AMERICAN JOURNAL OF EPIDEMIOLOGY
Vol. 127, No. 5
Printed in U.S.A.

THE ECOLOGICAL FALLACY

STEVEN PIANTADOSI,[1,2] DAVID P. BYAR,[1] AND SYLVAN B. GREEN[1]

The purpose of this paper is to emphasize for epidemiologists the possibility of serious errors resulting from inferences based on ecological analyses. Variables that describe groups of individuals, rather than the individuals themselves, are termed "ecological" and are often used when the analysis of individuals' data is not possible (1). Ecological analyses may be preferred when 1) variables are more conveniently defined or measured on groups because the analysis on individuals would require excessive time or extensive data gathering; 2) ecological analyses permit study of a wider range of values for the independent variable, as in international studies of diet; 3) the precision of aggregate measures like alcohol consumption is likely to be higher for groups than for individuals; and 4) population responses such as smoking quit rates may be of primary interest. Frequently, more than one reason applies. For example, some of the evidence favoring environmental and dietary causes of cancer comes from the comparison of incidence or mortality rates with average levels of risk factors measured on culturally or geographically defined groups of individuals. The first three reasons are relevant to this type of study.

We assume in this paper that measurements on individuals are not available, as in the diet and cancer example, since when this information is known, it might be used in place of, or to correct for biases in, the ecological analysis. Serious errors can result when an investigator makes the seemingly natural assumption that the inferences from an ecological analysis must pertain either to the individuals within the groups or to individuals across groups. A frequently cited early example of an ecological inference was Durkheim's study of the correlation between suicide rates and religious denominations in Prussia (2) in which the suicide rate was observed to be correlated with the number of Protestants. However, it could as well have been the Catholics who were committing suicide in largely Protestant provinces. The potential falsity of ecological inferences, at least in the case of simple correlations, was pointed out by Robinson (3), who gave it the name "ecological fallacy" and provided the mathematical relation, without proof, between the ecological correlation and the individual correlation across all groups. Duncan et al. (4) have extended the equations to include simple linear regression coefficients. The dangers of inferences about individuals from ecological studies have been emphasized by some investigators (5–7), while others (8–11) have sought to minimize the concern over the possible biases in ecological analyses, proposing alternatives or delineating circumstances in which ecological inferences are justified (e.g., certain linear regression models when data on individuals are available). Firebaugh (11) gives a particularly thorough discussion and list of references related to this aspect of the problem.

Although there has been a persistent interest in the problems associated with ecological analyses in the social science literature, the impression seems to remain, even among seasoned epidemiologists, that ecological analyses may not have large biases,

[1] Biometry Branch, Division of Cancer Prevention and Control, NCI, NIH, Bethesda, MD.

[2] Current address: The Johns Hopkins University, Department of Oncology/Biostatistics, 550 N. Broadway, Suite 1103, Baltimore, MD 21205. (Reprint requests to Dr. Steven Piantadosi at this address.)

The authors thank Dr. Sholom Wacholder, Dr. Charles Brown, and Dr. Peter Greenwald for helpful comments and Jennifer Gaegler for manuscript assistance.

at least in certain cases. Such impressions result, in part, from the nonintuitive sound of serious disparity between group level and individual level statistics. Our goal is to provide convincing evidence, intuitively, mathematically, and empirically, of the possibility of important bias in ecological analyses and to clarify some recent work on this topic. We present a hypothetical example of the ecological fallacy and a simple derivation of the relation between the individual correlation and the ecological correlation. We extend this derivation to outline the relation between the individual regression slope and the ecological regression slope. In addition, theoretical observations are supported by an ecological analysis of correlations and regression coefficients for a set of real data, using variables often encountered in epidemiologic practice.

HYPOTHETICAL EXAMPLE

The ecological fallacy is illustrated in a simple case by considering the hypothetical data in table 1. Although these data are contrived, they are useful since the relevant correlations are evident by inspection or simple calculation. Here, N individuals are classified into r groups of equal size $n_i = k$. A variable, X, is assigned to each individual and has the values of the consecutive integers from $k(i-1)+1$ to ki in the ith group. Within each group, the values of a second variable, Y, are chosen to be the descending consecutive integers from ki to $k(i-1)+1$ in the ith group (i.e., the reverse order of the X's). These data have the following characteristics. The overall means of X and Y, $\bar{\bar{X}}$ and $\bar{\bar{Y}}$, are equal, as are the within-group means, $\bar{X}_i$ and $\bar{Y}_i$. Similarly, the overall variances are equal, as are the within-group variances. The within-group correlation coefficient, ρ_i, equals -1 for all i, but since the group means for X and Y are identical, the between-group or ecological correlation, ρ_e, equals 1. It is easily shown (see Appendix 1), however, that for these data, the overall correlation between X and Y, ignoring groups, is

$$\rho = \frac{N^2 + 1 - 2k^2}{N^2 - 1}.$$

Thus, $\rho = \rho_i$ only when $k = N$ (i.e., a single group), and $\rho = \rho_e$ when $k = 1$ or as $N \longrightarrow \infty$ and $k \ll \infty$. In instances like this, however, in which there is an appreciable group effect (i.e., the expected value of Y given X, $E(Y \mid X)$, is not the same in all groups), the correlation of interest is neither ρ nor ρ_e but the average within-group correlation, ρ_w (defined later), which in the present example equals -1 since all ρ_i are the same. For this example, the value of ρ always exceeds 0.5, 0.92, and 0.98, respectively, for two, five, and 10 groups, whatever the number of subjects in each group. Furthermore, since the overall and within-group variances for X and Y are equal, the linear model

$$Y = \alpha + \beta X$$

and its ecological counterpart

$$\bar{Y}_i = \alpha_e + \beta_e \bar{X}_i$$

would give estimates $\hat{\beta} = \rho$ and $\hat{\beta}_e = \rho_e$. This simple example demonstrates maximal disparity between the ecological correlation or the ecological regression slope, and the corresponding overall or within-group estimates.

TABLE 1
Hypothetical data illustrating the ecological fallacy

X	Y	Group
1	k	1
2	$k-1$	1
3	$k-2$	1
.	.	.
.	.	.
.	.	.
k	1	1
$k+1$	$2k$	2
$k+2$	$2k-1$	2
$k+3$	$2k-2$	2
.	.	.
.	.	.
.	.	.
N	$N-k+1$	N/k

Derivation of general relations

To understand the source of the ecological fallacy in the general case, consider N individuals classified into r groups of size n_i, where $i = 1, 2, \ldots, r$, and a single covariate, X, linearly related to a response Y for each individual. The relation between ecological analyses and other analyses based on individuals can be seen by appropriate partitions of the total sums of squares and cross-products. Define

$$X_{\cdot\cdot} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} X_{ij} = N\bar{\bar{X}},$$

and an analogous definition for Y. For notational convenience, sums of squares and cross-products are denoted by T (total), W (within groups), or E (ecological or between groups), with subscripts indicating the variables involved. Thus, in the usual analysis of covariance (12) (table 2), T_{xx} is the total sum of squares of X, W_{xy} is the within-groups sum of cross-products of X and Y, and E_{yy} is the between-groups sum of squares of Y. The total sum of cross-products about the mean across all individuals may be partitioned into within-group and between-group components as follows:

$$T_{xy} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - \bar{\bar{X}})(Y_{ij} - \bar{\bar{Y}})$$

$$= W_{xy} + E_{xy},$$

where

$$W_{xy} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(Y_{ij} - \bar{Y}_i),$$

and

$$E_{xy} = \sum_{i=1}^{r} n_i(\bar{X}_i - \bar{\bar{X}})(\bar{Y}_i - \bar{\bar{Y}}).$$

The corresponding formulas for the sum of squares of X (or Y) are obtained by replacing Y with X (or X with Y) in these equations. The between-group or ecological sums of squares and cross-products are weighted by the number of individuals in each group. The correlation coefficient between X and Y, ignoring groups, is

$$\rho = \frac{T_{xy}}{\{T_{xx}T_{yy}\}^{1/2}}$$

$$= \frac{E_{xy}}{\{T_{xx}T_{yy}\}^{1/2}} + \frac{W_{xy}}{\{T_{xx}T_{yy}\}^{1/2}}. \qquad (1)$$

Rewriting equation 1 as

$$\rho = \frac{\{E_{xx}E_{yy}\}^{1/2}}{\{T_{xx}T_{yy}\}^{1/2}} \cdot \frac{E_{xy}}{\{E_{xx}E_{yy}\}^{1/2}} + \frac{\{W_{xx}W_{yy}\}^{1/2}}{\{T_{xx}T_{yy}\}^{1/2}} \cdot \frac{W_{xy}}{\{W_{xx}W_{yy}\}^{1/2}}, \qquad (2)$$

we may now define the *ecological correlation*

$$\rho_e = \frac{E_{xy}}{\{E_{xx}E_{yy}\}^{1/2}},$$

and the *average within-group correlation*

$$\rho_w = \frac{W_{xy}}{\{W_{xx}W_{yy}\}^{1/2}},$$

so that

$$\rho = \left\{\frac{E_{xx}E_{yy}}{T_{xx}T_{yy}}\right\}^{1/2} \rho_e + \left\{\frac{W_{xx}W_{yy}}{T_{xx}T_{yy}}\right\}^{1/2} \rho_w. \qquad (3)$$

Note that ρ_w is not influenced by group effects. Some additional definitions will show that this last result is equivalent to that given originally by Robinson (3). Details are given in Appendix 2.

In an analogous fashion, we may define

Table 2

A general analysis of covariance

Source	df	Sum of squares and products		
		$\Sigma\, x^2$	$\Sigma\, xy$	$\Sigma\, y^2$
Between groups	$r - 1$	E_{xx}	E_{xy}	E_{yy}
Within groups	$\Sigma n_i - r$	W_{xx}	W_{xy}	W_{yy}
Total	$\Sigma n_i - 1$	T_{xx}	T_{xy}	T_{yy}

the regression coefficients

$$\beta = \frac{T_{xy}}{T_{xx}},$$

$$\beta_e = \frac{E_{xy}}{E_{xx}},$$

and

$$\beta_w = \frac{W_{xy}}{W_{xx}},$$

and show that

$$\beta = \frac{E_{xx}}{T_{xx}} \beta_e + \frac{W_{xx}}{T_{xx}} \beta_w, \qquad (4)$$

where β, β_e, and β_w are the overall regression slope, *ecological regression* slope, and the *average within-group regression* slope, respectively. This can be written

$$\beta_e = \frac{T_{xx}}{E_{xx}} \beta - \left\{ \frac{T_{xx}}{E_{xx}} - 1 \right\} \beta_w$$

$$= \beta_w + \frac{T_{xx}}{E_{xx}} (\beta - \beta_w), \qquad (5)$$

with the first equality being the relation given by Duncan et al. (4).

Analysis of general relations

In the absence of group effects, the regression coefficient of interest is β, whereas when group effects are present, β_w is a more appropriate description of the data. In fact, when group effects are absent, $\beta = \beta_w$, so that β_w is always the regression coefficient of interest. The ecological fallacy consists of incorrectly assuming that, when group effects are present, $\beta_e = \beta_w$.

We can see immediately from equation 4 that the coefficients of β_e and β_w sum to 1, so that β is always a weighted average of the ecological and within-group regression slopes. The consequence of this is that β either lies between β_e and β_w (although the order of β_e and β_w cannot be predicted) or, when there are no group effects, $\beta = \beta_e = \beta_w$. More generally, however, there are group effects so that β and β_w are unequal, and thus β_e and β_w are also unequal. For regression coefficients, the notion of separating "cross-level" bias into aggregation bias (the difference between β and β_e) and specification bias (the difference between β and β_w) (1) is not meaningful since both biases either occur together or not at all, as implied by Firebaugh (11). In fact, it may easily be shown that

$$\beta_e - \beta = \frac{W_{xx}}{E_{xx}} (\beta - \beta_w),$$

further emphasizing that aggregation bias and specification bias do not occur separately.

The results for correlation coefficients are similar. The multipliers of ρ_e and ρ_w in equation 3 do not, however, sum to 1, so that ρ is not always constrained as β was. For correlations, the ecological fallacy consists of incorrectly assuming that ρ_e estimates either ρ or ρ_w.

The point has been correctly made (1, 10, 11) that ecological correlations are likely to be poorer estimates of their individual counterparts than ecological regression slopes. This is because correlations depend on the relative dispersions of X and Y and thus are determined by the design of the experiment. Note that

$$\rho_w = \beta_w \left\{ \frac{W_{xx}}{W_{yy}} \right\}^{1/2} \text{ and } \rho_e = \beta_e \left\{ \frac{E_{xx}}{E_{yy}} \right\}^{1/2}.$$

Selecting groups specifically because they differ in $\bar{X}_i$ will tend to increase the variance of $\bar{X}_i$ compared with the variance of $\bar{Y}_i$, and therefore increase the ecological correlation. In the absence of group effects in regression ($\beta = \beta_e = \beta_w$), the correlations can nevertheless differ, and the relation of ρ_e and ρ_w will depend on how the groups are chosen. Although demonstrating that $\rho_e \neq 0$ can be useful, its actual value seems quite arbitrary. If the goal is to make statements about Y on the basis of X, then in this situation, β_e is a more useful quantity than ρ_e.

It has been stated by some writers (see for example Stavraky (13) and Kleinbaum et al. (14)) that ecological associations are frequently an overestimate of the magnitude of the underlying individual effects. While this is always possible, we can find no justification for assuming that it is more likely than the alternative. As noted above, in the absence of group effects, the selection of groups will affect ρ_w (and may well tend to increase correlations). When group effects do, however, exist (which is quite likely in epidemiologic studies), the bias can be in either direction.

The concept of group effects in regression can be expressed in other ways. As stated above, by group effects we mean that $E(Y | X)$ is not the same in all groups or, equivalently, that $\beta \neq \beta_e \neq \beta_w$ (i.e., there is cross-level bias). Thus, group membership is related to some confounding factor(s) which affects the observed relation between X and Y. Firebaugh (11) addresses this situation by considering a regression of the form $Y = \alpha + \beta X + \gamma \bar{X}_i$; the presence of group effects implies that $\gamma \neq 0$. Without data on individuals, this situation cannot be detected, yet it is quite likely to exist in epidemiologic studies. Of course, if an investigator is aware of potential confounding variables measured on the groups, these can be included in the ecological regression to decrease the bias (10). The issue to consider is how likely are the groups to differ by other (unmeasured) variables.

A confounder can exist on two levels. A variable could be confounding only on the group level, and thus affect both β (the overall slope for X) and β_e (the ecological slope for X), but not β_w. For example, the variable could be a characteristic of the whole group rather than the individual (e.g., geographic latitude) or the variable could be independent of X within groups, but correlated with X across groups. If such a variable were known, it could be properly incorporated in an ecological regression. The problem is that variables such as this could well exist but not be identified. If individual level data on X and Y were available, individual level regression adjusted for group effect would permit unbiased estimation of the desired effect of X on Y.

Alternatively, a variable could be confounding on both the individual and group levels (e.g., age confounding the effect of diet). In this situation, information on the confounder would have to be incorporated into the regression whether at the individual level or the ecological level.

In theory, there is a third alternative in which a variable is confounding on the individual level but is not confounding in linear regression at the group level (because the variable is uncorrelated with group membership). For example, in an investigation of the relation of diet to the risk of colon cancer, sex is a possible confounding factor, but it is conceivable that all groups have essentially the same sex ratio. In this situation, ecological regression might be preferable to unadjusted individual level regression.

Empirical study of ecological correlations

We now consider an ecological analysis of data from the Second National Health and Nutrition Examination Survey (NHANES II) in which we can compare individual level and grouped estimates. This study, conducted by the National Center for Health Statistics between 1976 and 1980, was intended to assess the health and nutritional status of the general civilian noninstitutionalized population of the United States (15). Initially, a nationwide probability sample of approximately 28,000 persons was taken with oversampling in those groups thought to be at high risk of malnutrition (low income, preschool children, and the elderly). A 24-hour dietary recall questionnaire was given to a subset of 13,820 adults. From these records, we selected 11 variables for analysis describing or derived from health history, food frequency, and anthropometry (16). Measurements were on continuous scales for dietary measures, height, weight, age, and body mass index (weight in kilograms divided by the square of height in meters), ordered categories for income and education, and binary categories for sex and race (table 3). These variables were selected not because

TABLE 3
Coding scheme for binary and ordered category variables

Variable	Code
Sex	
Male	1
Female	2
Race	
White	1
Other	2
Education	
None	0
1*	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
1†	13
2	14
3	15
4	16
>5	17
Income‡	
<1	11
1–2	12
2–3	13
3–4	14
4–5	15
5–6	16
6–7	17
7–9	18
10–15	19
15–20	20
20–25	21
>25	22

* Grade.
† Year of college.
‡ ×1,000.

of pathologic behavior but because they represent the type of measurements likely to appear in ecological analyses conducted by epidemiologists. Three levels of aggregation were studied. Because individuals were originally selected for the NHANES II study through one of 64 primary sampling units consisting of counties or county aggregates, the primary sampling unit made a convenient grouping variable and was the lowest level of aggregation used. Individuals also represented 34 different states which were used as a second level of aggregation. Finally, states were classified into one of six regions (northeast, north central, etc.) for the highest level of aggregation. Primary sampling units contained an average of 216 people with extremes of 130 and 343. Aggregation into states resulted in an average group size of 406 with extremes of 132 and 1,250. The six region sizes were 3,078, 2,869, 3,395, 2,333, 763, and 1,382 (mean = 2,303).

Table 4 shows a partition of sums of squares and cross-products for the variables income and education (highest grade attained). The between-group sums of squares are weighted by the number of individuals in each area. These quantities are presented here because of their relevance to the derivation of equations 3 and 4 and could be used to calculate the necessary statistics. In practice, however, it is easier to calculate the ecological correlation coefficients and regression parameters directly from suitable statistical software packages without using the intermediate sums of squares. The quantity ρ_e is a weighted correlation of $\bar{X}_i$ and $\bar{Y}_i$ with weights equal to the number of individuals in each group. Similarly, β_e can be calculated as the weighted regression of $\bar{Y}_i$ on $\bar{X}_i$ with weights equal to the number of individuals in each group. The quantity ρ_w may be conveniently calculated from the within-group sums of squares. These same quantities may be used to calculate β_w. Alternatively, β_w can be calculated as the weighted average of within-group regression slopes; the weights for β_w are the within-group sums of squares about the mean for the independent variable X. Alternatively, in a linear analysis of covariance model that fits a separate intercept for each group, the common slope is β_w.

The behavior of ecological correlations in the NHANES II data can be seen by examining the results in table 5. Significance levels for these estimated correlation coef-

TABLE 4

Analyses of covariance for the NHANES data (x = income, y = education)

Source	df	Σx^2	Σxy	Σy^2
Between primary sampling units	63	7,761	5,468	10,609
Individuals within primary sampling units	13,756	92,523	30,769	143,669
Between states	33	5,675	4,544	8,130
Individuals within states	13,786	94,609	31,693	146,148
Between regions	5	2,634	3,567	4,905
Individuals within regions	13,814	97,650	32,670	149,373
Total	13,819	100,284	36,237	154,278

ficients were calculated by comparing $(N - 2)^{1/2}\,\hat{\rho}/(1 - \hat{\rho}^2)^{1/2}$ with the t distribution with $N - 2$ df (17), where N is the number of observations on which the correlation is based. It is apparent that correlations among group averages not only inadequately estimate both the correlation among individuals and the average within-group correlation but also show no discernible qualitative consistency to this effect. The ecological correlations may under- or overestimate the correlation among individuals, with equally unpredictable significance levels. The average within-area correlations in table 5 are, however, essentially identical to the overall correlation that ignores groups, indicating that for these data, most of the total variation is due to within-group variation.

Of the 13 variable pairs examined in table 5, four show increasing positive or negative correlation as the level of aggregation increases, three show the reverse pattern, and six are mixed. The levels of statistical significance for testing the hypothesis that the correlation is zero (not shown) may either increase or decrease with higher aggregation. It is not surprising that aggregation fails to preserve the statistical significance of some overall correlations (e.g., race-income) because of reduced degrees of freedom at higher levels of aggregation, but it is worrisome that nonsignificant individual level statistics can produce significant ecological estimates (e.g., height-body mass index).

The behavior of ecological regression coefficients in the NHANES II data can be seen by examining table 6. Approximate 95 per cent confidence limits were calculated as ± $t_{0.975}$ ($N - 2$ df) times the estimated standard error of each regression coefficient. These confidence bounds are useful for testing the hypothesis that a coefficient differs from zero, but they are not as useful for testing the equality of coefficients across levels of grouping because the covariance of two coefficients is not zero. Here again, there appears to be no constant pattern to the relation among the β's, except that β lies between β_e and β_w as required by expression 4. The order of β_e and β_w is, however, not consistent: The ecological regression coefficients may fluctuate (weight-body mass index), increase (income-education), or decrease (protein-fat) with respect to the overall regression. As for the correlation coefficients, nonsignificant individual level effects may produce significant ecological regressions (height-body mass index) and, conversely, significant individual level effects may produce nonsignificant ecological regressions (race-income). This correspondence between the regression and correlation occurs because the significance of the coefficients is assessed using equivalent tests. We note the relative consistency of the dietary ecological regressions in table 6, except at the regional level for two of them, but can offer no theoretical explanation for the observed behavior. For most variables in table 6, β_w

TABLE 5

Selected Pearson correlation coefficients for NHANES data

Variable pair	Individual (N = 13,820)	Primary sampling unit (N = 64)		State (N = 34)		Region (N = 6)	
	ρ	ρ_e	ρ_w	ρ_e	ρ_w	ρ_e	ρ_w
Height-weight	0.502	0.320	0.504	0.272*	0.504	0.442*	0.502
Weight-body mass index	0.855	0.769	0.856	0.792	0.855	0.440*	0.855
Height-body mass index	−0.007*	−0.356	−0.003*	−0.369	−0.004*	−0.610*	−0.006*
Age-body mass index	0.329	0.532	0.326	0.637	0.325	0.853	0.328
Calories-protein	0.831	0.822	0.831	0.797	0.831	0.516*	0.831
Calories-fat	0.898	0.898	0.898	0.857	0.899	0.206*	0.899
Protein-fat	0.785	0.711	0.787	0.658	0.787	−0.140*	0.787
Income-education	0.291	0.603	0.267	0.669	0.270	0.992	0.271
Race-income	−0.179	−0.076*	−0.200	−0.040*	−0.194	−0.656*	−0.169
Race-education	−0.131	−0.054*	−0.147	0.015*	−0.146	−0.596*	−0.121
Race-weight	0.003*	−0.233*	0.014*	−0.349	0.013*	0.098*	0.002*
Sex-weight	−0.314	−0.088*	−0.315	−0.018*	−0.315	0.025*	−0.314
Sex-height	−0.635	−0.377	−0.638	−0.426	−0.637	−0.555*	−0.636

* $p > 0.05$ (nonsignificant).

is close to β, indicating that most of T_{xx} is due to W_{xx}. Even when β_w is close to β, the ratio T_{xx}/E_{xx} can be large, making β_e and β_w different (equation 5).

Examination of the confidence bounds reveals that, in some cases such as height-weight and income-education, the values of β_e are not consistent with the values of β_w. In other words, the disparity between the ecological and individual regression estimates is not due solely to the larger variability of the former but is due, presumably, to a confounding factor. For the protein-fat pair, the ecological regression on regions not only is inconsistent with the individual coefficient but also has the wrong algebraic sign. We note the particularly wide confidence limits for ecological regression coefficients when the independent variable is the binary variable sex. This occurs because the sex ratio differs little across groups, suggesting that in such a situation ecological regression is particularly unwise.

Finally, the results for regressions reversing the role of the dependent and independent variables may be deduced from tables 5 and 6 since the correlation coefficient is the geometric mean of the regression coefficient for Y on X and that for X on Y. The significance levels for the two regressions are the same.

DISCUSSION

This paper shows that the relation between β_e and the individual level measurements is not qualitatively constant but depends on the effects of grouping. The case $\beta = \beta_e = \beta_w$ could arise only when the grouping has no effect on the regression. The emphasis here is, however, on observational studies in which individual level data are unavailable, and hence the assumption of no group effects cannot be verified.

In the case of linearly related variables, the ecological fallacy can be understood as the incorrect equating of between-group ratios of the sums of squares and cross-products with the relations between the totals (or within groups). When data are

TABLE 6

Selected regression coefficients for NHANES data

Independent variable	Dependent variable	Individual (N = 13,820) (β)	Primary sampling unit (N = 64)		State (N = 34)		Region (N = 6)	
			β_e	β_w	β_e	β_w	β_e	β_w
Height	Weight	0.821* (0.797, 0.845)†	0.384 (0.095, 0.672)	0.827 (0.803, 0.851)	0.345 (−0.094, 0.784)	0.825 (0.802, 0.849)	0.318 (−0.578, 1.215)	0.822 (0.799, 0.846)
Weight	Body mass index	0.272 (0.269, 0.274)	0.293 (0.231, 0.355)	0.271 (0.269, 0.274)	0.310 (0.224, 0.397)	0.271 (0.269, 0.274)	0.188 (−0.344, 0.720)	0.272 (0.269, 0.274)
Height	Body mass index	−0.004 (−0.012, 0.005)	−0.163 (−0.271, −0.054)	−0.001 (−0.010, 0.007)	−0.183 (−0.349, −0.017)	−0.002 (−0.011, 0.007)	−0.188 (−0.526, 0.151)	−0.003 (−0.012, 0.005)
Age	Body mass index	0.084 (0.080, 0.088)	0.091 (0.054, 0.128)	0.084 (0.080, 0.088)	0.105 (0.059, 0.150)	0.083 (0.079, 0.087)	0.108 (0.016, 0.200)	0.084 (0.080, 0.088)
Calories	Protein	0.036 (0.036, 0.037)	0.039 (0.032, 0.046)	0.036 (0.036, 0.037)	0.043 (0.031, 0.055)	0.036 (0.036, 0.037)	0.034 (−0.044, 0.112)	0.036 (0.036, 0.037)
Calories	Fat	0.044 (0.044, 0.045)	0.047 (0.041, 0.052)	0.044 (0.044, 0.045)	0.043 (0.034, 0.053)	0.044 (0.044, 0.045)	0.007 (−0.039, 0.053)	0.044 (0.044, 0.045)
Protein	Fat	0.886 (0.874, 0.897)	0.776 (0.581, 0.971)	0.887 (0.876, 0.899)	0.620 (0.364, 0.875)	0.888 (0.877, 0.900)	−0.072 (−0.782, 0.637)	0.888 (0.876, 0.899)
Income	Education	0.361 (0.341, 0.381)	0.705 (0.468, 0.942)	0.333 (0.312, 0.353)	0.801 (0.479, 1.121)	0.335 (0.315, 0.355)	1.354 (1.121, 1.587)	0.335 (0.315, 0.354)
Race	Income	−1.403 (−1.532, −1.274)	−0.346 (−1.508, 0.815)	−1.714 (−1.854, −1.573)	−0.202 (−2.021, 1.616)	−1.594 (−1.728, −1.459)	−6.768 (−17.566, 4.031)	−1.320 (−1.448, −1.192)
Race	Education	−1.277 (−1.438, −1.116)	−0.291 (−1.651, 1.069)	−1.567 (1.744, −1.390)	0.091 (−2.087, 2.269)	−1.494 (−1.663, −1.326)	−8.390 (−24.067, 7.288)	−1.167 (−1.327, −1.007)
Race	Weight	0.121 (−0.656, 0.897)	−1.943 (−4.006, 0.121)	0.727 (−0.155, 1.608)	−3.129 (−6.155, −0.103)	0.636 (−0.199, 1.470)	0.783 (−10.243, 11.810)	0.110 (−0.672, 0.893)
Sex	Weight	−10.042 (−10.548, −9.537)	−3.877 (−15.088, 7.334)	−10.066 (−10.572, −9.560)	−0.913 (−19.663, 17.837)	−10.060 (−10.565, −9.554)	1.271 (−70.616, 73.158)	−10.044 (−10.550, −9.538)
Sex	Height	−12.417 (−12.669, −12.166)	−13.941 (−22.633, −5.249)	−12.412 (−12.662, −12.161)	−17.513 (−30.923, −4.103)	−12.408 (−12.658, −12.157)	−39.929 (−122.980, 43.125)	−12.413 (−12.664, −12.161)

* Regression coefficient.

† 95 per cent confidence limits. In some instances, the bounds are asymmetric because of roundoff error.

available on individuals and the groups to which they were assigned, the ecological analysis is seen to be a part of the usual analysis of covariance. Ecological analyses are, however, incomplete analyses of covariance since, if the information on individuals were available, it could be used to avoid these problems, although covariance adjustment on unplanned experimental data has its own difficulties (18). We believe that the investigator is never justified in interpreting the results of ecological analyses in terms of the individuals who give rise to the data. This may seem to many readers to be an overstatement; however, our theoretical and empirical analyses offer no consistent guidelines for the interpretation of ecological correlations or regressions when data on individuals are unavailable.

We note that the literature on ecological analysis, as well as our derivation above, generally neglects the possibility of interactions. If, after adjustment for confounding, the regression coefficient for X is found to differ significantly according to some other variable, that variable and X are said to have an interaction (on the scale of measurement being used). In such a situation, one may well be interested in more than the overall slope (whether individual level or ecological) or the average within-group slope. For example, if the effect of X on Y were significantly different for males compared with females, then in addition to the coefficient for X, the interaction term would also be important. This could only be determined if sex were included in the regression. While in theory this could be done for both individual level and ecological regression, in the latter situation the groups would have to have different sex ratios as well as different values of $\bar{X}$ in order to fit the full regression.

Ecological analyses become flawed in exactly the same circumstances that individual level analyses do, i.e., in the presence of confounding. The consequences of confounding bias in the ecological analysis are more severe, however. With respect to inferences about individuals, the proper role of ecological analyses is to generate new hypotheses which must then be tested using more appropriate experimental or observational methods. To interpret ecological analyses sensibly, the investigator should use outside information to judge the likelihood of serious errors. Additionally, inferences should be confined to the level of observation (or experimentation). These conclusions apply both to simple correlation coefficients and to linear regression slopes. While we are unaware of theory for nonlinear response models, it seems likely that similar problems might arise, and the same caution should be used.

References

1. Morgenstern H. Uses of ecologic analysis in epidemiologic research. Am J Public Health 1982;72:1336–44.
2. Durkheim E. Suicide: a study in sociology. New York: The Free Press, 1951:153.
3. Robinson WS. Ecological correlations and the behavior of individuals. Am Sociol Rev 1950;15:351–7.
4. Duncan OD, Cuzzort RP, Duncan B. Statistical geography. New York: The Free Press, 1961.
5. Connor MJ, Gillings D. An empiric study of ecological inference. Am J Public Health 1974;74:555–9.
6. Kalimo E, Bice TW. Causal analysis and ecological fallacy in cross-national epidemiologic research. Scand J Soc Med I 1973;1:17–24.
7. Thind IS. Diet and cancer—an international study. Int J Epidemiol 1986;15:160–3.
8. Goodman LA. Ecological regression and the behavior of individuals. Am Soc Rev 1953;13:663–4.
9. Goodman LA. Some alternatives to ecological correlation. Am J Sociol 1959;64:610–25.
10. Langbein LI, Lichtman AJ. Ecological inference. Beverly Hills, CA: Sage Publications, 1978.
11. Firebaugh G. A rule for inferring individual level relationships from aggregate data. Am Sociol Rev 1978;43:557–72.
12. Ostie B, Mensing RW. Statistics in research. Chap 13. Ames, IA: The Iowa State University Press, 1975.
13. Stavraky PM. The role of ecologic analysis in studies of the etiology of disease: a discussion with reference to large bowel cancer. J Chronic Dis 1976;29:435–44.
14. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: principles and quantitative methods. Belmont, CA: Lifetime Learning Publications, 1982:81.
15. National Center for Health Statistics. Plan and operation of the Second National Health and Nutrition Examination Survey 1976–1980. Washington, DC: US GPO, 1981. (DHHS publication no. (PHS)81-1317).

16. National Center for Health Statistics. Public use data tape documentation: Health History Tape no. 5305, Food Frequency Tape no. 5701, Anthropometry Tape no. 5301. Washington, DC: US GPO, 1984.

17. Brownlee KA. Statistical theory and methodology. New York: John Wiley, 1965:413–14.

18. Snedecor GW, Cochran WG. Statistical methods. Chap 18. Ames, IA: The Iowa State University Press, 1980.

Appendix 1

Theorem: For the data in table 1,

$$\rho = \frac{N^2 + 1 - 2k^2}{N^2 - 1}. \tag{A.1}$$

Proof: By definition, the overall correlation coefficient is

$$\rho = \frac{\sum \sum X_{ij} Y_{ij} - N\bar{\bar{X}}\bar{\bar{Y}}}{\{(\sum \sum X_{ij}^2 - N\bar{\bar{X}}^2)\ (\sum \sum Y_{ij}^2 - N\bar{\bar{Y}}^2)\}^{½}}.$$

Since $\sum \sum X_{ij}^2 = \sum \sum Y_{ij}^2$ and $\bar{\bar{X}} = \bar{\bar{Y}}$,

$$\rho = \frac{\sum \sum X_{ij} Y_{ij} - N\bar{\bar{X}}^2}{\sum \sum X_{ij}^2 - N\bar{\bar{X}}^2}. \tag{A.2}$$

We begin by calculating

$$\sum \sum X_{ij}\ Y_{ij} = \sum_{i=1}^{N/k} \sum_{j=1}^{k} ((i-1)k + j)(ik + 1 - j).$$

Expanding the product and using the relations

$$\sum_{i=1}^{m} i = \frac{m(m+1)}{2}$$

and

$$\sum_{i=1}^{m} i^2 = \frac{m(m+1)(2m+1)}{6},$$

yields

$$\sum \sum \mathrm{X}_{ij}\mathrm{Y}_{ij} = \frac{2N^3 + 3N^2 - k^2N + 2N}{6}. \tag{A.3}$$

Similarly,

$$N\bar{\bar{X}}^2 = \frac{N^3 + 2N^2 + N}{4}, \tag{A.4}$$

and

$$\sum \sum X_{ij}^2 = \frac{2N^3 + 3N^2 + N}{6}. \tag{A.5}$$

Substituting expressions A.3–A.5 into A.2 yields equation A.1.

Appendix 2

To show that equation 3 is Robinson's result (3), define

$$\eta_y^2 = \frac{E_{yy}}{T_{yy}},$$

and

$$\eta_x^2 = \frac{E_{xx}}{T_{xx}}.$$

These quantities, termed the correlation ratios, measure the degree of clustering in X and Y among areas (3, p. 355). We can write equation 3 in terms of η_x^2 and η_y^2 by noting

$$(1 - \eta_y{}^2) = \frac{T_{yy} - E_{yy}}{T_{yy}} = \frac{W_{yy}}{T_{yy}},$$

and

$$(1 - \eta_x{}^2) = \frac{W_{xx}}{T_{xx}}.$$

Therefore, equation 3 becomes

$$\rho = \eta_x \, \eta_y \, \rho_e + (1 - \eta_x{}^2)^{1/2} \, (1 - \eta_y{}^2)^{1/2} \, \rho_w . \tag{B.1}$$

Solving equation B.1 for ρ_e yields

$$\rho_e = (1/\eta_x\eta_y) \; \rho - \left\{ \frac{(1 - \eta_x{}^2)^{1/2} \; (1 - \eta_y{}^2)^{1/2}}{\eta_x \; \eta_y} \right\} \rho_w, \tag{B.2}$$

which is the relation given by Robinson (3) without proof.