

## A review of multilevel theory for ecologic analyses

Sander Greenland\*

*Department of Epidemiology, UCLA School of Public Health and Department of Statistics, UCLA College of Letters and Science, Los Angeles, CA 90095-1772, U.S.A.*

### SUMMARY

This paper reviews multilevel and conventional models for the analysis of ecologic (group, aggregate) data. It emphasizes the non-separability of contextual (group-level) effects and individual-level effects that arises from the multilevel structure of the underlying effects. Contrary to common misperceptions, this problem afflicts ecologic studies in which the sole objective is to estimate contextual effects, as well as studies in which the objective is to estimate individual effects. Multilevel effects also severely complicate causal interpretations of model coefficients. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: ecologic analysis; multilevel theory

### INTRODUCTION

The terms *ecologic inference* and *ecologic analysis* have come to refer to any analysis based on data that are limited to characteristics of groups (aggregates) of individuals [1]. Typically, these aggregates are geographically defined populations, such as counties, provinces or countries. There is an enormous literature on ecologic inference extending at least as far back as the early 20th century. See for example Iverson [2] and Achen and Shively [3] for reviews of the social-science literature, and Morgenstern [4] for a review of the health-sciences literature.

The objectives of ecologic analyses may be roughly classified into two broad categories:

- (i) *individual*, to examine relations of group and individual characteristics to the outcomes of the individuals composing the groups;
- (ii) *purely contextual*, to examine relations of group characteristics to group outcomes.

A study may of course have both objectives. Most of the literature has focused on studies with only individual-level objectives, and emphasizes the obstacles to *cross-level inference* from group-level (ecologic) observations to individual-level relations. This focus has somehow led to a common misperception that individual-level relations can be safely ignored if only contextual objectives are of interest. Another common problem is the frequent cavalier use of strong (and sometimes even implausible) modelling assumptions to generate cross-level inferences [2–5].

---

\*Correspondence to: Sander Greenland, 22333 Swenson Drive, Topanga, CA 90290-3434, U.S.A.

It seems infrequently recognized that the two problems (the need to consider individual-level relations in contextual inferences, and the sensitivity of cross-level inferences to modelling assumptions) are intimately related. In a companion paper [5] I provide a non-technical narrative overview of the literature surrounding these problems and the often-heated controversy they generate. Here, I review some basic multilevel (hierarchical) structural-model theory to display the problems clearly. By ‘structural’ I mean underlying regression structure, apart from issues of random error or sampling variation. Further statistical theory including fitting methods can be found in references [6–9]. The discussion section will point out some difficulties that arise when attempting causal interpretations of multilevel-model parameters; see Robins *et al.* [10] for details.

### MULTILEVEL MODELS AND THEIR IMPLICATIONS FOR ECOLOGIC INFERENCE

Most of the points discussed here follow immediately from standard multilevel modelling theory [11]. Suppose  $Y$  and  $X$  are an outcome (response) variate and a covariate vector defined on individuals, and  $R$  and  $Z$  are an outcome and covariate vector defined directly on aggregates (clusters);  $X$  may contain products among individual and aggregate covariates. For example,  $Y$  could be an indicator of being robbed within a given year;  $X$  could contain age, income and other variables; with state as the aggregate,  $R$  could be the proportion robbed in the state over the year (which is just the average  $Y$  in the state), and  $Z$  could contain state average age and income, and an indicator of whether hypodermic needles are available without prescription in the state.

Let  $\mu_{ik}$  and  $x_{ik}$  be the expected outcome and  $X$ -value of individual  $i$  in aggregate  $k$ , with  $\bar{\mu}_k$  and  $\bar{x}_k$  their averages and  $z_k$  the  $Z$ -value for aggregate  $k$ ,  $k = 1, \dots, K$ . One generalized-linear model (GLM) for individuals is

$$\mu_{ik} = f(\alpha + x_{ik}\beta + z_k\gamma) \quad (1)$$

If  $R = \bar{Y}$ , as when  $R$  is a rate, proportion or average lifespan, model 1 induces the aggregate-outcome model

$$E_k(R) = \bar{\mu}_k = \sum_{i=1}^{N_k} f(\alpha + x_{ik}\beta + z_k\gamma) / N_k \quad (2)$$

where  $N_k$  is the size of aggregate  $k$  [12, 13]. If  $\beta \neq 0$ ,  $\bar{\mu}_k$  will depend on the within-aggregate  $X$ -distribution as well as the individual-specific effects  $\beta$  and the contextual effects  $\gamma$ .

#### *Special cases*

If the individual model is linear (so that  $f(u) = u$ ) and if  $Z = \bar{X}$ , models 1 and 2 become

$$\mu_{ik} = \alpha + x_{ik}\beta + \bar{x}_k\gamma \quad (3)$$

and

$$\bar{\mu}_k = \alpha + \bar{x}_k(\beta + \gamma) \quad (4)$$

Equation (4) exhibits the complete confounding of the individual effects  $\beta$  and the aggregate (contextual) effects  $\gamma$  in ecologic data [14, 15]. Although the induced regression is linear and can be fit with ecologic data, the identified coefficient  $\beta + \gamma$  represents a blend of individual and contextual effects, rather than just contextual effects as often supposed. To see this point, let  $Y$  again be a robbery indicator, let  $X$  be income (in thousands of dollars), and let  $k$  index states. Then:  $\beta$  represents the difference in individual robbery risks associated with a \$1000 increase in individual income  $X$  within a given state;  $\gamma$  represents the difference in individual robbery risks associated with a \$1000 increase in state-average income  $\bar{X}$  among individuals of a given income  $X$ ; and  $\beta + \gamma$  represents the unconditional difference in average state risk associated with a \$1000 increase in average state income. In other words,  $\beta$  represents an intrastate association of personal income with personal robbery risk,  $\gamma$  represents an interstate association of average (contextual) state income with personal robbery risk within personal-income strata, and  $\beta + \gamma$  represents an unconditional interstate association of average income with average risk, blending the conditional associations with robbery risk of personal and state-average income differences.

Linear models usually poorly represent risk and rate outcomes because of their failure to constrain predicted outcomes to non-negative values. If we drop the linearity requirement but maintain  $Z = \bar{X}$ , model (1) can be rewritten

$$\mu_{ik} = f[\alpha + (x_{ik} - \bar{x}_k)\beta + \bar{x}_k(\beta + \gamma)] \quad (5)$$

showing that  $\beta$  may also be interpreted as the within-aggregate (intracluster) association of the transformed individual expected outcome  $f^{-1}(\mu)$  with departures of  $X$  from the aggregate-specific average  $\bar{X}$ ;  $\beta + \gamma$  is then the complementary association of  $\bar{X}$  with  $\bar{\mu}$  across aggregates [7]. The departure  $X - \bar{X}$  is however a contrast of individual and contextual characteristics  $X$  and  $\bar{X}$ , rather than a purely individual characteristic  $X$  as in model (1), and  $\beta + \gamma$  remains a mixture of individual and contextual effects.

Unfortunately, not even the mixed effect  $\beta + \gamma$  can be unbiasedly estimated without further assumptions if the model is non-linear and the only available covariates are aggregate means  $\bar{X}$  and purely contextual variables. Richardson *et al.* [16] and Dobson [17] discussed assumptions for unbiased ecologic estimation of  $\beta + \gamma$  under log-linear models, in which  $f(u) = e^u$ , although they implicitly assumed  $\gamma = 0$  (no contextual effect) and so described them as assumptions for estimating  $\beta$ ; see also Sheppard [7], Prentice and Sheppard [12] and Greenland [18]. From a scientific perspective these assumptions can appear highly artificial (for example,  $X$  multivariate normal with constant covariance matrix across aggregates, which cannot be satisfied if  $X$  has discrete components that vary in distribution across aggregates), although the bias from their violation might not always be large [13, 16, 17]. None the less, even if the bias in estimating  $\beta + \gamma$  is small,  $\beta$  and  $\gamma$  remain inseparable without further assumptions or individual-level data.

#### *Extensions of the basic model*

Model (1) is a bilevel model, in that it incorporates covariates for individual-level characteristics (the  $X$ ) and for one level of aggregation (the  $Z$ ). One can extend the model to incorporate characteristics of multiple levels of aggregation [11]. For example, the model could include a covariate vector  $Z_c$  of census-tract characteristics such as the average income in the tract, and another vector  $Z_s$  of state characteristics such as the law indicators and enforcement variables

(for example, conviction rates and average time served for various felonies). The above points can be roughly generalized by saying that absence of data on one level should be expected to limit inferences about effects at other levels. The term ‘cross-level bias’ is sometimes used to denote bias in coefficient estimates on one level that results from incorrect assumptions about effects on other levels (for example, linearity or absence of effects); similarly, ‘cross-level confounding’ is sometimes used to denote inseparability of different levels of effect, as illustrated above.

An important extension of multilevel models adds between-level covariate products (‘interactions’) to model (1), thus allowing the individual effects  $\beta$  to vary across groups. Even with correct specification and multilevel data, the number of free parameters that results can be excessive for ordinary methods. One solution is to constrain the variation with hierarchical coefficient models, which induces a mixed model for individual outcomes [11]. For example, adding products among  $X$  and  $Z$  to model (1) is a special case of replacing  $\beta$  by  $\theta_k$  with  $\theta_k$  constrained by the second-stage model

$$\theta_k = \beta + A_k \delta \quad (6)$$

where  $A_k$  is a known fixed matrix function of  $z_k$ . The ensemble of estimates of the vectors  $\theta_1, \dots, \theta_K$  can then be shrunk toward a common estimated mean  $\tilde{\beta}$  by specifying a mean-zero, variance  $\tau^2$  distribution for  $\delta$ . This process is equivalent to (empirical) Bayesian averaging of model (1) with the expanded model [19]

$$\mu_{ik} = f(\alpha + x_{ik}\beta + z_k\gamma + x_{ik}A_k\delta) \quad (7)$$

Unfortunately, the non-identifiability and confounding problems of ecologic analysis only worsen in the presence of cross-level interactions [20].

## ECOLOGIC REGRESSION

Many if not most ecologic researchers do not consider the individual or aggregated models (1) or (2) but instead directly model the aggregated outcome  $\bar{Y}$  with an ecologic regression model. If  $Z = \bar{X}$ , an ecologic GLM is

$$E_k(\bar{Y}) = \bar{\mu}_k = h(a + \bar{x}_k b) \quad (8)$$

With  $h(u) = u$  and individual linearity (model (3)), the results given above imply that  $b$  will be a linear combination of individual and contextual effects. Given individual non-linearities, model (8) will at best be an approximation to a complex aggregate regression like model (2), and  $b$  will represent a correspondingly complex mixture of individual and contextual effects [7]. These results also apply when purely contextual variables are included with  $\bar{X}$  in  $Z$  and  $\bar{x}_k$  is replaced by  $z_k$  in model (8).

The classic ‘ecologic fallacy’ (perhaps more accurately termed cross-level bias in estimating individual effects from ecologic data) is to identify  $b$  in model (8) with  $\beta$  in model (1). This identification is fallacious because unless  $\gamma = 0$  (contextual effects absent), even linearity does not guarantee  $b = \beta$ . With individual-level non-linearity, even  $\gamma = 0$  does not guarantee  $b = \beta$ . Ecologic researchers interested in estimating contextual effects should note the symmetrical results that apply when estimating contextual effects from ecologic data. Even with linearity,

$b = \gamma$  will require  $\beta = 0$ , which corresponds to no individual-level effects of  $X$  even when the aggregate summaries of  $X$  in  $Z$  (such as  $\bar{X}$ ) have effects. In other words, for the ecologic coefficient  $b$  to correspond to the contextual effect  $\gamma$ , any within-group heterogeneity in individual outcomes must be unrelated to the individual analogues of the contextual covariates. For example, if  $X$  is income and  $Z$  is mean income  $\bar{X}$ ,  $b = \gamma \neq 0$  would require individual incomes to have no effect except as mediated through average area income. Such a condition would rarely if ever be plausible.

Another fallacy involves the use of non-comparably standardized or restricted covariates in the ecologic model [18, 21]. Typically this occurs when  $\bar{Y}$  is age-standardized and specific to sex and race categories, but  $\bar{X}$  is not; for example, disease rates are routinely published in age-standardized sex-race specific form, but most summary covariates (for example, alcohol and tobacco sales) are not. Such non-comparability can lead to bias in  $b$  even as a representation of the mixed effect  $\beta + \gamma$ , because the coefficients of these crude covariate summaries will not capture the rate variation associated with variation in alcohol and tobacco use by age, sex and race.

## DISCUSSION

We should expect problems arising from within-group heterogeneity to diminish as the aggregates are made more homogeneous on relevant covariates. One may also reasonably expect heterogeneity on social factors to decline as the aggregates become more restricted (for example, census tracts tend to be more homogeneous on income and education than counties). None the less, the benefits of such restriction may be nullified if important covariate measures (for example, alcohol and tobacco) are available only for the broader aggregates. Problems in causal interpretation of model coefficients may also worsen.

Like most of the literature, I have been rather loose in my use of the word 'effect'. Strictly speaking, the models and results used here apply only to interpretations of coefficients as measures of conditional associations. Their formal interpretation as measures of causal effects in potential-outcomes (counterfactual) models [22–24] raises some problems not ordinarily encountered in individual-level modelling.

To avoid technical details (which would require extensive notation), I will use the linear model example; see Robins *et al.* [10] for a general theoretical treatment. A causal interpretation of  $\beta$  in model (3) is that it represents the change in an individual's risk that would result from increasing that person's income by \$1000. Apart from the ambiguity of this intervention (does the money come from a pay raise or direct payment?), one might note that it would also raise the average income value  $\bar{x}_k$  by  $\$1000/N_k$ , and so produce an additional change of  $\gamma/N_k$  in the risk of the individual. This apparently contradicts the original interpretation of  $\beta$  as *the* change in risk, whence we see that  $\beta$  must be interpreted carefully as the change in risk given that average income is (somehow) held constant, for example, by reducing everyone else's income  $\$1000/(N_k - 1)$  each.

If  $N_k$  is large this problem is trivial, but if the aggregates are small (for example, census tracts) it may be important. One solution with a scientific basis is to recognize  $\bar{X}$  as a contextual surrogate for socio-economic environment, in which case the contextual value  $\bar{x}_k$  in model (1) might be better replaced by an individual-level variable  $\bar{x}_{(-i)k}$ , the average income of persons in aggregate  $k$  apart from individual  $i$ . Note that the mean of  $\bar{x}_{(-i)k}$  is  $\bar{x}_k$

and so its effect could not be separated from that of personal income when only aggregate means are available; that is,  $\beta$  and  $\gamma$  remain completely confounded.

Causal interpretation of the contextual coefficient  $\gamma$  is symmetrically problematic. It represents the change in risk from a \$1000 increase in average income  $\bar{x}_k$ , while keeping the person's income  $x_{ik}$  the same. This can be done in many ways, for example, by giving everyone but person  $i$  a  $\$1000N_k/(N_k - 1)$  raise, or by giving one person other than  $i$  a  $\$1000N_k$  raise. The difference between these two interventions is profound and might be resolved by clarifying the variable for which  $\bar{X}$  is a surrogate. Similar comments apply to causal interpretation of  $b$  in the ecologic model (model (8)).

Problems of causal interpretation only add to the difficulties in the analysis of contextual causal effects, but unlike confounding may be as challenging for individual-level and multilevel studies as they are for ecologic studies [2]. Addressing these problems demands that detailed subject-matter considerations be used to construct explicit multilevel causal models in which key variables at any level may be latent (not directly measured).

Another serious and often overlooked conceptual problem involves the special artificiality of typical model specifications above the individual level. Ecologic units are usually arbitrary administrative grouping, such as counties, which may have only weak relations to contextual variables of scientific interest such as social environment [3]. Compounding this problem is the potentially large sensitivity of aggregate-level relations to the grouping definitions [25]. In light of these problems it may be more realistic to begin with at least a trilevel specification in which one aggregate level is the one of direct scientific interest (for example, neighbourhood or community) and the other is the level for which data are available (for example, census tracts or counties). This larger initial specification would at least clarify the proxy-variable (measurement surrogate) issues inherent in most studies of contextual effects.

#### REFERENCES

1. Langbein LI, Lichtman AJ. *Ecological Inference*. Series/No. 07-010. Sage: Thousand Oaks, CA, 1978.
2. Iversen GR. *Contextual Analysis*. Sage: Thousand Oaks, CA, 1991.
3. Achen CH, Shively WP. *Cross-Level Inference*. University of Chicago Press, Chicago, 1995.
4. Morgenstern H. Ecologic studies. In *Modern Epidemiology*, 2nd edn, Rothman KJ, Greenland S (eds). Lippincott: Philadelphia, 1998, 459–480.
5. Greenland S. Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *International Journal of Epidemiology* 2001; **30** (in press).
6. Navidi W, Thomas D, Stram D, Peters J. Design and analysis of multilevel analytic studies with applications to a study of air pollution. *Environmental Health Perspective* 1994; **102** (Suppl 8):25–32.
7. Sheppard L. Insights on bias and information in group-level studies. *Biostatistics* (to appear).
8. Guthrie KA, Sheppard L. Overcoming biases and misconceptions in ecologic studies. *Journal of the Royal Statistical Society, Series A* (in press).
9. Wakefield J, Salway R. A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society, Series A* (in press).
10. Robins JM, Murphy S, Greenland S. Towards a formal theory of causation in ecologic and multilevel studies. *Journal of the Royal Statistical Society, Series A* (to appear).
11. Goldstein H. *Multilevel Models*, 2nd edn. Edward Arnold: London, 1995.
12. Prentice RL, Sheppard L. Aggregate data studies of disease risk factors. *Biometrika* 1995; **82**:113–125.
13. Sheppard L, Prentice RL. On the reliability and precision of within- and between-population estimates of relative rate parameters. *Biometrics* 1995; **51**:853–863.
14. Firebaugh G. A rule for inferring individual-level relationships from aggregate data. *American Sociology Review* 1978; **43**:557–572.
15. Firebaugh G. Assessing group effects. *Aggregate Data: Analysis and Interpretation*, Borgatta EF, Jackson DJ (Eds). Sage: Thousand Oaks, CA, 1980; Chapter 2, 13–24.
16. Richardson S, Stücker I, Hémon D. Comparison of relative risks obtained in ecological and individual studies: Some methodological considerations. *International Journal of Epidemiology* 1987; **16**:111–120.

17. Dobson AJ. Proportional hazards models for average data for groups. *Statistics in Medicine* 1988; **7**:613–618.
18. Greenland S. Divergent biases in ecologic and individual-level studies. *Statistics in Medicine* 1992; **11**: 1209–1223.
19. Greenland S. Multilevel modelling and model averaging. *Scandinavian Journal of Work Environment and Health* 1999; **25**(supplement 4):43–48.
20. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. *International Journal of Epidemiology* 1989; **18**:269–274.
21. Rosenbaum PR, Rubin DB. Difficulties with regression analyses of age-adjusted rates. *Biometrics* 1984; **40**: 437–443.
22. Rubin DB. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science* 1990; **5**:472–480.
23. Greenland S, Rothman KJ. Measures of effect and measures of association. In *Modern Epidemiology*, 2nd edn, Rothman KJ, Greenland S (Eds). Lippincott: Philadelphia, 1998.
24. Pearl J. *Causality*. Cambridge University Press: New York, 2000.
25. Openshaw S, Taylor PH. The modifiable area unit problem. *Quantitative Geography*, In Wrigley N, Bennett RJ (Eds). Routledge: London, 1981.