Methodological Problems Arising from
the Choice of an Independent
Variable in Linear Regression, with
Application to an Air Pollution
Epidemiological Study ¹

MASTER

Inge F. Goldstein,² Martin Goldstein,³
Joseph L. Fleiss,⁴ and Leon Landovitz⁵

NOTICE
This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

¹ 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.

² Division of Epidemiology, Columbia University School of Public Health,
600 West 168 Street, New York, N. Y. 10032

³ Department of Chemistry, Yeshiva University

⁴ Division of Biostatistics, Columbia University

⁵ Computer Center, Graduate Center of the City University of New York

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

Prepared For
THE U.S. ENERGY RESEARCH AND DEVELOPMENT ADMINISTRATION
UNDER CONTRACT NO. EY-76-S-02-2874

950 6450

702 6000
per C. Fox

950 6449

CB

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Abstract

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, especially if the regression of dependent on practical variable explains only a small part of the variance in the dependent variable. These points are illustrated with data and 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 ^{factors of} 15 to 150. 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.

linear regression, selection of independent variables, air pollution, acute health effects.

METHODOLOGICAL PROBLEMS ARISING FROM THE CHOICE
OF AN INDEPENDENT VARIABLE IN LINEAR REGRESSION,
WITH APPLICATION TO AN AIR POLLUTION EPIDEMIOLOGICAL STUDY
I.F. Goldstein, M. Goldstein, J.L. Fleiss, L. Landovitz

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 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 smoke-shade (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 (7,8,9), 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 smokes shade 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 $\alpha_{yx'}$?

Calculation of upper and lower bounds on the regression coefficient

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

$$\alpha_{yx} = \frac{\sigma_y}{\sigma_x} \rho_{yx} \quad (1)$$

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

The basis of our estimate of the relation between the observed (α_{yx}) and desired ($\alpha_{yx'}$) regression coefficients will be a formula

relating the correlation coefficients between three variables taken in pairs. Consider the partial correlation coefficient

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

Because a partial correlation coefficient must lie between the limits -1 and +1, the desired correlation coefficient $\rho_{yx'}$ must lie in the interval

$$\rho_{yx} \rho_{xx'} \pm (1 - \rho_{yx}^2)^{\frac{1}{2}} (1 - \rho_{xx'}^2)^{\frac{1}{2}}. \quad (2)$$

Substituting equation 1 into this result, we obtain an interval for all mathematically possible values of the desired regression coefficient $\alpha_{yx'}$:

$$\frac{\sigma_x \rho_{xx'}}{\sigma_{x'}} \alpha_{yx} \pm \frac{\sigma_y}{\sigma_{x'}} (1 - \rho_{yx}^2)^{\frac{1}{2}} (1 - \rho_{xx'}^2)^{\frac{1}{2}}. \quad (3)$$

The leading term of expression 3 has been proposed as a "corrected" regression coefficient, with the additional term following the symbol $\pm$ omitted (3). Such use is clearly inappropriate unless certain assumptions are made which require empirical test (see below). In general, it is necessary to bear in mind that the desired coefficient cannot be uniquely determined by the additional information about the relation between the practical and ideal variables, but can lie anywhere within a possibly wide interval about the "corrected" coefficient.

The width of the interval depends on the magnitudes of the correlation coefficients $\rho_{xx'}$ and ρ_{yx} . As $\rho_{xx'}$ approaches unity, the multiplicative factor $(1 - \rho_{xx'}^2)^{\frac{1}{2}}$ approaches zero, and the single "corrected" coefficient indeed becomes 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.

When ρ_{xx} is as large as 0.8 or 0.9, the interval of uncertainty would still be narrow if ρ_{yx} were close to unity. If, however, ρ_{yx} itself is small in magnitude, we can see that even an "excellent" correlation between practical and ideal independent variables can be consistent with considerable uncertainty in the regression coefficient α_{yx} , as estimated from the "corrected" value.

The quantity ρ_{yx}^2 represents the fraction of the variance in the dependent variable explained by the regression. We may sum up the above discussion by stating that the uncertainty in the "corrected" regression coefficient can be large either if the practical and ideal independent variables are imperfectly correlated, or if the regression of the dependent on the practical independent variable accounts for only a small part of the variance in the dependent variable, or, of course, both.

Application

As an example of the degree of uncertainty that can be introduced into an estimated regression coefficient, we give the results of applying the above formulae to the regression coefficients obtained in the air pollution study referred to earlier (3) (see Table 1). In this table we give the "corrected" regression coefficients for each of three mortality variates on the two pollution variates, sulfur dioxide and smoke-shade, and in addition, the upper bounds (+ sign from expression 3) and

lower bounds (- sign from expression 3) on the desired coefficient. As an estimator of the uncertainty that exists in the "corrected" coefficient we give the ratio of the range of uncertainty (upper bound minus lower bound) to the regression coefficient itself. It can be seen that the range of uncertainty of the regression coefficients greatly exceeds the regression coefficients themselves, by factors ranging from 15 to nearly 150. This should not be surprising in view of the above discussion. The correlation coefficients between practical and ideal variables are appreciable, but are far from perfect, and the regressions found in the study explain only a very small amount of the variance in daily mortality, although by the usual tests they are statistically significant.

The latter point is demonstrated in Table 2, where the six regression coefficients for the three mortality variates on each of the two pollution variates are tabulated; the squares of the six correlation coefficients are calculated from the regression coefficients and the standard deviations. These squared correlation coefficients, equal to the fractions of the variances explained by the regressions, range from .00013 to .0175, and are thus quite small in all cases.

The Appropriateness of the "Corrected" Coefficient

In an unpublished appendix to their air pollution - acute health effects study (3), available on request from them, its authors 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 expression 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. 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 geographical 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 it something other than merely an unbiased estimator of city-wide pollution. This is made additionally plausible by the facts that pollution at the central station tends to be greater than the city-wide average, and has changed over time in a different way from the average.

This a priori argument for the existence of a non-random 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 clear-cut 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 this assumption.

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 be necessarily perfectly correlated with the exposure to sulfur dioxide of the individual inhabitants of the city, some of whom spend most of their time indoors and others 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 smokes shade correlates with it. We must acknowledge therefore that as wide as the intervals of uncertainty are 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 a well-known relation among pairwise correlation coefficients, both a "corrected" regression coefficient and its range of uncertainty, applicable to the situation where practical considerations dictate a choice of independent variable other than the independent variable 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 the uncertainty is particularly enhanced when the fraction of the variance in the dependent variable explained by the regression is small. In the example we have considered in this paper, a study of the effect of air pollution on health, the uncertainty in the 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.

TABLE 1

Limits of Uncertainty About Six "Corrected" Regression Coefficients

POLLUTION VARIATE (x)	MORTALITY VARIATE (y) *	"CORRECTED" REGRESSION COEFFICIENT ($\alpha_{yx'}$)	UPPER BOUND ON $\alpha_{yx'}$	LOWER BOUND ON $\alpha_{yx'}$	RATIO OF INTERVAL LENGTH TO "CORRECTED" COEFFICIENT
<u>Sulfur Dioxide</u>					
(p _{xx'} =0.776)	M1	0.326	6.55	-5.89	38.2
	M2	0.0185	1.34	-1.30	143.
	M3	-0.158	3.76	-3.44	45.6
<u>Smokeshade</u>					
(p _{xx'} =0.652)	M1	0.427	4.21	-3.35	17.7
	M2	0.0142	0.823	-0.794	114.
	M3	0.251	2.44	-1.93	17.4

* Three separate city-wide daily mortality variables were examined (3):

M1: Total Mortality, M2: Respiratory Disease Mortality, and M3:
Heart Disease Mortality.

TABLE 2

Squared Regression and Correlation Coefficients of City-wide
Mortality Variates (y) on Centrally Measured Pollution Variates (x)

<u>POLLUTION VARIATE (x)</u>	<u>MORTALITY VARIATE (y)</u>	<u>σ_y^2</u>	<u>α_{yx}^2</u>	<u>ρ_{yx}^2</u>
<u>SO₂</u> ($\sigma_x^2=12.53$)	M1	329.42	.0476	.00181
	M2	14.82	.0001534	.00013
	M3	110.2	.01119	.00127
<u>Smokeshade</u> ($\sigma_x^2=56.13$)	M1	329.42	.09922	.0169
	M2	14.82	.000110	.00042
	M3	110.2	.03437	.0175

References

1. Glass, M., Greenburg, L.: Air pollution, mortality and weather. Arch. Environ. Health 22:334-343, 1971
2. Schimmel, H., Murawski, R.: SO₂-harmful pollutant or air quality indicator. J. Air Pollut. Cont. Assoc. 25: No.7, 739-744, 1975
3. Schimmel, H., Murawski, R.: The relation of air pollution to mortality. J. Occup. Med. 19:316-333, 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.
5. Buechley, R.W., Riggan, W.B., Hasselblad, V., Vanbruggen, J.S.: SO₂ levels and perturbations in mortality: a study in the New York - New Jersey metropolis. Arch. Environ. Health 27: 134-137, 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., Block, G.: Air pollution patterns in New York City. J. Air Pollut. Cont. Assoc. 24: No.2, 148-152, 1974
8. Goldstein, I.F., Landovitz, L.: Analysis of air pollution patterns in New York City: I. Can one station represent the large metropolitan area? Atmos. Environ. 11:47-52, 1977
9. Goldstein, I.F., Landovitz, L.: Analysis of air pollution patterns in New York City: II. Can one aerometric station represent the area surround-it? Atmos. Environ. 11:53-57, 1977

(continued)

References (continued)

10. Goldstein, I.F., Landovitz, L.: Sulfur dioxide: Harmful pollutant or air quality indicator? Air Pollut. Cont. Assoc. 24: No.12, 1195, 1975
11. Glasser M., Greenburg, L.: Air pollution mortality and weather. Arch. Environ. Health 22: 334-343, 1971
12. Goldstein, I.F., Landovitz, L., Fleiss, J.L.: Analysis of air pollution patterns in New York City: III. Distributions of air pollutants over the city. (Submitted to Atmos.Environ.)