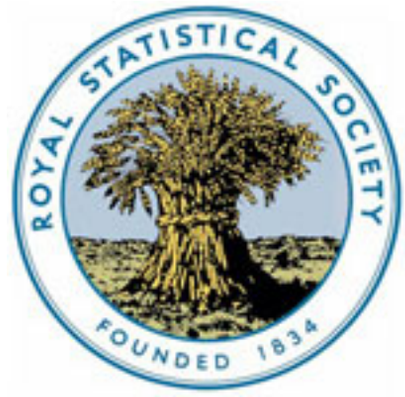

J. R. Statist. Soc. A (1994)
157, *Part* 3, *pp.* 433–440

A Conditional Approach to Point Process Modelling of Elevated Risk

By PETER J. DIGGLE† and BARRY S. ROWLINGSON

Lancaster University, UK

[Received March 1993. Revised September 1993]

SUMMARY

We consider the problem of investigating the elevation in risk for a specified disease in relation to possible environmental factors. Our starting point is an inhomogeneous Poisson point process model for the spatial variation in the incidence of cases and controls in a designated geographic region, as proposed by Diggle. We develop a conditional approach to inference which converts the point process model to a non-linear binary regression model for the spatial variation in risk. Simulations suggest that the usual asymptotic approximations for likelihood-based inference are more reliable in this conditional setting than in the original point process setting. We present an application to some data on the spatial distribution of asthma in relation to three industrial locations.

Keywords: BINARY REGRESSION; ENVIRONMENTAL EPIDEMIOLOGY; NON-LINEAR MODELLING; RELATIVE RISK; SPATIAL POINT PROCESS

1. INTRODUCTION

There has been much work in recent years on the development of statistical models for investigating possible elevation in the risk of various diseases near putative sources of environmental pollution. See, for example, Bithell and Stone (1989), Cook-Mozaffari *et al.* (1989), Muirhead and Darby (1989) and Diggle (1990). Several of these papers take the approach of a case–control study, in which the spatial distribution of the disease in a region A containing the putative source is compared with the spatial distribution of a control sample selected at random from the population at risk.

Diggle (1990) proposed an inhomogeneous Poisson point process model for the spatial distribution of disease around a single point source, as follows. He assumed a multiplicative model for the intensity function $\lambda(\mathbf{x})$, of the form

$$\lambda(\mathbf{x}) = \rho\, \lambda_0(\mathbf{x})\, f(\mathbf{x}-\mathbf{x}_0; \theta). \tag{1}$$

In model (1), $\lambda_0(\mathbf{x})$ is the intensity function of the population at risk, i.e. the number of people at risk per unit area in the neighbourhood of the point $\mathbf{x}$, and $f(\mathbf{u}, \theta)$ describes the elevation in risk as a function of location relative to a single point source $\mathbf{x}_0$. The quantity ρ is a scaling parameter which reflects the overall prevalence of the disease relative to the number of controls. To the extent that the number of controls is determined by the investigator, the value of ρ in a given application is a by-product of the sampling protocol and has no scientific interest. Diggle developed approximate likelihood-based methods for fitting model (1), using

†*Address for correspondence*: Department of Mathematics, Lancaster University, Lancaster, LA1 4YF, UK.
E-mail: maa026@uk.ac.lancs.cent1

 0035–9238/94/157433

a Gaussian kernel estimator (Silverman, 1986) in place of the unknown function $\lambda_0(\mathbf{x})$. He also made the pragmatic choice

$$f(\mathbf{u};\theta) = 1 + \theta_1 \exp(-\theta_2 \mathbf{u}'\mathbf{u}). \tag{2}$$

In this paper, we propose a new method of fitting model (1), in which we condition on the locations of the cases and controls. This converts the statistical model to a non-linear binary regression, avoids the problem of estimating $\lambda_0(\mathbf{x})$ and allows a straightforward extension of model (1) to incorporate whatever additional risk factors are considered relevant to each application.

Section 2 sets out the details of the statistical analysis for the non-linear binary regression. Section 3 presents an application to data on the spatial distribution of asthma.

2. FITTING THE MODEL

2.1. *Formulation*

The available data are a set of case locations $\mathbf{x}_i$, $i=1, \ldots, n$, and a set of control locations $\mathbf{x}_i$, $i=n+1, \ldots, n+m$, within a designated region A. We assume that the control locations are a realization of a Poisson process on A, with intensity $\lambda_0(\mathbf{x})$, and that the case locations are a realization of an independent Poisson process with intensity $\lambda(\mathbf{x})$ given by model (1). Then, conditional on the $n+m$ locations $\mathbf{x}_i$ which we call *events*, the *labels* of these $n+m$ events are a set of mutually independent Bernoulli random variables. Specifically, if $p(\mathbf{x})$ denotes the probability that an event at $\mathbf{x}$ is a case, it follows from model (1) that

$$p(\mathbf{x}) = \rho f(\mathbf{x}-\mathbf{x}_0;\theta)/\{1+\rho f(\mathbf{x}-\mathbf{x}_0;\theta)\}. \tag{3}$$

The conditioning eliminates $\lambda_0(\mathbf{x})$ from the model, as with the proportional hazards modelling of survival data introduced in Cox (1972).

One way to extend model (1)–(3) to accommodate two or more point sources is to assume a multiplicative elevation in risk from each source. Thus, if $g(\mathbf{u};\theta) = 1 + \theta_1 \exp(-\theta_2 \mathbf{u}'\mathbf{u})$ represents the elevation in risk from a single source, as in equation (2), and there are sources at locations $\mathbf{x}_{0k}$, $k=1, \ldots, q$, then we replace $f(\mathbf{x}-\mathbf{x}_0;\theta)$ in equation (3) by

$$f(\mathbf{x};\theta) = \prod_{k=1}^{q} g(\mathbf{x}-\mathbf{x}_{0k};\theta).$$

A further extension would be to allow separate parameter vectors $\boldsymbol{\theta}_k$ to operate on the function $g(\ )$ associated with each location $\mathbf{x}_{0k}$, and multiplicative risk factors associated with other explanatory variables, $z_j(\mathbf{x})$ say, to give

$$f(\mathbf{x};\theta_1, \ldots, \theta_k, \phi) = \prod_{k=1}^{q} g(\mathbf{x}-\mathbf{x}_{0k};\theta_k) \exp\left\{\sum_{j=1}^{r} \phi_j z_j(\mathbf{x})\right\}. \tag{4}$$

The $z_j(\mathbf{x})$ might correspond to environmental features such as land use or topography, or to subject-specific characteristics like age or sex. As with the formulation of the function $g(\ )$, we must ask in any particular application whether a log-linear formulation makes good sense.

2.2. *Inference*

The binary regression model implied by equation (3) cannot be a generalized linear model. This is because any sensible model for the elevation in risk due to the point source $\mathbf{x}_0$ will require that $f(\mathbf{x}-\mathbf{x}_0;\theta) \to 1$ as $(\mathbf{x}-\mathbf{x}_0)'(\mathbf{x}-\mathbf{x}_0) \to \infty$, implying that $p(\mathbf{x}) \to \rho/(1+\rho)$, an unknown value strictly between 0 and 1, whereas in any generalized linear model the link function which transforms the linear predictor to the mean response must not depend on unknown parameter values. Nevertheless, the log-likelihood function for ρ and θ takes the standard Bernoulli form,

$$L(\rho,\theta) = \sum_{i=1}^{n} \log p(\mathbf{x}_i) + \sum_{i=n+1}^{n+m} \log\{1-p(\mathbf{x}_i)\},$$

and we can use standard numerical methods to maximize $L(\rho,\theta)$. For ease of illustration, consider the form of $p(\mathbf{x})$ given by the basic model (3). Substitution from model (3) gives

$$L(\rho,\theta) = n\log\rho + \sum_{i=1}^{n} \log f(\mathbf{x}_i-\mathbf{x}_0;\theta) - \sum_{i=1}^{n+m} \log\{1+\rho f(\mathbf{x}_i-\mathbf{x}_0;\theta)\}. \qquad (5)$$

It is convenient to arrange the parameterization of the model so that, when $\theta=0$, $f(\mathbf{x}-\mathbf{x}_0;\theta)=1$ for all $\mathbf{x}$ as is the case when $f(\ )$ is specified by equation (2). Then, the log-likelihood with $\theta=0$ reduces to

$$L_0(\rho) = n\log\rho - (n+m)\log(1+\rho),$$

which is maximized at $\hat{\rho}_0 = n/m$. More generally, parameter estimates $\hat{\rho}$ and $\hat{\theta}$ and their standard errors can be obtained by numerical maximization of equation (5) and subsequent evaluation of the observed information matrix. Tests of hypotheses about θ can be carried out by the usual generalized likelihood ratio method; for example, if θ has r elements, a test of $\theta=0$ would involve a comparison between the test statistic $D=2\{L(\hat{\rho},\hat{\theta})-L_0(\hat{\rho}_0)\}$ and critical values of the χ^2-distribution on r degrees of freedom. The same general method can be applied to the extended model (4) and its associated log-likelihood $L(\rho,\theta_1,\ldots,\theta_k,\phi)$.

For likelihood-based inference, we require the function $f(\ )$, and therefore any explanatory variables $z_j(\mathbf{x})$ in the extended model (4), to be evaluated at all the events $\mathbf{x}_i$, but not necessarily in the remainder of the region A.

2.3. *Simulations*

Diggle (1990) found that the usual asymptotic properties of the likelihood ratio test did not hold in his particular application, and he used a Monte Carlo test to confirm his substantive conclusions. One of several possible explanations for the failure of the asymptotic distribution theory is the insertion of a nonparametric estimate in place of the unknown function $\lambda_0(\mathbf{x})$. The approach which we are now advocating avoids this complication, and it is therefore of interest to establish whether the usual asymptotic theory now holds to a better approximation. To achieve this, we ran a simulation experiment, generating data from the model given by equations (1) and (2). Throughout the experiment, we used a disc of unit radius for the region A, with the point source $\mathbf{x}_0$ at the centre of the disc. We also fixed $\rho=0.1$, $\lambda_0(\mathbf{x})=\lambda_0$, a constant, and $n+m=1000$.

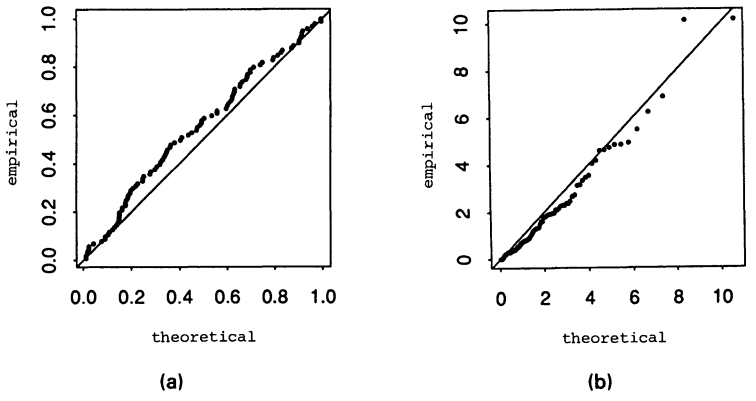

Fig. 1. Comparison between the empirical distribution of the generalized likelihood ratio statistic and the χ^2-distribution on 2 degrees of freedom: (a) *P–P* plot; (b) *Q–Q* plot

In the first set of 100 simulations, we fixed $\theta_1 = \theta_2 = 0$ and calculated the corresponding 100 values of the generalized likelihood ratio statistic D to test $\theta_1 = \theta_2 = 0$. Fig. 1 shows *P–P* and *Q–Q* plots of the resulting sample of D-values, with χ^2_2 as the reference distribution (Chambers *et al.*, 1983). The Kolmogorov–Smirnov goodness-of-fit statistic applied to the *P–P* plot is 0.115, corresponding to a p-value greater than 0.1. The *Q–Q* plot does not suggest any major discrepancy in the upper tail. We conclude that the χ^2_2-approximation to the null sampling distribution of D is satisfactory in this case.

In subsequent sets of 100 simulations we used non-zero values of θ, to enable us to investigate the adequacy of using the inverse of the observed information matrix as an approximation to the variance matrix of $\hat{\theta}$. In most cases, this appeared slightly to overestimate the true variances of $\hat{\theta}_1$ and $\hat{\theta}_2$. Fig. 2 summarizes the results for two slices through the (θ_1, θ_2) plane: one varying θ_1 with fixed $\theta_2 = 5$; the other varying θ_2 with fixed $\theta_1 = 6$. The vertical lines in the diagrams represent 95% confidence limits for the true variances of the $\hat{\theta}_j$ expressed as proportions of the average nominal variances derived from the observed information matrix. They were calculated as follows. For each combination of θ_1 and θ_2, the 100 simulations generated 100 values $\hat{\theta}_{ij}$, $i = 1, \ldots, 100$, $j = 1, 2$, and associated approximate variances v_{ij}. From these, we calculated s_j^2, the sample variance of the $\hat{\theta}_{ij}$, and $\bar{v}_j$, the sample mean of the v_{ij}. The sampling distribution of s_j^2 is proportional to χ^2 on 99 degrees of freedom. The end points of the vertical lines in the diagrams are therefore taken as $99s^2/\bar{v}_j c_1$ and $99s^2/\bar{v}_j c_2$, where c_1 and c_2 are the 0.975 and 0.025 quantiles of χ^2_{99}.

2.4. *Cancers of Larynx and Lung in South Lancashire*

The larynx and lung cancer data, from Diggle (1990), concern the spatial distribution of cancers of the larynx in the neighbourhood of a now disused industrial incinerator in the Chorley–Ribble area of Lancashire, England. The data set consists of 58 cases, together with 978 cases of cancer of the lung which are used as controls. Diggle fitted the model specified by equations (1) and (2), and obtained parameter estimates $\hat{\theta}_1 = 23.67$ and $\hat{\theta}_2 = 0.91$ with estimated standard errors of 24.69 and 0.91, and an estimated correlation of 0.83. There was also a significant

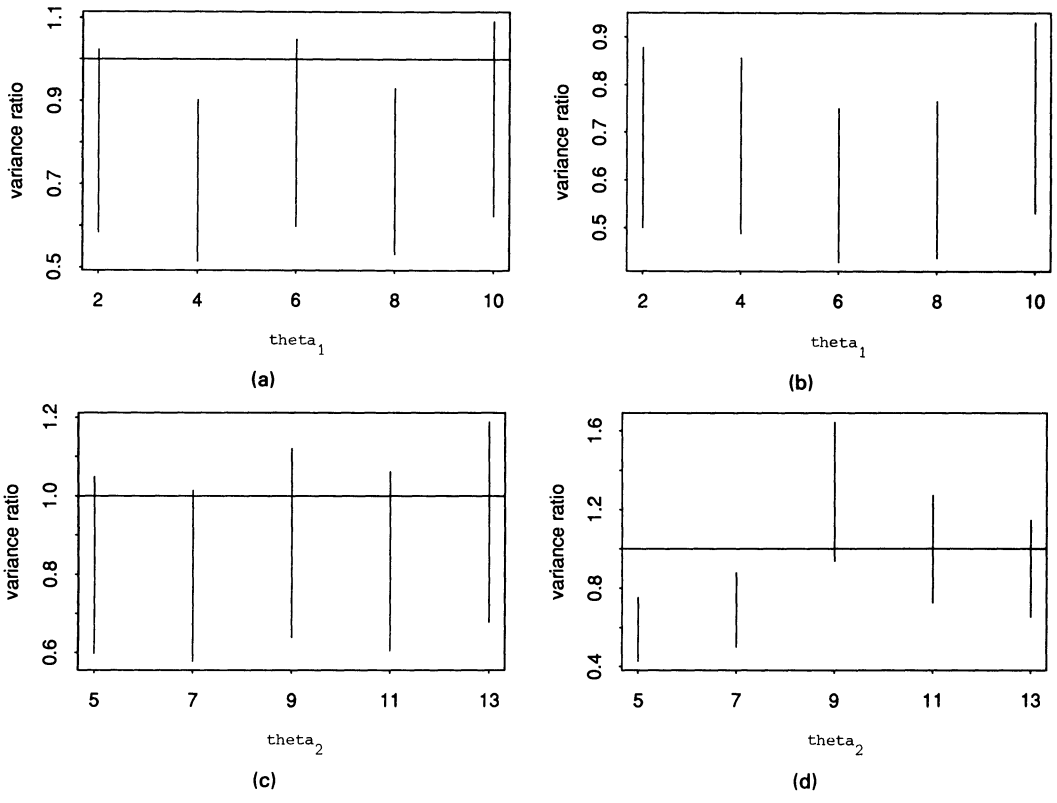

Fig. 2. Comparison between empirical and nominal variances of the maximum likelihood estimators $\hat{\theta}_j$: (a) variance ratio for $\hat{\theta}_1$, for varying θ_1 and fixed $\theta_2 = 5$; (b) variance ratio for $\hat{\theta}_1$, for varying θ_2 and fixed $\theta_1 = 6$; (c) variance ratio for $\hat{\theta}_2$, for varying θ_1 and fixed $\theta_2 = 5$; (d) variance ratio for $\hat{\theta}_2$, for varying θ_2 and fixed $\theta_1 = 6$

association with the location of the former incinerator; the generalized likelihood ratio test of the hypothesis $\theta = 0$ gave a nominal p-value of 0.008 using the suspect χ^2-approximation, or $p = 0.01$ by using a Monte Carlo test.

Our reanalysis, using the same model but the conditional analysis described in Section 2, gave parameter estimates $\hat{\theta}_1 = 33.69$ and $\hat{\theta}_2 = 1.11$, with estimated standard errors of 54.23 and 0.97, and an estimated correlation of 0.90. The generalized likelihood ratio test of $\theta = 0$ gave a χ^2-statistic $D = 8.66$ on 2 degrees of freedom, corresponding to $p = 0.013$. These results are in qualitative agreement with the earlier analysis.

3. ASTHMA IN NORTH DERBYSHIRE

We now use the extended model (4) to investigate the possible association between asthma and three industrial plants within the North Derbyshire Health Authority. In 1992, a survey was conducted of 2133 school-children in 10 schools within the area, five of which were chosen because the head teacher had previously expressed concern about apparently high asthma levels among the children. Parents completed a questionnaire for each child, which included the question

has the child ever suffered from asthma?

This is obviously less satisfactory than an objective clinical diagnosis, but it is all that is available. We therefore proceed, using this binary response variable to identify 'cases'. This produced $n = 216$ asthma cases and $m = 1076$ children who did not suffer from asthma, and whom we treat as controls. Fig. 3 shows the residence locations of the cases and controls, the network of major roads in the area and the three point sources. Source 1 is a coking works, source 2 a chemical plant and source 3 a waste treatment centre. The data also provide for each child the values of the following binary covariates:

(a) $z_1(\mathbf{x})$ — whether the household included at least one cigarette smoker;
(b) $z_2(\mathbf{x})$ — whether the residence suffered from a dust problem;
(c) $z_3(\mathbf{x})$ — whether the child suffered from hay fever;
(d) $z_4(\mathbf{x})$ — whether the child's school has previously expressed concern about high asthma levels.

We fitted model (4) to the data, incorporating multiplicative risk factors for the three sources, for distance from the relevant main road (also modelled by using the function $g(\ )$), and for each of the four binary covariates modelled as log-linear terms. Table 1 gives the maximized log-likelihood associated with various sub-models. In all cases, we include an adjustment for $z_4(\mathbf{x})$ since this corresponds to a prior stratification of the data, although its effect turned out to be small. Raw asthma rates in the two strata were 0.17 and 0.16 respectively. In the first four rows of Table 1 we consider possible associations with each of the three point sources, ignoring possible covariate effects apart from the prior stratification. There is modest evidence for an association with source 1 ($\chi^2_2 = 5.22$; $p = 0.074$). When we

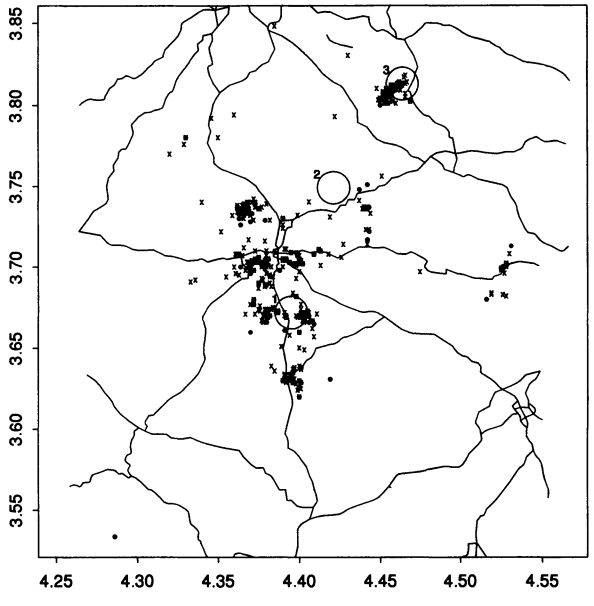

Fig. 3. Asthma in north Derbyshire: ●, cases; ×, controls; ∿∿, road network (the three point sources are labelled 1, 2 and 3; see text for details; the axis labels are Ordnance Survey eastings and northings, divided by 10000)

TABLE 1
Maximized log-likelihoods for various submodels, all including adjustment for the prior stratification of schools into two groups

Model	*Maximized log-likelihood*	*No. of parameters*
Null	−582.95	2
Source 1	−580.34	4
Source 2	−582.95	4
Source 3	−582.87	4
Source 1 + roads	−580.31	6
Source 1 + smoking	−579.71	5
Source 1 + dust	−580.02	5
Source 1 + hay fever	−563.79	5
Hay fever only	−566.24	3

added either or both of sources 2 and 3 to the model with source 1 included, we obtained negligible increases in the log-likelihood, and these values are therefore not shown. In the remainder of Table 1 we consider the effects of covariate adjustment. The associations with distance from a main road, cigarette smoking and dust were all non-significant, but the association with hay fever was highly significant ($\chi_1^2 = 33.10$; $p < 0.001$), as expected. More interestingly, having adjusted for the association with hay fever the evidence for association with source 1 is much the same as before ($\chi_2^2 = 4.90$; $p = 0.086$).

In conclusion, there is no evidence for association with point sources 2 or 3, or with distance from the main road network. There is a possible association with the coking works, which is known to produce particulate emissions of various kinds, and this association may be worth further investigation.

4. DISCUSSION

The conditional formulation developed in this paper greatly simplifies the analysis and interpretation of the model originally presented in Diggle (1990). Our reanalysis of the data from Diggle (1990) illustrates that in circumstances where the old method is applicable, i.e. a single point source and no covariate information, the new method gives similar results. One advantage of the new method is that the asymptotic approximation to the distribution of the likelihood ratio statistic appears to be considerably improved. More importantly, there are no longer any *a priori* constraints on the functional form which can be assumed for the elevation in risk near a point source, and covariate adjustments can be incorporated in a standard log-linear fashion. Indeed, if we chose to approximate the function $g(\ )$ by a function which is log-linear in some prescribed transformation of the distance variable, or by a step function with discontinuities at *known* distances, the model would reduce to a standard logistic regression. However, we believe that our inherently non-linear formulation is preferable, as it is only by this means that we can obtain a model with a genuinely spatial interpretation.

ACKNOWLEDGEMENTS

We thank Paul Elliott and Tony Gatrell for helpful discussions, and Jo Briggs for providing the asthma data.

The work was supported financially by the Economic and Social Research Council, research grant R000232547.

REFERENCES

Bithell, J. F. and Stone, R. A. (1989) On statistical methods for analysing the geographical distribution of cancer cases near nuclear installations. *J. Epidem. Commty Hlth*, **43**, 79–85.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983) *Graphical Methods for Data Analysis*, ch. 6. Belmont: Wadsworth.

Cook-Mozaffari, P. J., Darby, S. C., Doll, R., Forman, D., Hermon, C. and Pike, M. C. (1989) Geographical variation in mortality from leukaemia and other cancers in England and Wales in relation to proximity to nuclear installations, 1969–78. *Br. J. Cancer*, **59**, 476–485.

Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187–220.

Diggle, P. J. (1990) A point process modelling approch to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *J. R. Statist. Soc.* A, **153**, 349–362.

Muirhead, C. and Darby, S. (eds) (1989) Royal Statistical Society meeting on Cancer near nuclear installations. *J. R. Statist. Soc.* A, **152**, 305–384.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.