

Ecological Inference in Epidemiology

Jon Wakefield,

Departments of Statistics and Biostatistics,
University of Washington, Seattle.

jonno@u.washington.edu

Outline:

- Nature of ecological data in epidemiology, illustrative examples.
- Sources of bias in ecological studies: pure specification bias, confounding.
- Difficulties of contextual effect estimation.
- More general designs.
- Multilevel modeling:
 - Use of individual-level data.
 - Spatial dependence.
- Discussion

1

Health Disparities, May 7th, 2003

Ecological Studies (Morgenstern 1998)

In ecological studies, rather than data at the level of the individual, we analyze data at the group level – in this talk the groups will be areas.

Most graphical summaries are ecological!

Ecological correlation studies compare aggregate health summaries against aggregate predictor variables. Examples:

- Water constituents.
- Air pollution.
- Dietary variables.
- Socio-economic indicators.

Advantages of such studies: data availability, increased exposure contrasts.

Disadvantages: beyond the usual difficulties of observational studies, a number of additional biases, an umbrella term for which is *ecological bias*.

2

Health Disparities, May 7th, 2003

Sources of ecological bias (Greenland, 1992)

- Pure specification bias – change in form of individual-level model under aggregation. Effect measures (correlations, relative risks) aren't consistent across levels of aggregation.
- Between-area confounding.
- Within-area confounding.
- Contextual effects.
- Mutual standardization.
- Measurement error.
- Effect modification.

3

Health Disparities, May 7th, 2003

Infant Mortality Example:

Data supplied by Richard Hoskins of the Washington State Department of Health.

We have number of births and number of deaths in the first year of life in 252 census tracts of King County, in the period 1988–1992.

There were 794 deaths in this period out of a total of 102,043 births, median of 2 deaths, mean of 3.2 per census tract.

Poverty data from the 1990 census, proportion below the poverty level.

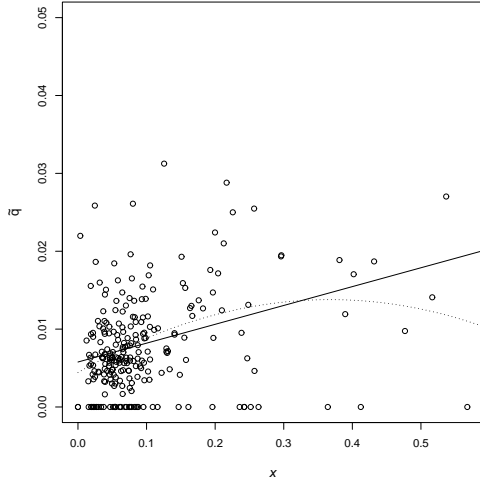
4

Health Disparities, May 7th, 2003

Area-Level Inference?

Area-level risk of infant mortality increases with increasing poverty, but what is this reflecting?

- Between-area confounding?
- Contextual effect?
- Within-area confounding?



5

Health Disparities, May 7th, 2003

Magnesium Example (Wakefield 2003)

Carried out in the Northwest of England. Health data: all deaths from myocardial infarction in the period 1990–1992.

Exposure data: time series of water constituents magnesium, calcium, fluoride from “water-zones”.

Confounder data: age, gender and a census-defined socio-economic measure – all at the enumeration district.

Hypothesis: Higher levels of magnesium in drinking water are associated with lower mortality from acute myocardial infarction.

Issues:

- Within-area variability in exposures.
- Joint distribution of confounders and exposures.
- Spatial dependence in residuals due to unmeasured variables.
- Poor exposure measure/measurement error.

6

Health Disparities, May 7th, 2003

Pure Specification Bias

Notation: Y_{ij} and x_{ij} denote the individual binary outcome and exposure of individual j in area i , $i = 1, \dots, m$, $j = 1, \dots, N_i$.

Individual-level model for a rare non-infectious outcome: $Y_{ij}|x_{ij} \sim \text{Bern}(p_{ij})$ where

$$p_{ij} = \exp(\alpha + \beta x_{ij}).$$

Suppose we observe exposure/confounder summaries ϕ_i : $Y_{ij}|\phi_i \sim \text{Bern}(q_i)$ where average risk:

$$q_i = \exp(\alpha) \int \exp(\beta x) f_i(x|\phi_i) \, dx.$$

Example: (e.g. Richardson and Montfort 2000)

$$x_{ij}|\phi_i \sim N(x_i, \sigma_i^2) \rightarrow q_i = \exp(\alpha + \beta x_i + \beta^2 \sigma_i^2 / 2).$$

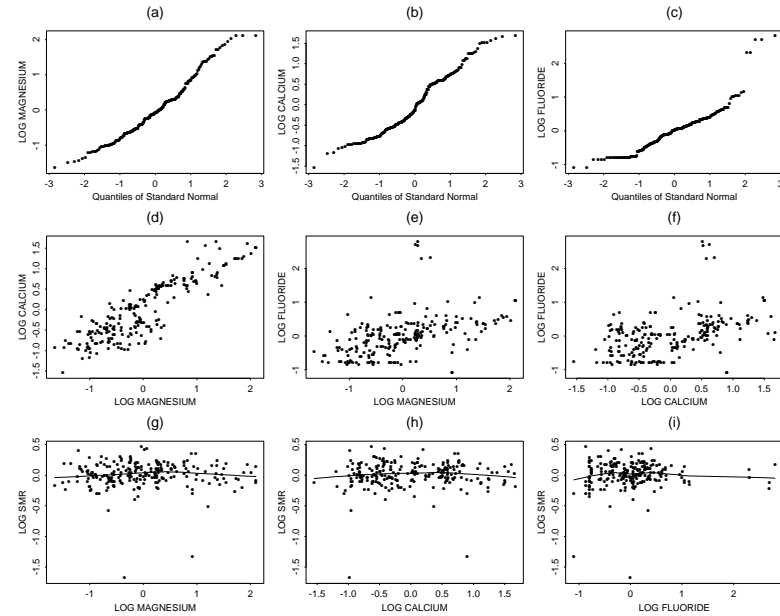
The variance is acting like a confounder – no bias if within-area variance is independent of mean exposure in area.

7

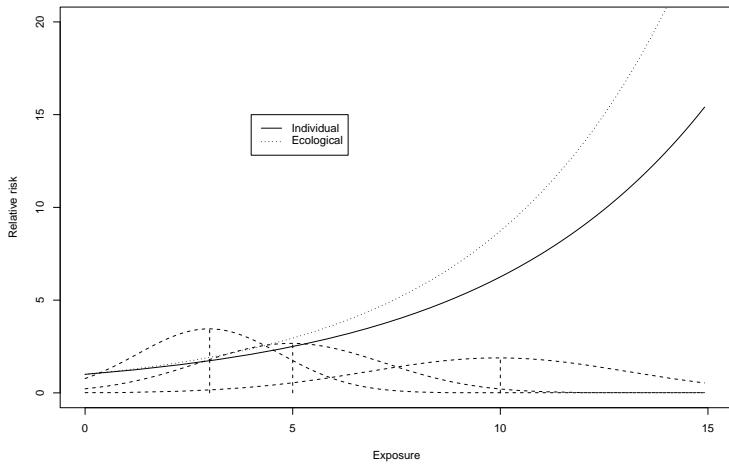
Health Disparities, May 7th, 2003

8

Health Disparities, May 7th, 2003



Effect of Pure Specification Bias



Contextual Effects

In measurement of health disparities contextual effects are of great relevance. A contextual variable is a characteristic of individuals in a shared neighborhood.

In the social sciences this aspect has been emphasized, less so in environmental epidemiology (more so in social epidemiology).

In an aggregate study we can't distinguish between individual-level and contextual effects (or a combination). This is well-documented in the social sciences (Freedman et al. 1991, Achen and Shively, 1995).

Confounding

Between-area confounding is analogous to conventional confounding (since the area is the unit of analysis).

Within-area confounding is complex since:

- In an ecological study we need to control for the complete within-area *distribution* of exposures and confounders – marginal prevalences of characteristics (for example) are not sufficient.
- Shape of risk model will in general change as we aggregate (so we have pure specification bias in the confounders too).

Statistical Models (Wakefield, 2004)

Let x denote a binary variable, for illustration we take above/below the poverty line.

Ecological regression:

Individual-level model:

$$Y_{ij}|x_{ij} \sim \text{Bern}\{\exp(\alpha + \beta x_{ij})\},$$

where i indexes areas and j individuals within areas.

Aggregate-level model:

$$Y_i|x_i \sim \text{Binomial}\left\{N_i, \frac{\exp(\alpha)}{N_i} \sum_{j=1}^{N_i} \exp(\beta x_{ij})\right\},$$

or, if q_i is the area-level “average” risk:

$$q_i = \exp(\alpha)\{(1 - x_i) + x_i \exp(\beta)\} = a + bx_i,$$

where x_i is the proportion below the poverty line.

Contextual (neighborhood) model:

Additive individual-level model (rare disease):

$$Y_{ij}|x_i \sim \text{Bern}\{a^* + b^*x_i\}.$$

Aggregate-level model:

$$Y_i|x_i \sim \text{Binomial}\{N_i, a^* + b^*x_i\},$$

or

$$q_i = a^* + b^*x_i,$$

which is indistinguishable from the individual poverty effect model.

With the multiplicative individual-level model

$$Y_{ij}|x_i \sim \text{Bern}\{\exp(\alpha^* + \beta^*x_i)\},$$

then the forms are distinguishable due to the nonlinearity but this will lead to unstable inference – analogous to Copas and Li (1987) criticism of Heckman (1979) selection model.

Extended ecological regression

Assume

$$p_{ji} = a_j + b_jx_i$$

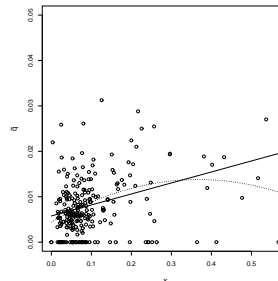
so that we have both individual and contextual effects, so that b_0 is the contextual effect for individuals below the poverty line, and b_1 is for above the poverty line.

Leads to

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

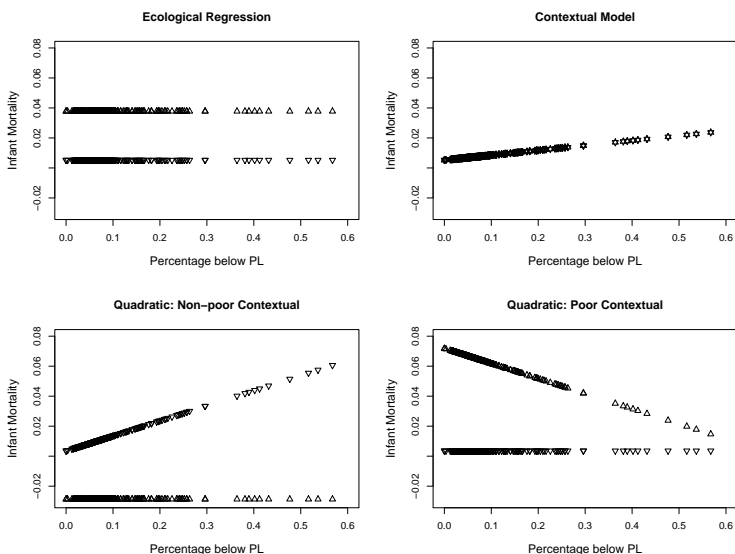
but we cannot estimate all four parameters – usual approach to set $b_0 = 0$ or $b_1 = 0$.

Linear and quadratic models for King County data:



Estimates for infant mortality (D) as a function of poverty level (PL), under different assumptions.

$\Delta = \Pr(D|\text{below PL})$, $\nabla = \Pr(D|\text{above PL})$.



A more plausible model for a rare disease

Individual-level model:

$$Y_{isj}|x_{isj}, x_i \sim \text{Bern}\{\exp(\alpha + \beta x_{isj} + \gamma_s + \delta x_i)\},$$

so that the contextual effect is the same for poor and not-poor (in the quadratic formulation this leads to a zero coefficient for the quadratic term).

Aggregate-level model: let $r_i = E[Y_i/E_i]$ with $E_i = \sum_{j=1}^{N_{is}} N_{is}p_s$. Then

$$r_i = \exp(\alpha + \delta x_i)\{(1 - x_i) + x_i \exp(\beta)\}.$$

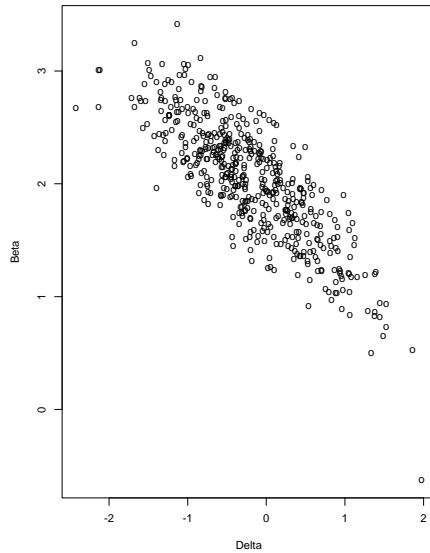
So identifiable...

Poverty Data

Strong dependence between individual poverty effect (β) and contextual poverty effect (δ) in the model

$$r_i = \exp(\alpha + \delta x_i) \{ (1 - x_i) + x_i \exp(\beta) \}.$$

Samples from the bivariate posterior $p(\beta, \delta|y)$:



17

Health Disparities, May 7th, 2003

Summary: poverty data

All models fitted using quasi-likelihood with variance \propto mean.

Without (with) control for gender and race: area level relative risk is 8.0 (3.8).

Both can be interpreted as individual-level or contextual-level effect:

- Without control the risk of infant mortality is 8 times greater if child is below the poverty level, as compared to above this level (individual-level).
- Without control, the risk of infant mortality is 8 times greater if in an area in which everyone is below the poverty level, as compared to someone living in an area with everyone above the poverty level (contextual).

With more plausible individual-level model, effects are highly unstable.

Identifiability from a non-linearity is not comforting.

18

Health Disparities, May 7th, 2003

Contextual Effect Estimation

Contextual effects may be induced by unmeasured confounding. For example, if we have the true model:

$$E[Y_{ij}|X_{ij}, Z_{ij}] = \exp(\alpha + \beta X_{ij} + \gamma Z_{ij}),$$

where within-areas:

$$\begin{bmatrix} X_{ij} \\ Z_{ij} \end{bmatrix} \sim N \left(\begin{bmatrix} X_i \\ Z_i \end{bmatrix}, \sigma_2^2 \begin{bmatrix} 1 & c_2 \\ c_2 & \omega_2^2 \end{bmatrix} \right)$$

and between areas:

$$\begin{bmatrix} X_i \\ Z_i \end{bmatrix} \sim N \left(\begin{bmatrix} X \\ Z \end{bmatrix}, \sigma_1^2 \begin{bmatrix} 1 & c_1 \\ c_1 & \omega_1^2 \end{bmatrix} \right).$$

If Z_{ij} and Z_i are unmeasured we obtain

$$E[Y_{ij}|X_{ij}] = \exp \left(\alpha^* + X_{ij} \{ \beta + c_2 \gamma \} + \bar{X}_i \frac{\sigma_1^2}{\sigma_2^2} (c_1 - c_2) \gamma \right),$$

so that a contextual effect has been induced.

Analogous to unmeasured confounding in longitudinal studies (Palta and Yao, 1991).

19

Health Disparities, May 7th, 2003

Summary for Contextual Effects

Contextual effects cannot in general be estimated from ecological data (due to confounding).

With the addition of individual-level data, multilevel models are needed to give appropriate standard errors – **Multilevel models don't alleviate confounding!**

Illustration with the linear model

$$E[Y_{ij}|X_{ij}] = \alpha + \beta_W(X_{ij} - \bar{X}_i) + \beta_B \bar{X}_i,$$

the effect estimates will be confounded if the baseline risk in area i is correlated with \bar{X}_i . The use of the model

$$E[Y_{ij}|X_{ij}] = \alpha_i + \beta_W(X_{ij} - \bar{X}_i) + \beta_B \bar{X}_i, \quad (1)$$

with α_i as fixed effects is not sufficient since we have no degrees of freedom for estimation of β_B .

As an alternative, we may consider (1) with $\alpha_i \sim_{iid} N(0, \sigma_\alpha^2)$, for example. Cannot control for confounding though. For spatial effects see later.

Parameter Interpretation in (1)

If in a particular area we increase $\bar{X}_i \rightarrow \bar{X}_i + 1$, then under a causal interpretation, the average risk will change by $\beta_W + \beta_B$.

Two people in the same area who differ by one unit of X will have a difference in risk of β_W .

20

Health Disparities, May 7th, 2003

Multilevel Modeling

Multilevel models allow the dependence of data at different levels to be acknowledged.

For example, we would expect residual rates in “close-by” areas to be correlated – ignoring this aspect leads to inappropriate standard errors.

“Overdispersion” in disease counts can also occur due to data anomalies in the numerator (under-registration) and denominator (migration, census under-enumeration).

Multilevel models also allow smoothing of rates – for rare diseases SMRs may be highly unstable due to small numbers.

Study Design Summary

Very useful categorization (Sheppard 2003):

		Exposure	
		Individual	Ecological
Disease	Individual	<i>Individual</i>	<i>Semi-individual</i>
	Ecological	<i>Aggregate</i>	<i>Ecological</i>

Prentice and Sheppard (1995) proposed the aggregate design in which individual-level survey data, $X_{ij}, j = 1, \dots, m_i$, are combined with ecological rates via the model:

$$q_i = \frac{1}{m_i} \sum_{j=1}^{m_i} e^{\alpha + \beta X_{ij}}.$$

Again it is theoretically possible to add a contextual effect but in practice identifiable through a nonlinearity.

In a common version of the semi-individual design individual-level data are available on disease status and confounders, but the exposure of interest is at the ecological level (for example from an air pollution monitor).

Semi-individual Design

Individual-level model:

$$E[Y_{ij}|X_{ij}, Z_{ij}] = \exp(\alpha + \beta X_{ij} + \gamma Z_{ij}),$$

where Y_{ij} is disease status, and X_{ij}, Z_{ij} are the exposure and confounders of individual j in area i .

This design is beneficial in allowing confounder adjustment, but may still lead to specification bias. For example suppose that we have a surrogate exposure measure in area i , W_i with

$$X_{ij} = W_i + U_{ij}, \quad U_{ij} \sim N(0, \sigma_u^2 W_i),$$

to give

$$E[Y_{ij}|W_i, Z_{ij}] = \exp(\alpha + W_i\{\beta + \beta\sigma_u^2/2\} + \gamma Z_{ij}),$$

so that we would overestimate the exposure effect.

Spatial Dependence

For geographical correlation studies with a rare outcome, Clayton et al. (1992), following Besag et al. (1991), suggested the model

$$Y_i|r_i \sim \text{Poisson}(E_i \times r_i),$$

where

$$\log r_i = \alpha + \beta x_i + T_i + S_i,$$

x_i is an area-level exposure summary, and T_i and S_i are random effects without and with spatial structure, the latter:

- to prevent “confounding by location”, and
- give appropriate standard errors.

If estimate of β changes greatly when spatial random effects are added then interpretation is difficult.

Guthrie, Sheppard and Wakefield (2003) extended the aggregate design to include spatial random effects within a Bayesian hierarchical model.

Magnesium Example

Random effect models: $T_i \sim_{iid} N(0, \sigma_T^2)$ and

$$S_i|S_{-i} \sim N(\bar{S}_i, \sigma_S^2/m_i),$$

where \bar{S}_i is the mean of, and m_i the number of neighbors.

Also the joint specification $cov(S_i, S_j) = \sigma^2 e^{-\phi d_{ij}}$ where d_{ij} is the distance between the centroids of areas i and j .

MODEL	MAGNESIUM			
	$\hat{\beta}_1$	Standard Error	Non-spatial $\hat{\sigma}$	Spatial $\hat{\sigma}$
Poisson	0.0075	0.0069	—	—
Quasi-Likelihood	0.0075	0.0113	$\hat{\kappa} = 2.72$	—
Non-spat r.e.	0.0066	0.0118	0.1100	—
Non-spat+ICAR r.e.	0.0033	0.0147	0.0225	0.1542
Non-spat+MVN r.e.	-0.0108	0.0160	0.0227	0.1284

MODEL	NORTHINGS			
	$\hat{\beta}_1$	Standard Error	Non-spatial $\hat{\sigma}$	Spatial $\hat{\sigma}$
Poisson	-0.0079	0.0169	—	—
Quasi-Likelihood	-0.0079	0.0278	$\hat{\kappa} = 2.72$	—
Non-spat r.e.	0.0061	0.0288	0.1101	—
Non-spat+ICAR r.e.	0.0463	0.0867	0.0222	0.1167
Non-spat+MVN r.e.	0.0433	0.1943	0.0546	0.1769

The posterior median for ϕ was 3.4 which corresponds to the spatial correlation dropping to 0.5 at a distance of 5.4 miles. A similar strength was found for the Northings analysis.

Discussion

- Need individual-level data to characterize within-area distribution of confounders and exposures.
- Ecological data should only be analyzed when $\beta_B = \beta_W$ (environmental exposures?). Small areas are better.
- Causal modeling needs work to extend to ecological settings.
- To tease out contextual effects is very difficult without relevant individual-level data.
- What is the appropriate geographical scale of the contextual effect?
- In geographical correlation studies more effort should be placed on confounding/within-area modeling than spatial dependence, the latter will be of secondary importance.
- Multilevel models acknowledge the dependencies in the data, but they don't sort out confounding.
- Ecological studies can add to the totality of evidence but alone are very susceptible to ecological bias. Semi-individual studies are more informative but still suffer from specification bias.
- Hybrid ecological designs.

References

Achen, C.H. and Shively, W.P. (1995). *Cross-Level Inference*, Chicago, University of Chicago Press.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics*, **43**, 1–59.

Clayton, D., Bernardinelli, L. and Montomoli, C. (1993). Spatial correlation in ecological analysis. *International Journal of Epidemiology*, **22**, 1193–1202.

Copas, J.B. and Li, H.G. (1997). Inference for non-random samples (with discussion). *Journal of the Royal Statistical Society, Series B*, **59**, 55–95.

Freedman, D.A., Klein, S.P., Sacks, J., Smyth, C.A. and Everett, C.G. (1991). Ecological regression and voting rights (and discussion), *Evaluation Review*, **15**, 673–816.

Greenland, S. (1992). Divergent biases in ecologic and individual-level studies, *Statistics in Medicine*, **11**, 1209–23.

Guthrie, K.A., Sheppard, L. and Wakefield, J. (2002). A hierarchical aggregate data model with spatially correlated disease rates. *Biometrics*, **58**, 898–905.

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, **47**, 153–161.

Morgenstern, H. (1998). Ecologic Studies. In Rothman,

K.J. and Greenland, S. (Eds.), *Modern Epidemiology, Second Edition*, pp. 459–480. Lipincott-Raven.

Palta, M. and Yao, T.-J. (1991). Analysis of longitudinal data with unmeasured confounders. *Biometrics*, **47**, 1355–1369.

Prentice, R.L. and Sheppard, L. (1995). Aggregate data studies of disease risk factors. *Biometrika*, **82**, 113–25.

Richardson, S. and Montfort, C. (2000). Ecological correlation studies. In *Spatial Epidemiology: Methods and Applications*. Eds: Elliott, P., Wakefield, J.C., Best, N.G. and Briggs, D.B, pp. 205—220. Oxford University Press, Oxford.

Sheppard, L. (2003). Insights on bias and information in group-level studies. *Biostatistics*, **4**, 265–278.

Wakefield, J.C. (2003). Sensitivity analyses for ecological regression. *Biometrics*, **59**, 9-17.

Wakefield, J.C. (2004). Ecological inference for 2×2 tables. To appear in *Journal of the Royal Statistical Society, Series A*.