# Ecological Inference and the Ecological Fallacy

by David A. Freedman
Department of Statistics
University of California
Berkeley, CA 94720
Prepared for the International Encyclopedia
of the Social & Behavioral Sciences
Technical Report No. 549
15 October 1999

In 19th century Europe, suicide rates were higher in countries that were more heavily Protestant, the inference being that suicide was promoted by the social conditions of Protestantism (Durkheim 1897; also see Neeleman and Lewis 1999). According to Carroll (1975), death rates from breast cancer are higher in countries where fat is a larger component of the diet, the idea being that fat intake causes breast cancer. These are 'ecological inferences,' that is, inferences about individual behavior drawn from data about aggregates.

To continue with Durkheim, the Protestant countries were different from the Catholic countries in many ways besides religion (the problem of 'confounding'). Moreover, Durkheim's data do not tie individual suicides to any particular religious faith. The first problem, of confounding, must be dealt with in any observational study. But the second problem—that exposure and response are measured only for aggregates rather than for individuals—is specific to ecological studies. If there is no confounding, the expected difference between effects for groups and effects for individuals is 'aggregation bias'; in general, the difference is partly attributable to confounding and partly to aggregation bias.

The ecological fallacy consists in thinking that relationships observed for groups necessarily hold for individuals: if countries with more Protestants tend to have higher suicide rates, then Protestants must be more likely to commit suicide; if countries with more fat in the diet have higher rates of breast cancer, then women who eat fatty foods must be more likely to get breast cancer. These inferences may be correct, but are only weakly supported by the aggregate data.

Ecological studies in epidemiology yielded important insights for Snow (1855), Finlay (1881), Goldberger (Terris 1964), and Dean (1938) among others. However, it is all too easy to draw incorrect conclusions from aggregate data. Greenland and Robins (1994) review the issues. For one example, recent studies of individual-level data cast serious doubt on the link between breast cancer and fat intake (Holmes et al. 1999). Another well-known example, on the sources of popular support for the Nazi party in pre-war Germany, is discussed by Lohmoller et el. (1985).

## 1. Ecological Correlation

Robinson (1950) discusses ecological inference, stressing the difference between ecological correlations and individual correlations. One striking example is the relationship between nativity and literacy. For each of the 48 states in the USA of 1930, Robinson computes two numbers: the percent of the population who are foreign-born, and the percent who are literate. (The base for each percentage is the number of residents of the state who are 10 or older; data are from the Census of 1930.) The correlation between the 48 pairs of numbers is .53. This is an 'ecological' correlation, because the unit of analysis is not an individual person but a group of people—the residents of a

state. The ecological correlation suggests a positive association between foreign birth and literacy: the foreign-born are more likely to be literate (in American English) than the native-born. In reality, the association is negative: the correlation computed at the individual level is −.11. The ecological correlation gives the wrong inference. The sign of the correlation is positive because the foreign-born tend to live in states where the native-born are relatively literate.

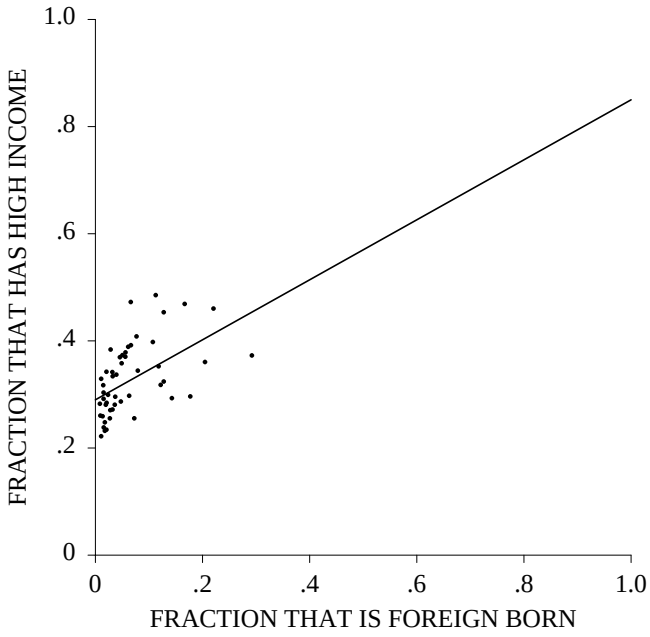


*Figure 1*
For each state, the horizontal axis shows the fraction of the population aged 25 and over that is foreign-born; the vertical axis shows the fraction having high incomes; the regression line is shown too; data from the Current Population Survey

The example can be replicated using data from the March 1995 Current Population Survey, restricted to persons age 25 and over. Literacy is not reported, but incomes are. Say that a person has 'high' income if family income in 1994 was $50,000 or more. Figure 1 shows for each of the 50 states the fraction of persons who are foreign-born, and the fraction of persons who have high incomes. The regression line is shown too. The ecological correlation (across the 50 areas) is .52, suggesting that the foreign-born have higher incomes on the whole than the native-born. The truth is the opposite. About 35% of the native-born have high incomes, compared to 28% for the foreign-born. The correlation at the individual level is −.05. In Fig. 1, there is no state where the fraction of foreign-born approaches 1. However, the same sort of reversal can occur even with a full spectrum of $x$-values. The issue is the disaggregation, not the range of the data.

## 3. Ecological Regression

Under certain restrictive assumptions, individual behavior can be inferred from aggregate data using a technique called 'ecological regression.' The unit of analysis in the regression is a group of people, typically defined by geography (as in Fig. 1). The technique has been widely used in voting rights litigation in the USA. For discussion from various perspectives, see Freedman et al. (1991), Grofman and Davidson (1992), Achen and Shively (1995), or Cho (1998).

It may, for instance, be desired to estimate the support for a particular candidate among Hispanics and non-Hispanics. For each precinct in an electoral contest, suppose the fraction $x$ of voters who are Hispanic is known; the fraction $y$ of votes obtained by the candidate of interest is also known. By contrast, the fraction of Hispanic voters who voted for the candidate is unknown, due to the secrecy of the ballot.

A regression equation can be fitted to the data:

$$y_i = a + bx_i + \epsilon_i. \tag{1}$$

In equation (1), the subscript $i$ indexes precincts; $x_i$ is the fraction of voters in precinct $i$ who are Hispanic; $y_i$ is the fraction of votes cast in that precinct for the candidate of interest. (More exactly, $y_i$ is the number of votes for that candidate, divided by the number of voters.) The precinct-level disturbance terms $\epsilon_i$ would usually be assumed to be independent and identically distributed with mean 0, although this is seldom made explicit.

The parameters $a$ and $b$ in (1) can be estimated by ordinary least squares; call the estimates $\hat{a}$ and $\hat{b}$, respectively. Then $\hat{a}$ would be interpreted as the fraction of non-Hispanic voters who supported the candidate in question: after all, $\hat{a}$ is the height of the regression line at $x = 0$, corresponding to precincts with no Hispanic voters. Likewise, $\hat{a} + \hat{b}$ (the height of the regression line at $x = 1$) would be interpreted as the fraction of Hispanic voters who supported the candidate.

The data are available for groups defined by area of residence; the inference is to groups defined by ethnicity. The statistical logic connecting the inference to the data rests on the 'constancy assumption,' that voting preferences within ethnic groups do not systematically depend on the ethnic makeup of the area of the residence (Goodman 1953, 1959). In Fig. 1, the regression equation is $y = .29 + .56x$, leading to the inference that 29% of the native-born have high incomes, compared to $29 + 56 = 85\%$ of the foreign-born. In reality, as noted above, the percentages are 35% and 28%. The ecological inference is in error, because the constancy assumption fails: the incomes of the native-born increase systematically with fraction of foreign-born in the state. Immigration to the USA tends to concentrate in richer states—California, Hawaii and New York rather than Kentucky, Tennessee and West Virginia.

## 4. The Method of Bounds

In 1995 in the state of Washington, 7.9% of the population was foreign-born, while 34.4% had high incomes. These are known. Let $p$ be the fraction of the foreign-born who had high incomes, and let $q$ be the corresponding fraction for the native-born. Investigators making ecological inferences would generally not know the fractions $p$ and $q$. The two unknowns satisfy an equation:

$$.079p + (1 - .079)q = .344. \tag{2}$$

This is one equation with two unknowns, which is the problem with ecological inference. However, (2) does contain some information, because $q = (.344 - .079p)/(1 - .079)$. Since $p$ is bounded between 0 and 1, it follows that $q$ is trapped between $.344/(1 - .079) \doteq .374$ and $(.344 - .079)/(1 - .079) \doteq .288$. This is 'the method of bounds' (Duncan and Davis 1953).

*Table 1*
The method of bounds; nativity (native- or foreign-born) by income ('high' means $50,000 a year or more) for persons age 25+ in the state of Washington; data from the Current Population Survey

| | Native | Foreign | Total |
|---|---|---|---|
| Low | ??? | ??? | 2,182,000 |
| High | ??? | ??? | 1,146,000 |
| Total | 3,066,000 | 263,000 | 3,329,000 |

The idea is shown in numerical form, for the Washington data (Table 1). In applications, the margins of the $2 \times 2$ table would be known, but the interior cells would not be known. The number of foreign-born persons with high incomes must be positive, but cannot exceed 263,000 (the column total). Therefore, the number of native-born persons with high incomes must be between 1,146,000 and $1{,}146{,}000 - 263{,}000 = 883{,}000$. So the percentage of native-born persons with high incomes is bounded between $1{,}146{,}000/3{,}066{,}000 = 37.4\%$ and $883{,}000/3{,}066{,}000 = 28.8\%$, as noted earlier. Numerical bounds can be computed for each study area, and then aggregated. However, in many applications, the bounds are too broad to be informative.

## 5. Ecological Regression with Random Coefficients

Models with random coefficients have been used to make ecological inferences. By way of illustration, consider a model for nativity and income. Index the study areas by $i$. As before, let $x_i$ be the fraction of the population in area $i$ that is foreign-born, and $y_i$ the fraction with high incomes. These fractions are known, as is the total population $n_i$ of area $i$. Let $p_i$ be the fraction of the foreign-born with high incomes, and let $q_i$ be the corresponding fraction for the native-born, so

$$y_i = p_i x_i + q_i(1 - x_i). \tag{3}$$

The problem is to make inferences about $p_i$ and $q_i$. To solve such problems, King (1997) assumes the pairs $(p_i, q_i)$ to be independent and identically distributed across the study areas. This is his version of the constancy assumption: the statistical behavior of a demographic group is not allowed to depend on area of residence. The parent distribution is taken to be bivariate normal, conditioned to lie in the unit square so that $p_i$ and $q_i$ are both between 0 and 1. The five parameters of this parent normal distribution can be estimated from the data by maximum likelihood, and estimates $(\hat{p}_i, \hat{q}_i)$ can be derived from (3). Rates for each demographic group can then be estimated by addition over areas, as $\sum_i n_i \hat{p}_i / \sum_i n_i x_i$ and $\sum_i n_i \hat{q}_i / \sum_i n_i (1 - x_i)$ respectively. More complex models have been developed too.

*Table 2*
Comparison of methods for estimating the percentage of population groups with incomes of $50,000 a year or more, from the data in Fig. 1; truth is known from the Current Population Survey; 'nbd' is the neighborhood model; 'ecoreg' is ecological regression; and 'random' is King's random-coefficients model, as implemented in his EZIDOS software version 1.31

| | Native-born | Foreign-born |
|---|---|---|
| Truth | 35% | 28% |
| Nbd | 34% | 36% |
| Ecoreg | 29% | 85% |
| Random | 30% | 72% |

To show the force of the constancy assumption, the 'neighborhood model' can be used. According to the neighborhood model, behavior is determined by geography not demography. Then $\hat{p}_i = \hat{q}_i = y_i$ for each study area $i$. (In the example of nativity and income, if 33% of the residents of a particular study area have high incomes, this percentage applies equally to the foreign-born and the native-born in that area.) The neigborhood model turns the constancy assumption on its head. In a variety of test applications where all the data are available, the neighborhood model gives more accurate estimates for demographic groups than ecological regression or random-coefficients models. Table 2 gives one example, which is not atypical—although the standard errors for the random-coefficients model are large, because the number of study areas is relatively small. In some contexts, of course, the neighborhood model proves deficient. If the data are incomplete so estimation is needed, it will be unclear which model if any is giving the right answers. For other examples and discussion from various perspectives, see Freedman et al. (1991, 1998), Achen and Shively (1995), King (1997), Cho (1998), or Schuessler (1999).

## 6. Summary and Conclusions

Aggregate data are often easier to obtain than data on individuals, and may offer valuable clues about individual behavior. Ecological inferences will therefore continue to be made. The problems of confounding and aggregation bias, however, are unlikely to be resolved in the proximate future.

## Bibliography


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

Carroll K 1975 Experimental evidence of dietary factors and hormone-dependent cancers. *Cancer Research* 35: 3374–83

Cho, W K Tam 1998 Iff the assumption fits: a Comment on the King ecological inference solution. *Political Analysis* 7: 143-163.

Duncan O D, Davis B 1953 An alternative to ecological correlation. *American Sociological Review* 18: 665–66

Dean, H T 1938. Endemic fluorosis and its relation to dental caries. *Public Health Reports* 53: 1443–52

Durkheim E 1897 *Le suicide.* F. Alcan, Paris. English translation by J A Spalding, 1951, Free Press, Collier-MacMillan, Toronto, Canada

Finlay, C J 1881. El mosquito hipoteticamente considerado como agente de transmision de la fiebre amarilla. (The mosquito hypothetically considered as the agent of transmission of yellow fever.) *Anales de la Academia de Ciencias Medicas, Físicas y Naturales de Habana* XVIII. English translation by Finlay, reprinted in Buck C, Llopis A, Nájera E, Terris M 1989. *The Challenge of Epidemiology: Issues and Selected Readings,* Scientific Publication No. 505, World Health Organization, Geneva

Freedman D A, Klein S P, Sacks J, Smyth C A, Everett C G 1991 Ecological regression and voting rights. *Evaluation Review* 15: 659–817 (with discussion).

Freedman D A, Klein S P, Ostland M, Roberts M R 1998 Review of *A Solution to the Ecological Inference Problem*. *Journal of the American Statistical Association* 93: 1518–22; with discussion, vol. 94 (1999) pp. 352–57.

Goodman L 1953 Ecological regression and the behavior of individuals. *American Sociological Review* 18: 663–64

Goodman L 1959 Some alternatives to ecological correlation. *American Journal of Sociology* 64: 610–25

Grofman B, Davidson C 1992 *Controversies in Minority Voting: The Voting Rights Act in Perspective.* Brookings Institution, Washington, D.C.

Greenland S, Robins J 1994 Invited commentary: ecologic studies—biases, misconceptions, and counterexamples. *American Journal of Epidemiology* 139: 747–60

Holmes M D, Hunter D J, Colditz G A, Stampfer M J, Hankinson S E, Speizer F E, Rosner B, Willett W C 1999 Association of dietary intake of fat and fatty acids with risk of breast cancer. *Journal of the American Medical Association* 281: 914–920

King G 1997 *A Solution to the Ecological Inference Problem.* Princeton University Press

Lohmoller J-B, Falter J, Link A, de Rijke J 1985 Unemployment and the rise of national socialism: contradicting results from different regional aggregations. In P Nijkamp, ed. Measuring the Unmeasurable. Martinus Nijhoof, Den Haag, pp357-70.

Neeleman J, Lewis G 1999. Suicide, religion, and socieconomic conditions. An ecological stuy in 26 countries, 1990. *Journal of Epidemiology and Community Health* 53: 204–210.

Robinson W S 1950. Ecological correlations and the behavior of individuals. *American Sociological Review* 15: 351–57

Schuessler, A A 1999. Ecological inference. *Proceedings of the National Academy of Sciences USA* 96: 10578–10581.

Snow, J 1855 *On the Mode of Communication of Cholera.* Churchill, London. Reprinted by Hafner, New York, 1965

Terris M 1964 *Goldberger on Pellagra.* Louisiana State University Press, Baton Rouge

David A. Freedman
Department of Statistics
University of California
Berkeley, CA 94720
Prepared for the International Encyclopedia
of the Social & Behavioral Sciences
Technical Report No. 549
15 October 1999