# A Poisson Process Approach for Recurrent Event Data with Environmental Covariates

Anup Dewanji Suresh H. Moolgavkar

# NRCSE

Technical Report Series

NRCSE-TRS No. 028

July 28, 1999

# A POISSON PROCESS APPROACH FOR RECURRENT EVENT DATA WITH ENVIRONMENTAL COVARIATES

By

Anup Dewanji
Applied Statistics Unit
Indian Statistical Institute
203 B. T. Road, Calcutta, India

and

Suresh H. Moolgavkar
Fred Hutchinson Cancer Research Center
Seattle, Washington 98109-1024

Address for correspondence:
Suresh H. Moolgavkar
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue North, MP-665
PO Box 19024, Seattle, Washington 98109-1024
Tel.: 206 667 4273
FAX: 206 667 7004
e-mail: smoolgav@fhcrc.org

# A POISSON PROCESS APPROACH FOR RECURRENT EVENT DATA WITH ENVIRONMENTAL COVARIATES

**Abstract.** We present a Poisson process formulation for studying the association between environmental covariates and recurrent events. The standard methods, which compare the covariate values (at event times) of the individuals having the event with those of the individuals at risk at that time, cannot accommodate environmental covariates, because they are identical for all individuals at risk. We suggest a flexible parametric model and a conditional likelihood analysis. We illustrate our method through an analysis of data on multiple hospital admissions for chronic respiratory disease in King County in relation to air pollution indices.

**Keywords.** Recurrent event data, Environmental covariates, Conditional likelihood analysis, Chronic respiratory disease, Case-crossover design.

# 1 Introduction

Recurrent events (or failures) occur frequently in studies in which the failures are not necessarily fatal (Kalbfleisch and Prentice, 1980, p179-182). Examples include asthmatic attacks, epileptic seizures, hospital admissions, etc.. A point process formulation is commonly used to describe and analyse such data. Regression analysis in this framework, in which the intensity rate of the point process under consideration is modeled as a function of covariates or covariate processes (time dependent covariates), has attracted a lot of attention since the first paper by Prentice et al. (1981), who considered a partial likelihood approach with arbitrary baseline intensity rates. There has been work on robust regression analyses of recurrent event data using the point process formulation, which, instead of modeling the intensity rates, models some other "marginal" quantities thus avoiding strong assumptions on the recurrent event process. See Wei et al. (1989), Pepe and Cai (1993) and Lawless and Nadeau (1995) for such work. These approaches focus on subject specific covariates requiring multiple subjects and fail for environmental covariates as they are the same for all the subjects at any event time.

Recent work on environmental covariates has focused on the use of generalized additive models for investigating associations between indices of air pollution measured at central monitoring locations and daily counts of events such as death or hospital admissions (Schwartz, 1993, 1994; Moolgavkar et al., 1999a). This approach fails to incorporate between subjects variation in the baseline parameters and subject specific covariates, if any.

Our work here focuses on a model that views the data on each subject as the realization of a point process, the intensity rate of which depends on the environmental covariates. This point process formulation allows us to incorporate subject specific covariates and also the previous history of the process. Specifically, we make a Poisson process assumption in section 2 and derive relevant likelihood functions for estimating the regression parameters under different assumptions on the baseline Poisson intensity. In section 3, we illustrate our method by means of an example of air pollution and hospital admissions for chronic respiratory disease as primary diagnosis. Section 4 ends with discussion.

## 2 Maximum Likelihood Estimation

First we describe a general development of the point process approach for analysing recurrent event data. For this development, let the covariate process consist of all the subject and environment specific covariates, some of which may be time dependent and random as well. We denote this, in general, by $X_t$, the vector of values of the covariates at time $t$, with $x_t$ being the corresponding observed values. Let the point process be denoted by $\{N(t),\ 0 \leq t \leq \tau\}$, where the process is assumed to be observed over the period $(0, \tau]$, and define $H_t = \{N(s),\ s \leq t\}$ as the "history" of the process up to time $t$ including information on the covariate process. The intensity of the process $\lambda(t, H_t)$ is defined by

$$\lambda(t, H_t) = \lim_{dt \downarrow 0} \left[ \frac{Pr\{dN(t) = 1 | H_t\}}{dt} \right] , \tag{1}$$

where $dN(t)$ denotes the number of events over the small interval $[t, t + dt)$. The probability distribution of the point process can be given in terms of $\lambda(t, H_t)$. In particular, the likelihood contribution from each subject (process) having $d$ events at times $t_1 < \cdots < t_d$ over the period $(0, \tau]$ is

$$\left[ \prod_{j=1}^{d} \lambda(t_j, H_{t_j}) \right] \times exp \left[ - \int_0^{\tau} \lambda(t, H_t) dt \right] . \tag{2}$$

The central aspect of analyzing recurrent event data by point process approach is to model the intensity $\lambda(t, H_t)$ judiciously to be able to make inference on the effect of $X_t$ on the occurrence of recurrent events without losing much flexibility in the baseline intensity. Let us first consider a simple model, that of a non-homogeneous Poisson process, for which the intensity for the $i$th process (subject) with $X_t = x_{it}$ is

$$\lambda(t, H_t) = \lambda(t, x_{it}) = \lambda_i exp[x_{it}^T \beta] , \tag{3}$$

so that the baseline intensity $\lambda_i$ varies from subject to subject but is independent of time and the relative risk parameter $\beta$, which is of primary interest, remains the same over all subjects. Using (2), this leads to the likelihood function

$$\prod_{i=1}^{n} \left\{ \lambda_i^{d_i} \left( \prod_{j=1}^{d_i} exp \left[ x_{it_{ij}}^T \beta \right] \right) exp \left( -\lambda_i \int_0^{\tau_i} exp[x_{it}^T \beta] dt \right) \right\} , \tag{4}$$

where $n$ is the number of subjects under study, with the $i$th subject being observed over the period $(0, \tau_i]$, $d_i$ is the number of events on the $i$th subject at times $t_{i1} < \cdots < t_{id_i}$, for $i = 1, \cdots, n$. This results in the estimate of $\lambda_i$ for a given $\beta$ as

$$\hat{\lambda}_i(\beta) = \frac{d_i}{\int_0^{\tau_i} exp[x_{it}^T \beta] dt} \ ,$$

for $i = 1, \cdots, n$. Putting this back in (4), we get the score equation for $\beta$ by differentiating

$$\prod_{i=1}^{n} \left\{ \frac{\prod_{j=1}^{d_i} exp\left[x_{it_{ij}}^T \beta\right]}{\left(\int_0^{\tau_i} exp[x_{it}^T \beta] dt\right)^{d_i}} \right\} \ . \tag{5}$$

Thus, maximization of (5) gives the maximum likelihood estimate of $\beta$, namely $\hat{\beta}$. In the presence of large number of nuisance parameters ($\lambda_i$'s), large sample properties of $\hat{\beta}$ may be in question. However, one can view (5) as a conditional or partial likelihood of the data given $(d_1, \cdots, d_n)$ in which case the large sample properties are similar to those from the full likelihood (Kalbfleisch and Sprott, 1970; Cox, 1975). Note that individual subjects can enter the study at different times and the likelihood (5) remains the same except the range of integration in the denominator changing accordingly.

Note that the likelihood (5) incorporates environmental covariates sucessfully, but it depends on the model (3) which assumes time independent baseline intensity $\lambda_i$ although allowing for heterogeneity between subjects. Interestingly, the Poisson process formulation allows one to incorporate some limited amount of time dependence in the $\lambda_i$'s. Assume the $\lambda_i$'s, as function of time $t$, to be piecewise constant over the period $(0, \tau_i]$ as in the following:

$$\lambda_i(t) = \lambda_{il} \text{ for } \ t \in I_{il} = (\tau_{i,l-1}, \tau_{il}], \text{ for } \ l = 1, \cdots, K_i, \tag{6}$$

with $0 = \tau_{i0} < \tau_{i1} < \cdots < \tau_{iK_i} = \tau_i$ being prespecified. Let $d_{il}$ denote the number of events for the $i$th subject in $I_{il}$. We shall call these intervals $I_{il}$'s 'strata' from now on as they describe a partition of the observation period in such a way that the baseline intensity for an individual is different in different strata, but the same within a stratum. Since the events in disjoint strata $I_{il}$'s are independent (because of the Poisson process assumption), one can form a conditional liklihood, given $d_{il}$, of the form given inside the braces in (5) corresponding to each of the strata $I_{il}$'s, and then take product over all

the $(i,l)$'s. This results in the likelihood given by

$$\prod_{i=1}^{n}\prod_{l=1}^{K_i}\left\{\frac{\prod_{j=1}^{d_{il}} exp\left[x_{it_{ilj}}^T\beta\right]}{\left(\int_{I_{il}} exp[x_{it}^T\beta]dt\right)^{d_{il}}}\right\} \ , \tag{7}$$

where $t_{ilj}$ is the occurrence time for the $j$th event in $I_{il}$ for the $i$th subject. Note that, although this kind of limited time dependence of the baseline intensity may not be satisfactory in light of their total arbitrary nature in the work of Prentice et al. (1981) and others dealing with subject specific covariates, this approach is flexible enough to allow between subject heterogeneity.

It is interesting to note that the likelihood (7) can also be derived, as for (5), by first finding estimates of $\lambda_{il}$'s from the full likelihood (see (2) and (4)), for a given $\beta$, and then putting them back in the full likelihood. This approach is reminiscent of that of Breslow (1974) for Cox's proportional hazards model, which coincides with Cox's likelihood for regression analysis with simple survival data (Kalbfleisch and Prentice, 1980, p76-79). In this approach, the stratification is done based on the observed failure times. With environmental covariates, this approach leads to a reasnable likelihood to base inference on, whereas Cox's partial likelihood fails, and this can be easily applied to recurrent event data, as seen in the derivation of (5). This approach also extends easily to some non-Poisson processes as will be discussed in section 4.

In special case when $\tau_i$ is same for all the subjects and the strata $I_{il}$'s are also the same (the $\lambda_{il}$'s need not be same), the likelihood (7) takes a simpler form given by

$$\prod_{l=1}^{K}\left\{\frac{\prod_{i\in D_l} exp\left[x_{it_{il}}^T\beta\right]}{\prod_{i\in D_l}\left(\int_{I_l} exp[x_{it}^T\beta]dt\right)}\right\} \ , \tag{8}$$

where $K$ is the number of strata, $D_l$ denotes the set of subjects having events (one subject may have more than one event) in the $l$th common stratum $I_l$ at times $t_{il}$'s. More specifically, if there are no subject specific covariates (only environmental covariates), then, writing $x_{it}$ as $x_t$ for the covariates at time $t$, (8) reduces to

$$\prod_{l=1}^{K}\left\{\frac{\prod_{i\in D_l} exp\left[x_{t_{il}}^T\beta\right]}{\left(\int_{I_l} exp[x_t^T\beta]dt\right)^{d_l}}\right\} \ , \tag{9}$$

where $d_l = |D_l|$, for $l = 1, \cdots, K$.

The independence of events in disjoint strata can also be used to deal with missing data in the covariates. The idea is to form a collection of disjoint intervals in such a way that the data is available in this collection and missing outside it. One can decide on a partition with a certain number of strata beforehand, and then, within each stratum, a collection of intervals, as above, is considered. The Poisson process assumption allows one to consider the conditional probability of observed data in a collection like this in each stratum, given the total number of events in the collection. In the example in section 3, we considered different types of partitioning or stratification and, in each stratum, formed the collection of intervals in this way (as described above) to deal with missing data, and used the likelihood (9) with $I_l$ denoting the collection in the $l$th stratum.

Although the above modeling of time dependent baseline intensity seems simple and natural, it is difficult in practice to decide on the partitions, $I_l$'s. There is clearly a trade-off between model flexibility in going to a large number of strata and loss of information. From the numerical work in section 3, the regression parameter estimates seem to be sensitive to the choices of partitions. Clearly, optimal partitioning will depend on temporal and cyclical trends in the covariates.

# 3 Example: Air Pollution and Hospital Admissions in King County

We illustrate the method by applying it to the analysis of hospital admissions for chronic respiratory disease in King County over the period 1990-1995. We considered the cohort of individuals admitted in 1990 to a King County hospital with a primary diagnosis of chronic respiratory disease (ICD-9 codes 490-496). For this cohort of individuals, we constructed a history of hospitalizations over the entire period 1990-1995. Our data then consisted of 5362 admissions for 1867 individuals. We obtained air pollution and weather information on a daily basis over this period of time from central monitoring locations. Details of this data can be found in recent publications (Sheppard et al., 1999; Moolgavkar et al., 1999a,b). Among the air pollution variables, we were particularly interested in carbon monoxide, $PM_{10}$ (particulate matter less than 10 microns in diameter) and an index of light scattering (LS) measured by nephelometry, which is a surrogate measure of fine particles.

We chose these indices of air pollution for our analyses because they had been shown in previous analyses (Sheppard et al., 1999; Moolgavkar et al., 1999a,b) to be associated with respiratory admissions.

We analyzed the data by maximizing the likelihood (9) using five distinct stratifications of the time period: a single stratum over the entire period (same as Navidi, 1998), 6 strata corresponding to each of the years 1990-1995, 25 strata representing distinct seasons as defined below, 72 strata corresponding to each month in the six-year period of interest, and 144 strata, two for each month. The seasonal strata were defined as follows. Winter was December together with January and February of the next year. Spring was defined as March, April and May. Summer was June, July and August. Fall was September, October and November. In addition to the pollution covariates with various lag times, we also included temperature and day of week in our analyses. Day of week has been shown in previous analyses to be an important predictor of hospital admissions. We used six indicator variables for day of week. We controlled temperature in one of two ways: either by a linear model, with a distinct slope for each season, or as a cubic polynomial.

We tried various lags for the covariates. Temperature with a three day lag appeared to be the strongest predictor of hospital admissions and all our subsequent analyses were done with this lag for temperature. Previous analyses of these data had indicated that the strongest effects were seen with lags of 3, 1 and 0 days for $PM_{10}$, LS and CO, respectively. We present our results in tables 1 and 2 with these lags for the pollutants. We found both day of week effects and temperature effects in these data. Table 1 presents the results of models with temperature controlled linearly. Table 2 presents the results of analyses with temperature controlled using a cubic polynomial.

In single pollutant analyses, it is clear from tables 1 and 2 that both measures of particulate matter ($PM_{10}$ and LS) and CO are associated with hospital admissions. Although the results are somewhat sensitive to the method used for controlling temperature, in general the effect of particulate matter is stronger than that of CO in the multipollutant models. This result is inconsistent with other recent analyses of the same data (Moolgavkar et al., 1999a,b), which find that the effect of particulate matter becomes insignificant when CO is simultaneously considered in the analyses. This inconsistency is troubling because it precludes any conclusions regarding the specific component of air pollution that may be responsible for the association with hospital admissions. The only conclusion that can be drawn collectively from these various analyses is that air pollution is associated with hospital

admissions for chronic respiratory disease.

Table 1. Parameter estimates with temperature controlled linearly

| Covariates | Stratification Type | | | | |
|---|---|---|---|---|---|
| | One stratum | 6 strata | 25 strata | 72 strata | 144 strata |
| LS (Light scattering) | .3134* (.0318) | .1033* (.0322) | .1166* (.0344) | .1009* (.0356) | .1075* (.0398) |
| $PM_{10}$ | .0085* (.0009) | .0029* (.0009) | .0021* (.0009) | .0025* (.0010) | .0027* (.0011) |
| CO | .2657* (.0219) | .0620* (.0230) | .0528* (.0238) | .0543* (.0247) | .0513 (.0264) |
| LS | .1797* (.0371) | .0842* (.0360) | .1053* (.0378) | .0891* (.0387) | .0983* (.0417) |
| CO | .2016* (.0271) | .0335 (.0279) | .0210 (.0289) | .0232 (.0296) | .0232 (.0309) |
| $PM_{10}$ | .0069* (.0009) | .0027* (.0009) | .0019* (.0010) | .0024* (.0010) | .0027* (.0011) |
| CO | .2070* (.0250) | .0337 (.0260) | .0321 (.0270) | .0363 (.0277) | .0322 (.0298) |

Note: The figures in parentheses are standard errors and * indicates significance at 5% level.

The results are sensitive to the particular stratification used. It is clear that temporal trends and seasonal variations in the covariates, which the stratification attempts to control, will affect the results of analyses. It is not clear what the optimal stratification scheme should be. The baseline intensity, which includes the effects of covariates not considered in an analysis, should remain constant within strata. Thus, it is likely that different schemes will be optimal for different pollutants and for different geographic locations.

Table 2. Parameter estimates with temperature controlled by a cubic polynomial

| Covariates | Stratification Type | | | | |
|---|---|---|---|---|---|
| | One stratum | 6 strata | 25 strata | 72 strata | 144 strata |
| LS (Light scattering) | .3007* (.0320) | .1349* (.0325) | .1229* (.0343) | .1153* (.0359) | .0998* (.0399) |
| $PM_{10}$ | .0077* (.0009) | .0035* (.0009) | .0025* (.0009) | .0023* (.0010) | .0021 (.0011) |
| CO | .2303* (.0205) | .0708* (.0218) | .0632* (.0234) | .0564* (.0247) | .0532* (.0264) |
| LS | .1699* (.0377) | .1076* (.0363) | .0999* (.0374) | .0946* (.0387) | .0862* (.0416) |
| CO | .1715* (.0247) | .0435 (.0253) | .0423 (.0268) | .0416 (.0285) | .0351 (.0297) |
| $PM_{10}$ | .0058* (.0009) | .0032* (.0009) | .0022* (.0009) | .0021* (.0010) | .0022* (.0011) |
| CO | .1965* (.0225) | .0504* (.0234) | .0582* (.0250) | .0435 (.0262) | .0373 (.0282) |

Note: The figures in parentheses are standard errors and * indicates significance at 5% level.

# 4 Discussion

The likelihood (5) compares the covariate values at failure times of an individual with those at other times, as is intuitively required for environmental covariates. In practice, the covariate data is not collected continuously in time and, therefore, the integrals in (5), (7) and (9) need to be approximated by an appropriate sum. With only one failure for each subject and no time dependence in baseline intensity, the approximate form of the likelihood (5) coincides with the likelihood for the case-crossover design (Navidi, 1998). However, for multiple failures per subject, Navidi's likelihood is numerically more difficult than (5).

Because of the Poisson process assumption, the events in successive strata are independent and, therefore, effectively treated as independent processes or individuals with different baseline intensities. In practice, it may be rea-

sonable to allow for some dependence between events in successive strata corresponding to a particular process or individual. For example, the events in a particular stratum, given the events in earlier strata, may be assumed to follow a Poisson process with intensity depending on the events (e.g., total number) in earlier strata. If this dependence can be modeled through only the baseline intensity, then we still have the likelihoods (7)-(9) with terms corresponding to successive strata being interpreted as conditional probabilities.

However, Breslow's approach, referred to in section 2, can be applied for some non-Poisson processes. For example, if the baseline intensity $\lambda_0(t)$ is modeled as a linear birth rate, $\lambda_0(t) = (n+1)\lambda_i$, for the $i$th individual, where $n = N_i(t-)$, the number of events before time $t$ in $i$th individual and $\lambda_i > 0$. Then, following the same derivation as (5), one can obtain the likelihood for the regression parameters as proportional to

$$\prod_{i=1}^{n} \left\{ \frac{\prod_{j=1}^{d_i} exp\left[x_{it_{ij}}^T \beta\right]}{\left(\sum_{j=1}^{d_i+1} j \int_{t_{i,j-1}}^{t_{ij}} exp[x_{it}^T \beta] dt\right)^{d_i}} \right\} ,$$

where the notation is as in (5) with $t_{i,d_i+1} = \tau_i$. If $\lambda_i$ can be assumed to be equal for all $i$, then this likelihood takes the form

$$\frac{\prod_{i=1}^{n} \prod_{j=1}^{d_i} exp\left[x_{it_{ij}}^T \beta\right]}{\left(\sum_{i=1}^{n} \sum_{j=1}^{d_i+1} j \int_{t_{i,j-1}}^{t_{ij}} exp[x_{it}^T \beta] dt\right)^{\sum_{i=1}^{n} d_i}} .$$

One can also deal with some semi-Markov models using the above approach. Suppose $\lambda_0(t)$ is modeled as

$$\lambda_0(t) = \lambda(t - t_{N_i(t-)}) = \lambda_{il}, \text{ if } t - t_{N_i(t-)} \in I_l,$$

for $l = 1, \cdots, K$, and for the $i$th individual, where the $I_l$'s represent a partition of the range of interoccurrence times and $t_{N_i(t-)}$ denotes the last event time before $t$ in $i$th individual. Then the likelihood for the regression parameters is

$$\prod_{l=1}^{K} \prod_{i=1}^{n} \left\{ \frac{\prod_{j \in D_{il}} exp\left[x_{it_{ijl}}^T \beta\right]}{\left(\int_{S_{il}} exp[x_{it}^T \beta] dt\right)^{d_{il}}} \right\} ,$$

where $D_{il}$ denotes the set of $d_{il}$ events for the $i$th individual having time since last event in $I_l$ with the event times $t_{ijl}$, $j \in D_{il}$, and $S_{il} = \{0 < t \leq \tau_i :$

$t - t_{N_i(t-)} \in I_l\}$, the subset of $(0, \tau_i]$ where the time since last event for the $i$th individual lies in $I_l$. As before, if $\lambda_{il}$ can be assumed to be equal for all $i$, then this likelihood takes the form

$$\prod_{l=1}^{K} \frac{\prod_{i=1}^{n} \prod_{j \in D_{il}} exp\left[x_{it_{ij}}^T \beta\right]}{\left(\sum_{i=1}^{n} \int_{S_{il}} exp[x_{it}^T \beta] dt\right)^{\sum_{i=1}^{n} d_{il}}} \ .$$

In the above semi-Markov model, we need to assume that an event occurs at time 0, or the time since last event is known at time 0. Since events in non-overlapping intervals are generally not independent (as in Poisson process), dealing with missing data is difficult with non-Poisson models as above. Also, the above likelihoods for non-Poisson models cannot be viewed as conditional likelihood as those in section 2.

**Acknowledgement.** This research was supported by EPA grant R825266-01-0. The authors are grateful to Drs W. D. Hazelton and E. G. Luebeck for helpful discussions.

**References:**

1. Breslow, N. (1974). "Covariance analysis of censored survival data". *Biometrics* **30**, 89-99.

2. Cox, D. R. (1975). "Partial likelihood". *Biometrika* **62**, 269-276.

3. Kalbfleisch, J. D. and Sprott, D. A. (1970). "Application of likelihood methods to models involving large number of parameters". *Journal of the Royal Statistical Society, Series B,* **32**, 175-208.

4. Kalbfleisch, J. D. and Prentice, R. L. (1980). *"The Statistical Analysis of Failure Time Data".* John Wiley& Sons, New York.

5. Lawless, J. F. and Nadeau, J. C. (1995). "Some simple robust methods for the analysis of recurrent events". *Technometrics* **37**, 158-168.

6. Moolgavkar, S. H., Hazelton, W. D., Luebeck, E. G., Levy, D. and Sheppard, L. (1999a). "Air pollution, pollens,and respiratory admissions for chronic obstructive pulmonary disease in King County". *Inhalation Toxicology (In Press).*

7. Moolgavkar, S. H., Dewanji, A., Hazelton, W. D. and Luebeck, E. G. (1999b). "Case-crossover analysis of air pollution and multiple hospital admissions for chronic respiratory disease in King County". *Submitted.*

8. Navidi, W. (1998). "Bidirectional case-crossover designs for exposures with time trends". *Biometrics* **54**, 596-605.

9. Pepe, M. S. and Cai, J. (1993). "Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates". *Journal of the American Statistical Association* **88**, 811-820.

10. Prentice, R. L., Williams, B. J. and Peterson, A. V. (1981). "On the regression analysis of multivariate failure time data". *Biometrika* **68**, 373-379.

11. Schwartz, J. (1993). "Air pollution and daily mortality in Birmingham, Alabama". *The American Journal of Epidemiology* **137**, 1136-1147.

12. Schwartz, J. (1994). "Air pollution and hospital admissions for the elderly in Birmingham, Alabama". *The American Journal of Epidemiology* **139**, 589-598.

13. Sheppard, L., Levy, D., Norris, G., Larson, T. V. and Koenig, J. Q. (1999). "Effects of ambient air pollution on nonelderly asthma hospital admissions in Seattle, Washington, 1987-1994". *Epidemiology* **10**, 23-30.

14. Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989) "Regression analysis of multivariate incomplete failure time data by modeling marginal distributions". *Journal of the American Statistical Association* **84**, 1065-1073.