Session: Ecological Analysis

Ecological Analysis concerns making inference about the relation between explanatory variables (usually at an aggregated level) and geographical disease incidence. Consider the simple linear model:

$$y_i = f(\beta_0 + \beta_1 x_i) + e_i,$$

where y_i is the incidence rate and x_i the explanatory variable, each measured at location $i.$, and $f(.)$is a link function. In simple linear regression we seek to fit the model and estimate $\beta_0, \beta_1.$ (This is a simplified version of the more complex log-linear model (with Poisson likelihood) which is usually employed for this task.)

In this simple example some features are shared with ecological analysis: we my want to apply the model to unobserved data within the data itself (interpolation of the relation); we may want to extend the model beyond the data (extrapolation of the relation)

In ecological analysis, there are a number of issues which arise:

- if data are aggregated, i.e. spatially, then how do we make inferences about individuals. Classic example: high correlation between individual covariates and incidence (eg smoking and lung cancer) may lead to a strong spatial association due to there being concentrations of people who smoke in some areas. However, this is an individual effect and hence has little implication for non-smokers. On the other hand, there might be a high correlation between water hardness and CHD over a geographic region, and this will have an impact on individual CHD risk.

- Important to separate out individual covariates (which may happen to have a geographic expression) from covariates common to a geographic area.
- Even when these are separated out, it is still a considerable task to ascribe aggregate relations to individuals within areas.

This is the *ecological* fallacy: the attribution of aggregate relations to individuals

The opposite of this is the *atomistic* fallacy: the attribution of individual characteristics to aggregates of populations

Note that the ecological fallacy often relates to count data (e.g. tract level or above), while the atomistic fallacy often applies to case event data.

- What is the need to carry out geographic

  studies of aggregate relations?
  1) direct spatial hypotheses: e.g. a) putative sources of hazard are an example of ecological analyses where the explanatory variables are exposure surrogates such as distance and direction, b) the relation of disease clusters and spatial covariates (e.g. environmental gradients)
  2) spatial relations can yield observations not available in conventional studies. For example, cohort studies can suffer from censoring and so some parts of the relationship may be missed.
  3)Geographical studies can be designed to assess non-geographic hypotheses concerning aetiology.

- The main issues which arise in these studies are:

1) *scale aggregation and inference at different levels*

2) *unmatched scale components (interpolation)*

3) *individual level inference with covariates*

Scale Aggregation

- Does our analysis apply at different scales?

Usually this is not true, but we want to know at which scales it is reliable. For example, if we find a positive relation between the proportion of car owners in census tracts of a city district and CHD numbers, then is that relation also true if we looked at country wide health authorities?

We may want to quantify the changes which scale makes on the analysis. In the Geostatistics area, this idea is called ’change of support’ and in geography as the ’modifiable areal unit problem’ (MAUP). It has been addressed by introducing a ’geography’ variable into the analysis i.e. including some measure of the scale e.g. a binary (0/1) variable denoting a two scale problem(Cressie, 1996). Another possibility, may be to include a *random* efffect in the analysis which can be estimated by aggregation upwards in the data.

Changes in Scale

Components

- What if components in the problem are measured on different scale units?

For example, in the Armadale putative source example the case disease was available as residential address locations, but we also had available expected rates for the case disease only in 18 census tracts (within the study area). Hence the expected rates were aggregated above the level of the case data.

Such mismatches can be handled by using special methods, including interpolation of covariates to the location of the count (centroid) or case events. Assuming that we have covariates at an aggregated level, then we would have two operations involved in interpolating to the case locations: 1) the aggregated covariate (e.g. expected rate) is already smoothed (due to aggregation) and then 2) must be smoothed under a interpolation model to the locations. This smoothing will be with error and so this should be added to the model. Special hybrid models can be used for this (see e.g. Lawson and Williams, 1994) or a general approach which is useful is that of Bayesian Hierarchical modelling where each level of data and parameters can be a level in the hierarchy.

An example of mismatched units of observation would be disease counts and pollution measurements made on a network. Here there are covariates at locations other than the tracts. We need also to interpolate here but may also need to estimate the total pollution over each tract.

Individual Level Inference

The ecological fallacy in spatial data

The regression analyses based on geographically collected data are subject to bias due to their aggregate nature and to the potential presence of spatial autocorrelation among the responses: these two aspects are related to each other, due to the fact that aggregation and scale change can lead to autocorrelation due to smoothing. Autocorrelation is also found due to unobserved confounder variables. In this way, making inference on the basis of ecologic associations to individual level behaviour could have serious pitfalls.

The problem, known as the *ecological fallacy* (also named *ecological bias*), was first pointed out by Robinson (1950), who demonstrated that the total correlation between two variables as measured at an ecologic level can be expressed as the sum of a within-group and a between-group component. Later Duncan et al. (1961) extended this result deriving the relationship between the regression coefficients in a linear model. The sources of ecological bias have been investigated by many authors (see for example Richardson et al., 1987; Piantadosi, 1988; Greenland and Morgenstern, 1989; Greenland, 1992; Greenland and Robins, 1994).

In addition to the individual level sources (misspecification, within-group confounding, no additive effects, misclassification) special attention has been given to the bias due to grouping individuals. In particular Greenland and Morgenstern (1989) analyzed how grouping influences associations of exposure factors to disease, they pointed out that ecological bias may also arise from confounding by group and effect modification by group.

Now consider some ecological groups indexed by k and let p_k be the proportion of exposed subjects (a dichotomous variable), r_{0k} the individual rate in unexposed and r_{1k} the individual rate in exposed at the site k. The crude rate in group k is given by:

$$r_{+k} = r_{0k}(1 - p_k) + r_{1k}p_k$$

$$= r_{0k} + D_k p_k,$$

where $D_k = r_{1k} - r_{0k}$ is the individual rate difference.

Consider a linear regression model of average disease level on the average exposure level in groups:

$$E(r_{+k}) = \alpha + \beta p_k,$$

then $1 + \frac{\beta}{\alpha}$ is the ecological rate ratio estimate. Greenland and Morgenstern demonstrated that the ecological regression coefficient β can be viewed as the expected rate difference at individual level plus two bias terms. The mathematical relationship is given by:

$$\beta = E(D_k) + \frac{cov(p_k; r_{0k})}{var(p_k)} + \frac{cov([p_k - E(p_k)]p_k; D_k)}{var(p_k)},$$

The first bias component

$$\frac{cov(p_k; r_{0k})}{var(p_k)}$$

is present when the unexposed rate is associated with the level of exposure in the group, and it may be viewed as a bias term due to confounding by group. It is plausible that such confounding acts because some external factor causing the disease is

associated with groups having higher level of exposure factor. The second bias component

$$\frac{cov([p_k - E(p_k)]p_k; D_k)}{var(p_k)}$$

is present when the risk difference in a group is associated with the level of exposure and it may be viewed as a bias term due to effect modification by group. Hence one commits ecological fallacy if one assumes that the ecological rate ratio estimate $1 + \frac{\beta}{\alpha}$ is only determined by the individual rate difference effect when, in fact, it may be also caused by the two bias components effect. Several strategies can be adopted to tackle the potential flaws of ecological modelling.

First one could try to estimate the joint distribution of outcome and explanatory variables within areas using a sample drawn from the populations investigated, and use the information collected to adjust the ecological regression coefficient and standard errors. This approach has been proposed by Plummer and Clayton (1996), and Prentice and Sheppard (1996). It can also be viewed as an example of multilevel model with individual and ecological variables (see Lawson and Williams(1994) for an example of multiple level exposure risk modelling).

When sampling within areas is not feasible, a second strategy could be to adjust for the correlation between area prevalence of the exposure variable and baseline rate of disease, provided no effect modification has occurred. If the level of aggregation is sufficiently thin a regression model for autocorrelated data would result in a sort of stratification by spatial closeness, where the baseline rates would be expected not to vary. Clayton et al. (1993) gave a justification of this approach in term of a hidden spatially structured confounder. Indeed, where the spatial variation of the risk factor is similar to that of disease, geographical location may act as a confounder. *Thus introducing in ecological models a component accounting for spatial interaction may produce control of bias due to the confounding effect of geographical location*.

Moreover unknown confounders are likely to be present in ecological data since factors which are not at individual level can be confounders at aggregate level. Unstructured hidden confounders result in a certain degree of extra-variability which in this analysis should be taken into account. The control of small level spatial variation has many similarities to that developed in time series analysis, the main differences in geographical epidemiology is that the focus of the research is on the regression coefficient and not primarily in the interaction component, which are in most applications regarded as nuisance.

Poisson regression models.

Before introducing spatial models we first consider the Poisson regression model that represents the starting point of statistical methods in ecological analysis. Let $\{Y_i, i = 1, \ldots, n\}$ be the set of observed number of events of a certain disease and $\{E_i, i = 1, \ldots, n\}$ the set of expected number under a reference set of age-specific rates for n areas of the region of interest. Then Y_i follows a Poisson distribution with expectation:

$$\mu_i = \theta_i E_i,$$

where θ_i is the relative risk for the site i. The maximum likelihood estimates of θ_i, under a saturated model, are given by the standardized mortality ratios:

$$SMR_i = \frac{Y_i}{E_i}.$$

This model can be extended to a set of explanatory variables $X_1, X_2, \ldots, X_H$ in a log-linear formulation:

$$\log \mu_i = \log E_i + \sum_{h=1}^{H} \beta_h x_{ih}.$$

Maximum likelihood estimates of the coefficients β_h can be obtained in a generalized linear models framework.

A Poisson regression model can include the influences on disease of many ecologic factors (the covariates X_h) but it does not control for the autocorrelation and for the extra-Poisson variability, which may arise due to, for example, unobserved confounder variables. Some authors argued that non linear ecological models give biased estimates of the individual level coefficients. This bias is negligible for moderately large risk ratios (see e.g. Richardson et al., 1987 and Greenland, 1992).

Bayesian mixed models.

Unstructured and structured extra-Poisson sources of variability can be further considered by the following generalized linear mixed model:

$$\log \theta_i = t_i + u_i + v_i,$$

where

$$t_i = \sum_{h=1}^{H} \beta_h x_{ih}$$

denote the fixed regression component, u_i is the random unstructured terms (named *heterogeneity*) and v_i the random spatial structured terms (named *clustering*).

Introducing the heterogeneity and clustering terms represents a way of taking account of unmeasured covariates. Defining appropriate prior distribution on the hyperparameter involved in the model, a Bayesian inference using the posterior distribution of θ_i can be done. In particular estimates of the relative risks can be computed by running MCMC algorithms. This model was introduced by Clayton and Kaldor (1987) in disease mapping framework and then it was developed by Besag et al. (1991) and Clayton et al. (1993) in ecological analysis.