The Use of Doubly Stochastic Poisson Processes in Estimating Health Effects Due to Air Pollution*

Jack M. Cuzick**

This paper looks at some questions associated with statistical inference of doubly stochastic Poisson processes. This general model appears to be of value in studying a number of problems in environmental health in which the factors which affect the rate at which certain diseases occur cannot be deterministically characterized. In particular, we consider the model's applicability to a study of air pollution and acute respiratory disease in the New York City area which is being conducted by Inge Goldstein. We shall focus our attention on problems which are suggested by this study. We begin by briefly describing this study.

The study will examine the effects of air pollution on the incidence of acute respiratory diseases, especially asthma. The ultimate goal is to predict the rates of occurrence of these disorders as a function of air pollution concentrations. The present study involves recording the number of emergency room visits for respiratory complaints at two major hospitals located in different parts of New York City. Daily temperature, humidity, and concentrations of $\text{SO}_2$ and smokeshade are recorded from the 40 station aerometric network for


**Department of Mathematical Statistics of Columbia University, New York, New York 10027.
DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.
DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.
York City. The regions covered by the hospitals are about 8 miles apart and their weather conditions are similar. However, the correlation between their pollution (SO₂) readings is quite low so that if pollution affects the rate of occurrence of acute asthma, there should be differences in the daily number of visits at the two hospitals. Hopefully these variations can be related to pollution levels and other ephemera. Attention is restricted to the differences in asthma attack rates within a single city and on the same day to minimize the number of confounding factors.

A model considered for the problem is the doubly stochastic Poisson process. In such a model we hypothesize a general positive stochastic process λₙ to model the rate of occurrence of events on day n. Such a model can allow for such things as correlation between days and different distributions of λₙ on different days. Then, conditional on λₙ, the number of events on day n is Poisson with rate λₙ. For practical purposes the model is made tractable by specifying some simple structure for the λₙ rate process, e.g. it is a Markov process. In our case we let λₙ denote the susceptibility of the population to attack on day n and Nₙ denote the actual number of patients seen at a single hospital on day n. The intensity process λₙ can be assumed to depend on pollution, temperature, and other factors. The most basic model is the linear relation
\[ \lambda_n = \alpha p_n + \varepsilon_n \]

where \( p_n \) is the pollution level on day \( n \), \( \alpha \) is a constant, and \( \varepsilon_n \) is a positive random variable, independent of all other variables. For such a model it is possible to estimate \( \alpha \) given the values of \( N_n \) and \( p_n \) by least squares or Bayesian techniques along the lines of Grandell (1972). More complicated models would allow regression on lagged variables, correct for temperature, season, etc.

Our early work suggests that regression models of this sort are unsuitable for the following reasons:

1. Major nonlinear dependence of \( \lambda_n \) on temperature, in particular there are effects due to cold spells, heat waves, etc.
2. Interaction between pollution levels and temperature patterns.
3. Large number of correction factors – seasonal trend, day of week, special time periods due to influenza outbreaks, holiday periods, etc.
4. Inaccuracy and lack of correlation in pollution measurements.

This last item has caused serious concern. The ability to characterize the pollution levels in a region by selected monitoring stations is the subject of a separate investigation. Because of our concern over the validity of the pollution data, we are studying the hospital data for the two hospitals separately, without trying to
relate them to outside factors, e.g. pollution levels and temperature patterns. Thus we shall be interested in doubly stochastic Poisson processes in which we have no outside information about the underlying intensity process. The following items are of interest.

A basic problem is that of inference about the intensity process from observations of the counting process. Under the assumptions of stationarity and ergodicity, one can ask for estimates of the mixing distribution $\lambda = \lambda_n$ which is now independent of $n$. Such an estimate will be based on the observed counts $N_n$, where one-dimensional distributions will again be independent of $n$. This is exploratory information and gives the investigator some knowledge about the type of variation to be expected in the intensity process. The method of Tucker (1963) can be used to estimate $\lambda$. If we let $N = N_n$, he has shown that if one has consistent estimates of $P(N = k)$ for each $k$, then one can construct a consistent estimate for the distribution of $\lambda$. Nothing is known about the rate of convergence of this estimate or its asymptotic properties, even if it is assumed that all the observations $N_i$ are independent. Such results would be of major practical interest. The difficulties are best illustrated by a brief description of Tucker's method.
We first need some notation. Let $\lambda$ have distribution function $F(\lambda)$. Then

$$P(N=k) = \int_0^\infty \frac{\lambda^k}{k!} e^{-\lambda} dF(\lambda).$$

Let

$$\mu_k = k! P(N=k) = \int_0^\infty \lambda^k dH(\lambda)$$

where

$$H(\lambda) = \int_0^\lambda e^{-\tau} dF(\tau).$$

Then the $\{\mu_k\}$ form a Stieltjes moment sequence for the distribution function $H(\lambda)$. These sequences are characterized by the fact that the determinants

$$\Delta_n = \det \begin{pmatrix} \mu_0 & \mu_1 & \cdots & \mu_{n-1} \\ \mu_1 & \mu_{n-1} \\ \vdots & \ddots & \vdots \\ \mu_{n-1} & \cdots & \mu_{2n-2} \end{pmatrix}$$
and

\[
\Delta_n' = \text{det} \begin{pmatrix}
\mu_1 & \mu_2 & \cdots & \mu_n \\
\mu_2 & \mu_n & & \\
& \ddots & \ddots & \\
& & \mu_n & \cdots & \mu_{2n-1}
\end{pmatrix}
\]

are non-negative for all \( n \). Further, if \( \Delta_n \) or \( \Delta_n' \) equal zero, then \( \Delta_m = \Delta'_m = 0 \) for all \( m > n \) and the distribution function \( H \) has no more than \( n \) points of increase. We shall denote by a hat the sample values of the above quantities so that

\[
\hat{\mu}_k = k! \frac{\# \text{days with } N=k}{\text{total } \# \text{days}}.
\]

We can now describe the algorithm. For a fixed sample size compute the \( \hat{\mu}_k \) and from these find the largest \( n \) so that both \( \hat{\Delta}_n \) and \( \hat{\Delta}_n' \) are positive.

The estimate of \( \hat{H}(\lambda) \) will then have \( n \) mass points \( x_i \) with mass \( a_i \) calculated from \( \{\hat{\mu}_0, \ldots, \hat{\mu}_{2n-1}\} \) as follows:

Let the polynomial \( P_n(x) \) be given by
Then $P_n(x)$ will always have exactly $n$ real positive distinct roots which become the mass points $x_i$. The weights $a_i$ are found by solving the linear equations

$$
\hat{\mu}_i = \int_0^\infty \lambda^i d\hat{H}(\lambda) = \sum_{j=1}^n a_j x_j^i, \quad i = 0, \ldots, n-1i,
$$

which always have a unique solution.

Finally we obtain $\hat{F}(\lambda)$ from $\hat{H}(\lambda)$ via

$$
\hat{F}(\lambda) = \int_0^\lambda e^T d\hat{H}(\tau).
$$

Tucker shows that if $\hat{\mu}_k \to \mu_k$ pointwise with increasing sample size, then $\hat{F} \to F$ in distribution. Recently, Simar has considered maximum likelihood estimates for $F(\lambda)$ and has established their existence and uniqueness. His approach warrants further attention.

We next consider relationships between second order properties of the $\lambda$ process and the $N$ process. Difficulties arise because
Var(N) = Var(\lambda) + E(\lambda)
so that even though \( E(\lambda_n \lambda_{n+j}) = E(N_n N_{n+j}) \) for \( j \neq 0 \) the correlation coefficient for the intensity process will be larger in magnitude than that for the counting process. For example, if \( \lambda \) has an exponential distribution, then

\[
\frac{\text{Cor}(\lambda_n, \lambda_{n+j})}{\text{Cor}(N_n, N_{n+j})} = 2 \quad j \neq 0
\]

and the counting process must have correlation less than one-half.

If we try to estimate \( \text{Var}(\lambda) \) by

\[
\hat{\sigma}^2(\lambda) = \hat{\sigma}^2(N) - \overline{N}
\]

then we have the usual problem when estimating a component of variance that it may be negative. This can be partially overcome by a Bayesian approach but the results are not totally satisfactory.

We shall close by considering two types of autoregressive schemes that can be considered for \( N_i \), namely the doubly stochastic model and the self-exciting model. We derive a few properties of each model in the simplest case which serves to contrast the two types of models.

**A. Doubly stochastic Poisson model.** Here \( \lambda_n \) depends only on the previous \( \lambda_i, i \leq n \), and there is no dependence on the actual values \( N_i \). This is a reasonable model when there is no contagious
factor, but only varying environmental factors. The simplest example is

\[ \lambda_n = \alpha \lambda_{n-1} + \varepsilon_n \]

with \( 0 < \alpha < 1 \) and \( \varepsilon_n \) positive i.i.d. random variables. If we write \( E(\varepsilon_n) = \mu \) and \( \text{Var}(\varepsilon_n) = \sigma^2 \), then in the steady-state we have

\[ E(\lambda_n) = E(N_n) = \frac{\mu}{1-\alpha} \]

\[ \text{Var} \lambda_n = \frac{\sigma^2}{1-\alpha^2} \]

\[ \text{Var}(N_n) = \text{Var} \lambda_n + E\lambda_n = \frac{\sigma^2}{1-\alpha^2} + \frac{\mu}{1-\alpha} \]

and finally

\[ \text{Cov}(N_n, N_{n+k}) = \begin{cases} \text{Var} \lambda + E(\lambda), & k = 0 \\ \alpha^k \text{Var}(\lambda), & k \neq 0 \end{cases} \]

B. Self-exciting model. In this case \( \lambda_n \) depends on the previous counts \( N_i \), not the previous rates \( \lambda_i \). This gives a more variable process for the same parameters. This model is more appropriate if there is a contagious aspect to the disease. The simplest example is

\[ \lambda_n = \beta N_{n-1} + \varepsilon_n, \quad 0 < \beta < 1 \]

with the same restrictions on \( \varepsilon_n \) as above.
For this model in the steady state we have

\[ E(\lambda_n) = E(N_n) = \frac{\mu}{1-\beta}, \]

\[ \text{Var}(N_n) = \frac{E(N_n) + \sigma^2}{1-\beta^2} = \frac{\sigma^2}{1-\beta^2} + \frac{\mu}{(1-\beta)(1-\beta^2)}. \]

Some calculation shows that

\[ \text{Cov}(N_n, N_{n+k}) = \beta^k \text{Var}(N_n). \]

A notable difference between the models is the extra delta function at the origin found in the doubly stochastic Poisson model. With enough data this may be helpful in deciding if an environmental disease also has a contagious aspect.

REFERENCES
