

PART FOUR

13 A Common Framework for Ecological Inference in Epidemiology, Political Science, and Sociology

Ruth Salway and Jonathan Wakefield

ABSTRACT

Ecological studies arise within many different disciplines. This chapter describes common approaches to ecological inference in an environmental epidemiology setting, and compares these with traditional approaches in political science and sociology. These approaches vary considerably, both in their use of terminology and notation, and in the relative importance of the various issues that make ecological analyses problematic. The aims of this chapter are twofold. Firstly, we describe ecological inference in an epidemiology setting, where the interest is in the relationship between disease status and exposure to some potential risk factor. We concentrate on those issues which are of particular concern in epidemiology, for example the presence of additional (possibly unmeasured) covariates, termed confounders. Secondly, we seek to unite the current work in epidemiology, political science, and sociology by clarifying differences in terminology, by describing commonly used approaches within a common statistical framework, and by highlighting similarities and differences between these approaches. Often different models can be attributed to different sets of underlying assumptions; we emphasize that such assumptions are crucial in the conclusions drawn from ecological data, and their appropriateness should be carefully considered in any specific situation. Combining approaches from all three disciplines gives a broad range of possible assumptions and available techniques from which to choose.

13.1 INTRODUCTION

Ecological studies arise within many different disciplines; in this chapter we consider ecological inference from an epidemiology perspective and compare it with approaches in political science and sociology. Although all three disciplines utilize ecological data, commonly used approaches vary considerably, and there is very little communication between epidemiology, on the one hand, and sociology and political science, on the other. Although methods in one discipline may not always be applicable in another, in general many of the concerns addressed are common to all.

When using ecological data to make inference about individuals, bias may occur due to aggregating data within areas; this bias is known as *aggregation bias* (Achen and Shively, 1995) in the social sciences, and as *ecological bias* (Richardson and Monfort, 2000) in epidemiology. A more general term used in sociology is *cross-level bias* (Firebaugh, 1978), which refers to bias that occurs when data are available at one level but inference is made at a different level (so it can also refer to bias that arises from using individual data to make inference about areas).

In Section 13.2 we describe how ecological studies arise in epidemiology. The common problem is to make individual-level inference in the presence of possible ecological bias; one of the ways in which epidemiological approaches differ is in explicitly concentrating

on the underlying causes of bias. We follow such an approach throughout this chapter; first, we identify the main sources of ecological bias, and then we consider the effect on inference of each source separately. We introduce a general statistical model in Section 13.3 that allows us to explicitly model three of the main sources of bias. These are bias due to unmeasured covariates, bias due to the presence of contextual effects, and bias due to the model parameters varying between areas, and are discussed in Section 13.4. This model will provide the framework for this chapter and will be used to link approaches to ecological inference in the three disciplines.

Section 13.5 discusses some of the primary issues in ecological analysis in epidemiology. We discuss the aims of ecological analyses, and describe the general approaches that are used to tackle the sources of bias. Section 13.5.2 introduces the use of hierarchical models to model overdispersion, and Sections 13.5.3 and 13.5.4 consider two ideas from general epidemiology that may be applied to ecological analyses when the availability of data is limited. The first is that of choosing between competing explanations, in the presence of little information from the data themselves, on the basis of *plausibility*. The second demonstrates the use of a sensitivity analysis to investigate the possible effects on inference of unobserved confounding.

We compare and contrast the epidemiology and social science approaches to ecological inference in Section 13.6. We focus on a scenario in which the data are discrete and consist of one 2×2 table for each area; the links between this model and the more general framework of Section 13.3 are highlighted in Section 13.6.1. One of the important differences between epidemiology and political science is that in the latter it may be the unobserved individual data that are of interest rather than underlying probabilities (which is essentially a difference between prediction and causality); the differences are discussed in Section 13.6.2. Section 13.6.3 focuses on the situation in which the underlying probabilities vary between areas. Section 13.7 considers some common models used in social science and relates them to those used in epidemiology, and Section 13.8 provides a concluding discussion.

Throughout this chapter we will use terminology from epidemiology, so that the outcome of interest is a disease indicator; an individual with the disease is known as a *case*, and a disease-free individual a *noncase*. The covariate of interest, a potential risk factor for the disease, is called the *exposure* variable, and we are specifically interested in the nature of the relationship between the disease indicator and exposure to the risk factor, after control for confounding variables. Roughly speaking, a confounder is a variable that is related to both the response and the exposure, does not lie on the causal pathway between exposure and response, and is not caused by the response (see Rothman and Greenland, 1998, for more discussion). So for example, the disease indicator may be whether an individual has a respiratory disease, such as asthma. The exposure may be discrete, for example a genetic trait, or it may be continuous, for example the sulfur dioxide concentration in ambient air in the neighborhood of the individual. Similarly confounders may be discrete (for example gender) or continuous (for example, dietary measures such as fat consumption). For both confounders and exposures, continuous variables may be artificially discretized. Although this reduces information, a large number of categories allows for flexible risk–exposure–confounder relationships, and for interactions to be considered.

In a political science application, the response may correspond to voter turnout, with a “case” being a voter and a “noncase” being a nonvoter. The “exposure” is race, which is discrete for example, with two categories, black and white. A possible confounder in this example might be income, which we might expect to be related to both voter turnout and race. The problem of determining causality between exposure and disease in the presence of confounding variables is central to epidemiology. The equivalent problem in political science

might be to assess the causal relationship between race and voting behavior, controlling for income.

13.2 ECOLOGICAL INFERENCE IN EPIDEMIOLOGY

Determining causality in any observational study is problematic, since exposures are not randomly assigned to individuals. A major cause of bias in observational studies is that due to confounding; consequently, many epidemiologic analytical techniques are concerned with controlling for confounding factors. This general philosophy extends to ecological studies; the aim is usually to make inference for individuals concerning the relationship between disease and exposure in the presence of confounding. By many epidemiologists ecological studies are viewed with skepticism. This view would seem too pessimistic, however. Ecological studies are not only useful hypothesis-generating mechanisms, but can also add to the totality of evidence when building a case for a disease risk–exposure relationship (Morgenstern, 1998). The appeal of ecological studies is that they can utilize routinely available data (and so are relatively inexpensive to carry out) and can cover a broad geographical area, thus taking advantage of large exposure contrasts and large populations; both of these factors increase power.

Historically, epidemiologists have concentrated on methods developed for contingency tables. A typical analysis with binary exposure and disease variables would initially examine the marginal observed association (that is, collapsing across confounder stratum), before examining the effects of stratification (confounder) variables such as age and sex, perhaps following a test for heterogeneity of the association across stratifying variables, that is, a test for effect modification. Chapter 3 of Breslow and Day (1987) and Chapter 4 of Breslow and Day (1980) provide an excellent introduction to such approaches. More recently, a more explicit model-based approach has grown in popularity (Clayton and Hills, 1993). The advantages of such an approach are that universal statistical principles can be followed in a more general modeling setting, and problems of small cell counts can be avoided to some extent by smoothing across cells. In addition the assumptions that lead to particular estimators can be made explicit, which is particularly important in ecological studies. Stratified analyses can often be viewed as a special case of the more general framework.

Hence the current focus in ecological models in epidemiology is on explicitly modeling the risk–exposure relationship and estimating effect parameters. The disease–exposure relationship is often nonlinear. Diseases are usually rare in a statistical sense. Ecological studies are particularly important in environmental epidemiology in examining the effects of air pollution (Pope and Dockery, 1996) and water quality (see, for example, Maheswaran et al., 1999). Often the scale is international; for example, Yasui et al. (2001) examine the ecological association between incidence of breast cancer and incidence of non-Hodgkins lymphoma, and Prentice and Sheppard (1990) describe an ecological approach to studying the relationship between total dietary fat intake and incidence of breast cancer. Numerous studies have also examined the ecological association between measures of socioeconomic status and different health outcomes; see Singh and Siahpush (2002) and the references contained therein. Richardson and Monfort (2000) and Wakefield (2003) provide further examples of ecological studies.

In a model-based approach we are interested in the underlying individual parameters (and derived probabilities of disease). This is consistent with the search for causal relationships and may be contrasted with a predictive approach in which it is not estimates of parameters that are of concern, but rather imputation in which the missing cell entries are the target of inference. The objective of most epidemiological studies is to estimate the change in

disease risk attributable to a specific factor for a rare disease. This can be expressed on various scales, for example, as a *risk difference*, or as a *relative risk* of disease for each area. Typically a risk difference is used with a linear model and represents the additive difference between the disease risk of, for example, an exposed individual and an unexposed individual. Relative risks are more natural in a log-linear framework and represent the multiplicative increase in disease risk in the exposed population relative to the unexposed population. No exposure effect (that is, when the risk of an exposed and unexposed individual is the same) corresponds to a risk difference of 0, or a relative risk of 1.

13.3 STATISTICAL MODEL

In this section we introduce notation that allows us to separate different sources of ecological bias. We begin by describing an explicit model at the level of the individual (following Richardson, Stucker, and Hémon, 1987; Prentice and Sheppard, 1995; and Wakefield and Salway, 2001). We are interested in examining individual relationships; it is advantageous to specify models in terms of individual parameters, as this links ecological inference to individual inference. By stating an underlying model we are also better equipped to make explicit the assumptions of any analysis and to identify the plausibility of such assumptions. This approach is of particular benefit when attempting to identify causal relationships.

The notation in this chapter differs from that used in the Introduction to this book, for the latter does not extend easily to an epidemiology context. However, it is consistent with that used in Chapters 1 and 12.

Suppose we have a study area partitioned into a disjoint set of m areas, with area i containing N_i individuals, $i = 1, \dots, m$. The response is a Bernoulli random variable Y_{ij} representing the disease outcome of individual j in area i , $i = 1, \dots, m$, $j = 1, \dots, N_i$, over a specific time period, with $Y_{ij} = 1$ corresponding to a case and $Y_{ij} = 0$ a noncase. Similarly, we let X_{ij} represent the univariate exposure of individual j in area i . In our general formulation, X_{ij} may be a discrete variable with two or more categories, or it may be continuous. It is straightforward to extend this model to consider multiple exposures (for example, three different air pollutants), but for notational simplicity we will assume it is univariate. We begin with an individual risk–exposure model for a noninfectious disease (so that outcomes on different individuals within an area may assumed to be independent, after controlling for risk factors). The model takes the form

$$Y_{ij}|q_{ij} \sim_{\text{ind}} \text{Bernoulli}(q_{ij}),$$

where

$$q_{ij} = P(Y_{ij} = 1 | X_{ij}, Z_{ij}, X_i, \text{area } i),$$

and

$$g(q_{ij}) = \beta_{0i} + \beta_{1i}(X_{ij} - X_i) + \beta_2 X_i + \gamma Z_{ij} + \delta_i, \tag{13.1}$$

where “ind” is an abbreviation for *independently distributed*, and we have assumed linearity on a scale determined by a link function $g(\cdot)$. Here X_i is the mean exposure in area i , and Z_{ij} is a univariate individual-level confounder (again the extension to multiple confounders is straightforward).

The model 13.1 allows for the possibility of confounding, contextual effects, effect modification, and overdispersion, including spatial dependence:

- The effect of an individual's exposure relative to the area-level average exposure is given by β_{1i} ; this is the effect parameter of interest, and is an area-specific exposure effect, so that we have effect modification (also known as interaction) by area.
- We also allow the baseline risk parameter β_{0i} to vary between areas.
- The parameter β_2 measures a contextual effect due to exposure, that is, an effect due to the overall average exposure in the area beyond the effect of an individual's personal exposure. Contextual effects in the exposure will be present if $\beta_{1i} \neq \beta_2$.
- The presence of confounding is represented through the covariate Z_{ij} and the associated nuisance parameter γ . This may be a within-area confounder (a variable measured at the level of the individual – for example, a behavioral variable), or a between-area confounder, in which case $Z_{ij} = Z_i$ (a characteristic of the area – for example, income disparity or access to health services). The model does not allow confounder effects to vary by region, or for a contextual effect in the confounder, but it is general enough to allow a number of possible sources of bias to be illustrated.
- Finally, the random effect error term δ_i may or may not have spatial structure.

The model 13.1 incorporates a link function $g(\cdot)$, which allows for a nonlinear risk–exposure relationship. Suitable link functions include a logit link, which constrains the probabilities q_{ij} to lie between 0 and 1, or a log link as an approximation when q_{ij} is small (that is, the disease is rare). The latter model is frequently used in epidemiology, where disease counts in an area are small relative to the population size. This corresponds to a multiplicative risk–exposure relationship, with

$$E[Y_{ij}|X_{ij}, X_i, Z_{ij}] = \exp \{ \beta_{0i} + \beta_{1i}(X_{ij} - X_i) + \beta_2 X_i + \gamma Z_{ij} + \delta_i \}.$$

Great care is required to interpret the parameters of this model (since we cannot increase an individual's exposure and keep the average the same). If we increase the exposure for all individuals in area i by one unit, then for each individual in area i we have

$$\frac{P[Y_{ij} = 1|X_{ij} = x + 1, X_i = \bar{x}_i + 1, Z_{ij}]}{P[Y_{ij} = 1|X_{ij} = x, X_i = \bar{x}_i, Z_{ij}]} = \exp(\beta_{1i} + \beta_2),$$

which gives an interpretation to the sum of the two effect parameters when Z is kept constant for all individuals and the two individuals that are compared are in the same area (since we need the parameters β_{0i} to cancel).

We are concentrating upon a linear link function in this chapter, in which case, if we increase the exposure for all individuals in area i by one unit, the risk difference is given by

$$\begin{aligned} &P[Y_{ij} = 1|X_{ij} = x + 1, X_i = \bar{x}_i + 1, Z_{ij}] - P[Y_{ij} = 1|X_{ij} = x, X_i = \bar{x}_i, Z_{ij}] \\ &= \beta_{1i} + \beta_2. \end{aligned}$$

Again, under a causal interpretation we may consider two individuals j and j' in the same area i whose exposures differ by one unit and who have the same value of Z , to obtain

$$\begin{aligned} &P[Y_{ij} = 1|X_{ij} = x + 1, X_i = \bar{x}_i, Z_{ij} = z] - P[Y_{ij'} = 1|X_{ij'} = x, X_i = \bar{x}_i, Z_{ij'} = z] \\ &= \beta_{1i}. \end{aligned}$$

Suppose that neither the baseline risk parameter β_{0i} nor the effect parameter β_{1i} depend on i . Now, consider two areas i and i' whose mean exposures differ by one unit, and consider

two individuals, j in area i , and j' in area i' , who have the same value of Z but whose exposures differ by one unit. Then

$$P[Y_{ij} = 1 | X_{ij} = x + 1, X_i = \bar{x}_i + 1, Z_{ij} = z] - P[Y_{i'j'} = 1 | X_{i'j'} = x, X_i = \bar{x}_i, Z_{i'j'} = z] = \beta_2.$$

The above considerations illustrate the care that must be taken in parameter interpretation.

The linear model is somewhat unrealistic in that no constraints are placed on q_{ij} , which must be between 0 and 1, since it is a probability. In practice such a simplistic model will be suitable in only a few situations, for example as an approximation for low levels of exposure and small exposure effects. In general, fitting a linear model when the true relationship is nonlinear can introduce serious bias (Greenland, 1992). However, nonlinearity introduces additional problems of mathematical complexity which make the exposition less easy to follow. Many of the results follow for a log-linear model, and where differences occur, they are noted in the text.

In an ecological study only area-level data are available. Typically these data consist of the area means X_i , Y_i , and perhaps Z_i . In epidemiology the ecological data usually consist of counts of cases within each area, $Y_{i+} = \sum_j Y_{ij}$; since $Y_i = Y_{i+}/N_i$, these are interchangeable, for the population sizes are known (at least in principle, although data anomalies may be problematic; see Wakefield and Elliott, 1999). The disease counts Y_{i+} and mean Y_i correspond to T'_i and T_i respectively in the notation of the Introduction.

Here we will formulate the ecological model in terms of disease counts, Y_{i+} . We derive the model induced at the ecological level in Equation 13.1 by aggregating over individuals within each area. In general the form of this ecological model will depend on the joint within-area distribution of X_{ij} , Z_{ij} :

$$E[Y_{i+} | X_{ij}, Z_{ij}] = N_i E_{X_{ij}, Z_{ij}}[q_{ij} | X_i, Z_i],$$

where the expectation is with respect to the joint distribution of $(X_{ij}, Z_{ij}) | X_i, Z_i$. Here, we assume that both exposure and covariate are discrete binary variables, with

$$X_{ij} | \pi_{xi} \sim_{\text{ind}} \text{Bernoulli}(\pi_{xi}),$$

and

$$Z_{ij} | \pi_{zi} \sim_{\text{ind}} \text{Bernoulli}(\pi_{zi}),$$

so that, for example, the probability of an individual in area i being exposed is given by $P(X_{ij} = 1 | \pi_{xi}) = \pi_{xi}$. Under these circumstances,

$$Y_{i+} | \pi_{xi}, \pi_{zi} \sim_{\text{ind}} \text{Binomial}\{N_i, q_i\},$$

with

$$q_i = \beta_{0i} + \beta_2 \pi_{xi} + \gamma \pi_{zi}. \tag{13.2}$$

The term $\beta_{1i}(X_{ij} - X_i)$ has disappeared, since $E[X_{ij} - X_i] = 0$ when we average over an area. Equation 13.2 depends only on the area-specific probabilities (π_{xi}, π_{zi}) because we have a linear relationship; in general, a nonlinear relationship will result in an additional term involving the joint probability $\pi_{xzi} = P(X_{ij} = 1, Z_{ij} = 1 | \pi_{xzi})$ (Lasserre,

Guihenneuc-Joyaux, and Richardson, 2000). Similarly, the joint distribution needs consideration when we have effect modification by Z .

In the more general situation where exposures may be continuous and the link function is nonlinear, the expression for the marginal risk q_i will depend on higher moments of the within-area exposure–confounder distribution, such as the variance–covariance matrix (Richardson et al., 1987; Wakefield and Salway, 2001).

Typically we do not know the underlying population parameters π_{xi} and π_{zi} and must use estimates, for example the means X_i and Z_i . In this case, $X_{ij}|X_i$ are not independent and the binomial distribution is an approximation to the true distribution of disease counts (although with large N_i this should not be a problem). We have

$$E[Y_i|X_i, Z_i] = q_i = \beta_{0i} + \beta_2 X_i + \gamma Z_i, \tag{13.3}$$

but the variance will be smaller than under a binomial model. The true distribution is a convolution of binomials (Chapter 1).

The main difference between this model and that Introduction is that here we have explicitly described the individual-level relationship. The focus is clearly on estimating the individual-level parameters β_{0i} , β_{1i} (and hence the underlying individual probabilities), rather than the unobserved cell proportions. The relationship between these two depends on the form of the individual model (via the choice of link function and whether confounding, contextual effects or effect modification is present). This is discussed further in Section 13.6.1.

The simplest case of the model 13.2 (which is unrealistic in practice) is when we have no confounding (so $\gamma = 0$), no contextual effects (so $\beta_{1i} = \beta_2$), and nonvarying baseline risk and effect estimates (so the parameters β_{0i} , β_{1i} do not vary between areas). Then Equation 13.3 becomes

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i, \tag{13.4}$$

or, written in terms of the relative risk $\theta = p_1/p_0$, where $p_x = P(Y_{ij} = 1|X_{ij} = x)$,

$$E[Y_i|X_i] = p_0 + p_0(\theta - 1)X_i \tag{13.5}$$

(also considered by Plummer and Clayton, 1996: 116). In this special case, the ecological model takes the same form as the individual-level model, with the same parameters. It is well documented (for example, Piantadosi, Byar, and Green, 1988) that in this case estimates derived from the ecological model will be unbiased estimates of the underlying individual-level parameters. This result requires a linear relationship between X_{ij} and Y_{ij} ; with other link functions the ecological model will not in general take the same form (Richardson et al., 1987), even in the absence of confounding, contextual effects, and effect modification, unless there is no within-area variability in areas (in a political science context, an example would be each area containing individuals of one race only). The distinction between linear and nonlinear forms is important in epidemiology, since risk–exposure relationships are often multiplicative. The bias that arises in fitting an ecological model of the same form as the individual-level model is known as *pure specification bias* (Greenland, 1992).

13.4 SOURCES OF ECOLOGICAL BIAS

We will now see how the ecological parameter estimates behave when the simple case of no confounding, no contextual effects, and no effect modification, as in Equation 13.4, does

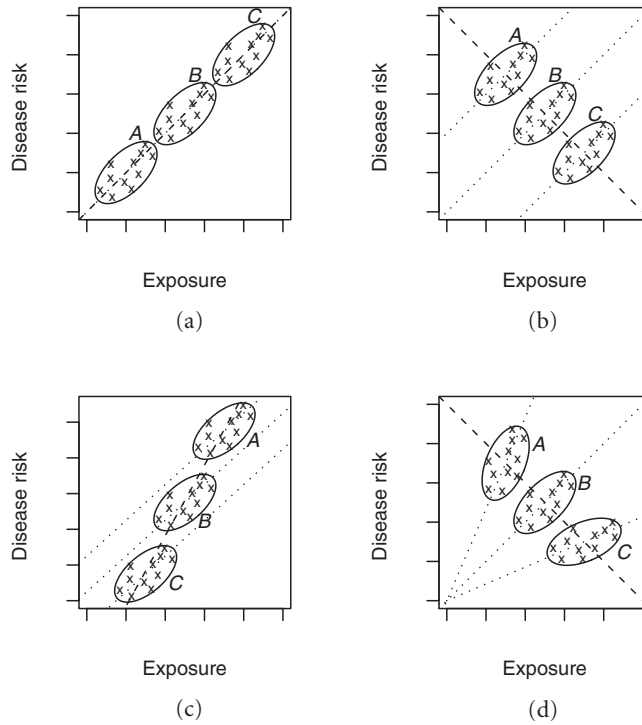


Figure 13.1. Types of ecological bias for a linear model. In each case, the dotted lines represent the relationship within areas, and the dashed line represents the ecological relationship. (a) No ecological bias. (b) Negative bias due to between-area confounding, within-area confounding, confounding by group, or contextual effects. (c) Positive bias due to between-area confounding, within-area confounding, confounding by group, or contextual effects. (d) Negative bias due to effect modification.

not hold. We will consider three situations: when unmeasured confounders are present, when contextual effects are present, and when the parameters β_{0i} , β_{1i} vary between areas. Throughout we assume a linear link.

Figure 13.1 illustrates the effect of different types of ecological bias for three areas, A , B , and C . In each case, the dotted lines represent the individual relationships between disease and exposure within each area, and the dashed line represents the ecological relationship. The figure is based on the individual model in Equation 13.1, with a linear link function; that is,

$$E[Y_{ij}|X_{ij}, X_i, Z_{ij}] = \beta_{0i} + \beta_{1i}(X_{ij} - X_i) + \beta_2 X_i + \gamma Z_{ij}.$$

Figure 13.1a illustrates the straightforward case where there is no confounding or contextual effect and the parameters do not vary between areas. In this case, the individual-level model is

$$E[Y_{ij}|X_{ij}] = \beta_0 + \beta_1 X_{ij},$$

and the ecological model is

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i.$$

In this case there is no bias, and the individual and ecological regression lines are equal.

Figure 13.1b and 13.1c illustrate two situations where there is bias. In both cases, the individual effect parameter is the same for each area (so the dotted lines are parallel), but the baseline risk (where each line crosses the Y axis) differs. As a result the ecological estimate of the exposure effect is biased. In case (b) the relationship between disease and exposure within areas is positive, but the ecological effect is negative (so that what is sometimes called a *protective effect* has been induced). In case (c) both relationships are positive, but the ecological effect is an overestimate of the individual effect.

These situations can arise for a linear model in one of four ways. Firstly, they can arise when the baseline risk parameter β_{0i} varies between areas, that is, the individual-level model is

$$E[Y_{ij}|X_{ij}] = \beta_{0i} + \beta_1 X_{ij}.$$

The ecological-level model becomes

$$E[Y_{ij}|X_i] = E[\beta_{0i}|X_i] + \beta_1 X_i,$$

so bias is caused when there is correlation between X_i and β_{0i} . For ease of explanation suppose

$$E[\beta_{0i}|X_i] = a + bX_i,$$

so that

$$E[Y_{ij}|X_i] = a + (b + \beta_1)X_i.$$

If the true individual effect is positive ($\beta_1 > 0$), a negative correlation ($b < 0$) will result in underestimating the effect parameter; in extreme cases ($b < -\beta_1$) this will cause a negative ecological effect as in case (b). A positive correlation ($b > 0$) will cause the ecological effect parameter to be greater than the individual effect, as in case (c). That the baseline risk appears to vary by area could be due to data anomalies in the population or disease counts.

Secondly, such bias can be caused by an unmeasured confounder, acting either between or within areas. For a between-area confounder the individual-level model is

$$E[Y_{ij}|X_{ij}, Z_i] = \beta_0 + \beta_1 X_{ij} + \gamma Z_i,$$

and so we have

$$E[Y_{ij}|X_i] = \beta_0 + \beta_1 + \gamma E[Z_i|X_i],$$

leading to bias as with the previous example. If we write $\beta_{0i} = \beta_0 + \gamma E[Z_i|X_i]$, then we have

$$E[Y_{ij}|X_i] = \beta_{0i} + \beta_1 X_i,$$

showing how the variation of baseline risk by area has been induced; thus β_{0i} and X_i will always be correlated (since Z_i is a between-area confounder), and bias will result as described above.

Similarly, for a within-area confounder the individual-level model is

$$\begin{aligned} E[Y_{ij}|X_{ij}, Z_{ij}] &= \beta_0 + \beta_1 X_{ij} + \gamma Z_{ij} \\ &= \beta_{0i} + \beta_1 X_{ij}, \end{aligned}$$

with $\beta_{0i} = \beta_0 + \gamma E[Z_{ij}|X_i]$. In this case, there will again be bias, since Z_{ij} and X_{ij} are correlated, leading to β_{0i} and X_i being correlated. It is possible for a within-area confounder Z_{ij} to be correlated with X_{ij} (since it is a within-area confounder) without the averages Z_i being correlated with X_i (that is, it need not also be a between-area confounder). Hence an unmeasured within-area confounder may not cause ecological bias, although often Z will be a confounder at both levels. If Z_{ij} is both a within-area confounder and a between-area confounder and if Z_i is measured and Z_{ij} unmeasured, then no bias will result with a linear model. This is not the case for a nonlinear model, where omission of a within-area confounder will lead to bias, even if the area-level confounder is measured.

Finally, this type of bias may be due to contextual effects, where the individual-level model is

$$\begin{aligned} E[Y_{ij}|X_{ij}, X_i] &= \beta_0 + \beta_1(X_{ij} - X_i) + \beta_2 X_i \\ &= \beta_{0i} + \beta_1 X_{ij}, \end{aligned} \tag{13.6}$$

with $\beta_{0i} = \beta_0 + (\beta_2 - \beta_1)X_i$, so that

$$E[Y_{ij}|X_i] = \beta_{0i} + \beta_1 X_i,$$

with β_{0i} depending on X_i . So a contextual effect acts in the same way as a between-area confounder, since β_{0i} and X_i will always be correlated and thus cause ecological bias. For this model we end up with

$$E[Y_{ij}|X_i] = \beta_0 + \beta_2 X_i, \tag{13.7}$$

so only the contextual effect can be estimated. If we rewrite Equation 13.6 in a form that is more familiar in the social sciences,

$$E[Y_{ij}|X_{ij}, X_i] = \beta_0 + \beta_1 X_{ij} + \beta_2^* X_i,$$

then we obtain

$$E[Y_{ij}|X_i] = \beta_0 + (\beta_1 + \beta_2^*)X_i,$$

illustrating that we are estimating the combined effects of individual and contextual effects in an ecological study.

Figure 13.1d illustrates a different source of bias, due to effect modification, that is, when the effect parameter β_{1i} is different in each area (dotted lines all have different slopes). Here the individual model is

$$E[Y_{ij}|X_{ij}] = \beta_0 + \beta_{1i} X_{ij}.$$

An individual study would calculate separate estimates for each effect parameter, corresponding to the slopes of the dotted lines; each of these effects is positive. However, an ecological study estimates the dashed line, and concludes that there is a protective effect.

This occurs because β_{1i} is negatively related to X_i ; here we have decreasing slopes with increasing exposure. In this example each area has the same baseline parameter β_0 , and so all three dotted lines cross at the same point on the Y axis (at $Y = \beta_0$).

If we have (say) $\beta_{1i}|X_i \sim N(\beta_1, \sigma^2)$ and a linear model, then an ecological analysis with a constant effect parameter across areas would provide an unbiased estimator of β_1 .

We now describe in more detail each of unmeasured confounding, contextual effects, and parameters that vary across areas.

13.4.1 Unmeasured Confounding

As described earlier, a confounder is a covariate that is related to both the outcome and the exposure (and is not on the causal pathway, and is not affected by the response). Suppose we have two populations in the age range 20–45, one that is exposed to air pollution and has a high rate of asthma, and the second that is unexposed and has a low rate of asthma. Suppose we also know that the exposed population contains more smokers than the unexposed population, and that smoking is associated with asthma. In this situation we do not know if there is a true association between air pollution and asthma or whether it is just due to differences in smoking behavior; smoking is said to be a confounder. In ecological studies confounders can act within areas, between areas, or both. In any observational study it is always possible that an observed association is due to unmeasured confounding, and (as already mentioned) much of analytical epidemiology is concerned with designs and analysis strategies that attempt to minimize bias due to confounding.

Bias due to confounding arises from omitting either within-area or between-area confounders from the model. Greenland and Robins (1994) give examples that demonstrate such biases; it is possible for the individual disease–exposure relationship to act in one direction, while the ecological data indicate a relationship in the opposite direction (this is illustrated in Figure 13.1b). A within-area confounder would also act as a confounder in an individual study. In a linear model, the aggregated within-area confounder will appear in the ecological model as an area-level confounder, unless the average exposure and the average confounder are uncorrelated across areas. As discussed by Greenland and Robins (1994), within-area confounding becomes more problematic when dealing with a nonlinear risk–exposure relationship, since within-area variability in confounders means that including a simple average value for an area is not sufficient to control for confounding. Instead we need to control for the *within-area confounder distribution* (there is a further increase in complexity when we have multiple confounders).

Between-area confounding is analogous to the usual confounding in an individual-level study, since the area is the level of analysis. Between-area confounders include covariates that represent characteristics of the area such as whether the area is urban or rural, or the average income in the area. In both these cases the confounder is often acting as a surrogate for the area average of important unmeasured individual-level characteristics. However, identifying such confounders may not be straightforward. Often it is the case that a variable is a confounder at both levels. If the variables are known, they may be included as between-area confounders in the ecological model 13.3. The specific form of the model should always be assessed, since, as always, there is no a priori reason why the effects should be additive. Unfortunately, checking the form of the ecological risk model is impossible when only ecological-level data are available (though a plausible form may be known from individual-level studies, in animal or man).

It is of theoretical interest to examine the likely size and direction of bias due to omitted confounders, and in practice methods for addressing the sensitivity to unmeasured

confounding are of interest (see Section 13.5.4). For the remainder of this section, we will assume that there are no contextual effects, and that parameters do not vary between areas.

In the model 13.1 confounding is introduced through the final term with $\gamma \neq 0$, and with Z_{ij} and X_{ij} correlated. From Equation 13.3 we can see that if the area averages for the confounder, Z_i , are known, they can be included in the model and the parameter estimates will be unbiased. This is not the case for other link functions, where typically the ecological model involves terms which are unobserved in practice (Wakefield, 2003). If the average area-level confounder is not included in the model, then the estimates of the parameters (β_0, β_1) will be biased. This bias will depend on the strength of the relationship between the confounder and the outcome (γ) and on the extent of dependence between the confounder and the exposure as measured through $E[Z_i|X_i]$.

The true individual model with no contextual effects or effect modification is

$$E[Y_{ij}|X_{ij}, Z_{ij}] = \beta_0 + \beta_1 X_{ij} + \gamma Z_{ij}. \tag{13.8}$$

Omitting the confounder gives the “true” ecological model:

$$E[Y_i|X_i] = \beta_{0i} + \beta_1 X_i + \gamma E[Z_i|X_i],$$

where β_1 is the effect parameter which is of interest. Suppose that the confounder, Z_{ij} , is binary. Then we can write this model in terms of the probabilities $q_x = P(Z_{ij} = 1|X_{ij} = x)$, that is, the distribution of the confounder given the exposure variable, assuming for simplicity that this relationship is constant across areas. So

$$E[Z_i|X_i] = q_0 + (q_1 - q_0)X_i,$$

and the ecological model becomes

$$E[Y_i|X_i] = (\beta_0 + \gamma q_0) + \{\beta_1 + \gamma(q_1 - q_0)\} X_i. \tag{13.9}$$

Suppose we fit the ecological model assuming that there is no confounding. So we obtain estimates of the parameters, β^* , from

$$E[Y_i|X_i] = \beta_0^* + \beta_1^* X_i. \tag{13.10}$$

Then from Equation 13.9,

$$\beta_0^* = \beta_0 + \gamma q_0, \tag{13.11}$$

$$\beta_1^* = \beta_1 + \gamma(q_1 - q_0). \tag{13.12}$$

Note that if $q_1 = q_0$ (so that X and Z are independent) or if $\gamma = 0$ (so Z_{ij} is not associated with Y_{ij}), there is no bias in the effect parameter β_1^* . Otherwise the effect parameter is biased by a component that depends on γ , the relationship between disease and confounder, and q_0, q_1 , the relationship between exposure and confounder. If the confounder is positively associated with both the disease and the exposure, then the bias will be positive and the estimator $\hat{\beta}_1^*$ will overestimate β_1 . If they are of opposite signs, then the bias in $\hat{\beta}_1^*$ will be negative; so if β_1 is positive, then it is possible for $E[\hat{\beta}_1^*]$ to be negative. The larger the true effect β_1 , the less likely this change of sign is to occur. Wakefield (2003) discusses the above and more general situations with a log link risk model and continuous exposures and confounders.

13.4.2 Contextual Effects

Contextual effects are area-level summary variables, such as the average exposure in an area, that affect the individual's outcome in addition to the individual-level variable. For example, an individual's health might be affected both by their own level of poverty and also by the general level of poverty in the area in which they live (sometimes such effects are known as neighborhood effects). In epidemiology contextual effects are often surrogates for combinations of unmeasured risk factors. In other disciplines contextual effects arise unambiguously; for example, in education the class IQ as well as individual IQ may be predictive of performance. In this example it is clear that the class IQ is a potentially relevant variable. In epidemiology, the area or neighborhood over which the contextual variable should be calculated is less clear.

Studies at the level of the individual can include a group average term in the analysis. However, at an ecological level it is not possible to distinguish between the effect of the term representing the aggregated individual variable and the contextual effect. This is illustrated by the comparison of the ecological regression and linear neighborhood models; see for example Chapter 1 and Equation 13.6, leading to Equation 13.7. This demonstrates that even if the interest is in the contextual effect $\beta_2 - \beta_1$, this cannot be estimated from the ecological data alone (Greenland, 2002).

As discussed above and illustrated in Figure 13.1, contextual effects can be considered as a special case of a between-area confounder (writing $Z_{ij} = X_i$), and so much of the discussion of between-area confounding above is applicable to contextual effects. The main difference is that no matter how many ecological data are available, the individual and contextual effects cannot be estimated separately. This is a fundamental difficulty of ecological inference, on which the social sciences literature concentrates.

13.4.3 Parameters That Vary between Areas

If one or more of the parameters vary between areas, in an individual-level analysis we can fit a separate model for each area or include area as a covariate in the model. However, ecological data do not contain enough information to estimate separate effect parameters for each area without imposing additional assumptions (King, 1997), since there are more parameters to be estimated than there are data points. If the baseline risk β_{0i} varies between areas, then the underlying risk for an unexposed individual will depend on the area in which the individual belongs. For example, unexposed individuals in different areas are at different baseline risk of asthma, due both to differences in unmeasured individual-level risk factors, and to true area effects such as different levels of health care in different areas. When the effect parameter β_{1i} varies between areas, the effect of being exposed is different for different areas. So, for example, the effect of air pollution on an individual's risk of asthma will depend on where they live.

When the baseline risk β_{0i} varies randomly (so that the coefficient is uncorrelated with X_i) due to unmeasured factors between areas, then it can be modeled as a random effect as discussed in Section 13.5.2. Such variation can arise through unmeasured variables that have no association with the exposure, or through data anomalies, again without association with the exposure. When the parameter β_{0i} varies systematically between areas, it is sometimes referred to as *confounding by group* (Greenland and Morgenstern, 1989). This can then be considered as a special case of an unmeasured between-area confounder, by writing $\beta_{0i} = \beta_0 + \gamma Z_i$. The inclusion of random effects cannot in general control for confounding.

Effect modification occurs when β_{1i} varies between areas. It may arise from a multiplicative interaction term at the individual level. This could be due to the presence of an unmeasured variable that changes across areas and has an interaction at the individual level, or to data anomalies that are associated with exposure (Greenland, 1992). Hence effect modification is distinct from confounding: a confounder is a nuisance variable which causes bias in the effect estimate (and can theoretically be controlled for), while effect modification is a property of the effect of interest.

In an ecological study, it might be thought that the ecological model would estimate the average effect parameter across all areas, that is, $\bar{\beta}_1 = E[\beta_{1i}]$. However, if confounding by group or effect modification which is dependent on X_i is present, this is not the case. Greenland and Morgenstern (1989) partition the ecological estimate into components due to confounding by group and to effect modification. Assuming no confounders and no contextual effects in the model 13.1 for simplicity, we have

$$E[Y_{ij}|X_{ij}] = \beta_{0i} + \beta_{1i}X_{ij}.$$

If we then fit the ecological model $E[Y_i|X_i] = \beta_0^* + \beta_1^*X_i$, we have (Greenland and Morgenstern, 1989)

$$\begin{aligned} \hat{\beta}_1^* &= \frac{\text{cov}(X_i, Y_i)}{\text{var}(X_i)} = \frac{\text{cov}(X_i, \beta_{0i} + \beta_{1i}X_i)}{\text{var}(X_i)} \\ &= E[\beta_{1i}] + \frac{\text{cov}(X_i, \beta_{0i})}{\text{var}(X_i)} + \frac{\text{cov}(\{X_i - E[X_i]\}X_i, \beta_{1i})}{\text{var}(X_i)}, \end{aligned} \quad (13.13)$$

using the identity

$$\text{cov}(XZ, X) = E[Z]\text{var}(X) + \text{cov}(X\{X - E[X]\}, Z).$$

The first term in Equation 13.13 is the average parameter across areas; $\hat{\beta}_1^*$ will be an unbiased estimator for the average effect $\bar{\beta}_1$ if the remaining terms are zero. The second and third terms can be viewed as bias components. The second term is attributable to confounding by group; it will be zero if $\beta_{0i} = \beta_0$, that is, if the baseline risk does not vary between areas, or if β_{0i} is uncorrelated with X_i . The third term is due to effect modification, and it will be zero if $\beta_{1i} = \beta_1$, or if β_{1i} is uncorrelated with X_i . So in a linear model it is when the area-specific parameters depend on the mean area exposure that problems arise.

It can be seen from this partition that the smaller the between-area variance in exposure means, $\text{var}(X_i)$, the larger the bias in the estimate $\hat{\beta}_1^*$, and so the bias is theoretically unbounded. Similar results can be obtained for other link functions, although in these cases there is an additional term representing pure specification bias. This additional term arises because of within-area variability in the exposure and is present even if parameters do not vary between areas. In these cases, bias can also occur due to effect modification even when β_{1i} is uncorrelated with X_i .

13.5 ISSUES IN ECOLOGICAL INFERENCE IN EPIDEMIOLOGY

13.5.1 Ecological Bias in Epidemiology

It is worth reiterating that determining causality in any observational study is problematic, since the existence of unmeasured variables (confounders) that induce bias in the observed

association can never be disproved. The interpretation of ecological results in the presence of unmeasured confounding is thus of central importance, and the approaches described in subsequent sections reflect this.

The environmental epidemiology literature in general is less concerned with contextual effects. Contextual effects in the exposure variable are less common in chronic disease epidemiology, in contrast to infectious disease epidemiology, in which an individual's risk of disease may depend both on personal immunity and on the immunity of those around. In noninfectious-disease epidemiology, contextual effects are more likely to occur in confounders (such as deprivation) than in environmental exposures. There is a large literature in social epidemiology, especially on the effect of living in an area of low socioeconomic status, beyond that of a person's own socioeconomic status (see Smith, 2000; Singh and Siahpush, 2002; and the references therein). Contextual effects are often a proxy for other unmeasured variables, but, as illustrated, contextual effects and individual-level effects cannot both be estimated if aggregate data only are available. In the example given above, the socioeconomic status of an area is considered a surrogate for other characteristics of the area or individuals within it. If all these characteristics were available, the contextual effect would disappear. Sheppard (2003) discusses various issues relating to the estimation of contextual effects in an environmental epidemiology context.

In epidemiology we would always expect effect modification to be present, but it is usual to assume that the variability in effects is small. Sufficient data to estimate area-specific effects are generally not available, since in a typical study diseases are rare.

A common assumption in environmental epidemiology is that the exposure effect is constant across both area and confounders. Stratifying the analysis by one or more confounders allows a separate effect for each confounder group. A major disadvantage is the unavailability of ecological data at the levels necessary; for example, stratification by age would require incidence rates and exposure variables for each age group in each area, and the latter are unlikely to be available. This is closely related to the consideration of mutual standardization, in which (for example) age-standardized disease rates must be regressed on age-standardized exposures; otherwise bias will result (Rosenbaum and Rubin, 1984). It is usually assumed that the exposure distributions are at least approximately constant across strata.

13.5.2 Overdispersion and Random Effects

Overdispersion occurs when the variance of the response exceeds that predicted from the model. Model-based standard errors will be inappropriate if the model does not allow for overdispersion. Overdispersion can arise for a variety of reasons, including the omission of important variables, errors in the data (including the response, the population counts, and exposures and confounders; see Wakefield and Elliott, 1999), and misspecifying the functional form of the mean. Often the first explanation will be the main source, and if overdispersion is found, it is an indication that variables associated with the outcome are unmeasured; if these variables are confounders, then estimators will be biased. The level of overdispersion can therefore be used as an informal indicator of the extent of unmeasured confounding, and a large value for the overdispersion parameter suggests that caution should be exercised when interpreting observed associations.

The introduction of random effects to represent the unexplained sources of variation between areas can help to address the problem of overdispersion by giving more appropriate standard errors, though it cannot control for unmeasured confounding. Wakefield (2004a) gives a more detailed discussion of the role that random effects play in an ecological study. Spatial as well as unstructured random effects may be included.

Data anomalies are an example of features that may be accommodated for using nonspatial random effects, and many unmeasured risk factors, such as environmental exposures, will have spatial structure. Besag, York, and Mollie (1991) originally introduced a model with both unstructured and spatial random effects, in the context of disease mapping, and Clayton, Bernardinelli, and Montomoli (1993) included such effects in an ecological regression setting. Following these authors, we write the residual relative risk (on the linear predictor scale) as $\delta_i = V_i + U_i$ in the model 13.1. The component V_i represents unstructured effects which are independent and identically distributed from some distribution, typically the normal. The component U_i represents spatially structured area-specific random effects which display dependence between U_i and $U_{i'}$, $i' \neq i$. The choice of this model is more difficult than for the independent random effects, and inference is much more likely to be influenced by the specific choice made. One possibility is a conditional autoregressive (CAR) model, with the limiting *intrinsic* form being a common choice. Richardson and Monfort (2000) offer a review of the use of Bayesian hierarchical models in an ecological setting, and include a description of this choice. Chapter 12 of this book describes their use to allow for spatial dependence between areas in a political science context. Clayton et al. (1993) state that the U_i terms are an attempt to control for “confounding by location.” The estimated regression coefficient may change from those obtained from a model containing nonspatial random effects only, and one never knows whether a genuine part of the exposure effect has been erroneously removed. Both estimates may be reported, and it is a judgment call whether the effect should be estimated from local or global exposure contrasts (corresponding to the inclusion and exclusion of spatial random effects, respectively).

Suppose that in the linear model there is a single unmeasured variable, Z . To illustrate how random effects might take account of unmeasured variables, suppose the ecological model is

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i + \gamma E[Z_i|X_i].$$

If X and Z are independent (so Z is not a confounder), then there is no bias in estimation of the effect parameter, and we have

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i + \gamma E[Z_i] \tag{13.14}$$

and

$$Y_i = E[Y_i|X_i] + \delta_i,$$

with $\delta_i = \gamma\{Z_i - E[Z_i]\}$. If Z_i is a confounder, then there will be bias in the estimator of the effect of X . Particular distributional assumptions for the random effects distribution correspond to different assumptions about the distribution of Z_i across areas.

An advantage of hierarchical models is that they allow strength to be borrowed from other areas in a structured way, thus smoothing rates which in areas with low populations may be highly unstable (Clayton and Kaldor, 1987). The assumption of constant baseline probabilities across areas is also avoided. As for any Bayesian approach, the choice of priors is important, particularly for the variance components in the random effects distributions and for spatial dependence parameters. Ecological studies are particularly sensitive to prior choice (Chapters 1 and 12 of this book). Since nonhierarchical models for ecological data have shown themselves to be highly sensitive to the choice of model, it is no surprise that hierarchical models behave similarly.

13.5.3 Plausibility

Causality cannot be proved from observational studies, but conclusions can be reached on the basis of *plausibility*. In epidemiology, rather than relying on a single study, evidence about a potential relationship is collected from a variety of different studies, each with its own strengths and weaknesses. Biological mechanisms and animal studies are also important. Thus evidence is gradually built up, and is combined in the hope of providing a consistent story. An obvious example is the causal relationship now known to exist between smoking and lung cancer. No single observational study provides incontrovertible evidence of this association, but many different observational studies in humans, and experimental studies in animals, have shown associations. One of the important early arguments was given by Cornfield et al. (1959) who used a sensitivity study to show that the strength of unmeasured confounding required to explain away the observed association was highly implausible. For a review of the health effects of smoking, see Doll (1998).

Another term for plausibility is *coherence*. See Rosenbaum (2002, Chapter 9), in which the author states, “A coherent pattern of associations is one that is, at each of many points, in harmony with existing knowledge of how the treatment should behave if it has an effect.”

One drawback of ecological studies is that if we do not obtain individual-level data, then we have no way of checking the form of the individual-level model. Conclusions will depend on this underlying model, but there is nothing in the ecological data to help us distinguish between competing explanations. Often, however, it can be argued that one explanation is more plausible than another. For example, ecological data cannot say whether a linear or a log-linear model is more suitable. However, information from other sources might suggest that the relationship is more likely to be multiplicative, and so a log link is the more appropriate choice. One of the advantages of taking the approach advocated in this chapter, of stating methods in terms of an explicit underlying individual-level model, is that it allows the explicit statement and critique of required assumptions.

13.5.4 Sensitivity to an Unmeasured Confounder

In this section we describe a sensitivity analysis approach to address the problem of unmeasured confounding, illustrating the idea of plausibility discussed in the previous section. Rothman and Greenland (1998) provide background on the use of sensitivity analyses in general in epidemiology, and Wakefield (2003) specifically in the area of ecological studies. The basic idea can also be applied to other situations, such as sensitivity to contextual effects, pure specification bias, spatial dependence, classification errors, and selection bias. Unfortunately, there is no information in the data about the extent of bias from these sources, though external sources can provide invaluable information. For example, local knowledge can inform us on the accuracy of population counts and of disease registries. If the observed association is large, then, as we shall see, it is more difficult to explain it away with unmeasured confounding.

Here we concentrate on sensitivity to an unmeasured confounder. The approach for an ecological analysis is as follows. We start from the proposed individual-level model, which includes the unmeasured confounder; for the purposes of the sensitivity analysis we assume that this model is correct, and so includes the “true” parameter. We derive the appropriate ecological level model and compare this with the fitted model to give an expression for the estimated parameter in terms of the “true” parameter. We can then use this expression with various assumptions about the unobserved confounder to give estimates of the “true” parameter. Finally, we consider how plausible such assumptions are. For example,

we may conclude that unmeasured confounding cannot plausibly explain a result (as with the smoking–lung-cancer example).

Suppose we fit the simple model 13.10, assuming no confounding, contextual effects, or effect modification:

$$E[Y_i|X_i] = \beta_0^* + \beta_1^* X_i,$$

and obtain an observed effect estimate $\hat{\beta}_1^*$. Suppose, however, that there is evidence of overdispersion; this suggests that unmeasured confounding is present and the true individual model is Equation 13.8:

$$E[Y_{ij}|X_{ij}, Z_{ij}] = \beta_0 + \beta_1 X_{ij} + \gamma Z_{ij},$$

with corresponding ecological model

$$E[Y_i|X_i] = \beta_{0i} + \beta_1 X_i + \gamma E[Z_i|X_i].$$

From Section 13.4.1 (Equation 13.12) we have

$$\beta_1^* = \beta_1 + \gamma(q_1 - q_0).$$

We know that $\hat{\beta}_1^*$ is a potentially biased estimator of β_1 ; we are interested in the extent to which this is a real relationship rather than due to ecological bias. In particular:

- Is the association of sufficient size that it cannot be plausibly explained away by ecological bias?
- Can we obtain an approximate adjusted estimate (or a range) of relative risk?

We use the expression for the bias given above to examine these questions.

We will consider a simple artificial example, comparing the incidence of respiratory diseases in boys under five and the average air pollution (low or high) for a set of areas. We let $Y_{ij} = 0(1)$ denote respiratory-disease nonincidence (incidence), and $X_{ij} = 0(1)$ represent unexposed (exposed), for the j th individual in area i . Hence Y_i and X_i are the average disease incidence and exposed, respectively, in area i . Across areas, the incidence rates vary between 0% and 24%, and the proportion exposed to high air pollution varies between 6% and 99%. This example should be viewed as merely illustrative, for studies such as this are far more complicated (for example, involving more than two levels of the primary exposure of specific interest and multiple secondary exposures).

We fit the basic ecological model 13.10 and obtain

$$E[Y_i|X_i] = 0.06 + 0.04 X_i,$$

so $\hat{\beta}_1^* = 0.04$. If we accept this result at face value, we have a risk difference of 0.04, or a relative risk of 1.6; so this evidence suggests that a child exposed to high air pollution has a 60% greater risk of respiratory disease than an unexposed child. However, we do not believe here the assumption of no confounding; there are many potentially important missing variables, such as genetic components and child, parent, household, and lifestyle characteristics. An obvious confounder that has been overlooked is poverty (which is a surrogate for lifestyle and behavioral characteristics and is related to many diseases). We are interested in knowing if air pollution really causes the observed increased incidence of asthma, or whether it is due to the differing levels of poverty in the study.

Table 13.1 Sensitivity analysis results

		θ				
		1.1	1.2	1.5	2	4
γ	0.1	–	–	–	0.40	0.13
	0.2	–	–	0.40	0.20	0.07
	0.5	0.80	0.40	0.16	0.08	0.03

Note: The strength of the linear relationship between disease and confounder is represented by γ , and q_0 and q_1 are the probabilities of poverty for unexposed and exposed individuals, respectively, that is, $q_x = P(Z_{ij} = 1 | X_{ij} = x)$, $x = 0, 1$. For given γ and $\theta = q_1/q_0$, the table shows the necessary value of q_0 for a confounder to explain away the observed association. Inadmissible q_x , in which probabilities outside of (0, 1) are obtained, are shown as dashes.

If the observed effect were entirely due to poverty and not to air pollution, then the true effect would be $\beta_1 = 0$. Substituting $\beta_1 = 0$ and $\beta_1^* = 0.04$ in Equation 13.12 gives

$$0.04 = \gamma(q_1 - q_0).$$

We will look at a range of possible values for γ , q_0 , and q_1 which satisfy this expression. In this example both γ and $q_1 - q_0$ must have the same sign (since we know their product is positive). We will assume that they are both positive, so the confounder, poverty, is more likely among those exposed to air pollution, and the confounder is positively associated with respiratory disease risk (which, as just stated, is typical of a variable such as poverty). We will also write $q_1 = \theta q_0$; so θ is the relative risk of poverty for exposed individuals relative to unexposed, so that θ represents how much more prevalent the confounder is among exposed than unexposed individuals. This reformulation aids interpretation of the sensitivity analysis; however, it should be noted that θ must be constrained so that both q_0 and q_1 lie between 0 and 1.

For a range of values of γ and θ , Table 13.1 illustrates some possible values that q_0 could take for us to observe a coefficient of 0.04, when the real coefficient is 0; inadmissible solutions (where the probabilities q_x are not between 0 and 1) are shown as dashes. Values for γ represent a range of beliefs about the strength of the relationship between disease and confounder on an additive scale.

This table gives some idea of the characteristics of a missing confounder that would be wholly responsible for the observed effect. In this example, the observed risk difference of 0.04 is less likely to be caused solely by a relatively weak confounder (with $\gamma = 0.1$), since, for example, this would require the probability of poverty for an unexposed individual to be 0.13, and the probability for an exposed individual to be four times as likely, that is 0.52. If a moderate unmeasured confounder ($\gamma = 0.2$) is responsible for the association between respiratory disease and air pollution, it would need to be around 1.5–4 times more prevalent among exposed children than among unexposed children. A stronger confounder (with $\gamma = 0.5$) would need very little difference in poverty between exposed and unexposed groups.

The final step is to interpret this in terms of the study and form a conclusion as to how likely it is that such a confounder exists. We know from other studies that poverty has a

reasonably strong effect on most diseases, at least compared to most environmental sources of pollution (Carstairs, 2000). If we assume that respiratory diseases follow a similar pattern, then it is likely that poverty is at least a moderate confounder. We then have to consider how much more prevalent we would expect poverty to be among children exposed to high air pollution than among those not exposed; this is measured by θ . It depends very much on the study design. If, for example, the main sources of air pollution for the study area were major locations of heavy industry, then we would expect poorer areas to be close to the industry, and $\theta = 1.5$ or higher would be reasonable. In this case we would conclude that although our study suggests a link between pollution and asthma, it can easily be explained by unmeasured confounding. If the source of air pollution in the study was pollution from proximity to major roads, depending on the area in which the study is based, we may expect less difference in poverty between exposed and unexposed areas. In this hypothetical example we would conclude that the true effect is unlikely to be as great as was observed.

If, as is always the case in practice, there are multiple confounders, then it is far easier to create plausible scenarios which explain away observed associations. Of course, if such confounders are negatively associated with either disease or exposure, they could also be masking a true association. Such issues lead naturally into the planning of a study that is carefully designed to examine this relationship. If only small risk differences are envisaged in an ecological study, then the study should not be carried out, since biases due to within-area variability in exposures and confounders, and pure specification bias, are likely to dominate the observed association.

Approaches to sensitivity in the same spirit have been considered in the social sciences, but in a less formal manner. See for example, Flanigan and Zingale (1985) and Achen and Shively (1995 Chapter 8).

13.6 ISSUES IN ECOLOGICAL INFERENCE IN SOCIOLOGY AND POLITICAL SCIENCE

In the previous section we have described a number of issues relating to ecological inference in epidemiology, with an emphasis on spatial epidemiology and in particular on issues of confounding. We have kept this discussion as general as possible; a link function enables a range of choices for suitable models, and exposures and confounders may be either discrete or continuous.

In this section, we look at how the issues in epidemiology discussed above fit in the wider picture and compare the results with those in other disciplines. We consider some of the specific concerns of sociology and political science and see how they relate to the model that we have described, and we demonstrate the links between these approaches and those in epidemiology. We also identify differences between the disciplines.

13.6.1 Ecological Inference for 2×2 Tables

In the social sciences data often consist of discrete outcome and predictor variables, and so the ecological data consist of a series of cross-classified data that may be represented by a set of 2×2 , or more generally, $r \times c$ tables, as described in the Introduction to this book. In this chapter we have used slightly different notation, concentrating on the underlying probabilities rather than the unobserved cell entries. Table 13.2 establishes notation for the data on the left, and for the underlying probability model on the right; the ecological data consist of the margins of the left table only. The left-hand table may be compared to Table 0.1 in the Introduction; here we have represented the actual numbers in each cell, rather than the proportions. The proportions β_i^b and β_i^w defined in the Introduction correspond to the fractions $n_{11i}/\{N_i X_i\}$ and $n_{01i}/\{N_i(1 - X_i)\}$, respectively, in our notation.

Table 13.2 Cell counts for the individual data (left) and the underlying probabilities (right) for a generic 2×2 table in area i

		Y		
		0	1	
X	0	n_{00i}	n_{01i}	$N_i(1 - X_i)$
	1	n_{10i}	n_{11i}	
		$N_i - Y_{i+}$	Y_{i+}	N_i

		Y		
		0	1	
X	0	$1 - p_{0i}$	p_{0i}	$1 - \pi_{xi}$
	1	$1 - p_{1i}$	p_{1i}	
		$1 - q_i$	q_i	1

We now focus on the underlying marginal probabilities in the case of a binary exposure $P(Y_{ij} = 1|X_{ij} = x) = p_{xi}$, where i indexes areas, $i = 1, \dots, m$; j indexes individuals within areas, $j = 1, \dots, N_i$; and $x = 0, 1$ (these probabilities are marginal because we have averaged over contextual effects and confounders). The ecological model for the disease rate Y_i in terms of these probabilities is given by

$$\begin{aligned}
 E[Y_i|X_i] &= P(Y_{ij} = 1|X_{ij} = 0)P(X_{ij} = 0|X_i) + P(Y_{ij} = 1|X_{ij} = 1)P(X_{ij} = 1|X_i) \\
 &= p_{0i} + (p_{1i} - p_{0i})X_i;
 \end{aligned}
 \tag{13.15}$$

that is, a linear ecological model with intercept p_{0i} and slope $p_{1i} - p_{0i}$. The ecological relationship between disease counts Y_{i+} and proportion exposed X_i will be linear regardless of the form of the individual model, even if confounding or contextual effects are present. In the latter case, although the model is still of this form, the difference is in interpretation and imputed cell entries will depend greatly on whether contextual effects are assumed to be present or absent.

The ecological model 13.15 is in terms of the probabilities (p_{0i}, p_{1i}) , rather than the regression parameters (β_{0i}, β_{1i}) which we have focused upon in previous sections. However, these parameters are related. In the case of a linear link function we have

$$\begin{aligned}
 p_{0i} &= \beta_{0i} + (\beta_2 - \beta_{1i})X_i + \gamma E[Z_{ij}|X_{ij} = 0], \\
 p_{1i} &= \beta_{0i} + \beta_{1i} + (\beta_2 - \beta_{1i})X_i + \gamma E[Z_{ij}|X_{ij} = 1].
 \end{aligned}
 \tag{13.16}$$

In general the relationship between the probabilities and the regression parameters is not straightforward, depending on contextual effects and on the relationship between the confounder and the predictor X . The interpretation of the probabilities is complicated when confounding or contextual effects are present. For a linear link function, and in the absence of additional variables, Z (so that $\gamma = 0$), and contextual effects (so that $\beta_2 = \beta_{1i}$), the probabilities and parameters are simply related by

$$\begin{aligned}
 p_{0i} &= \beta_{0i}, \\
 p_{1i} &= \beta_{0i} + \beta_{1i}.
 \end{aligned}
 \tag{13.17}$$

13.6.2 Fractions versus Probabilities

In this chapter we have so far concentrated on estimating the parameters of a hypothetical model in which there is an infinite population within each area, in which case these probabilities correspond to the proportion of category x whose response is $Y = 1$. Many

methods in the political science literature – for example, the method of bounds (Duncan and Davis, 1953) and King’s EI method (King, 1997) – are concerned with estimating (*predicting*) the unobserved cell entries (or equivalently, the fractions β_i^b and β_i^w as defined in the Introduction). Which of these is of interest will depend upon the particular application; it is important to distinguish between the two, however, since they are not interchangeable. We note, however, that building a causal model will generally aid in producing a good predictive model. In an individual-level study where the proportions are observed, they can be used as estimates of the underlying probabilities ($\hat{\beta}_i^w = \hat{p}_{0i}$ and $\hat{\beta}_i^b = \hat{p}_{1i}$), and if the number of individuals in each group-by-area margin is large, these estimates will be accurate, regardless of the existence of contextual effects and confounding. By contrast, ecological estimates of the fractions will only be accurate under very strict conditions.

Estimating the unobserved proportions is a missing-data–imputation problem. The fractions are of interest when the actual numbers in the table are required – for example, in court cases concerning voting rights of minorities (see, for example, Freedman et al., 1991), or in a public health context where, for example, the actual numbers of elderly people with a disease in an area might be required in order to determine allocation of health resources). In these cases, the data in the table represent the entire population of interest; we are only concerned with the individuals eligible to vote in that specific election, or with the diseased individuals in that specific public health area. If the missing data were available, we would report the numbers in the table, and would not typically be interested in further statistical analysis.

The underlying probabilities are of interest when we are concerned with examining causal relationships between variables. In this case, the data in the table represent samples, and we wish to extrapolate to a wider population. For example, if a study of air pollution and asthma is conducted in a particular study area, we will generally be interested in applying the conclusions to a wider region. In this situation, if we had the individual data, we would model the observed data as a function of exposure and confounders in order to obtain estimates (with associated interval estimates) of the risk attributable to exposure.

Whether predictive or causal inference is required, the use of a causal model in which variation is modeled in terms of the primary predictor X , confounders, and contextual effects is likely to be advantageous. In many areas of political science the fractions have traditionally been taken as the primary target of interest. In historical voting studies, such as determining voting patterns for Hitler’s National Socialist German Worker’s Party in 1930 (Hamilton, 1982) we may want to know the probabilities in an underlying model for political theory, to examine how different demographic, religious, and occupational groups were voting. Usually in epidemiology and sociology underlying causal relationships are of interest, and these may be addressed by estimating regression coefficients in a probabilistic model. We reiterate that most applications would benefit from thinking in terms of an individual-level model, since this allows one to think about variables that may be distorting relationships that are of interest.

13.6.3 Probabilities That Vary between Areas

In social science applications the probabilities vary between areas, and this must be acknowledged to obtain accurate area-level estimates. (When estimating average causal relationships, however, we may not need to acknowledge such variability under certain assumptions, such as a linear model with randomly varying coefficients; see Equation 13.13.) However, without additional data or assumptions, it is not possible to estimate separate probabilities for each area, since we have $2m$ quantities of interest and just m observed data points. In this section

we see how nonconstant probabilities arise as a result of the sources of bias that we have considered.

From Equation 13.16 we can see that there are three ways in which the probabilities p_{0i} , p_{1i} may vary between areas:

- if one or both of the parameters (β_{0i}, β_{1i}) vary between areas;
- if contextual effects are present;
- if there is unmeasured confounding.

The last two ways correspond to the conditions given by Firebaugh (1978) for *cross-level bias* to be present. If contextual effects or confounding is present, this will result in the probabilities being dependent on *context*; in general, we might expect probabilities to vary as a result of all three causes.

Probabilities that depend on context have been modeled in sociology and political science by

$$\begin{aligned} p_{0i} &= a_0 + b_0 X_i, \\ p_{1i} &= a_1 + b_1 X_i. \end{aligned} \tag{13.18}$$

We will show how this assumption corresponds to different assumptions about contextual effects and confounding in the general individual model with a linear link function, that is,

$$E[Y_{ij}|X_{ij}, X_i, Z_{ij}] = \beta_{0i} + \beta_{1i}(X_{ij} - X_i) + \beta_2 X_i + \gamma Z_{ij}.$$

For this model, as derived previously, the marginal probabilities are given by

$$\begin{aligned} p_{0i} &= \beta_{0i} + (\beta_2 - \beta_{1i})X_i + \gamma E[Z_{ij}|X_{ij} = 0], \\ p_{1i} &= \beta_{0i} + \beta_{1i} + (\beta_2 - \beta_{1i})X_i + \gamma E[Z_{ij}|X_{ij} = 1] \end{aligned}$$

(from Equation 13.16).

Suppose that we have constant regression probabilities, so that $\beta_{0i} = \beta_0$ and $\beta_{1i} = \beta_1$ (and that the contextual effects of interest are $\beta_2 - \beta_1$). Then the marginal probabilities in terms of the regression parameters are given by

$$\begin{aligned} p_{0i} &= \beta_0 + (\beta_2 - \beta_1)X_i, \\ p_{1i} &= \beta_0 + \beta_1 + (\beta_2 - \beta_1)X_i. \end{aligned}$$

Here it is clear that if contextual effects are present, both probabilities will vary linearly with X_i . This corresponds to the model 13.18 with

$$\begin{aligned} a_0 &= \beta_0, \\ a_1 &= \beta_0 + \beta_1, \\ b_0 &= \beta_2 - \beta_1, \\ b_1 &= \beta_2 - \beta_1. \end{aligned} \tag{13.19}$$

The *extended ecological regression* model given by Equation 13.18 is discussed extensively by Achen and Shively (1995). At the ecological level its use leads to

$$q_i = a_1 + (a_0 + b_1 - a_1)x_i + (b_0 - b_1)x_i^2,$$

and so an additional constraint must be imposed for identifiability. A common choice is to set either b_0 or b_1 to zero. Here, we see that the model we have derived in terms of the regression parameters (as summarized in Equation 13.19) corresponds to the alternative constraint that $b_0 = b_1$, and to an assumption that the contextual effect is the same for both exposure groups (and this parameter is not identifiable from the quadratic model).

The identifiability constraint should be chosen by consideration of what is appropriate for the data in hand, rather than on tractability of the model. For example, in an epidemiology context in which X represents poverty (with $X = 0$ and 1 representing poor and nonpoor), the constraint $b_0 = 0$ means that a poor individual's risk does not depend on the average poverty in the area, while $b_1 = 0$ means that a nonpoor individual's risk does not depend on the average poverty in the area. Carrying out analyses with several values of b_0 and b_1 has the same flavor as the sensitivity analyses described in Section 13.5.4.

The marginal probabilities may also vary as a result of an unmeasured confounder. Assuming that $\beta_{0i} = \beta_0$ and $\beta_{1i} = \beta_1$, we have

$$\begin{aligned} p_{0i} &= \beta_0 + \gamma E[Z_{ij}|X_{ij} = 0], \\ p_{1i} &= \beta_0 + \beta_1 + \gamma E[Z_{ij}|X_{ij} = 1]. \end{aligned}$$

These probabilities will be correlated with X_i if the values of $q_x = E[Z_{ij}|X_{ij} = x]$, $x = 0, 1$, are correlated with X_i . For example, if q_0, q_1 are linearly related to the mean exposure X_i , then

$$\begin{aligned} E[Z_{ij}|X_{ij} = 0] &= c_0 + d_0 X_i, \\ E[Z_{ij}|X_{ij} = 1] &= c_1 + d_1 X_i. \end{aligned}$$

This gives Equation 13.18 with

$$\begin{aligned} a_0 &= \beta_0 + \gamma c_0, \\ b_0 &= \gamma d_0, \\ a_1 &= \beta_0 + \beta_1 + \gamma c_1, \\ b_1 &= \gamma d_1. \end{aligned}$$

Confounding is not often discussed explicitly in the sociology and political science literature. The general term *specification bias* is sometimes used in the sociology literature to refer to incorrect specification of the individual model; an aggregate model which ignores within-area confounding is a particular instance of this in which the model is incorrect because it omits important covariates. Between-area confounders may arise as a result of the way groups are formed. Thus between-area confounders can be seen in terms of *aggregation bias* in the sociology literature (for example, Langbein and Lichtman, 1978), in which the allocation of people into areas may depend on the response, the exposure, or both, possibly through the effect of other variables.

Although the sociology and political science literatures are much concerned with probabilities that vary between areas, an investigation into this variation is not the usual approach. Thus, although the underlying reasons for varying probabilities may differ, the method of analysis will be the same. This is in contrast to the approach in epidemiology (for example, Greenland and Morgenstern, 1989), where the source of variation is identified and influences the choice of approach. All three types of bias cause the probabilities to vary between

areas, but the implications for analysis are different. For example, Section 13.4.3 showed that effect modification cannot be removed by controlling for confounders. Hence it is beneficial to identify the reasons for the variability in probabilities across areas.

13.7 RELATIONSHIP BETWEEN MODELS IN EPIDEMIOLOGY AND SOCIAL SCIENCE

A common approach to inference in the social sciences is *ecological regression* (Goodman, 1953, 1959). We let \tilde{p}_{xi} denote the fractions responding for $X = x$. If the fractions are constant in expectation ($E[\tilde{p}_{0i}|X_i] = \tilde{p}_0$, $E[\tilde{p}_{1i}|X_i] = \tilde{p}_1$), which would arise if the underlying probabilities were common across areas ($p_{0i} = p_0$, $p_{1i} = p_1$), then

$$E[Y_i|X_i] = \tilde{p}_0 + (\tilde{p}_1 - \tilde{p}_0)X_i. \tag{13.20}$$

This model now has only two parameters, and can be fitted with ecological data. This is Equation 13.4 with $\beta_0 = \tilde{p}_0$ and $\beta_1 = \tilde{p}_1 - \tilde{p}_0$; in the epidemiology literature the parameters may also be written in terms of the relative risk (as we did in Equation 13.5). Goodman discussed fitting this model, conditional on many caveats, to obtain $\tilde{p}_0 = \hat{\beta}_0$ and $\tilde{p}_1 = \hat{\beta}_0 + \hat{\beta}_1$; Achen and Shively (1995) give expressions for the standard errors of these estimates.

A least squares approach to estimation in this model is often used, which implicitly assumes that the variance is constant; as noted earlier in Section 13.3, the true variance is nonconstant. While this has been considered, it is of secondary importance compared to other assumptions such as the existence of contextual effects. Achen and Shively (1995) argue that more sophisticated models allowing for nonconstant variance are not of practical importance, since over the typical ranges for a political science application the variances of \tilde{X}_i and Y_i are similar and vary so little as not to be a problem. In epidemiology diseases are typically rare and studies are based on small counts. In such cases, assuming a constant variance will give very poor estimates of the standard errors. To remedy this the log disease rate may be regressed on X_i , or (preferably) a Poisson log-linear model may be used (as described in Chapter 1); see Richardson and Monfort (2000) for further details.

The most serious drawbacks of ecological regression are the assumptions of constant probabilities across areas and of the absence of contextual effects. The former is unrealistic in practice because for most applications we expect demographic and area characteristics to modify the probabilities. In terms of the model 13.1, ecological regression means assuming that there are no contextual effects, no unmeasured confounders, and no effect modification, and these are again implausible assumptions in most situations. Goodman (1959) was aware that the constancy assumption would not be valid in general, but suggested that the method might be appropriate when the expected values of \tilde{p}_{0i} and \tilde{p}_{1i} are constant, and \tilde{p}_{0i} and \tilde{p}_{1i} do not systematically vary with X_i (this will occur if the parameters β_{0i} , β_{1i} vary randomly across areas), although predictions for particular areas may still be poor. This is consistent with Section 13.4.3, where we demonstrated that we have an unbiased estimate of the average $E[\beta_{1i}]$ if the parameters β_{0i} and β_{1i} (and hence the probabilities p_{0i} and p_{1i}) are uncorrelated with X_i . In such a case we would be estimating the average marginal probabilities p_{0i} , p_{1i} across areas. It is well known that least squares has robust estimation properties for regression parameters but is poor for prediction (since the distribution of the error terms is needed for this).

Freedman et al. (1991) proposed an alternative model, the (nonlinear) neighborhood model with the assumption that there is no difference in the two probabilities of disease in each area; that is, $p_{0i} = p_{1i} = q_i$. This corresponds to an assumption of no exposure effect,

that is, $\beta_1 = 0$. The marginal probability q_i may vary due to unmeasured characteristics, summarized at the area level. A special case, the linear neighborhood model, allows this common probability to vary between areas depending on the average exposure, that is, $q_i = a + bX_i$. The probability q_i may vary due to contextual effects or between-area confounding (with confounding by group being one potential explanation). The resulting ecological model is

$$E[Y_i|X_i] = a + bX_i,$$

which is indistinguishable from ecological regression (Equation 13.20), but the interpretation of the coefficients is very different. Freedman's assumption corresponds to assuming that there is no individual exposure effect, but that apparent differences in probabilities between areas are due to X_i .

The linear neighborhood model is not generally used in practice, since its assumptions are even more restrictive than Goodman's regression; indeed, this model was initially proposed to discredit Goodman regression by demonstrating that a different assumption (which is uncheckable from the data alone) gives rise to a different conclusion, but an identical ecological mean model. This further illustrates the fundamental difficulty in ecological analyses: assumptions are crucial and can drastically affect the conclusions of a study, and are often uncheckable from the ecological data alone. In Sections 13.5.3 and 13.5.4 we stressed the importance of choosing assumptions based on context and checking their importance via a sensitivity analysis.

In epidemiology, hierarchical models may be used to deal with spatial and residual variation, as described in Section 13.5.2. In the political science literature, King (1997) proposed the *ecological inference* (EI) method, a particular form of hierarchical model that addresses the problem of probabilities that vary between areas. Wakefield (2004b) this book, Chapter 1) describes the general use of hierarchical models. In the basic model it is assumed that the disease probabilities p_{0i} , p_{1i} are independent of X_i , so there are no contextual effects or unmeasured confounding, and probabilities vary only due to random effect modification (that is, they do not depend on exposure). Expressed in terms of the three types of ecological bias, it can be seen that this is a strong assumption, and substantial bias may arise if it is violated (e.g. Cho, 1998).

King's EI method is popular among political scientists because it provides estimation in the presence of random effect modification, and user-friendly software is available. However, it may produce poor estimates if confounding, contextual effects, or structured effect modification are present (Cho, 1998). The method can be extended to incorporate confounders (King, 1997: Chapter 9; this book, Introduction); the fractions \tilde{p}_{0i} , \tilde{p}_{1i} are regressed on an area-level variable Z_i . One specific case often considered is when $Z_i = X_i$; that is, the probabilities depend on the average exposure. However, strong prior information is required for stable information whenever such regressions are carried out (Wakefield, 2004b). The problem is also avoided if additional individual-level data are available.

An approach that has been taken (see for example, the references in Herron and Shotts, 2004), but is incorrect, is to obtain estimates of p_{0i} , p_{1i} and then to regress these on area-level variables, in a two-stage approach. An analogous approach in epidemiology would be to control for confounding variables, and then to regress area-level relative risks upon area-level variables of interest. The problem is that the effect of the latter variables may be distorted unless the stratification variables are independent of the variables of interest (that is, are not confounders). For example, suppose we wish to investigate the effect of an environmental exposure, and wish to control for the confounder, poverty (and for the

sake of exposition we assume that individuals in areas of high poverty are more likely to receive high exposure). If we control *a priori* for poverty using data from the study region (via internal standardization), then we will have removed some of the effect of exposure, and we then will overestimate the effect of poverty. In epidemiology external rates for the stratification variables are often used to avoid this problem, or simultaneous estimation of the exposures and confounders is carried out. This issue is closely related to that of mutual standardization.

13.8 CONCLUSIONS

In this chapter we have summarized some of the issues that are relevant to ecological inference in epidemiology, and shown how these relate to work in sociology and political science. Although the motivations for ecological studies differ, the approaches have much in common when one translates the different notations and languages used. The social sciences literature does not generally state a model at the level of the individual. By specifying a common framework, we have seen how different models actually correspond to different sets of underlying assumptions, and have identified similarities between approaches.

We have identified three main sources of bias: parameters that vary between areas, the presence of contextual effects, and the presence of confounding. Each discipline deals with these considerations in different ways and with different emphasis and terminology; however, we have seen that there is substantial overlap. In particular, although confounding is not explicitly considered in sociology and political science, we have seen how one of the main concerns, probabilities varying between areas, can be naturally interpreted as due to unmeasured confounding. We can thus borrow ideas from epidemiology to help deal with nonconstant probabilities from this source, by attempting to control for confounders.

There are differences in the context and the form of the data in different disciplines. In epidemiology the rarity of diseases allows a log-linear model to be used, which is more tractable than a logistic form. The sparsity of cases means that effect modification is rarely considered, and hierarchical models are often used for stable estimation. Nonrare outcomes obviously provide more information, but following the individual modeling approach described here is more difficult when using the logistic model. Another important difference is the emphasis on causality in epidemiology. Prediction does not require an explicit causal model, and in political science, when the interest is in the actual numbers of people voting, the problem becomes one of imputation of missing values, rather than estimation of underlying parameters. However, a modeling approach will often be beneficial for prediction. The extensions to continuous variables, which are more common in epidemiology, are not necessarily of use for political science problems, although they may be of some interest in sociology.

In sociology and political science, the focus on the unobserved fractions may obscure consideration of an underlying model, and does not explicitly allow identification of the reasons for variation between areas. Nonconstant probabilities can be due to contextual effects or to confounding or varying parameters. The main difference in the underlying approach to ecological inference is that political science is concerned with capturing the variation between areas, but is less interested in the actual source of the variation. On the other hand, epidemiology attempts to model the actual source of variation; this is better for making predictions for individuals in unobserved areas with particular exposure distributions, which is of more interest when the emphasis is on causality. There are other advantages to examining these different sources separately; for example, a method that reduces bias from confounding may not reduce bias from effect modification (Greenland and Morgenstern, 1989).

Additionally, since a major concern in epidemiology is controlling for potential confounders, it is important to be able to isolate bias from this source.

Assumptions in ecological inference are crucial. Often restrictive assumptions are necessary to be able to carry out an analysis, and ecological inference is highly sensitive to such assumptions. Since these assumptions cannot typically be checked from ecological data alone, we emphasize the idea of drawing conclusions on the basis of consistency of results across different modeling assumptions. In epidemiology, ecological inference is considered more plausible if consistent across different studies (which may correspond to different areas or different time periods) and if supported by biological mechanisms. In the social sciences, having different study areas with different distributions across the grouping variable is desirable, in particular where the proportions in each group, x_i , are nearly uniformly distributed across the interval (0, 1). When cell counts are of interest, studies in different areas are still useful for consistency arguments. A related idea is that of a sensitivity analysis, such as that described in Section 13.5.4, which enables us to see how sensitive results are to different assumptions and provide some insight into the importance of possible biases and the uncertainty of results.

In this chapter we have concentrated on highlighting areas in epidemiology that offer benefits to sociology and political science, in particular the specification of explicit models. However, there is much to be gained in the other direction. The study of contextual effects is becoming of increasing interest in epidemiology, although as yet very little consideration has been given to contextual effects in an ecological model in environmental epidemiology. This is one area where epidemiology can benefit from the existing work in sociology. Another such area is that of effect modification. Although rare diseases and small areas frequently mean that effect modification cannot be studied, it may be possible in some studies with nonrare diseases, such as asthma, to take advantage of the current work in this area. The simplifications when dealing with a single binary exposure (for example, the ecological regression model) are not widely exploited in epidemiology.

Hierarchical models have proved useful in all disciplines. They provide a flexible way of incorporating assumptions and prior knowledge into the analysis, allow probabilities to vary between areas (including the possibility of spatial variation), and can easily incorporate the explicit modeling of observed confounders. In particular the choice of model can be tailored to the particular study; the routine use of any single model in all situations is not a good strategy.

Not all approaches in one discipline are suitable for use in another. However, the problems are sufficiently similar that there is much to be gained by being aware of work in different areas. Identifying links between one discipline and another is not straightforward; this chapter has concentrated on making some of these links explicit through the use of a common individual model.

REFERENCES

- Achen, C. H. and W. P. Shively. 1995. *Cross-level Inference*. University of Chicago Press.
- Besag, J., J. York, and A. Mollie. 1991. "Bayesian Image Restoration with Two Applications in Spatial Statistics," *Annals of the Institute of Statistics and Mathematics*, 43: 1–59.
- Breslow, N. and N. Day. 1987. *Statistical Methods in Cancer Research, Volume 2 – The Analysis of Cohort Studies*. Oxford: Oxford University Press.
- Breslow, N. and N. E. Day. 1980. *Statistical Methods in Cancer Research, Volume 1 – The Analysis of Case-Control Studies*. Scientific Publications No. 32. Lyon: International Agency for Research on Cancer.

- Carstairs, V. 2000. "Socio-economic Factors at Areal Level and Their Relationship with Health." In P. Elliott, J. C. Wakefield, N. G. Best, and D. J. Briggs (eds.), *Spatial Epidemiology: Methods and Application*, Chapter 4. Oxford: Oxford University Press.
- Cho, W. K. T. 1998. "Iff the Assumption Fits...: A Comment on the King Ecological Inference," *Political Analysis*, 7: 143–163.
- Clayton, D. and M. Hills. 1993. *Statistical Models in Epidemiology*. Oxford Science Publications.
- Clayton, D. and J. Kaldor. 1987. "Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping," *Biometrics*, 43: 671–681.
- Clayton, D. G., L. Bernardinelli, and C. Montomoli. 1993. "Spatial Correlation in Ecological Analysis," *International Journal of Epidemiology*, 22: 1193–1202.
- Cornfield, J., W. H. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. 1959. "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions," *Journal of the National Cancer Institute*, 22: 173–203.
- Doll, R. 1998. "Uncovering the Effects of Smoking: Historical Perspective," *Statistical Methods in Medical Research*, 7: 87–117.
- Duncan, O. D. and B. Davis. 1953. "An Alternative to Ecological Correlation," *American Sociological Review*, 18: 665–666.
- Firebaugh, G. 1978. "A Rule for Inferring Individual-Level Relationships from Aggregate Data," *American Sociological Review*, 43: 557–572.
- Flanigan, W. and N. Zingale. 1985. "Alchemist's Gold: Inferring Individual Relationships from Aggregate Data," *Social Science History*, 9: 71–92.
- Freedman, D. A., S. P. Klein, J. Sacks, C. A. Smyth, and C. G. Everett. 1991. "Ecological Regression and Voting Rights," *Evaluation Review*, 15: 673–711.
- Goodman, L. A. 1953. "Ecological Regressions and the Behavior of Individuals," *American Sociological Review*, 18: 663–664.
- Goodman, L. A. 1959. "Some Alternatives to Ecological Correlation," *American Journal of Sociology*, 64: 610–625.
- Greenland, S. 1992. "Divergent Biases in Ecologic and Individual-Level Studies," *Statistics in Medicine*, 11: 1209–1223.
- Greenland, S. 2002. "A Review of Multilevel Theory for Ecologic Analyses," *Statistics in Medicine*, 21: 389–395.
- Greenland, S. and H. Morgenstern. 1989. "Ecological Bias, Confounding and Effect Modification," *International Journal of Epidemiology*, 18, 1: 269–274.
- Greenland, S. and J. Robins. 1994. "Invited Commentary: Ecologic Studies – Biases, Misconceptions and Counterexamples," *American Journal of Epidemiology*, 139, 8: 747–764.
- Hamilton, R. 1982. *Who Voted for Hitler?* Princeton, NJ: Princeton University Press.
- Herron, M. and K. W. Shotts. 2004. "Logical Inconsistency in King-Based Ecological Regressions," to appear in *American Journal of Political Science*.
- King, G. 1997. *A Solution to the Ecological Inference Problem*. Princeton University Press.
- Langbein, L. I. and A. J. Lichtman. 1978. *Ecological Inference*. Beverley Hills, CA: Sage Publications.
- Lasserre, V., C. Guihenneuc-Jouyau, and S. Richardson. 2000. "Biases in Ecological Studies: Utility of Including Within-Area Distribution of Confounders," *Statistics in Medicine*, 19: 45–59.
- Maheswaran, R., S. Morris, S. Falconer, A. Grossinho, I. Perry, J. Wakefield, and P. Elliott. 1999. "Magnesium in Drinking Water Supplies and Mortality from Acute Myocardial Infarction in North West England," *Heart*, 82: 455–460.
- Morgenstern, H. 1998. "Ecologic Study." In P. Armitage and T. Colton (eds.), *Encyclopedia of Biostatistics*, Vol. 2. Wiley, pp. 1255–1276.
- Piantadosi, S., D. P. Byar, and S. B. Green. 1988. "The Ecological Fallacy," *American Journal of Epidemiology*, 127, 5: 893–904.
- Plummer, M. and D. Clayton. 1996. "Estimation of Population Exposure," *Journal of the Royal Statistical Society, Series B*, 58: 113–126.
- Pope, C. A. and D. Dockery. 1996. "Epidemiology of Chronic Health Effects: Cross-Sectional Studies." In R. Wilson and J. Spengler (eds.), *Particles in Our Air: Concentrations and Health Effects*. Boston: Harvard University Press, pp. 149–167.

- Prentice, R. L. and L. Sheppard. 1990. "Dietary Fat and Cancer: Consistency of the Epidemiologic Data and Disease Prevention That May Follow from a Practical Reduction in Fat Consumption," *Cancer Causes Control*, 1: 87–97.
- Prentice, R. L. and L. Sheppard. 1995. "Aggregate Data Studies of Disease Risk Factors," *Biometrika*, 82, 1: 113–125.
- Richardson, S. and C. Monfort. 2000. "Ecological Correlation Studies." In P. Elliott, J. C. Wakefield, N. G. Best, and D. J. Briggs (eds.), *Spatial Epidemiology: Methods and Application*, Chapter 11. Oxford: Oxford University Press.
- Richardson, S., I. Stucker, and D. Hémon. 1987. "Comparison of Relative Risks Obtained in Ecological and Individual Studies: Some Methodological Considerations," *International Journal of Epidemiology*, 16, 1: 111–120.
- Rosenbaum, P. and D. Rubin. 1984. "Difficulties with Regression Analyses of Age-Adjusted Rates," *Biometrics*, 40: 437–443.
- Rosenbaum, P. R. 2002. *Observational Studies*, 2nd ed. New York: Springer-Verlag.
- Rothman, K. J. and S. Greenland. 1998. *Modern Epidemiology*. Lippincott-Raven.
- Sheppard, L. 2003. "Insights on Bias and Information in Group-Level Studies," *Biostatistics*, 4: 265–278.
- Singh, G. K. and M. Siahpush. 2002. "Increasing Inequalities in All-Cause and Cardiovascular Mortality among US Adults Aged 25–64 Years by Area Socio-economic Status, 1969–1998," *International Journal of Epidemiology*, 31: 600–613.
- Smith, G. D. 2000. "Learning to Live with Complexity: Ethnicity, Socioeconomic Position and Health in Britain and the United States," *American Journal of Public Health*, 90: 1694–1698.
- Wakefield, J. and P. Elliott. 1999. "Issues in the Statistical Analysis of Small Area Health Data," *Statistics in Medicine*, 18: 2377–2399.
- Wakefield, J. C. 2003. "Sensitivity Analyses for Ecological Regression," *Biometrics*, 59: 9–17.
- Wakefield, J. C. 2004a. "A Critique of Statistical Aspects of Ecological Studies in Spatial Epidemiology," *Ecological and Environmental Statistics*, to appear.
- Wakefield, J. C. 2004b. "Ecological Inference for 2×2 Tables," *Journal of the Royal Statistical Society, Series A*, to appear.
- Wakefield, J. C. and R. E. Salway. 2001. "A Statistical Framework for Ecological and Aggregate Studies," *Journal of the Royal Statistical Society, Series A*, 164: 119–137.
- Yasui, Y., J. Potter, J. Stanford, M. Rossing, M. Winget, M. Bronner, and J. Daling. 2001. "Breast Cancer Risk and "Delayed" Primary Epstein–Barr Virus Infection," *Cancer Epidemiology, Biomarkers and Prevention*, 10: 9–16.