Biostatistics (2000), **1**, 1, *pp*. 89–105
Printed in Great Britain

Point-source modelling using matched case-control data

PETER J. DIGGLE

Department of Mathematics and Statistics, Lancaster University, Lancaster, UK

SARA E. MORRIS AND JON C. WAKEFIELD

Department of Epidemiology and Public Health, Imperial College School of Medicine, London, UK

SUMMARY

We describe an extension to *matched* case-control studies of the parametric modelling framework developed by Diggle (1990) and Diggle and Rowlingson (1994) to investigate raised risk around putative sources of environmental pollution. We use a conditional likelihood approach for the family of risk functions considered in Diggle and Rowlingson (1994). We show that the likelihood surface that results from these models may be highly irregular, and provide a Bayesian analysis in which we investigate the posterior distribution using Markov chain Monte Carlo. An analysis of one-one matched data that were collected to investigate the relationship between respiratory disease and distance to roads in East London is presented.

Keywords: Bayes factors; Environmental epidemiology; Point process modelling; Point sources of risk; Matched case-control studies.

1. INTRODUCTION

This paper considers the analysis of individually matched case-control data in spatial epidemiology. The case-control study has a long and controversial history in epidemiology (Breslow, 1996) and is the subject of a number of book-length accounts (Breslow and Day, 1980; Schlesselman, 1982). Case-control data alone cannot generally be used to estimate probabilities of disease given exposure but, through the use of logistic regression models, can be used to estimate odds ratios of interest (e.g. Cornfield, 1951; Anderson, 1972; Prentice and Pyke, 1979).

One requirement for a case-control study is that the selection of cases and controls is independent of exposure status. In the most straightforward design, cases and controls are identified and their exposure status, along with the values of relevant confounder variables (e.g. age and sex), are determined, and the subsequent analysis adjusts for their effects by including them in the logistic regression model as additional terms in the linear predictor or by carrying out a stratified Mantel–Haenszel analysis. The design is potentially inefficient, since controls are effectively wasted if they fall within confounder-defined strata that contain few cases. Matched case-control designs seek to avoid this inefficiency by keeping the ratio of cases to controls constant within each sub-group defined by the values of relevant confounders, so-called matching variables. *Frequency* matching corresponds to maintaining constant ratios across broad strata, whilst *individual* matching selects a set of matched controls for each case. When such a design is followed it is vital to acknowledge the sampling scheme by introducing a parameter for each of the matching variables. In the case of individual matching, inference must proceed via conditional likelihood (e.g. Breslow and Day, 1980). In this approach the likelihood results from conditioning on the observed set of exposures within each matched set, without knowledge of which exposure is associated with each of the cases and controls.

© Oxford University Press (2000)

Downloaded from http://biostatistics.oxfordjournals.org/ at University of North Texas on December 31, 2016

Investigation of raised relative risk around putative point or line sources has been the subject of much recent attention, motivated in part by a greater public awareness of the possible effect of environmental factors on health. Area-level data, consisting of disease counts and associated covariate information aggregated over a set of geographic regions, are often routinely available from census information. Methods for the investigation of disease risk in relation to point or line sources of environmental pollution have been proposed by Stone (1988), Bithell and Stone (1989) and Diggle *et al.* (1997). These methods are susceptible to ecological bias: area-level relationships provide an accurate reflection of individual-level relationships only under strict conditions which depend upon both the underlying mathematical risk model and assumptions concerning confounder variables (Piantadosi *et al.*, 1988; Greenland and Morgenstern, 1989; Richardson, 1992).

Individual-level studies avoid the problem of ecological bias provided that exposure status is also measured at the individual level and not, for example, determined from the area in which an individual resides. Associated statistical methods have been proposed by Diggle (1990), Lawson (1993) and Diggle and Rowlingson (1994). Wakefield and Elliott (1999) discuss the statistical framework, and associated issues of interpretation, for both individual and area-level studies. Biggeri *et al.* (1996) describe a matched study investigating the relationship between air pollution and lung cancer in Italy. They used the point process framework of Diggle and Rowlingson (1994) but did not account for the matching, thus producing an inappropriate analysis. Dolk *et al.* (1998) used similar models in an investigation into increased risk of birth defects near landfill sites.

Our focus in this paper is on the analysis of individually matched data when the exposure of primary interest is defined by the spatial location of each individual relative to a point or line source of pollution. In Section 2 we review the point process framework of Diggle and Rowlingson (1994) and describe its extension to matched data. In Section 3 we describe our implementation of the matched case. Section 4 provides a Bayesian approach to this problem and in Section 5 we illustrate our methods via an investigation into the relationship between respiratory illness and distance to roads in East London. Section 6 contains a concluding discussion.

2. MODEL FORMULATION

2.1. The basic model

Suppose that the locations of individuals in the population at risk form an inhomogeneous Poisson process with intensity function $g(x)$, that the probability that an individual at location x becomes a case is $p^*(x)$ and that individuals do or do not become cases independently of each other. It follows that the point processes of cases and non-cases are independent Poisson processes with respective intensity functions $p^*(x)g(x)$ and $\{1 - p^*(x)\}g(x)$. To allow the estimation of relevant summary parameters we must model the spatially varying odds of disease, given by $r^*(x) = p^*(x)/\{1 - p^*(x)\}$. If we now take independent random samples of cases (ideally a complete census) and of non-cases (controls), the point process of sampled individuals again form independent Poisson processes but now with respective intensity functions $ap^*(x)g(x)$ and $b\{1 - p^*(x)\}g(x)$, where a and b are the proportions of cases and non-cases that are sampled. The probability that a *sampled* individual at location x is a case is given by

$$\begin{aligned} p(x) &= [ap^*(x)g(x)]/[ap^*(x)g(x) + b\{1 - p^*(x)\}g(x)] \\ &= ap^*(x)/[ap^*(x) + b\{1 - p^*(x)\}] = \frac{(a/b)r(x)}{1 + (a/b)r(x)} \end{aligned}$$

and the odds of disease amongst sampled individuals are

$$r(x) = p(x)/\{1 - p(x)\} = (a/b)r^*(x). \tag{1}$$

Downloaded from http://biostatistics.oxfordjournals.org/ at University of North Texas on December 31, 2016

Hence, logistic regression modelling of the sampled data, coding cases and controls as 1 and 0, respectively, allows estimation of the true disease odds, $r^*(x)$, up to a generally unknown constant of proportionality. In deriving the key equation (1), we choose to introduce the underlying Poisson process formulation for clarity of exposition and the form of this equation makes it clear that the odds scale is that which allows both case-control and cohort data to be analysed within the same framework. More generally, the validity of (1) rests on the weaker assumption that individual case and non-case outcomes are mutually independent (this may be shown directly following the Bayes theorem argument of Breslow and Day, 1980, p. 205). This would not apply, for example, to a study of the spatial distribution of an infectious disease. However, in general, it does seem appropriate for investigating the spatial distribution of chronic diseases suspected to be of environmental origin.

Specific models follow from (1) by the imposition of a parametric form for the odds function $r(x)$, leading in turn to a regression function $p(x) = r(x)/\{1 + r(x)\}$. For a general discussion, see Thomas (1981).

Suppose now that we have a putative point source at location x_0. Interest is on how the risk surface changes in relation to x_0. The obvious way to proceed would be to use a log-linear model for the odds of disease, $r(x) = \exp(\alpha + \beta d)$ where $d = ||x - x_0||$, leading in turn to a linear logistic regression model for the binary case-control indicators. The equivalent log-linear model as applied to area-level data was utilized by Cook-Mozaffari *et al.* (1989). A possible disadvantage of this model is that as $d \to \infty$, for $\beta < 0$ we have $p^*(x) \to 0$, and not background risk as required. This led Diggle (1990) and Diggle and Rowlingson (1994) to consider a non-linear logistic regression model with

$$r(x) = \rho f(x)$$

in which ρ is a nuisance parameter that is given by $(a/b)\rho^*$ where ρ^* is the background odds of disease, and $f(x)$ is chosen so that as $d \to \infty$, $f(x) \to 1$ and $p(x) \to \rho/(1 + \rho)$, the background risk. In particular, Diggle (1990) used (a reparameterized version of)

$$f(x) = 1 + \alpha \exp\{-(d/\beta)^2\}. \tag{2}$$

One appealing feature of (2) is that the parameters have a natural interpretation; α is the proportional increase in disease odds at the source whilst β measures the rate of decay with increasing distance from the source, in units of distance. A superficially simpler way to deal with non-zero background risk would be to use a non-linear transformation of distance in the log-linear model for $r(x)$, for example $r(x) = \exp\{\alpha + \beta/(c + d)\}$, where the positive constant c is introduced to prevent the converse problem of infinite $r(x)$ at the source. However, unless c is known, the resulting model is intrinsically non-linear and conveys no particular benefit over (2). We note that there are a number of models for atmospheric dispersion (e.g. Fisher, 1979 and Panofsky and Dutton, 1984). These generally contain a large number of parameters, however, and the sparsity of data will often prevent their use. We also note that if modelled or monitored air pollution measurements are available, then an obvious approach is to use these measurements within a log-linear model. Since the extremes of the measurements are finite, no problems occur at the endpoints.

Diggle *et al.* (1997) extended the Diggle (1990) model to allow for a plateau of constant elevated risk up to an unknown distance δ from the point source. This mimics the common epidemiological practice of classifying individual locations simply as 'close' or 'not close' to the putative source. Lawson (1993) introduced angular terms to model directional variation in $r(x)$. The extension to line sources follows immediately from the above if we assume that exposure is well approximated by, say, perpendicular distance to nearest road (though contributions from all 'close' roads could be considered).

Diggle and Rowlingson (1994) assume multiplicative risk factors for the combined effects of s sources, at locations x_{0l}, $l = 1, \ldots, s$, and allow for covariate adjustment of the odds via additional log-linear

Downloaded from http://biostatistics.oxfordjournals.org/ at University of North Texas on December 31, 2016

terms. The resulting model for the odds at location x, in the presence of s sources and q spatially referenced covariates $z_k(x)$, $k = 1, \ldots, q$, takes the form

$$r(x) = \rho \exp\left\{\sum_{k=1}^{q} \phi_k z_k(x)\right\} \prod_{l=1}^{s} f_l(x). \tag{3}$$

For data consisting of n cases at locations x_j, $j = 1, \ldots, n$, and m controls at locations x_j, $j = n+1, \ldots, n+m$, in an unmatched case-control study, and writing $p(x) = r(x)/\{1 + r(x)\}$, the log-likelihood is of the form $\sum_{j=1}^{n} \log p(x_j) + \sum_{j=n+1}^{n+m} \log\{1 - p(x_j)\}$.

2.2. The extension to individually matched data

Suppose initially that we observe J matched case-control pairs in a study region A. We denote the case locations by x_{j0}, $j = 1, \ldots, J$, and the corresponding control locations by x_{j1}, $j = 1, \ldots, J$. The underlying model for a single point source is now that the probability of disease for an individual at location x in stratum j is

$$p_j(x) = \frac{r_j(x)}{1 + r_j(x)} = \frac{\rho_j f(x)}{1 + \rho_j f(x)}.$$

The effect of location is therefore assumed to act proportionally, with the baseline odds ρ_j varying between matched pairs. This assumption is common in epidemiology (e.g. Breslow and Day, 1980) and is consistent both with our aim for a simple description and with the conventional use of log-linear adjustments for unmatched covariates, as in (3). We now proceed to derive the conditional likelihood; to do this we must condition on the set of 'exposures', i.e. locations, within each matched pair. Hence, conditional on there being a matched case-control pair at locations x_{j0} and x_{j1}, and writing f_{ji} for $f(x_{ji})$, the conditional probability that the case is at x_{j0} is

$$p_{j0} = \frac{\rho_j f_{j0}/[(1 + \rho_j f_{j0})(1 + \rho_j f_{j1})]}{\rho_j f_{j0}/[(1 + \rho_j f_{j0})(1 + \rho_j f_{j1})] + \rho_j f_{j1}/[(1 + \rho_j f_{j1})(1 + \rho_j f_{j0})]} = \frac{f_{j0}}{f_{j0} + f_{j1}}. \tag{4}$$

For data with $m > 1$ controls per case, given that the $m + 1$ members of the jth matched case-control set are at locations x_{ji}, $i = 0, 1, \ldots, m$, the conditional probability that the jth case is at location x_{j0} is $p_{j0} = f(x_{j0}; \theta)/\sum_{i=0}^{m} f(x_{ji}; \theta)$. As in the unmatched case, we can extend the expression for p_{j0} to include log-linear adjustments for unmatched individual-level covariates to give

$$p_{j0} = \frac{f(x_{j0}; \theta) \exp\{\sum_{k=1}^{q} z_k(x_{j0})\phi_k\}}{\sum_{i=0}^{m} f(x_{ji}; \theta) \exp\{\sum_{k=1}^{q} z_k(x_{ji})\phi_k\}}, \tag{5}$$

where $z_k(x_{ji})$ is the value of the kth covariate ($k = 1, \ldots, q$) associated with the event at x_{ji} and $\phi = (\phi_1, \ldots, \phi_q)^T$ is the corresponding vector of regression coefficients. In (5), we distinguish between covariate parameters ϕ and additional parameters θ which define the model for spatial variation in disease relative to the point source and, in the present context, are the quantities of primary interest.

Now, for data consisting of J cases at locations x_{j0} and m matched controls per case at locations x_{ji}, $i = 1, \ldots, m$, the conditional log-likelihood is of the form

$$L(\theta, \phi) = \sum_{j=1}^{J} \log p_{j0}. \tag{6}$$

Downloaded from http://biostatistics.oxfordjournals.org/ at University of North Texas on December 31, 2016

3. Likelihood approach

A basic issue for inference from case-control data is whether inference should be based on a conditional or an unconditional likelihood. The conventional wisdom (e.g. Clayton and Hills, 1993) is that if the number of strata is small in comparison with the number of cases, then the unconditional likelihood may be used, although from a classical perspective (e.g. Cox and Hinckley, 1974) the conditional likelihood is the 'correct' approach. In other cases, including individually matched studies, the conditional likelihood should be used, because in this case the number of nuisance parameters ρ_j, eliminated by the conditioning, tends to infinity with the number of matched sets. We shall use the conditional log-likelihood (6).

We consider the matched pairs case for presentational simplicity. All of the following applies equally to studies with $m > 1$ controls per case, or with variable numbers of controls per case.

Maximum likelihood estimates (MLEs) $\hat{\theta}$ and $\hat{\phi}$ can be found using a general numerical optimization routine. The null log-likelihood, denoted $L_0(\phi)$, is obtained by setting $f(\cdot) = 1$ in (5) whereupon (6) becomes

$$L_0(\phi) = -\sum_{j=1}^{J} \log\left[1 + \exp\left\{\sum_{k=1}^{q}(z_k(x_{j1}) - z_k(x_{j0}))\phi_k\right\}\right]. \tag{7}$$

This corresponds to the unconditional log-likelihood for a logistic regression model without an intercept, where the sampling unit is the matched pair and the regression variables are the differences in the case and control covariates (Breslow and Day, 1980, p. 251). Thus, any routine for fitting generalized linear models can be used to estimate the null values of the covariate parameters. This can also be useful in suggesting appropriate starting values for the parameters ϕ when maximizing the full conditional log-likelihood. The expression (7) also shows, as is well known, that matched covariates cannot be included in the model for the conditional likelihood. A variable used loosely in matching, say ± 10 years of age, *can* be included although the size of the coefficient is not interpretable as a main effect.

In testing a null model with p_0 parameters against an alternative with $p_0 + p_1$ parameters, the generalized likelihood ratio statistic, $D = 2\{L(\hat{\theta}, \hat{\phi}) - L_0(\hat{\phi}_0)\}$, has an asymptotic chi-squared distribution on p_1 degrees of freedom under the null hypothesis, provided that the usual regularity conditions are satisfied (Breslow and Day, 1980, p. 253). Under the same regularity conditions, approximate confidence regions for model parameters can be derived from the estimated Hessian matrix and the asymptotic multivariate normality of $\hat{\theta}$.

For the application to respiratory disease in East London, we shall use model (2). Unfortunately, this model has an irregularity at the null hypothesis of interest, since $f(x) = 1$ corresponds to $\alpha = 0$ with β indeterminate. Away from the null, our experience with this model has been that for the sample sizes we are considering in the current application, the log-likelihood for the parameters of interest, $\theta = (\alpha, \beta)$, is far from quadratic even when we reparameterize θ to the real line using logs. Standard likelihood-based asymptotics are therefore unreliable.

There is a large body of work on likelihood inference for non-regular problems (e.g. Self and Liang, 1987). Improved asymptotic approximations have been derived for a range of models but it is not clear at what sample size these results are appropriate in our context and so instead we choose to use simulation.

To test the null hypothesis that $f(x) = 1$ we use a Monte Carlo test, simulating data under the null and ranking the observed value of the likelihood ratio statistic D amongst the simulated values; if the observed value of D ranks kth largest amongst $N - 1$ simulated values, the p value of the test is k/N (Barnard, 1963). In the simple case where there are no covariates, the null hypothesis translates to $p_{j0} = 0.5$ for all j, i.e. a simple null hypothesis, and the Monte Carlo test is exact.

When there are unmatched covariates associated with each individual, the probabilities p_{j0} are no longer constant under the null hypothesis but depend on the unknown quantities ϕ, i.e. the null hypothesis is composite, and an exact Monte Carlo test is not available. For an approximate test in this situation,

Downloaded from http://biostatistics.oxfordjournals.org/ at University of North Texas on December 31, 2016

we estimate the vector of covariate parameters under the null model, and proceed as follows. Let $\hat{\phi}_0 = (\hat{\phi}_{10}, \ldots, \hat{\phi}_{q0})$ denote the vector of estimated values of the parameters under the null model, and write

$$\tilde{p}_{j0} = \left[1 + \exp\left\{\sum_{k=1}^{q}(z_k(x_{j1}) - z_k(x_{j0k}))\hat{\phi}_{k0}\right\}\right]^{-1}.$$

We then construct a simulated data set under the null by randomly selecting, for each pair, the location of the simulated case. The original case is chosen with probability $\tilde{p}_{j0}$ and the original control with probability $1 - \tilde{p}_{j0}$. When the likelihood terms (5) are evaluated, the covariates are those associated with the simulated case and control. In summary, the x and z values are tied together in the simulation. For the test, we again compare the observed value of D with a sample of values obtained from $N - 1$ data sets simulated in this way.

One way to construct a confidence region for θ is to use a profile log-likelihood. For each value of θ we maximize the log-likelihood with respect to ϕ, to give estimates $\hat{\phi}_\theta$, and define the confidence region for θ as the set of all values such that $2\{L(\hat{\theta}, \hat{\phi}) - L(\theta, \hat{\phi}_\theta)\} < c$. The conventional choice for c would be the appropriate critical value from the chi-squared distribution on two degrees of freedom. However, as the asymptotic justification for this choice is questionable, we also consider an alternative, Monte Carlo approach. We simulate data sets from the fitted model with parameter values $(\hat{\theta}, \hat{\phi})$ in (5). If we call these simulation probabilities $\tilde{p}_{j0}$, then we select a location from each pair (x_{j0}, x_{j1}) to be the case location with respective probabilities $\tilde{p}_{j0}$ and $1 - \tilde{p}_{j0}$. For each simulated data set, we re-estimate the model parameters (θ, ϕ) and use the resulting empirical distributions of parameter estimates as approximations to the corresponding true sampling distributions. In particular, we use quantiles of the empirical sampling distribution to define an interval estimate for any quantity of interest. Note that these interval estimates are not, strictly, confidence intervals, but they do provide sensible reference intervals to indicate the precision with which quantities of interest are estimated.

4. BAYESIAN APPROACH

A Bayesian approach provides an alternative to the likelihood-based inference of the previous section. Wakefield and Morris (1998) describe how a Bayesian area-level analysis for the models suggested by Diggle *et al*. (1997) may be implemented.

From a likelihood perspective the conditional approach is followed because of the poor properties of MLEs when there is a large number of nuisance parameters (relative to the number of data points). From a strictly Bayesian point of view, the 'correct' approach would be to use the unconditional likelihood and integrate out the nuisance parameters $\rho = (\rho_1, \ldots, \rho_J)^T$. However, the difficulties encountered by an (unconditional) likelihood approach would suggest that great care must be taken when specifying a prior distribution for ρ. In an early paper, Altham (1969) derived Fisher's exact test (which is based on the conditional likelihood) for the 2×2 table case, from a Bayesian perspective by placing a Dirichlet prior on the cell probabilities and then integrating out the nuisance parameters. Finding a prior distribution and obtaining the conditional likelihood as the resultant marginal distribution in the general case has remained elusive.

In practice, we would have considerable difficulty specifying a meaningful prior for ρ. We note that the choice of an exchangeable prior for the elements of ρ is not sensible since epidemiological knowledge will in most cases give, at the least, a ranking on the ρ_j (for example, j may index age). We therefore proceed with the conditional likelihood as our starting point. A key question to ask therefore, is to what extent conditioning to eliminate the nuisance parameters ρ also sacrifices information about the parameters of interest θ. Essentially we are acting 'as if' the likelihood factors into two parts and we are then concentrating on that part containing θ (the conditional likelihood). We note that the use of the conditional

Downloaded from http://biostatistics.oxfordjournals.org/ at University of North Texas on December 31, 2016

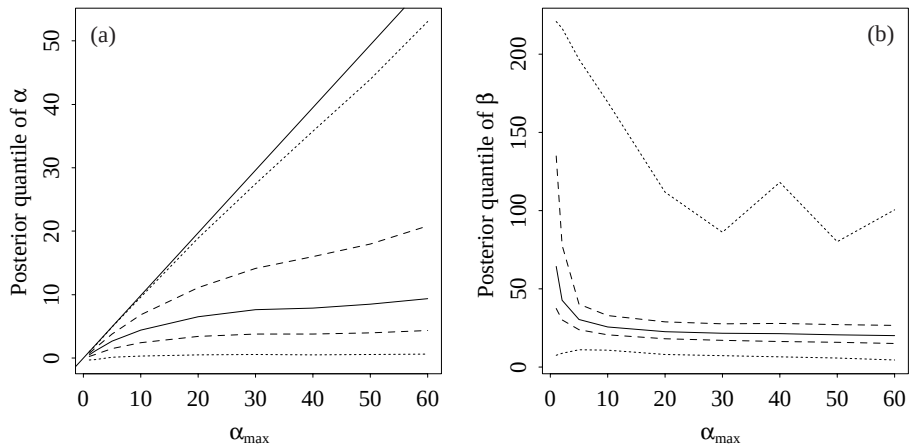

Fig. 1. Posterior 2.5%, 25%, 50%, 75% and 97.5% quantiles of α and β for different priors on α, indexed by $\alpha_{\max}$. The solid diagonal line on (a) marks $y = x$. Prior for β uses $\beta_{\max} = 230$ in all cases.

likelihood is consistent with independent priors for ρ and (θ, ϕ). This approach is consistent with other authors who have used a conditional likelihood in related contexts (e.g, Liao, 1999), and also the use of the partial likelihood rather than the full likelihood in survival analysis (Raftery *et al.* 1996).

We consider the following specification for the prior distributions of (θ, ϕ) in model (6) in conjunction with distance model (2). For the parameters of interest $\theta = (\alpha, \beta)$, we specify a uniform prior so as to focus on the likelihood.

Prior for ϕ. We use independent priors for the covariate coefficients, $\phi \sim \mathrm{N}(M_\phi, V_\phi)$. The mean parameter M_ϕ and diagonal variance–covariance matrix V_ϕ may be chosen on substantive grounds or, as in Section 5, a large value of V_ϕ can be used to provide an effectively flat prior for ϕ.

Prior for α. We use a uniform prior over the range $(-1, \alpha_{\max})$, the upper limit representing a plausible upper bound to the increase in odds at source relative to the background odds.

Prior for β. We use a uniform prior over the range $(0, \beta_{\max})$. The value of $\beta_{\max}$ was chosen to reflect the range of distances over which emissions are likely to act.

Figure 1 shows 20 simulations from the prior for $\theta = (\alpha, \beta)$, summarized through the function $f(d, \theta)$, with $\alpha_{\max} = 60$ and $\beta_{\max} = 0.2 \times d_{\max}$ (a) and $\beta_{\max} = 0.5 \times d_{\max}$ (b). These values were chosen to allow a reasonable range for the function $f(\cdot)$.

In this context, because the simple risk/distance model we consider is not linked to any specific biological mechanism, we would not want to restrict ourselves to a single prior distribution. Let

$$p(\theta, \phi | y) \propto l(\theta, \phi) \times \pi(\theta, \phi)$$

denote the unnormalized posterior, where l and π denote the (conditional) likelihood and prior, respectively. In general, an analytic approximation, for example the Laplace method (e.g. Tierney and Kadane, 1986), or a simulation-based approach may be used when the posterior distribution is, as here, not analytically tractable. We follow the latter course and use Markov chain Monte Carlo (MCMC), e.g. Smith and Roberts (1993) and Gilks *et al.* (1996). We do not go into the details of this approach but note that the Metropolis–Hastings algorithm (Smith and Roberts, 1993), implemented component-wise, may be

Downloaded from http://biostatistics.oxfordjournals.org/ at University of North Texas on December 31, 2016

used via the BUGS software (Spiegelhalter *et al.*, 1996). Given (dependent) samples from the Markov chain the posterior distribution may be summarized either via its moments or, more appropriately for the non-regular model considered here, via one- and two-dimensional marginal distributions.

As a Bayesian counterpart to a test of the null hypothesis that $f(x) = 1$, we consider Bayes factors (Kass and Raftery, 1995) which are defined as the ratio of posterior to prior odds. Specifically, the Bayes factor for comparing models M_0 (the null model) and M_1 (model(2)) is given by

$$B_{10} = \frac{\int p(y|\theta, \phi, M_1)\pi(\theta, \phi|M_1)d\theta d\phi}{\int p(y|\phi, M_0)\pi(\phi|M_0)d\phi} = \frac{p(y|M_1)}{p(y|M_0)}. \tag{8}$$

Kass and Raftery (1995) suggest calculating $2\log_e(B_{10})$ as a Bayesian analogue of a log-likelihood ratio statistic or deviance. Values greater than 2 indicate increasing evidence against M_0: between 2 and 6 is 'positive' evidence, 6 to 10 is 'strong' and over 10 is 'very strong' evidence against M_0.

The calculation of the numerator and denominator of (8) is not straightforward using MCMC and a number of alternatives have been proposed (DiCiccio *et al.*, 1997). For this low-dimensional problem we use the importance sampling estimator $\frac{1}{T}\sum_{t=1}^{T} l(\theta^t, \phi^t)\pi(\theta^t, \phi^t)/g(\theta^t, \phi^t)$, where (θ^t, ϕ^t), $t = 1, \ldots, T$, are sampled from the proposal distribution $g(\theta, \phi)$. For the proposal density $g(\cdot)$ we have used the prior for θ and standard normal densities (scaled to provide an efficient algorithm) for the covariate parameters ϕ.

5. APPLICATION: RESPIRATORY DISEASE IN TOWER HAMLETS

We illustrate the methods described in Sections 2–4 on two sets of data described more fully in Morris *et al.* (2000). The aim of the study was to examine the relationship between proximity of residence to main roads and increased risk of respiratory disease. The first data set consists of 125 asthma cases, and the second 124 chronic obstructive airways disease (COAD) cases. Cases were clear-cut diagnoses attending the casualty department at the Royal London Hospitals in the years 1991–1992. Controls were selected from non-chest related admissions. Each case was individually matched to a single control using sex, race and consultant team. Controls were also matched, where possible, on age to within ±5 years. When this was not possible, age ranges were relaxed incrementally until a control matching on the other variables was found. Approximately 80% of the pairs are matched to within ±5 years. Since the matching on age is not exact we include it as a covariate in the model.

A small-area measure of deprivation, the so-called Carstairs' index (Carstairs and Morris, 1991), was used to account for socio-economic status. The index is derived from four census variables concerning unemployment, overcrowding, car-ownership and social class. The index has been shown to be a powerful predictor of health outcomes (e.g. Jolley *et al.*, 1992) and is scaled to have zero mean over England and Wales with larger values indicating increased deprivation. For the asthma study, the empirical quartiles of the Carstairs index were 4.4, 6.5 and 10.8, illustrating that the study area is relatively deprived. Each individual's address was linked to the corresponding census enumeration district (ED), using the Geographical Information System (GIS) Arc-Info, in order to determine a deprivation score. Bartholomew's road map data and the GIS were used to calculate the shortest perpendicular distance between the residence of each event and the nearest motorway, primary road, A-road or B-road.

In the following sub-sections, we report those parts of our analysis that are based on the distance-odds model (2). We also investigated more general models, including the plateau model of Diggle *et al.* (1997) and a generalized linear model with reciprocal distance in the linear predictor, but omit the details as in neither case did the other models provide a demonstrably better fit to the data than did model (2). As a first exploratory step we used a generalized additive model (GAM, Hastie and Tibshirani, 1990) within a conditional likelihood framework. Figure 2 illustrates the spline fit to the distance-odds relationship

Downloaded from http://biostatistics.oxfordjournals.org/ at University of North Texas on December 31, 2016

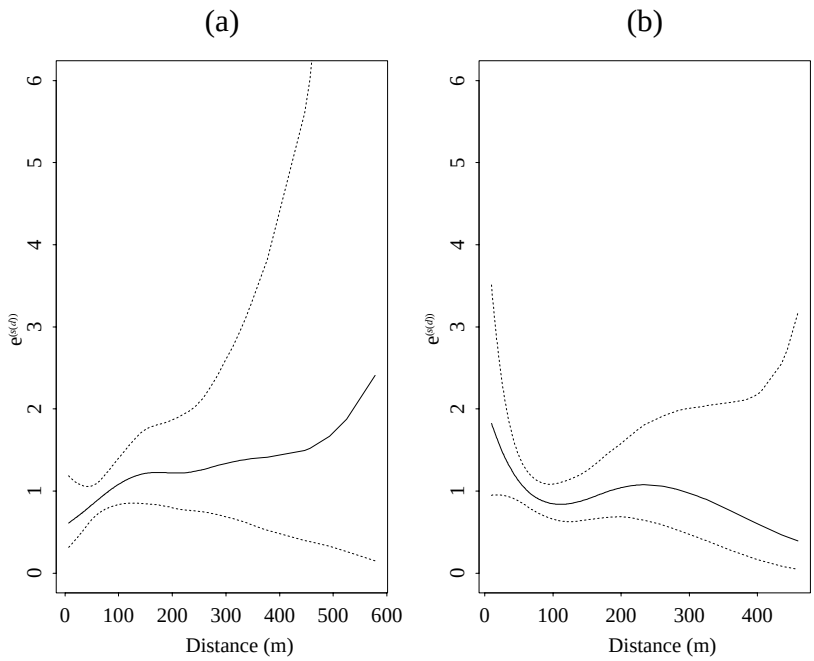

Fig. 2. Fitted B-splines, $s(d)$, for (a) asthma and (b) COAD data. Dashed curves denote approximate 95% confidence intervals.

for asthma (a) and for COAD (b). We see that there appears to be a monotonic relationship for COAD, with risk increasing with proximity to roads, but for asthma the relationship seems to be in the opposite direction. The logistic model used to construct this plot (and those in the following two sections) makes adjustment for age and deprivation (through the Carstairs index). The GAM model provides a useful initial look at the data, but formal inference concerning the relationship between distance and disease odds is not so straightforward.

5.1. Asthma data

We first use the observed distances of the asthma data to investigate the chi-squared approximation to the null sampling distribution of D in the test of the hypothesis that $f(x) = 1$ within model (2). The estimated true significance levels of nominal 10%, 5% and 1% tests, based on 1000 simulations, were 11.3%, 4.8% and 1.1%, respectively. Thus for these data, notwithstanding the theoretical irregularity, the standard asymptotic approximation appears to be adequate. We use the chi-squared approximation in all subsequent tests against the null.

Figure 3 shows the relationship between disease status and age, deprivation index and distance to nearest main road. The case-control labels are plotted as vertical ticks and have been 'jittered' in the x-axis, where necessary, to break ties. The solid lines are loess smoothers (Cleveland, 1979), using a span of 2/3. Figure 3(a) shows that there appears to be no remaining relationship between disease and age. Figure 3(b) indicates that, if anything, the probability of asthma is greater in less deprived EDs but Figure 3(d) shows very little relationship between distance and deprivation.

The plot of disease status against distance (Figure 3(c)) suggests a relationship between distance and

Downloaded from http://biostatistics.oxfordjournals.org/ at University of North Texas on December 31, 2016

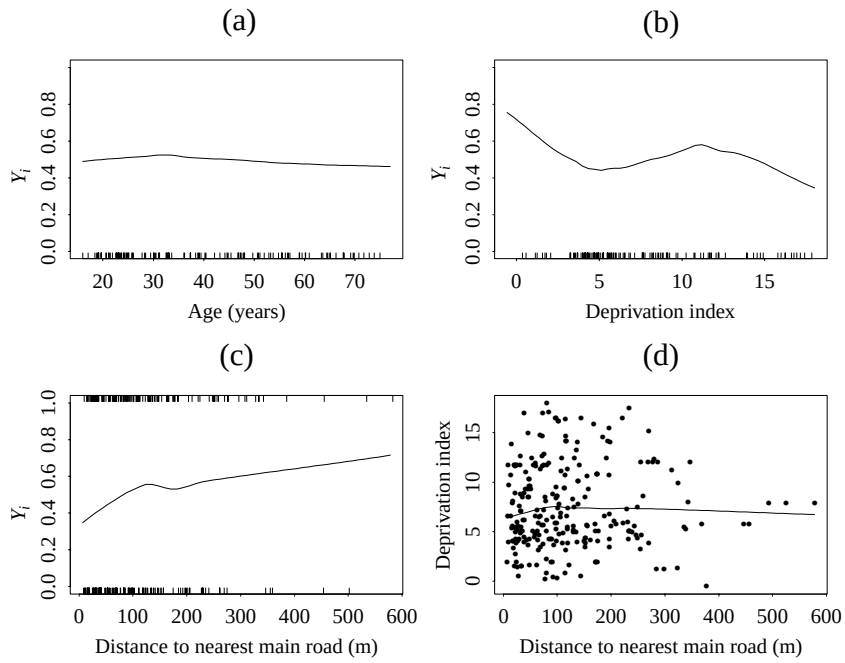

Fig. 3. Plots (a), (b) and (c), respectively, show relationships between asthma status ($Y_i = 1$ denotes asthma case) and age, deprivation and distance. Plot (d) shows the relationship between deprivation and distance. Solid curves denote loess smoothers.

asthma in the opposite direction to that hypothesized (in line with Figure 2). However, there is no evidence that model (2) fits better than the null model ($D = 3.44$, df = 2, $p = 0.18$). Note that the null model in this case includes log-linear terms for age and deprivation.

Turning to the Bayesian analysis, we used the priors described in Section 4 with $\alpha_{\max} = 60$ and $\beta_{\max} = 230$ (the maximum distance here is 459.6 m). This prior is displayed in the lower panel in Figure 1. For both data examples we used essentially flat priors for the two covariate parameters ϕ, namely independent normal with zero mean and variance 10^6. We obtained samples from the posterior using MCMC, ran the chain with a 500 iteration 'burn in', and judged convergence by visual examination of the chains for each of the parameters. Posterior quantiles (2.5%, 50%, 97.5%) for the parameters were based on a further 9500 iterations and were $(-0.806, -0.122, 55.6)$ for α and $(0.335, 47.34, 217.7)$ for β. The posterior distributions for α and β are highly skewed. Furthermore, the upper quantile in each case indicates that the likelihood has a heavy tail, and consequently the upper limit of the prior is very influential. The Bayes factor comparing model (2) against the null model is -5.67 indicating positive evidence *in favour of* the null model.

In summary, for these data we detect no association between asthma and distance of residence to main roads. However, we make the obvious point that proximity to main roads may be a very poor measure of personal exposure to traffic-related pollution. For example, no account is taken of individual daily changes in location.

Downloaded from http://biostatistics.oxfordjournals.org/ at University of North Texas on December 31, 2016

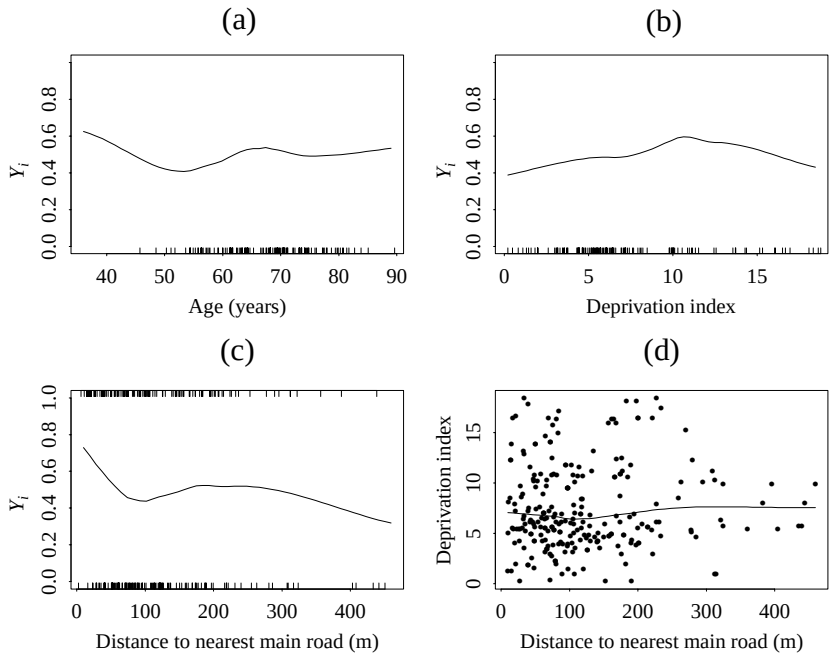

Fig. 4. Plots (a), (b) and (c), respectively, show relationships between COAD status ($Y_i = 1$ denotes COAD case) and age, deprivation and distance. Plot (d) shows the relationship between deprivation and distance. Solid curves denote loess smoothers.

Table 1. *MLEs and approximate* 95% *confidence intervals for model (2), COAD data*

Parameter	α		β		ϕ_{age}		ϕ_{dep}	
MLE	4.634		24.664		0.012		0.028	
Hessian-based	0.261	24.171	12.753	47.700	−0.050	0.075	−0.033	0.089
Profile-based	0.547	35.239	11.137	55.594	−0.050	0.079	−0.033	0.090
Simulation-based	1.033	71.087	11.107	123.75	−0.056	0.088	−0.035	0.097

5.2. COAD data

Figure 4 shows the relationships between disease status, age, deprivation and distance to nearest main road. In particular, Figure 4(c) appears to indicate that a model such as (2), in which the log odds of risk decrease with increasing distance, may be appropriate (again consistent with Figure 2). The null model, as before, includes terms for age and deprivation. Fitting untransformed distance as an additional predictor within the conditional logistic model gives no evidence that the distance term is a significant improvement over the null model ($D = 1.0802$, df = 1, $p = 0.299$). In contrast, there is evidence that model (2) is a better fit than the null model ($D = 7.472$, df = 2, $p = 0.0239$).

The log-likelihood surface is highly non-quadratic in α and β. Consequently, Hessian-based estimates of standard errors and confidence intervals are unreliable for these parameters, although they appear to be adequate for the covariate regression parameters ϕ. Working with $\log \alpha$ and $\log \beta$ gave a surface that

Downloaded from http://biostatistics.oxfordjournals.org/ at University of North Texas on December 31, 2016

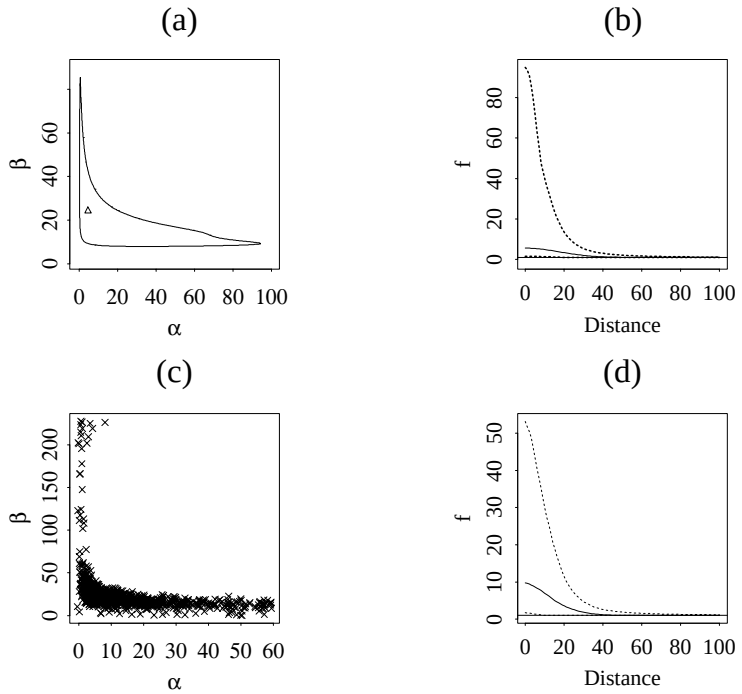

Fig. 5. Inference for COAD data. (a) 95% profile likelihood confidence region for (α, β). Position of the MLE denoted by $\triangle$. (b) The corresponding interval of $f(\alpha, \beta)$. The dashed curve represents the interval, the solid curve the MLE and the faint horizontal line represents $f = 1$. (c) Bivariate posterior sample, based on the prior $\alpha \sim U(-1, 60)$, $\beta \sim U(0, 230)$. (d) The corresponding 95% credible interval on $f(\alpha, \beta)$. The dashed curve represents the interval, the solid line the median and the faint horizontal line represents $f = 1$.

was closer to quadratic. The resulting interval estimates, along with those from profile log-likelihoods and those obtained by simulating 10 000 sets of data from the fitted model (via the likelihood procedure outlined in Section 3) are given in Table 1. We note that the Hessian and profile approaches underestimate the uncertainty for large values of α and β. Since α and β are likely to be highly correlated, we have constructed a two-dimensional profile confidence region, shown in Figure 5(a). This plot clearly shows the irregular behaviour of the likelihood. Figure 5(b) shows the corresponding interval for the estimated function $f(\cdot)$ evaluated over a fine grid of points contained within the confidence region. Note that computation of the likelihood simulation-based intervals can be both intricate and time consuming.

As we might expect from the likelihood analysis, a Bayesian analysis is highly sensitive to the choice of prior. Figure 5(c) shows a sample of 1000 points from the bivariate posterior $p(\alpha, \beta|y)$, based on the priors $\alpha \sim U(-1, 60)$, $\beta \sim U(0, 230)$. The support of the posterior distribution follows quite closely the shape of the profile likelihood confidence region. Given a sample from the posterior for (α, β), it is straightforward to make inference for other properties of the model by plugging the sampled values of (α, β) into the expression for any quantity of interest. Figure 5(d) gives quantiles of the posterior for $f(x, \theta) = 1 + \alpha \exp\{-(d/\beta)^2\}$ as a function of d, evaluated pointwise. Figure 6 shows the marginal posterior distributions for α and β based on two different priors. The posterior for α in particular has a very heavy tail. The posterior distributions under both priors are fairly similar. Figure 7 illustrates the effect of the prior for α on the posterior quantiles of α and β. The posterior for β is relatively robust over

Downloaded from http://biostatistics.oxfordjournals.org/ at University of North Texas on December 31, 2016

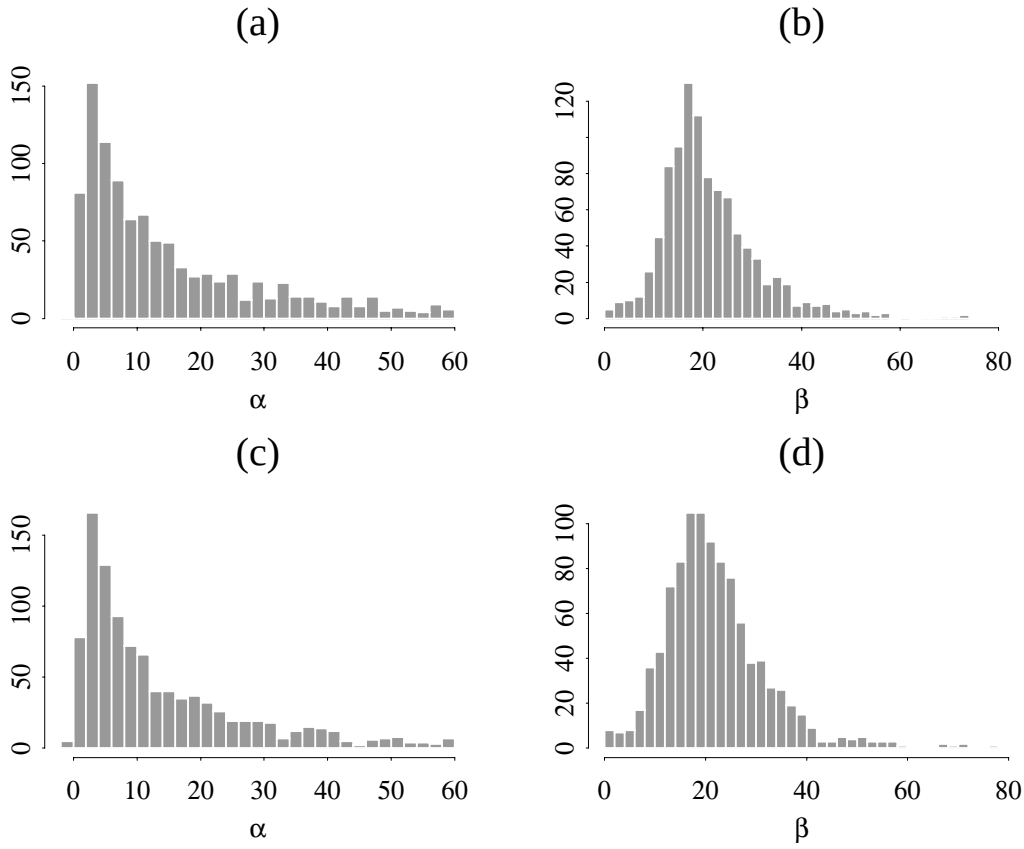

Fig. 6. Histogram representations of marginal posterior distributions $p(\alpha|y)$ and $p(\beta|y)$. Panels (a) and (b) correspond to the prior $\alpha \sim U(-1, 60)$, $\beta \sim U(0, 230)$ and (c) and (d) to the prior $\alpha \sim U(-1, 60)$, $\beta \sim U(0, 92)$. The x-axis of panel (b) has been truncated at 80 for comparison with (d), 3% of the points lay beyond the value 80.

Table 2. *The quantity* $2\log_e(B_{10})$, *as a function of* $\alpha_{\max}$ *and* $\beta_{\max}$ *where* B_{10} *denotes the Bayes factor in favour of model (2) versus the null model. All estimates based on sampling* $T = 10^6$ *points from the proposal densities*

	$\beta_{\max}$				
$\alpha_{\max}$	10	30	70	150	230
10	0.39	4.94	4.02	2.59	1.48
30	0.66	4.17	2.84	1.47	0.76
60	0.82	3.30	1.85	0.56	−0.36

Downloaded from http://biostatistics.oxfordjournals.org/ at University of North Texas on December 31, 2016

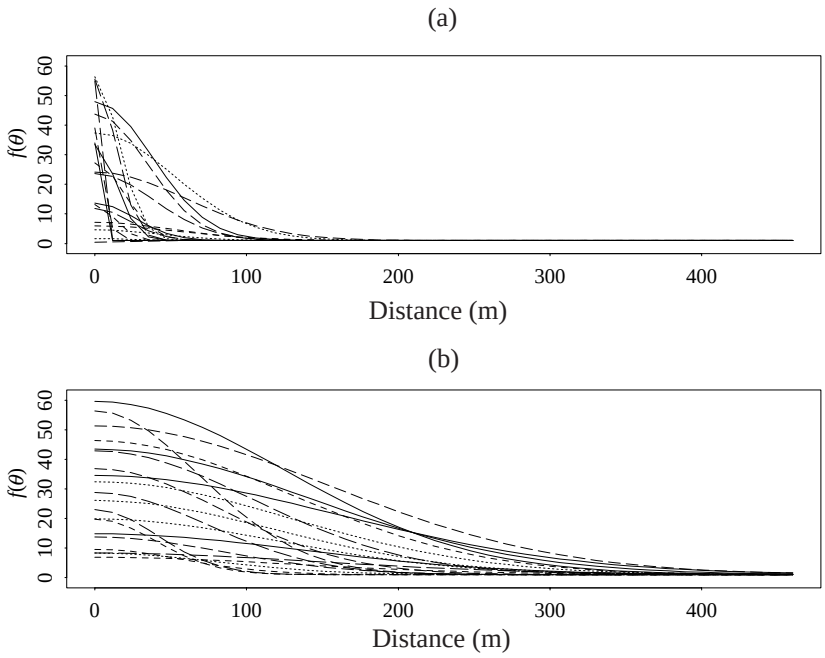

Fig. 7. Curves of $f(\theta)$ resulting from 20 simulations from the priors for α and β. Both priors for α use $\alpha_{\max} = 60$. Two different priors are used for β, namely $\beta_{\max} = 0.2 \times d_{\max}$ (a) and $\beta_{\max} = 0.5 \times d_{\max}$ (b).

a range of values of $\alpha_{\max}$, whereas that for α shows far greater sensitivity to this choice.

Table 2 shows, as expected, that the Bayes factor for comparing the null model with model (2) is sensitive to the choice of prior. We note that for large values of $\alpha_{\max}$ and $\beta_{\max}$ the Bayes factor favours the null model. This is consistent with Lindley's paradox (Lindley, 1957) since the Bayes factor averages the likelihood over the prior and, under model (2), with large values of $\alpha_{\max}$ and $\beta_{\max}$, much of the prior density is in areas in which the likelihood is close to zero. The sensitivity of Bayes factors to the prior is well documented. As possible solutions to this problem, several authors have proposed adjusted Bayes factors (e.g. O'Hagan, 1995; Berger and Pericchi, 1996) or 'reference' priors (Kass and Wasserman, 1995). Unfortunately, for the non-linear model considered here, these alternatives are not straightforward to apply. In this case we prefer to base inference on estimation rather than testing which, as we have seen, is more stable under different prior specifications. The sensitivity to the prior here indicates that there is limited information in the data and a more comprehensive conclusion could only be reached with additional data.

We conclude that for the COAD data there is reasonable evidence of elevated risk close to main roads, but the functional form of the variation in risk with distance is estimated very imprecisely. The possibility that this elevation in risk is due to residual confounding cannot, of course, be eliminated. The reader is referred to Morris *et al*. (2000) for further discussion of the substantive conclusions, including references to related studies.

Downloaded from http://biostatistics.oxfordjournals.org/ at University of North Texas on December 31, 2016

6. Discussion

In this paper we have demonstrated how matched case-control data, collected to investigate how disease risk varies in relation to a point or line putative source, may be modelled using a simple but reasonable distance/risk model. We have considered both likelihood-based and Bayesian approaches to inference for this problem. In the example, we demonstrated that for the particular model we use, the likelihood function was highly irregular and care must be taken when summaries are reported.

We have seen that both the likelihood ratio test and the Bayes factor can be difficult to interpret for the model considered here. Our preferred approach, which we have emphasized and which avoids this difficulty, is simply to fit the model and examine interval estimates of quantities of interest. In the Bayesian setting, this is a natural by-product of an MCMC implementation. In the likelihood-based setting it involves a similar, albeit more ad hoc, device of Monte Carlo sampling from the fitted model to estimate the sampling distribution of any quantity of interest.

In spatial epidemiological studies, the possibility of residual spatial variation always exists, perhaps because of an infectious agent or spatially varying unmeasured risk factors. Our methods do not allow for such a phenomenon, but one possible extension, in the spirit of Diggle *et al.* (1998), is to assume

$$\text{logit } p_j(x) = \log \rho_j + \log f(x;\theta) + \sum_{k=1}^{q} z_k(x)\phi_k + S(x),$$

where the residual spatial variability term $S(x)$ is modelled as a Gaussian random field. To implement this model the conditional likelihood may again be considered, and a Bayesian approach via MCMC is the obvious choice for inference.

Acknowledgements

We thank Barbara J. Boucher of the Royal London Hospital for the data and Paul Elliott for helpful comments. We gratefully acknowledge the support of the ESRC, via grant R000237819. The 1991 Census data was provided with the support of the ESRC. Digital map data copyright Bartholomew. This work was supported, in part, by an equipment grant from the Wellcome Trust (0455051/Z/95/Z) and is based on data supplied by Ordnance Survey. These data remain the property of the Crown and copyright of HMSO and Royal Mail. B-splines were fitted using T. Therneau's `survival5` routines from the Mayo Medical Center.

References

Altham, P. M. E. (1969). Exact Bayesian analysis of a 2 × 2 contingency table, and Fisher's 'Exact' significance test. *Journal of the Royal Statistical Society* B **31**, 261–269.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.

Barnard, G. A. (1963). Contribution to the Discussion of Professor Bartlett's paper. *Journal of the Royal Statistical Society* B **25**, 294.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109–122.

Biggeri, A., Barbone, F., Lagazio, C., Bovenzi, M. and Stanta, G. (1996). Air pollution and lung cancer in Trieste, Italy: spatial analysis of risk as a function of distance from source. *Environmental Health Perspectives* **104**, 750–754.

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

Downloaded from http://biostatistics.oxfordjournals.org/ at University of North Texas on December 31, 2016

BRESLOW, N. E. (1996). Statistics in epidemiology: the case-control study. In *Advances in Biometry*, eds Armitage, P. and David, H. Wiley.

BRESLOW, N. E. AND DAY, N. E. (1980). *Statistical Methods in Cancer Research. Vol I—The Analysis of Case-Control Studies*. IARC Scientific Publications No. 32.

CARSTAIRS, V. AND MORRIS, R. (1991). *Deprivation and Health in Scotland*. Aberdeen: Aberdeen University Press.

CLAYTON, D. AND HILLS, M. (1993). *Statistical Models in Epidemiology*. Oxford: Oxford University Press.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

COOK-MOZAFFARI, P., DARBY, S., 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. *British Journal of Cancer* **59**, 476–485.

CORNFIELD, J. (1951). A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute* **11**, 1269–1275.

COX, D. R. AND HINCKLEY, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

DICICCIO, T. J., KASS, R. E., RAFTERY, A. E. AND WASSERMAN, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association* **92**, 903–915.

DIGGLE, P. J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a pre-specified point. *Journal of the Royal Statistical Society* A **153**, 349–362.

DIGGLE, P. J., ELLIOTT, P., MORRIS, S. E. AND SHADDICK, G. (1997). Regression modelling of disease risk in relation to point sources. *Journal of the Royal Statistical Society* A **160**, 491–505.

DIGGLE, P. J., MOYEED, R. A. AND TAWN, J. A. (1998). Model-based geostatistics (with discussion). *Applied Statistics* **47**, 299–350.

DIGGLE, P. J. AND ROWLINGSON, B. S. (1994). A conditional approach to point process modelling of raised incidence. *Journal of the Royal Statistical Society* A **157**, 433–440.

DOLK, H., VRIJHEID, M., ARMSTRONG, B. G. AND THE EUROHAZCON COLLABORATIVE GROUP. (1998). Congenital anomalies near hazardous waste landfill sites in Europe. In *Disease Mapping and Risk Assessment for Public Health Decision Making*, ed. Lawson, A. Wiley (to appear).

FISHER, B. E. A. (1979). Use of the diffusion approximation to model atmospheric dispersion over short and long distances. In *Mathematical Modelling of Turbulent Diffusion in the Environment*, ed. Harris, C. J., pp. 237–275. London: Academic Press.

GILKS, W. R., RICHARDSON, S. AND SPIEGELHALTER, D. J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

GREENLAND, S. AND MORGENSTERN, H. (1989). Ecological bias, confounding and effect modification. *International Journal of Epidemiology* **18**, 269–274.

HASTIE, T. AND TIBSHIRANI, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

JOLLEY, D., JARMAN, B. AND ELLIOTT, P. (1992). Socio-economic confounding. In *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, eds Elliott, P., Cuzick, J., English, D. and Stern, R., pp. 115–124. Oxford: Oxford University Press.

KASS, R. E. AND RAFTERY, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

KASS, R. E. AND WASSERMAN, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* **90**, 928–934.

Downloaded from http://biostatistics.oxfordjournals.org/ at University of North Texas on December 31, 2016

LAWSON, A. B. (1993). On the analysis of mortality events associated with a pre-specified fixed point. *Journal of the Royal Statistical Society* A **156**, 363–377.

LIAO, J. G. (1999). A hierarchical Bayesian model for combining multiple 2×2 tables using conditional likelihoods. *Biometrics* **55**, 268–272.

LINDLEY, D. V. (1957). A statistical paradox. *Biometrika*, **44**, 187–192.

MORRIS, S. E., SALE, R. C., WAKEFIELD, J. C., FALCONER, S., ELLIOTT, P. AND BOUCHER, B. J. (2000). Hospital admissions for asthma and chronic obstructive airways disease in East London Hospitals and proximity of residence to major roads. *Journal of Epidemiology and Community Health*, to appear.

O'HAGAN, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society* B **57**, 99–138.

PANOFSKY, M. A. AND DUTTON, J. A. (1984). *Atmospheric Turbulence.* New York: Wiley.

PIANTADOSI, S., BYAR, D. P. AND GREEN, S. B. (1988). The ecological fallacy. *American Journal of Epidemiology* **127**, 893–904.

PRENTICE, R. L. AND PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411.

RAFTERY, A. E., MADIGAN, D. AND VOLINSKY, C. T. (1996). Accounting for model uncertainty in survival analysis improves predictive performance (with discussion). In *Bayesian Statistics 5*, eds Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., pp. 323–349. Oxford: Oxford University Press.

RICHARDSON, S. (1992). Statistical methods for geographical correlation studies. In *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, eds Elliott, P., Cuzick, J., English, D. and Stern, R., pp. 181–204. Oxford: Oxford University Press.

SCHLESSELMAN, J. J. (1982). *Case-control Studies: Design, Conduct and Analysis*. New York: Oxford University Press.

SMITH, A. F. M. AND ROBERTS, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society* B **55**, 3–23.

SPIEGELHALTER, D. J., THOMAS, A., BEST, N. AND GILKS, W. R. (1996). BUGS 0.5: *Bayesian Inference Using Gibbs Sampling—Manual (version ii)*. Medical Research Council Biostatistics Unit, Cambridge, 1996.

STONE, R. A. (1988). Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. *Statistics in Medicine* **7**, 649–660.

THOMAS, D. C. (1981). General relative-risk models for survival time and matched case-control studies. *Biometrics* **37**, 673–686.

TIERNEY, L. AND KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82–86.

WAKEFIELD, J. C. AND ELLIOTT, P. (1999). Issues in the statistical analysis of small-area health data. *Statistics in Medicine* **18**, 2377–2399.

WAKEFIELD, J. C. AND MORRIS, S. E. (1998). The Bayesian modelling of disease risk in relation to a point source. Under revision for *Journal of the American Statistical Association.*

[*Received April 26, 1999. Revised October 25, 1999*]

Downloaded from http://biostatistics.oxfordjournals.org/ at University of North Texas on December 31, 2016