# Some notes on Statistical Models

Case Event Data

- It is reasonable to assume that at a residential location $\mathbf{x}$ the probability of observing a case is independent of such a probability at other locations. This will at least hold true conditionally, given knowledge of a spectrum of ancillary information (covariates and other spatial structure information).This model assumption essentially regards individuals as having an independent probability of becoming a case.
- In addition to conditional independence of case events, it is possible to include both heterogeneous background and non-stationarity or long-range spatial trend components in our models by adopting a special type of Poisson process model.
- A heterogeneous Poisson process (HEPP) model is a simple extension of the Homogeneous Poisson process where first-order intensity $\lambda$, is allowed to be spatially dependent ($\lambda(\mathbf{x})$).

For this case, the expected number of events in an area $T$, say, is now :

$$E\{n(T)\} = \int_T \lambda(\mathbf{u}) d\mathbf{u}. \tag{1}$$

The definition of $\lambda(\mathbf{x})$ is quite flexible and allows the inclusion of a modulating function which can represent the heterogeneous (population) background, and also covariate information. In addition, any realization of $m$ events in $T$ has likelihood:

$$\prod_{i=1}^{m} \lambda(\mathbf{x}_i).e^{-\int_T \lambda(\mathbf{u})d\mathbf{u}}. \tag{2}$$

- This is the unconditional likelihood for a realization of $m$ events in $T$. The number of events ($m$) is Poisson distributed with parameter $\rho$. It is also important to note that the likelihood 2, can be simplified by conditioning on the realized value of $m$. This may be useful when we are only concerned with the *spatial* structure of events and not the overall intensity (which is characterised by the realized value of $m$).This conditioning leads to the likelihood:

$$\prod_{i=1}^{m} \lambda(\mathbf{x}_i).\{\int_T \lambda(\mathbf{u})d\mathbf{u}\}^{-m}. \tag{3}$$

Note that if a constant intensity parameter ($\rho$) is included in the parameterization of $\lambda(\mathbf{x})$, then this factors out of 3, and greater parsimony is a result.

- The inclusion of population background in the above models is usually achieved by defining an extra modulating component in $\lambda(\mathbf{x})$. A basic formulation for the modulated intensity is

$$\lambda(\mathbf{x}) = g(\mathbf{x}).m(F(\mathbf{x}).\alpha) \tag{4}$$

where $g(\mathbf{x})$ is a function of the 'at risk' population distribution, and $F(.)$ is an $n \times p$ (spatially-dependent) design matrix of spatial and non-spatial covariates, $\alpha$ is a $p \times 1$ vector of parameters. The function $F(\mathbf{x})$ represents the design matrix evaluated at the location $\mathbf{x}$.

- The function $m(.)$ is usually included to provide a flexible link between the background population-induced intensity and covariates included in the design matrix $F$. Some possibilities are defined in Table 1.

Table 1: Some link types for Hepp models

| $m(F(x).\alpha)$ | link |
|---|---|
| $F(x).\alpha$ | multiplicative-identity |
| $\exp(F(x).\alpha)$ | multiplicative-log |
| $1+F(x).\alpha$ | additive-identity |
| $1+\exp(F(x).\alpha)$ | additive-log |

Note that a scaling parameter can be included in the specification of $F$ which allows the covariate contribution to be separately scaled, from the background intensity.

- The link functions defined in Table 1 represent a range of possible effects which may be thought relevant in the relation of disease incidence to background rate. The multiplicative models represented by the first two entries require that $g(\mathbf{x})$ is directly related to any change in disease incidence, and further that the change is proportional to the background rate.
- For some applications this specification may not be realistic. In some cases where the disease concerned can be regarded as adding to the background propensity then the last two links may be more appropriate.
- In fact the additive-log link has a number of significant advantages in applications where it is important to maintain background risk where there is negligible excess risk predicted and the log component ensures positivity.

- This type of link has been applied in the analysis of putative sources of health hazard. It is not always clear *a priori,* however, which of these links is appropriate in any given situation, and in that case it may be appropriate to examine a range or family of link functions to determine the best specification. It may be appropriate to consider such a range of models in any particular application.

- In the original definition of $\lambda(\mathbf{x})$, the background $g(\mathbf{x})$ function appears in the likelihood, and hence must be estimable at the case event locations $\{\mathbf{x}_i\}$. This implies that the 'at-risk' population must be able to be interpolated to the case locations,. if not already available and measured at these sites. This assumption has implications for the epidemiological interpretation of this model.

- First, the assumption of a continuous $g(\mathbf{x})$ background over a study region may require re-specification if areas of no population occur within the study window. Although this consideration relates to the method of estimation of $g(\mathbf{x})$, the issue is related to the 'ecological fallacy'.

- The ecological fallacy can occur " when a suspected risk factor and disease are associated at the population level, but not at the individual subject level" .

- This can also apply to the use of a population background function $g(\mathbf{x})$ used to describe the probability of an individual case at $\mathbf{x}$.

- In general, the problem can be interpreted as the attribution of *average* characteristics to an individual within a region. Evidently, individuals rarely display such 'average' characteristics, but randomly varying ideographic features.

### 0.0.1 The $g(x)$ estimation problem

- The function $g(\mathbf{x})$,as defined here, is a spatially continuous function representing the propensity of the local population towards contraction of the given case disease. This is termed the 'at-risk' structure of the population. As this function appears within the intensity 4, it must be included in any analysis of this intensity function. Hence, either: 1) $g(\mathbf{x})$ must be estimated and this estimate must also be capable of interpolation to a variety of spatial locations (including the observed case locations $\{\mathbf{x}_i\}$); or 2) $g(\mathbf{x})$ must be removed from the problem. In the first case, $g(\mathbf{x})$ can be estimated prior to analysis of parameters in $m(.)$,in which case inference concerning these latter parameters would be made conditional on the estimated value of $g(\mathbf{x})$, $\widehat{g}(\mathbf{x})$ say. This could lead to a type of profile likelihood analysis of $m(.)$. An alternative approach could be to include $g(\mathbf{x})$ estimation within a general procedure which explores the interaction between $g(\mathbf{x})$ estimation and $m(.)$ estimation. The disadvantage of the profile approach is that it could lead to estimates of $\alpha$ which are sensitive to the value and variability of $\widehat{g}(\mathbf{x})$. These developments were in the analysis of small area health data around putative sources of health hazard, but the methods have wide applicability in situations where the 'at-risk' population related to a realization of case events has to be estimated.

- The second approach to the function $g(\mathbf{x})$, that of removal from the problem, can be accomplished in a variety of ways. First, it could be possible to integrate $g(\mathbf{x})$ out of the intensity and use the resulting integrated intensity ($\lambda^*(\mathbf{x})$) in further analysis. An alternative approach, which is only available when another case event map is used to estimate $g(\mathbf{x})$, is to condition on the realization of case/control marks on the two disease map locations. This leads to a binary logistic regression and $g(\mathbf{x})$ is factored out of the analysis. The advantage of this approach is that it does not require any knowledge of, or manipulation of the $g(\mathbf{x})$ function. The disadvantage is that it is limited to situations where two disease maps are available.

- Methods for the estimation of $g(\mathbf{x})$ require that data be available which describe the 'at-risk' structure of the population. Traditionally when examining counts of disease within small areas, use is frequently made of a standardized rate for each region, which is calculated from known regional or national rates for the case disease. This is usually scaled by the population structure of the region to allow for local effects. This standardization is readily available at census tract level in many countries. However, it is often only available at an aggregate level and hence at a level of aggregation *above* that of case event data. Instead of utilizing such data, it is possible to use a surrogate measure which is available at the case event resolution level. It has been proposed that a mapped realization of another disease could be used to represent the 'at-risk' population structure which must be controlled for in the analysis of case disease data. this additional disease map is used as a spatial 'control' for the case disease and in principal should be matched closely to the population affected by the case disease, but unaffected by the case effects under study. For instance, in a study of clustering of a cancer (case disease) it may be thought appropriate to use coronary heart disease (CHD) as a control disease. If the cancer affects similar ages and sexes in the population then any excess clustering in the cancer will be apparent *above* the local variation in CHD. In the original work, a two dimensional kernel density estimate was used to interpolate the control disease to the case data points. Subsequent inference was made conditional on the value of $\widehat{g}(\mathbf{x})$ found optimally by cross validation of the kernel bandwidth smoothing parameter. However,there are drawbacks to the use of such control diseases, which limit their usefulness as a general panacea in this case. First, the problem of false accuracy of the residential address of the control could lead to misinterpretation. For example, a control disease could be related to factors which are not strongly related to the spatial address structure of the case disease. Hence in this case the only argument for the use of such a control is the aggregate relevance of the spatial expression. In addition the idea that such controls can be interpolated to case data points is also an assumption which should be verified.

- Diggle and Rowlingson (1994) have suggested an approach which 'factors out' the $g(\mathbf{x})$ function from the analysis. This conditional approach directly models the probability of a location being a case rather than a control, given the joint realization of cases and controls. This leads to a different joint likelihood for the case and control data, but conditions the analysis on the observed pattern.

- Given the joint intensity of cases and controls is $g(\mathbf{x})+g(\mathbf{x}).m(F\alpha)$,define the probability of a case at $\mathbf{x}$ as :

$$P(\mathbf{x}) = \frac{g(\mathbf{x}).m(F\alpha)}{g(\mathbf{x})+g(\mathbf{x}).m(F\alpha)} = \frac{m(F\alpha)}{1+m(F\alpha)} \tag{5}$$

then the conditional likelihood of a joint realization of cases and controls is given by

$$L = \prod_{i=1}^{m}\{\frac{m(F_i\alpha)}{1+m(F_i\alpha)}\}.\prod_{j=m+1}^{m+n}\{1-\frac{m(F_j\alpha)}{1+m(F_j\alpha)}\} \tag{6}$$

where there are $m$ cases and $n$ controls.

- While there are many benefits to this approach, not least of which is the fact that $g(\mathbf{x})$ does not require to be estimated and window boundaries no longer need be considered, it remains limited by the fact that it requires the use of a control point map, which, as noted above, has a number of significant drawbacks. If, in addition, only aggregate level standardized rates are available then it cannot be used.

### 0.0.2 Matched case control modelling

- In most of the models considered above the 'at risk' population background was assumed to be represented by a continuous function $g(\mathbf{x})$. In that case the use of control diseases or other expected rate estimators does not allow the inclusion of information about individuals who are matched to the case on selected criteria but who have not expressed the disease.

- Such matching is fundamental to matched case control studies in epidemiology and the usefulness of such individual controls is clear.

- It is possible to define a conditional probability of a particular location, $x_{j0}$ being a case, given the occurrence of the case-control location pair $x_{j0}$ and $x_{j1}$. This probability is

$$p_{jo} = \frac{m(F(x_{j0})\alpha)}{m(F(x_{j0})\alpha)+m(F(x_{j1})\alpha)}.$$

- It is possible to construct a likelihood based on this derivation, and also to extend the derivation to multiple matched controls

Count Data

- We assume that given $m$ tracts, the count $n_j$ within each tract is observed for a fixed time period. Based on the assumptions used to define the case event Poisson process model, it is possible to derive basic model results for counts in tracts. Assuming an underlying modulated heterogeneous Poisson process model for case events,then it is known that for such a process, counts of events within disjoint subregions of the process are independent and the expected count in the $j$ th tract is

$$E\{n_j\} = \int_{a_j} \lambda(\mathbf{u})d\mathbf{u}. \tag{7}$$

- In addition, it is also the case that the counts in these regions are Poisson distributed with expectation given by 7.

- This model implies that within a realization of counts in $m$ regions, the tract counts are independent Poisson distributed with expectation and variance equal to 7. Define the integral in 7 as $\lambda_j$ for brevity. The likelihood of $m$ tract counts is then

$$L = \prod_{j=1}^{m} \left\{ \frac{\lambda_j^{n_j} . e^{-\lambda_j}}{n_j!} \right\} \tag{8}$$

and log-likelihood (bar a constant involving only the data), is

$$l = \sum_{j=1}^{m} n_j \log(\lambda_j) - \sum_{j=1}^{m} \lambda_j. \tag{9}$$

- This also implies that, conditional on $n_T = \sum_{j=1}^{m} n_j$, the total sum of the tract counts (the window or region total), that the counts in tracts have a multinomial distribution with likelihood given by

$$L_{cond} = \prod_{j=1}^{m} \left( \frac{\lambda_j}{\sum_{j=1}^{m} \lambda_j} \right)^{n_j} . \tag{10}$$

- These likelihoods (8), (10) mirror the unconditional and conditional likelihoods found for the case event situation. In principle it is possible to use these models as a basis for the analysis of count data found in arbitrary regions.

- Given the general availability of software for fitting discrete data likelihoods such as the Poisson (e.g. R, S-plus, Minitab), it is surprising that many examples of count data analysis employ approximations to the likelihoods.

- the favoured method is to assume $\lambda_j$ is a constant within tracts

The Parameterization of $\lambda_j$

- The definition of the $\lambda_j$ can follow as for case events. Assuming a constant rate:
$$\lambda_j = E(n_j\} = \rho . e_j . \theta_j$$
- Here $\theta_j$ is the relative risk in the $j$ th tract
- $e_j$ is the expected rate in the $j$ th tract
- $\rho$ is a constant overall rate.

# Log-linear Models

- What happens when we assume a $E(n_j\} = \rho.e_j.\theta_j$ and a Poisson likelihood?
- $l = \sum_{j=1}^{m} n_j \log(\lambda_j) - \sum_{j=1}^{m} \lambda_j$ gives

$$\begin{aligned} l &= \sum_{j=1}^{m} n_j \log(\rho.e_j.\theta_j) - \sum_{j=1}^{m} \rho.e_j.\theta_j \\ &= \sum_{j=1}^{m} n_j \log(\rho) + \sum_{j=1}^{m} n_j \log(e_j) + \sum_{j=1}^{m} n_j \log(\theta_j) \\ & \quad -\rho \sum_{j=1}^{m} e_j.\theta_j \end{aligned}$$

- If we redefine the intensity $\rho.e_j.\theta_j = e_j.\theta_j$ and allow the $\theta_j$ to include a constant term, then we can write:

$$\begin{aligned} & \sum_{j=1}^{m} n_j \log(e_j) + \sum_{j=1}^{m} n_j \log(\theta_j) \\ & -\sum_{j=1}^{m} e_j.\theta_j \end{aligned}$$

Models for the Relative Risk

1) $\theta_j = \exp(\beta_0)$ : constant risk model leads to

$$\begin{aligned} l = & \sum_{j=1}^{m} n_j \log(e_j) + \beta_0 \sum_{j=1}^{m} n_j \\ & - \sum_{j=1}^{m} e_j . \exp(\beta_0) \end{aligned}$$

2) $\theta_j = \exp(\beta_0 + \beta_1 x_j)$ where $x_j$ is a covariate. Then $\log(\theta_j){=}\beta_0 + \beta_1 x_j$.

3) In general, for a design matrix of $p$ covariates $\mathbf{x}$ and a $p \times 1$ parameter vector $\boldsymbol{\beta}$ then

$$log(\theta_j) = \mathbf{x}_j \boldsymbol{\beta}$$

where $\mathbf{x}_j$ is the $j$ th row of the design matrix. This is known as a *log-linear model.*

$$\begin{aligned} l = & \sum_{j=1}^{m} n_j \log(e_j) + \sum_{j=1}^{m} \mathbf{x}_j \boldsymbol{\beta} \\ & - \sum_{j=1}^{m} e_j . \exp(\mathbf{x}_j \boldsymbol{\beta}) \end{aligned}$$

This just like a simple regression model where $n_j$ is the dependent variable. In addition the $e_j$ is a fixed effect and we call $\log(e_j)$ the log offset. It is included but not with a parameter and is fixed.

4) On various packages we can fit these log-linear models as long as there is a Poisson regression or log-linear modeling or generalised linear modelling facility. On SAS Proc *Genmod* or *Catmod* can be used. On R or S-Plus the *glm* function can be used.