Basic Statistical Concepts

Case Event Data

- Define the case event locations found in a study region T as : $\{\mathbf{x}_i\},\ i = 1, \ldots, n$
- Define the first order intensity of cases at $\mathbf{x}$ as $\lambda(\mathbf{x})$. This is just a function which describes the local density of cases on the map.
- Assume that the case intensity contains a background modulating function $g(\mathbf{x})$, which represents the 'at risk' population
- We usually assume that the case incidence can be described by $\lambda(\mathbf{x};\theta) = g(\mathbf{x}).f(\mathbf{x};\theta)$, where the last function ($f(\mathbf{x};\theta)$) represents the effect of interest
- Often it is reasonable to assume that individuals have independent risk of disease (at least conditional upon knowledge of all relevant confounders/factors) for non-infectious diseases.
- A conditional model for independent point process events is the Heterogeneous Poisson process.
- This allows quite general models for the spatial distribution of disease (including clustering) and also allows for ecological analysis via the specification of $f(\mathbf{x};\theta)$.
- Further refinement of the model via the use of prior distributions for parameters and Bayesian methods can be made

Basic Statistical Concepts

Tract Count Data

- Define the number of tracts/regions within T as p.
- Define the count of disease within the i th tract/region as n_i
- As the tracts/regions are an *arbitrary non-overlapping* regionalisation of the study region then it is a well known result from point process theory, that counts in such regions, found by totalling events of a Poisson process, will be independent and Poisson distributed.

Some Useful Results For Tract Count Models

- If you assume that the relative risk (θ) is constant over the whole study region, then the maximum likelihood estimator of θ is given by

$$\hat{\theta} = \frac{\sum n_i}{\sum e_i}$$

this is the overall relative risk for the study region.

- If you assume that there is a different relative risk for each region i.e. $E(n_i) = e_i.\theta_i,$ then the maximum likelihood estimator of θ_i is the SMR for each region/tract:

$$\widehat{\theta_i} = \frac{n_i}{e_i}$$

- these are the minimal and maximal models for the relative risk (assuming *constant* tract rates)

Inference Issues

General Issues

- In most applications it is not possible to assume that the usual sampling distributions will apply (as the usual assumptions may not be met)

- This means that forms of Monte Carlo testing must be used to assess the significance of model parameters and goodness-of-fit measures.

- In general, extra variation in small area health data (e.g. overdispersion in Poisson counts) can arise due to unobserved confounding variables (**uncorrelated heterogeneity**) and can lead to extra variability in the intensity of the disease process. Because this extra variability affects the variability of parameter estimates, then it should be made allowance for in any analysis

- In general, spatial autocorrelation (**correlated heterogeneity**) in small area health data can *confound* effects and so should be *either* included in the model *or* residual autocorrelation should be assessed after model fitting. If any significant autocorrelation exists after fitting models without correlation, then a second fit should be made with autocorrelation included. This is approach is similar to REML estimation.

Inference Issues In Specific Applications

- In the analysis of putative pollution sources, it is often the case that a *supposed effect* (e.g. a cluster of disease) has been reported or claimed to exist at a location, and subsequent analysis is carried out with knowledge of this effect. This can lead to *a posteriori inference problems.*

- Multiple comparison problems can also arise in health status studies where *multiples of disease* are examined

- Disease mapping: interpretation of maps; discretised region rates; edge effects

- Ecological analysis: matching of region level covariates to incidence; confounding (Sardinia : Pascutto et al 1996); error in interpolation of covariates to data points

- In putative hazard assessment, it is often the case that inadequate exposure modelling is carried out: need to model the specific exposure paths e.g. air pollution is directional and distance-based, and hence need to examine polar coordinate systems around sources.

LIkelihood Models

Likelihood is the basis of much within statistics and so it is important that we exazmine it at some point during this course.

- Basic ingredients:

1. a probability model for the data considered (usually a distribution such as the Poisson or binomial or point process)
2. a set of parameters governing the behaviour of the model and describing the data
3. a likelihood is set up which describes how likely the data is given the parameters: this is usually defined (for one parameter θ) as

$$L(\mathbf{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

where there are n data items, $\mathbf{x}$ is the set of data items, $f(x,\theta)$is the probability model for the data. The product sign is used as it is assumed that tyhe data are independent (such as in a randon sample of values). That is $L(\mathbf{x};\theta) = f(x_1;\theta).f(x_2;\theta).f(x_3;\theta).\ldots\ldots\ldots\ldots f(x_n;\theta)$

Usually we work with the log of the likelihood as it is easier:

$$l(\mathbf{x};\theta) = \sum_{i=1}^{n} \ln f(x_i,\theta).$$

Notice how there is now a sum instead of a product.

Example: often count of disease within a map of n regions are assumed to have a Poisson distribution: let $\{y_i\}, i = 1,\ldots,n$ denote the count of disease in these small areas. Hence, $E(y_i) = \lambda_i$

$$f(y_i;\theta) = \lambda_i^{y_i}\exp(-\lambda_i)/y_i! \qquad\qquad \lambda > 0;$$

1. **a.** The log-likelihood for this model is

$$l(\mathbf{y};\theta) = \sum_{i=1}^{n} \ln f(x_i,\theta) = \sum_{i=1}^{n} \ln\{\lambda_i^{y_i}\exp(-\lambda_i)/y_i!\,\}$$
$$= \sum_{i=1}^{n} y_i \ln\lambda_i - \sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n}\ln y_i!$$

   b. This log likelihood tells us about how the parameters affects the data and also can show us how the data informs us about the parameters.

   c. Usually in spatial epidemiology the expectation (λ_i) is made up of two different components:

   d. 1) a function representing the population at risk of the disease (called the *at risk background* or *arb* for short)

   e. 2) a function where we *model* the variation in risk

   A typical defintion is

$$E(y_i) = \lambda_i = e_i\theta_i$$

Here , e_i could be an expected rate for the disease within the small area (based on the local *arb*) and θ_i is the *relative risk* (which we mentioned earlier. Generally we are interested in the relative risk part of the function and not very interested in the e_i. The e_i must be estimated (well) but is otherwise a *nuisance*.They are usually assumed to be constant. (*Any comments on that?*)

If we put all these ingredients into our log likelihood then we have

$$l(\mathbf{y};\theta) = \sum_{i=1}^{n} y_i \ln(e_i\theta_i) - \sum_{i=1}^{n}(e_i\theta_i) - C$$

(the last term C is a function of the data only and can be ignored in what follows).

The relative risk is the most important feature of this model and tells us about how much the local disease risk is elevated or decreased compared to the *arb*.

Estimation

The log-likelihood $l(\mathbf{y};\theta)$ can be seen as a function informing us about the parameters given the data we observ ed. Thats is , if we examine $l(\mathbf{y};\theta)$ we could find out what the best or most likely value of θ is given the data we observed. The method of *maximum likelihood* seeks to find the most likely value of θ by maximising the likelihood function treating the data as constant. Mathematically maximum likelihood involves taking the first derivative of the likelihood wrt the parameter (or parameters, if there is more than one) and setting this equal to zero. Then the solution of this equation is the maximum likelihood estimator of θ.

$dl/d\theta$ is defined to be the first derivative of $l(\mathbf{y};\theta)$ with respect to θ

set $dl/d\theta = 0$ and solve to give $\widehat{\theta}$, the maximum likelihood estimator of θ.

In our example, lets assume that there is only one θ :

$$l(\mathbf{y};\theta) = \sum_{i=1}^{n} y_i \ln(e_i\theta) - \sum_{i=1}^{n}(e_i\theta) =$$
$$= \sum_{i=1}^{n} y_i \ln e_i + \ln\theta \sum_{i=1}^{n} y_i - \theta \sum_{i=1}^{n} e_i$$

and, remeber that the sums of the data and *arbs* are constant and so:

$$\frac{dl}{d\theta} = \frac{\sum y_i}{\theta} - \sum_{i=1}^{n} e_i = 0$$

Solving for θ gives:

$$\widehat{\theta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} e_i}$$

this is just the total disease count divided by the total expected rate and is a an overall estimate of

the study region relative risk. Note that as we have only a single parameter, this is called a *minimal model*.

Confounders and Random Effects

Confounders

- The models described above, are observational in that they describe the probability model of the observations. However, it is clear that structure can remain in the data after likelihood models are fitted, due to the existence of explanatory variables which have *not been included* in the model.

- These variables may be measurable and available, and so should be included in any analysis so that the proper allowance is made for these effects. Inclusion of *deprivation indices* is an example of the use of extra explanatory variables. These variables are sometimes known as *known confounders.* The reason for the term confounder is that the variable in question impacts on the disease incidence and therefore confounds or alters the appearance of the effect of interest. For example, in a disease clustering study, low income households may be congregated in some areas and this will lead to apparent clusters of disease in these areas. The use of deprivation indices, which should make allowance for low income areas, should extract the confounder effect.

- In addition, there may be thought to be other variables affecting the spatial distribution of the disease which cannot be measured easily or are even unknown aetiological agents. These are *unknown confounders*, and they may be thought to exist in *any* disease mapping application (even when known confounders are included)

Methods for Inclusion of Confounders

- *Known confounders* can be included as regression variables in the specification of the intensity function, i.e.:

$$\lambda(\mathbf{x}_i;\theta) = g(\mathbf{x}_i).m\{F(\mathbf{x}_i).\beta\}$$

where $F(\mathbf{x}_i)$ is the i th row of the design matrix F, the k columns of which are regression variables, and β is a $k \times 1$ parameter vector, and $m\{\}$ is a suitable link function.

- *Known confounders* for count data can be included in a Poisson regression (log linear model)

- *Unknown confounders* may be included via *random effects*