Environmental Health Perspectives
Vol. 32, pp. 311-315, 1979

Methodological Problems Arising from the Choice of an Independent Variable in Linear Regression, with Application to an Air Pollution Epidemiological Study

by Inge F. Goldstein,* Joseph L. Fleiss,† Martin Goldstein,‡ and Leon Landovitz**

In epidemiological studies using linear regression, it is often necessary for reasons of economy or unavailability of data to use as the independent variable not the variable ideally demanded by the hypothesis under study but some convenient practical approximation to it. We show that if the correlation coefficient between the "practical" and "ideal" variables can be obtained, then a range of uncertainty can be obtained within which the desired regression coefficient of dependent on "ideal" variable may lie. This range can be quite wide, even if the practical and ideal variables are fairly well correlated. These points are illustrated with data on observed regression coefficients from an air pollution epidemiological study, in which pollution measured at one station in a large metropolitan area (containing 40 aerometric stations) was used as the practical approximation to the city-wide average pollution. The uncertainties in the regression coefficients were found to exceed the regression coefficients themselves by large factors. The problem is one that may afflict application of linear regression in general, and suggests caution when selecting independent variables for regression analysis on the basis of convenience, rather than relevance to the hypotheses tested.

Introduction

In the course of reviewing the literature on the acute health effects of air pollution (*1-5*) we became aware that too little attention has been paid to the problem of incomplete concordance between the population described by the health effects data and the population described by the air pollution data. We believe that this problem might have a much wider relevance than to the specific studies which suggested it to us.

The problem may be stated in general terms as follows. It is often necessary, for reasons of economy or availability of data, to use as an independent variable not the "ideal" variable demanded by the hypothesis under study but some more easily accessible "practical" variable which is believed to be an adequate approximation to the ideal one. If the ideal variable is really inaccessible, so that we have no quantitative information about the relationship between it and the practical variable, we have no choice but to rely on our intuitive judgment that the second is an adequate approximation to the first. However, it occasionally happens that some quantitative information, even if incomplete, can be obtained about their relationship — for example, their correlation. The question we consider in this paper is: given such information, can we estimate what kinds of errors exist in the regression or correlation coefficients between the dependent variable and the "practical" independent variable if these coefficients are regarded as estimators of the regression or

*Division of Epidemiology, Columbia University School of Public Health, 600 West 168th Street, New York, New York 10032.

†Division of Biostatistics, Columbia University, New York, New York 10032.

‡Division of Natural Sciences and Mathematics, Yeshiva University, New York, New York 10033.

**Computer Center, Graduate Center of the City University of New York, New York, New York 10036.

correlation coefficients between the dependent and "ideal" independent variables?

The problem may be made more concrete by consideration of a particular air pollution epidemiological study (3). In this study a multiple regression analysis was performed of daily city-wide mortality in New York City on daily values of sulfur dioxide and smokeshade (coefficients of haze) measured at a single centrally located monitoring station over a 10-year period. Excess deaths attributable to pollution were calculated by multiplying the regression coefficients by the average pollutant concentrations observed at the central station. It was found that the excess deaths attributable to sulfur dioxide, which greatly exceeded the share attributable to particulates, remained constant over the ten-year period, even though the average annual sulfur dioxide concentration decreased to one-third of its initial value over this time period in response to regulatory efforts. The conclusion drawn was that sulfur dioxide is not really the agent responsible for the excess mortality, but instead the agent is some component of urban pollution whose daily fluctuations are highly correlated with sulfur dioxide, but which has not changed over the ten-year period in which sulfur dioxide has decreased significantly. The conclusions of this study have been cited in the course of administrative and legislative hearings on air quality standards (6).

During the course of the ten-year study, a number of additional air pollution monitoring stations were established, and, from 1967 on, data from a network of 40 such stations have been available. On the basis of an extended analysis of the data provided by this network for the three-year period from 1969 to 1971 inclusive (7-9), we have found that correlation coefficients for daily pollution concentrations between pairs of stations, among them the central station used in the 10-year study, were quite low, averaging 0.4 for smokeshade and 0.5 for sulfur dioxide. We were led therefore to examine how the conclusions of the former study might be altered by this new information about the relation between pollution measured at the central station, regarded as the practical variable, and the city-wide daily average, regarded as the ideal variable.

The problem can be stated concisely in statistical terms. Given the regression coefficient α_{yx} of a dependent variable y on an independent variable x, and the correlation coefficient $\rho_{xx'}$ between x and a second variable x', what can be inferred about the regression coefficient $\alpha_{yx'}$? We assume throughout that the observable coefficients are based on sufficiently many observations that they may be taken as population parameters. Considering the coefficients as being subject to sampling variation would add yet further uncertainty to the determination of the desired coefficients.

Limits of Uncertainty on the Regression Coefficient

Using the symbols σ_x, $\sigma_{x'}$, and σ_y for the standard deviation of the variables x, x', and y, respectively, we have the well known relation between regression and correlation coefficients

$$\alpha_{yx} = (\sigma_y/\sigma_x)\,\rho_{yx} \tag{1}$$

where ρ_{yx} is the correlation coefficient between y and x.

The basis of our determination of the relation between the observed (α_{yx}) and desired ($\alpha_{yx'}$) regression coefficients will be the formulas for partial and multiple correlation coefficients, which relate the correlation coefficients between three variables taken pairwise. Consider the partial correlation coefficient

$$\rho_{yx'\cdot x} = \frac{\rho_{yx'} - \rho_{yx}\,\rho_{xx'}}{\sqrt{1-\rho_{yx}^2}\,\sqrt{1-\rho_{xx'}^2}} \tag{2}$$

and the squared multiple correlation coefficient

$$R_{y\cdot xx'}^2 = \rho_{yx}^2 + (1-\rho_{yx}^2)\,\rho_{yx'\cdot x}^2 \tag{3}$$

Assuming that the partial correlation coefficient [Eq. (2)] is nonnegative (this is a reasonable assumption in the present case, though not a necessary one), we find that the desired correlation coefficient is

$$\rho_{yx'} = \rho_{yx}\,\rho_{xx'} + \sqrt{1-\rho_{xx'}^2}\,\sqrt{R_{yxx'}^2 - \rho_{yx}^2} \tag{4}$$

and that the desired regression coefficient is

$$\begin{aligned}\alpha_{yx'} = {} & \alpha_{yx}\,(\rho_{xx'}\,\sigma_x/\sigma_{x'}) + \\ & (\sigma_y/\sigma_{x'})\,\sqrt{1-\rho_{xx'}^2}\,\sqrt{R_{y\cdot xx'}^2 - \rho_{yx}^2}\end{aligned} \tag{5}$$

If the partial correlation coefficient is negative, the second term on the right-hand side of Eq. (5) has a minus instead of a plus sign attached to it.

The leading term in Eq. (5),

$$\alpha_{yx'}^* = \alpha_{yx}\,(\rho_{xx'}\,\sigma_x/\sigma_{x'}) \tag{6}$$

has been proposed as a "corrected" regression coefficient, with the second term omitted (3). Such use is clearly inappropriate unless either $\rho_{xx'}^2 = 1$ (this may be checked empirically from data at hand) or $R_{y.xx'}^2 = \rho_{yx}^2$ (this assumption usually requires validation from other sources, because the squared multiple correlation coefficient is not likely to be available). Without these assumptions, it is necessary to bear in mind that the "corrected" coefficient cannot be simply assumed to equal the desired coefficient; i.e., that knowledge of the pairwise coefficients is not sufficient to determine the desired coefficient.

As $\rho_{xx'}$ approaches unity (i.e., as the practical variable converges to the ideal variable), the multiplicative factor $(1 - \rho_{xx'}{}^2)^{\frac{1}{2}}$ approaches zero, and the single "corrected" coefficient indeed approaches the correct one. If, however, $\rho_{xx'}$ is 0.8, a value generally taken to represent excellent correlation, the multiplicative factor is equal to 0.6, which is quite large. When $\rho_{xx'} = 0.9$, the factor is equal to 0.44, which is still appreciable.

The second factor contributing to uncertainty in the desired coefficient, $(R_{y.xx'}{}^2 - \rho_{yx}{}^2)^{\frac{1}{2}}$, measures the increase in the proportion of the variance in y associated with adding x' to the linear regression equation already containing x. If x' contains no information relevant to y not already contained in x, then there is no increase in the proportion of the y variance explained, and the second factor is zero. If, however, the addition of x' results in as much as a 10% increase in the proportion of the variance of y explained, the value of the factor is 0.32; even if its addition adds only 5% to the proportion of y's explained variance, the value of the factor is 0.22. Factors of these magnitudes may have appreciable effects, as shown below.

We may sum up the above discussion by stating that uncertainty in the "corrected" regression coefficient is a function both of imperfect correlation between the practical and ideal independent variables and of the existence of incremental information about the dependent variable in the ideal independent variable not contained in the practical one.

Application

As an example of the degree of uncertainty that can be introduced into determining a regression coefficient, we give some results of applying the above formulae to the regression coefficients obtained in the air pollution study referred to earlier (*3*) (see Table 1). In Table 1 we give the "corrected" regression coefficients [$\alpha^*_{yx'}$ from Eq. (6)] for each of three mortality variates on the two pollution variates, sulfur dioxide and smokeshade, measured by the aerometric network, and, in addition, intervals for he desired regression coefficients associated with three assumed increments in the proportion of variance in y explained by adding x' to the prediction equation already containing x. The upper limits are given by Eq. (5) by using the plus sign; the lower limits are obtained by using the minus sign.

Even when the increment is assumed to be as small as 1%, the upper bounds on the desired regression coefficients vary from nearly twice to eight times the "corrected" coefficients. When the increment in variance is assumed to be 10%, the upper bounds are greater yet. The lower bounds on the desired coefficients are associated with a negative partial correlation coefficient; these vary from slightly positive to appreciably negative values. While mathematically possible, a negative partial correlation coefficient for the variates under study does not make substantive sense. The value of $\rho_{yx'\cdot x}$ is negative if, on days when the city-wide average pollution level is higher (or lower) than would be expected from the pollution level measured at the central station, the city-wide mortality is lower (or higher) than would be expected from the level at the central station. We are unable to

Table 1. Observed regression coefficients of city-wide mortality variates (*y*) on centrally measured pollution variates (*x*), and ranges for regression coefficients on city-wide pollution averages (*x'*).

Pollution variate x	Mortality variate y[b]	α_{yx}	$\alpha^*_{yx'}$	Interval for $\alpha_{yx'}$ at different values of $R_{y\cdot xx'}{}^2 - \rho_{yx}{}^2$		
				0.01	0.05	0.10
Sulfur dioxide[c]						
	y_1	0.22	0.33	−0.30,0.95	−1.07,1.72	−1.64,2.30
	y_2	0.01	0.02	−0.11,0.15	−0.28,0.31	−0.40,0.44
	y_3	0.11	0.16	−0.20,0.52	−0.65,0.96	−0.98,1.30
Smokeshade[d]						
	y_1	0.32	0.43	0.05,0.81	−0.43,1.28	−0.78,1.63
	y_2	0.01	0.01	−0.07,0.09	−0.17,0.19	−0.24,0.27
	y_3	0.18	0.25	0.03,0.47	−0.24,0.74	−0.45,0.95

[a] The regression coefficients were calculated by using Table 7, 8, and 9, with the aid of Table 3 of Schimmel and Murawski (*3*). The periods covered by these tables do not coincide exactly with the period 1969-1971 for which we have air pollution data for the entire city. Therefore, we calculated weighted averages of Schimmel and Murawski's data for their periods of 1967-1969 and 1970-1972, assigning a weight of ⅓ to the earlier period, to obtain estimates for the 1969-1971 period.

[b] y_1 = total mortality ($\sigma_{y1}{}^2 = 329.4$), y_2 = respiratory disease mortality ($\sigma_{y2}{}^2 = 14.82$), y_3 = heart disease mortality ($\sigma_{y3}{}^2 = 110.2$).

[c] For sulfur dioxide: $\sigma_x{}^2 = 12.53$, $\sigma_{x'}{}^2 = 3.38$, $\rho_{xx'} = 0.78$.

[d] For smokeshade $\sigma_x{}^2 = 56.13$, $\sigma_{x'}{}^2 = 13.00$, $\rho_{xx'} = 0.65$.

identify any physically sensible set of circumstances which might give rise to such a state of affairs.

In summary, the uncertainty in the "corrected" coefficients is great, even though the correlation coefficients between the practical and ideal variables (0.78 for sulfur dioxide, 0.65 for smokeshade) are appreciable, and even though only modest increments in the proportions of explained variance were assumed. Further, because negative values of $\rho_{yx' \cdot x}$ are unreasonable in this particular case, we suggest that the "corrected" coefficient is likely to be an underestimate of the desired coefficient.

Appropriateness of the "Corrected" Coefficient

In an unpublished appendix to the air pollution-acute health effects study cited earlier (*3*), available on request from the authors, Schimmel and Murawski responded to our earlier criticisms (*10*) of the use of a single monitoring station to represent the whole metropolitan area. They made an estimate, using our published correlation coefficients for pairs of stations, of the effect their procedure had on their conclusions. As they have not yet published their analysis in the literature, we will not criticize it in detail. Their correction effectively amounts to the use of the first term on the right-hand side of Eq. (3), i.e., of the "corrected" coefficient; the implication is that there is no range of uncertainty in the "corrected" regression coefficient.

In our view their implication is valid only if the assumption is made that the differences between the practical and ideal variables are due solely to random errors in the practical variable (in statistical terms, only if $\rho_{yr'.x} = 0$ or, equivalently, that $R_{y.xx'}{}^2 = \rho_{yx}{}^2$, so that x' adds nothing to the prediction of y from x alone). The validity of this assumption is germane to the subject of this paper. The assumption that the errors in the practical variable are random ones is a very restrictive one, and should not be made without direct evidence that they are.

In the case of the relation between pollution at the central station and city-wide pollution there are *a priori* reasons for believing that at least some of the factors giving rise to differences are structured rather than random. The central station is located in a specific geographic area (Harlem) of the city, and has its own relation to geographical features such as hilly terrain, distance from the rivers surrounding Manhattan Island, prevailing wind directions, and so on. These, in turn, interact with the location of major pollution sources in such a way as to make it likely that there is a distinct pattern to pollution in the vicinity of this station that makes its daily variations something other than merely unbiased estimators of the daily variations of city-wide pollution. This is made additionally plausible by the facts that pollution at the central station tends to be considerably greater than the city-wide average, and its changes over long periods of time do not always parallel those of the city-wide average.

This *a priori* argument for the existence of a nonrandom component in the relation between the two variables is confirmed by our own studies of the data of the New York City aerometric network, from which we have concluded that poor correlation between stations is not solely due to random errors of measurement (*8*, *9*). On the other hand, our attempts to find an inner structure to the data, in terms of meteorological patterns, inter-station distances, proximity to pollution source, etc., have met with only modest success (*11*). For example, principal component analyses of the covariance matrix of daily pollution readings did not reveal clearcut systematic patterns in the data.

Our failure to discover such patterns may reflect a large component of random errors or it may reflect our own lack of ingenuity in the search. In any event, our efforts in this direction are continuing.

How Ideal is the Ideal Variable?

In the above analysis we have assumed that the city-wide average level of pollution is the ideal variable, to which pollution measured at a central station is a practical approximation. However, we must acknowledge that epidemiological considerations raise questions about even this conclusion.

The pollutants measured by the New York City aerometric network — sulfur dioxide and smokeshade — are only two of a great number of different pollutants in urban air that might have adverse health effects. Further, the average as we have defined it does not weight the individual stations according to the size of the populations in the areas in which they are located, nor does it take into account the demographic characteristics of these populations.

A population-weighted city average of, say, sulfur dioxide, need not necessarily be perfectly correlated with the exposure to sulfur dioxide of the individual inhabitants of the city, some of whom spend most of their time indoors while others do not, and some of whom spend their days in a different area of the city from where they spend their nights while others stay in one area all the time. Still further, it is not clear whether the focus of the health study should be on the population as a whole, or on susceptible subgroups, nor whether mortality is a better indicator of health effects of pollution than morbidity.

It should be clear from this discussion that what we have designated as the "ideal" variable is far from ideal. We do not have a solid empirical basis for deciding what the ideal independent variable in air pollution studies should be, nor, inevitably, any knowledge of how well the city-wide average of sulfur dioxide or smokeshade correlates with it. We must acknowledge therefore that as great as the degree of uncertainty is in our estimates of the regression coefficients of health outcomes on the city-wide averages, they are probably gross underestimates of the real uncertainty in our knowledge of the health effects of air pollution.

Conclusions

While we have discussed this problem using one particular air pollution study as an example, it should be obvious that it is a problem of much wider relevance in all areas of epidemiology, and, for that matter, whenever linear regression is used to provide clues to causal relationships.

We have calculated, using well-known relations among pairwise correlation coefficients, both a "corrected" regression coefficient and indications of its uncertainty, applicable to the situation where practical considerations dictate a choice of independent variable other than the one ideally demanded by the hypothesis under test and where the correlation coefficient between the two independent variables is known.

We have found that the range of uncertainty may greatly exceed the coefficient itself, even if the two independent variables are fairly well but not perfectly correlated. We have shown further that appreciable uncertainty exists even when only a small increment in the proportion of variance in the dependent variable is assumed to be associated with the ideal variable. In the example we have considered in this paper, a study of the effect of air pollution on health, the uncertainty in the "corrected" regression coefficient due to the imperfect correlation between the practical and ideal independent variables makes it quite unreliable as an estimator of health effects, in spite of the fact that it is statistically significant by the usual tests.

It is more commonly the case that the practical independent variable is recognized to be an uncertain measure of the ideal variable, but no quantitative information about their relationship is available. An awareness of how uncertain the observed regression coefficient can actually be under such conditions should lead to greater caution in interpretation of the results of a regression analysis.

This work was supported by the Study on Statistics and Environmental Factors in Health under the auspices of the SIAM Institute for Mathematics and Society and by Grant 00899 from the National Institute of Environmental Health Sciences.

REFERENCES

1. Glasser, M., and Greenburg, L. Air pollution, mortality and weather. Arch. Environ. Health 22: 334 (1971).
2. Schimmel, H., and Murawski, R. SO_2-harmful pollutant or air quality indicator. J. Air Pollut. Control Assoc. 25: No. 7, 739 (1975).
3. Schimmel, H., and Murawski, R. The relation of air pollution to mortality. J. Occup. Med. 19: 316 (1976).
4. Hodgson, R. A. The effect of air pollution on mortality in New York City, Report No. 1012, presented before the Statistics Section at the 96th Annual Meeting of the American Public Health Association, Detroit, Mich., Nov. 1968.
5. Buechley, R. W., Riggan, W. B., Hasselblad, V., and Vanbruggen, J. S. SO_2 levels and perturbations in mortality: a study in the New York - New Jersey metropolis. Arch. Environ. Health 27: 134 (1973).
6. Hearing before the New York State Department of Environmental Conservation and the New York City Environmental Protection Administration on the health effects of air pollution. 1976/1977.
7. Goldstein, I. F., Landovitz, L., and Block, G. Air pollution patterns in New York City. J. Air Pollut. Control Assoc. 24: No. 2, 148 (1974).
8. Goldstein, I. F., and Landovitz, L. Analysis of air pollution patterns in New York City: I. Can one station represent the large metropolitan area? Atmos. Environ. 11: 47 (1977).
9. Goldstein, I. F., and Landovitz, L. Analysis of air pollution patterns in New York City: II. Can one aerometric station represent the area surrounding it? Atmos. Environ. 11: 53 (1977).
10. Goldstein, I. F., and Landovitz, L. Sulfur dioxide: Harmful pollutant or air quality indicator? Air Pollut. Control Assoc. 24: No. 12, 1195 (1975).
11. Goldstein, I. F., Landovitz, L., and Fleiss, J. L. Analysis of air pollution patterns in New York City: III. Distributions of air pollutants over the city. Atmos. Environ., in press.