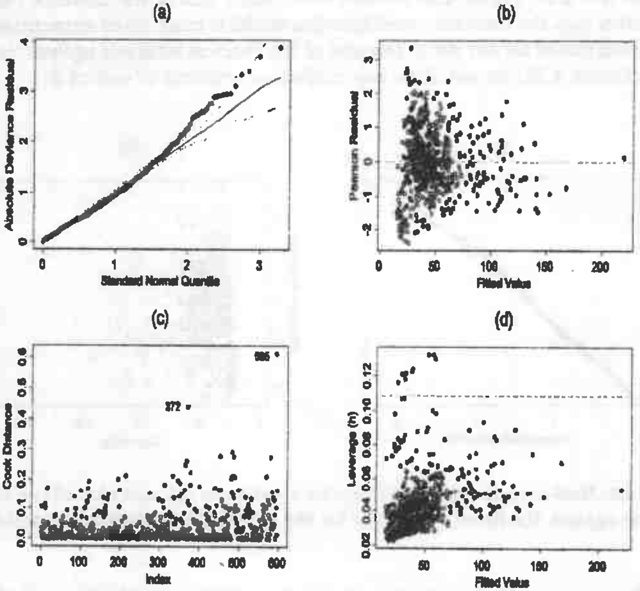

Figure 4.3. Half-normal plot of the deviance residuals (a), plot of the Pearson residuals against the fitted values (b), index plot of the Cook distance (c) and plot of the leverage against the fitted values (d) for the negative binomial model.

Positive association between the response variable and SO_2 was observed in both overdispersed models. This indicates that the higher the concentration of this pollutant, the higher the expected value for the number of daily emergency hospital visits for respiratory diseases. The relative risk associated with the pollutant was estimated for the increase of $1\mu g/m^3$ and also $11.93\mu g/m^3$, which corresponds to the interquartile range (75th-25th percentiles). This last measure was calculated aiming to represent the ratio of the expected number of daily visits for respiratory diseases in a high air pollution day compared with a day of low air pollution. These results were very similar in both overdispersed models, and are presented in Table 4.2. Notice that an increase in the pollutant concentration from the 25th to the 75th percentile is associated with an average increase of roughly 11% in the number of visits for respiratory diseases.

Table 4.2. Estimated relative risk (RR) associated with the pollutant.

Overdispersion	RR($\mu g/m^3$)		RR($11.93\mu g/m^3$)	
Models	Estimate	CI(95%)	Estimate	CI(95%)
Constant	1.009	[1.004;1.013]	1.108	[1.050;1.169]
Negative Binomial	1.009	[1.005;1.013]	1.113	[1.065;1.164]

5. Concluding remarks

The constant overdispersion and the negative binomial regression models are more appropriate than the Poisson and the homoscedastic normal linear regression models for the analysis of the association between the number of daily emergency hospital visits for respiratory diseases and the levels of air pollution in Vitória for the period from 1993 to 1997. There is some evidence that the constant overdispersion model is the most appropriate among those considered in this paper. Under both the overdispersed models we found positive association between the number of daily visits and the concentration of SO_2. Our results are in agreement with the concept that adequate statistical modelling of daily measures of pollution, weather and health outcomes represent a powerful tool to detect the adverse effects of air pollution on human health.

References

[1] Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989). *Statistical Modelling in GLIM*. Oxford: Oxford University Press.

[2] Atkinson, A. (1985). *Plots, Transformations and Regression*. Oxford: Clarendon Press.

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

[4] 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.

[5] Breslow, N. (1984). Extra-poisson variation in log-linear models. *Applied Statistics*, **33**, 38-44.

[6] Breusch, T.S. and Pagan, A.R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica* **47**, 1287-94.

[7] Cameron, A.C. e Trivedi, P.K. (1998).*Regression Analysis of Count Data*. Econometric Society Monographs 30. Cambridge: Cambridge University Press.

[8] Conceição, M.S.C., Miraglia, S.G.E.K., Kishi, H.S., Saldiva, P.H.N., Singer, J.M. (2001). Air Pollution and child mortality: a time-series study in São Paulo, Brazil. *Environmental Health Perspectives*, **109** (suppl 3), 347-350.

[9] Cook, R.D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society* B, **48**, 133-169.

[10] Dean, C. (1992). Testing overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association*, **87**, 451-457.

[11] Demétrio, C. and Hinde, J. (1997). Half-normal plots and overdispersion. *GLIM Newsletter*, **27**, 19-26.

[12] Hardin, J. and Hilbe, J. (2001). *Generalized Linear Models and Extensions.* College Station: Stata Press.

[13] Hinde, J. and Demétrio, C. (1998). *Overdispersion: models and estimation. Computational Statistics and Data Analysis*, **27**, 151-170.

[14] Lambert, D. and Roeder, K. (1995). Overdispersion diagnostics for generalized linear models. *Journal of the American Statistical Association*, **95**, 1225-1237.

[15] Lawless, J. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, **15**, 209-225.

[16] Lin, C.A., Martins, M.A., Farhat, S.C., Pope III, C.A., Conceição, G.M.S., Anastasio, V.M., Hatanaka, M., Andrade, W.C., Hamaue, W.R., Bhm, G.M., Saldiva, P.H.N. (1999). Air pollution and respiratory illness of children in São Paulo, Brazil. *Paediatric and Perinatal Epidemiology*, **13**, 475-488.

[17] McCullagh, P. and Nelder, J.A. (1989).*Generalized Linear Models* (The Monographs on Statistics and Applied Probability, Vol 37), 2nd edition. London: Chapman and Hall.

[18] McNeney, B. and Petkau, J. (1994). Overdispersed Poisson regression models for studies of air pollution and human health. *The Canadian Journal of Statistics*, **22**, 421-440.

[19] Morgan, G., Corbett, S.,Wlodarczyk, J. and Lewis, P. (1998). Air pollution and daily mortality in Sydney, Australia, 1989 through 1993. *American Journal of Public Health*, **88**, 759-764.

[20] Neter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W. (1996). *Applied Linear Statistical Models* (Irwin Series in Statistics), 4th edition. Chicago: Irwin.

[21] Pedan, A. (2001). Analysis of count data using the SAS system. *Proceedings of the 26th SAS Users Group International Conference*, P247-26, 1-6. Available at **www2.sas.com/proceedings/sugi26/p247-26.pdf**.

[22] Rustagi, J.S. (1994). *Optimization Techniques in Statistics*. San Diego: Academic Press.

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

[24] Saldiva, P.H.N., Pope III, C.A., Schwartz, J., Dockery, D.W., Lichtenfels, A.J.F.C., Salge, J.M., Barone, I.A. and Bõhm, G.M. (1995). Air pollution and mortality in elderly people: a time-series study in São Paulo, Brazil. *Archives of Environmental Health*, **50**, 159-163.

[25] Schwartz, J. (1994). Nonparametric smoothing in the analysis of air pollution and respiratory illness. *The Canadian Journal of Statistics*, **22**, 471-487.

[26] Schwartz, J., Spix, C. Touloumi, G., Bachárová, L., Barumamdzadeh, T., le Tertre, A., Piekarksi, T., Ponce de Leon, A., Põnkä, A. Rossi, G., Saez, M. and Schouten, J.P. (1996). Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. *Journal of Epidemiology and Community Health*, **50** (Suppl 1), S3-S11.

[27] Svetliza, C.F. and Paula, G.A. (2001). On diagnostics in log-linear negative binomial models. *Journal of Statistical Computation and Simulation*, **71**, 231-243.

[28] Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-Plus.* Third edition. New York: Springer.

[29] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, **61**, 439-447.

[30] WHO, World Health Organization. (1996a). *Linkage Methods for Environment and Health Analysis. General Guidelines.* HEADLAMP Project, Genéve.

[31] WHO, World Health Organization. (1996b). *Linkage Methods for Environment and Health Analysis. Technical Guidelines.* HEADLAMP Project, Genéve.

[32] Wong,C.-M., Ma, S., Hedley, A.J. and Lam, T.-H. (2001). Effect of air pollution on daily mortality in Hong Kong. *Environmental Health Perspectives*, **109**, 335-340.

Acknowledgment

We gratefully acknowledge financial support from the following Brazilian agencies: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Financiadora de Estudos e Projetos (FINEP), and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). We would like to thank Professor Paulo Hilario do Nascimento Saldiva and Alfésio Luís Ferreira Braga for their comments on the manuscript.

Departamento de Estatística, Universidade de São Paulo, Brazil
E-mail address: `sferrari@ime.usp.br`

Departamento de Estatística, Universidade de São Paulo, Brazil
E-mail address: `jackdavid@terra.com.br`

Laboratório de Poluição Atmosférica Experimental, Faculdade de Medicina, Universidade de São Paulo, Brazil
E-mail address: `pandre@osite.com.br`

Laboratório de Poluição Atmosférica Experimental, Faculdade de Medicina, Universidade de São Paulo, Brazil
E-mail address: `luiz@lim05.fm.usp.br`

ÚLTIMOS RELATÓRIOS TÉCNICOS PUBLICADOS

2002-01 - BUENO, V.C. Asymptotic reliability for a coherent system with an infinite number of components. 2002. 10p. (RT-MAE-2002-01)

2002-02 - TAVARES, H.R., ANDRADE, D.F. Item Response Theory for Longitudinal Data: Item and Population Ability Parameters Estimation. 2002. 20p. (RT-MAE-2002-02)

2002-03 - BOLFARINE, H., VALENÇA, D.M. Score tests for Weibull-regression models with random effect. 2002. 16P. (RT-MAE-2002-03)

2002-04 - BUENO, V.C. Using a second order stochastic dominance for active redundancy allocation in a K-out-of-n system. 2002. 7p. (RT-MAE-2002-4)

2002-05 - FERRARI, S.L.P., LUCAMBIO, F., CRIBARI-NETO, F. Adjusted and Bartlett-Adjusted profile likelihood ratio tests. 2002. 19p. (RT-MAE-2002-05).

2002-06 - PEREIRA, C.A.B., STERN, J.M. Full Bayesian Significance Test: Invariant Formulation. 2002. 10p. (RT-MAE-2002-06).

2002-07 - BUENO, V.C. Optimal arrangement of components for systems of dependent components. 2002. 13p. (RT-MAE-2002-07)

2002-08- KOLEV, N., PAIVA, D. Multinomial Latent Model for Random Sums. 2002. 9p. (RT-MAE-2002-08)

2002-09 - BUENO, V.C. A Note on a Hazard Processes Ordering: Application on a *K-out-of-N System*. 2002. 9P. (RT-MAE-2002-09)

The complete list of "Relatórios do Departamento de Estatística", IME-USP, will be sent upon request.

Departamento de Estatística
IME-USP
Caixa Postal 66.281
05311-970 - São Paulo, Brasil