Environmental Health Perspectives
Vol. 34, pp. 165-183, 1980

Air Pollution and Human Health: A Review and Reanalysis

by L. A. Thibodeau,* R. B. Reed,* Yvonne M. M. Bishop,* and L. A. Kammerman*

Since 1970, Lave and Seskin have published a series of articles dealing with the question, "Does air pollution shorten lives?" Their recent book reports revised and extended analyses of their previous studies emphasizing policy implications. We have undertaken a review of Lave and Seskin's book to evaluate the methodology used and hence gain some insight into the strength of the conclusions reached. This review concentrates on methodology and its application to establishing and quantifying the association between air quality and health. Beyond simply reviewing the analyses reported in Lave and Seskin's book, we have duplicated and expanded two of the reported analyses. Our detailed reanalysis is presented both to verify reported results, and to illustrate the difficulties encountered in such an analysis.

Our overall conclusion is that Lave and Seskin have done a thorough job of reporting and interpreting the various analyses that they performed. Lave and Seskin have made a pioneering effort in showing an association between mortality rates and air pollution. We do not disagree with the conclusion of the existence of an association but have some reservations about their methods of estimating its magnitude.

We were particularly concerned that Lave and Seskin did not fully investigate how well their models fit these data. Our reanalysis results in estimated effects which differ considerably from the values reported by Lave and Seskin. Thus, we conclude that the regression coefficients are quite unstable and so must be used with care. Assessing the relative costs and benefits of reducing air pollution without extensive sensitivity analysis could, therefore, be misleading.

Introduction

Since 1970, Lave and Seskin have published a series of articles dealing with the question, "Does air pollution shorten lives?" They explain that they started the book, *Air Pollution and Human Health* (*1*) with the intention of pulling together their previous studies which began in 1967, but that they have done more than that. Data sets used previously were revised and extended. All models have been reestimated by using a standard set of data, and new analyses are presented. They write in their preface, "We hope that these studies will prove valuable to those personnel in federal, state, and local environmental protection agencies who are responsible for decision making."

We have undertaken a review of Lave and Seskin's book to evaluate the methodology used and hence gain some insight into the strength of the conclusions reached. We present our review in three parts. The first part is a description of the contents of Lave and Seskin's book, together with comments on the methodology. In the second part, we describe the data set used by Lave and Seskin, report on our duplication of their 1960 total mortality analyses, and consider the robustness of their findings. Third, a summary of our major criticisms and comments appears after our reanalysis. We compared the data set used by Lave and Seskin with the original sources and report on this in Appendix I, where we describe each variable and list its abbreviated name. Appendix II contains some technical details of our robust analysis.

Review of Methodology

Data analysis is essentially open-ended, and thus, no matter where one chooses to stop, it is possible and probable that someone will argue that an additional step must obviously be taken. Despite

*Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115.

this, people continue to analyze data and to make and publish conclusions from the analyses. The essence of data analysis is that it is necessary to pursue a methodology which provides a complete analysis, that is, one which investigates reasonable alternatives to the hypothesized relationship. Lave and Seskin explain in their preface that this book is a compilation and reanalysis of earlier work. The result is that they have presented an often discontinuous report of applying selected disjoint techniques to the analysis of an important data set. We have the impression that Lave and Seskin considered and tried out every suggestion that had been given to them regarding alternative analyses.

Our criticism here is of methodology and of whether Lave and Seskin's conclusions were justified by the results presented. We found it hard to focus on conclusions as the authors' enthusiasm seems to change from chapter to chapter, although in their preface they write, "This book — rather than any of our previous results — reflects our current thinking as well as our most up-to-date estimates."

The authors divide the chapters of their book into five sections, and we follow this pattern in our discussion. The first section is concerned with background and theoretical framework and does not contain any analysis of data. The next section consists of Chapters 3 through 7 and is the largest section of the book. Four of the five chapters of Section II contain cross-sectional analyses of the relationship between mortality rates and air pollutants in the United States for the years 1960, 1961, and 1969. The remaining chapter, Chapter 6, reports analyses of venereal disease and crime rates.

Section III consists of Chapters 8 and 9, which deal with time series analyses of mortality rates and air pollutants. Benefits and costs of air pollution abatement are described in Chapter 10, and an overall summary and conclusions are given in Chapter 11. These last two chapters form Section IV on policy implications. The final Section V consists of seven appendices.

Most of the analyses reported, starting with Chapter 3, are of cross-sectional data. The very first analysis presented, Eq. 3.1-1, relates crude total mortality rates in 117 Standard Metropolitan Statistical Areas (SMSAs) or State Economic Areas (SEAs) in 1960 to 11 explanatory variables. These explanatory variables are three measures of ambient sulfate level, three measures of suspended particulate level, four socioeconomic variables, and total population. The three measures for each pollutant are the minimum, mean, and maximum of 26 biweekly readings, or where these were not available, quarterly readings.

Equation 3.1-1 is the focus of the authors' cost-benefit discussions in Chapter 10 (p. 225) and their conclusions regarding the effect on the mortality rate of reductions in particulates and sulfur oxides. The estimated coefficients specifying Eq. 3.1-1 were obtained by least-squares linear regression. From the summed elasticities of 5.04 for sulfates and 4.37 for particulates, they compute the effect of change on mortality; for example, a 50% reduction of air pollution corresponds to ½(5.04 + 4.37) and this gives us the 4.7% reduction in mortality shown in Table 10.1 (p. 218). Thus any error in the estimation of b values will cause a corresponding error in elasticities of the same relative magnitude, and it is this issue that is our primary concern.

Comments on Section I

Background and Theoretical Framework

The first section contains an introductory chapter and a chapter on theory and methods. The introduction highlights the motivation behind this study and some of the logic that necessarily underlies data analysis. In it, we learn that the primary goal of this book ". . . is to quantify the health benefits that would result from the abatement of air pollution" (p. 4). To achieve this goal requires the analysis of data, a first step of which is recognition of the "factors affecting the shape of the analysis." The factors considered are:

"Which pollutants are pernicious?" Lave and Seskin respond that data on sulfur oxides and suspended particulates are available, and moreover, that these pollutants have documented effects.

"How can health status be measured?" Lave and Seskin state that morbidity cannot be effectively measured and thus only mortality remains.

"What statistical methods are appropriate?" Lave and Seskin selected multivariable regression as an exploratory data technique as it is "robust" and "powerful."

"Translating effects into dollars." Lave and Seskin propose that, if a quantitative relationship between air pollution and mortality can be estimated, the cost of air pollution abatement and the benefit of reduced mortality must be translated into dollars as described in Chapter 10.

The next step towards data analysis is the formulation of hypotheses. Lave and Seskin decide on two competing hypotheses for the relationship between air pollution and mortality: H_1: "air pollution causes an increase in the mortality rate" (Both short and long term effects are included.) H_2: "there is a 'true'

factor, or set of factors, that causes both air pollution and increased mortality."

Lave and Seskin's goal was to distinguish between H_1 and H_2. Factors suggested as affecting mortality and thus to be used in their analysis are: physical, socio-economic, environmental, and personal.

The factors considered seem to be entirely appropriate, but the next step of determining how best to measure these factors creates some problems. Lave and Seskin give rationalizations for the use of available data which we find debatable. For example:

"Which pollutants are pernicious?" Lave and Seskin respond that data on sulfur oxides and suspended particulates are available. Note that this is not a direct response to the question, but rather a rationalization of the use of the 1960's data.

"How can health status be measured?" Lave and Seskin state that morbidity cannot be effectively measured and thus only mortality remains. Of course mortality is a measure of health status, but given the recognized measurement problems in adjusted mortality, it could be argued that, instead of an analysis of the 1960 mortality data, a carefully planned collection of data on morbidity would provide much clearer inferences.

"What statistical methods are appropriate?" Lave and Seskin selected multivariable regression as an exploratory data technique because it is "robust" and "powerful." The use of these terms to describe multivariable regression may be misleading, as multivariable regression is as "robust" and as "powerful" as the data analyst is adept.

Whether multivariable regression is appropriate depends on its use. Lave and Seskin seem to have two main purposes, exploring for relationships and prediction. Interrelationships between the explanatory variables do not invalidate building a regression model to estimate the association between air pollution and mortality, provided we condition on the values of these variables. However, changing one or more of the explanatory variables affects others, and thus standard prediction may be suspect. These concerns carry over to evaluation of estimated parameters and elasticities.

Theory and Method

In Chapter 2, Lave and Seskin compare laboratory and natural experiments in terms of informational quality and the goal of the analysis and decide in favor of natural experiments. Additionally, they review some previous studies and note that the statistical methods used in these studies were predominantly cross-tabulations and simple correlations. In reviewing these studies, they note several problems: (1) failure to control for important factors affecting health; (2) sample sizes too small to lend confidence to the results; (3) a reluctance to publish negative results, thus creating a potentially biased impression of the association between air pollution and health; (4) the controlled studies were not applicable to the population as a whole. That is, the inferences from such studies are either not applicable or are imprecise.

Complicating the data collection problems are estimation problems resulting from the quality of the data collected. Four more problems are discussed by Lave and Seskin: (4) missing variables; (5) badly measured variables; (6) current pollution levels being used as indicators for cumulative pollution over a lifetime; (7) extrapolation of single station measurements to a region.

In describing the complicated and serious data collection problems, Lave and Seskin state that, "These problems all tend to obscure the air pollution-mortality relationship and to bias the estimated relationship towards statistical nonsignificance" (p. 25). While they justify this statement with a few examples, they ignore the possibility of bias in the opposite direction. For instance, high mortality rates in large urban areas could be inflated if the census enumeration missed large numbers of persons — it is known that minorities and poor persons are consistently undercounted.

In general, Chapter 2 gives a good overview of the many problems that exist in trying to assess the health effects of air pollution and the authors' last statement in the chapter leads the reader to believe that they are well aware of the problems. They write, "We shall present the first of our estimates in Chapter 3. Needless to say, the estimated parameters are to be viewed with caution" (p. 25).

Comments on Section II

Cross-Sectional Analysis of U.S. SMSAs 1960, 1961, and 1969

Section II consists of five chapters, each analyzing the relationship between mortality and air quality in a different way. As Chapter 3 presents the analysis that they eventually use for cost-benefit analysis, we discuss it in some detail.

Total U.S. Mortality, 1960 and 1961.

Chapter 3 explores the relationship between the total mortality rates by use of the cross-sectional data on standard metropolitan statistical areas

(SMSA). Lave and Seskin estimate a linear relationship between the unadjusted total mortality rate and both air pollution and socio-economic variables for 1960. As a validity check, they fit the model to the 1961 data from the same SMSAs. Some different functional forms of the relationship, including blocking by census region, are fitted to detect departures from the linear model; no significant differences in the estimated pollution coefficients are obtained. A form of jackknifing coefficient estimates is used to show that the estimated pollution coefficients are not sensitive to outlying values. Collinearity between groups of the explanatory variables was assessed by a two-state fit — first fit all except air pollution variables, then fit air pollution variables to the residuals. No significant change in the estimated coefficients was observed.

Many methodological problems exist in this chapter. Some of these issues are addressed in later chapters but the reader is not given clear signposts at this point. Problems that appear in this chapter are collinearity, residual analysis, bias, and specification.

Collinearity. While Lave and Seskin recognize the collinearity between the air pollution measurements, they ignore the collinearity between the socioeconomic variables. Their own analyses point out this problem; for example, the small coefficients with changing sign for the variable labeled "percent poor" (p. 31, 35, 42). This may be due simply to sampling variability or it may indicate collinearity between the variables. It is not possible to distinguish between these two causes from the information presented.

Residual Analysis. No analysis of the residuals was reported in this chapter. This chapter is essentially an introduction to the more detailed analyses in later chapters. The analysis of gross aggregate data is thus a starting point. In such situations, a residual analysis is often used to direct the further analysis. Some key factors to consider are: influential points such as leverage points or outlying values, missing variables in the model, and other model problems. The approach used by the authors is different. The next chapter deals with different breakdowns of the mortality rate. We do not have any analysis of residuals until we reach Chapter 7.

Bias. The replication of the analysis with the 1961 data does not check on bias, as Lave and Seskin state (p. 41). If the 1960 data are biased by commission or omission, the 1961 data are almost surely biased in the same way. In fact, about half of the data on pollutants in the 1961 data set are actually 1960 values.

Specification. Despite the evidence that the linear specification consistently underestimated the effect of the mean particulate count at lower values [a conservative "pseudo" F-test showed significant differences in slope (pp. 47-48)]. Lave and Seskin chose to ignore the problem and stay with the single linear specification.

Some of these problems are addressed in Chapter 7. In the concluding sentences of this chapter (in marked contrast to the caution suggested in the preceding chapter) the authors conclude that their analyses "demonstrate a significant, robust association between both sulfates and suspended particulates, and the total mortality rate" (p. 52). The analyses of the remaining chapters of this section are all compared by the authors with the results of Chapter 3.

Other Mortality Rates, 1960 and 1961

Chapter 4 examines variations on the mortality rate: infant rates, age-sex-race adjusted rates, sex-race-specific rates, and disease-specific mortality. These rates are regressed on the same variables used in Chapter 3.

We found the analysis of infant mortality rates to be both of interest and also demonstrative of the problem of instability of estimates. If air pollution affects health throughout a person's lifetime, then analyses based on adult death rates in 1960 and pollution levels in the same year must be based on the assumptions that (a) the populations are nonmigratory and (b) the pollution levels have remained at the same relative levels over time. These assumptions are certainly controversial, and opposing theories exist.

One of the opposing theories is that persons in poor health move away from highly polluted areas and thus only those not susceptible to the pollution remain. Another hypothesis is that persons of low socioeconomic status who are in poor health drift to the urban areas in their search for employment and medical care. The first assumption would make differences in mortality between areas associated with pollution more difficult to detect, the second assumption would bring about differences not attributable to pollution. The problem of migration is not as great when infant deaths are considered, although not entirely absent because of the relationship between infant mortality and mother's health.

The authors do not view the analysis of infant deaths as a separate problem to be tackled with its own rationale and hypothesis. Instead they have included the same set of characteristics of the SMSAs as are included in the analyses for adult deaths. Had the authors selected a different set of variables for this analysis, and found a strong association with air pollution, this would have been more supportive of

their general hypothesis of association between mortality and air pollution than is the present analysis.

When the authors turn to the age-specific rates for newborns through adulthood, they believe that "air pollution has less effect on the young than on the old" (p. 63). The values of the coefficients for whites (the larger population) are as shown in Table 1 with *t* values in parentheses.

Table 1. Coefficients for whites.

	Male		Female	
	Min *S*	Mean *P*	Min *S*	Mean *P*
All ages (age-adjusted)	0.481 (1.78)	0.300 (1.52)	0.935 (3.81)	0.274 (1.53)
Newborn to 14	−0.043 (−0.36)	0.028 (0.31)	−0.032 (−0.42)	0.080 (1.43)
15-44	0.004 (0.04)	0.072 (1.06)	−0.025 (−0.46)	0.102 (2.57)
45-64	0.441 (0.85)	0.546 (1.43)	0.810 (2.70)	0.599 (2.74)
65+	4.530 (2.22)	1.687 (1.13)	6.641 (3.70)	0.706 (0.54)

If we think of possible cause-and-effect mechanisms, we could postulate that life-long exposure may be an important factor in determining health effects; thus in the older persons we see cumulative effects; alternatively older persons may be more susceptible to air pollution; or all persons may be equally susceptible but older persons are less able to move away before dying from the effects. Whatever the cause, these findings seem to indicate that the effect may not be linear with age.

Chapter 4 also contains an analysis of disease-specific mortality rates. The authors give a good discussion of why, under a Poisson model, the variance between SMSAs with low death rates would be such that a regression could only hope to account for a relatively small proportion of the total variance. They conclude that the low incidence and the problem of unmeasured factors such as smoking impede their ability to explain the disease-specific mortality rates.

For each analysis reported in Chapter 4, the authors have an explanation for the lack of association between mortality and air pollution variables. They are encouraged by the finding that the 1961 results are similar to those of 1960 and they reach the conclusion that these results support those of Chapter 3. The nature of the support is however quite limited. The fact that none of the variables considered change a great deal from year to year means that the similarity of the two analyses cannot be interpreted as an indication that the formulation of the underlying model is correct.

Other Explanatory Variables

The process of looking at different sorts of mortality rates does not in general seem to be very supportive of the findings of Chapter 3. In Chapter 5, the authors look instead at variations in the other side of the equation by introducing additional variables. They state, "we investigated a number of factors that had been suggested as the 'true' causes of these relationships" (p. 77). The relationships cited are the associations between 1960 mortality rates and sulfates and suspended particulates. The new variables are considered in two batches: "Occupation Mix" and "Climate and Home-Heating."

From the occupational breakdowns, the results that the authors remark on as being of interest generally seem to indicate a positive association between mortality and unemployment, and between mortality and use of public transportation, both of which might be regarded as socio-economic variables. There is also a tendency for agriculture, construction, and white collar work to be negatively associated with mortality.

In addition to these selected new variables showing relatively high associations, there are large fluctuations in the previous socioeconomic coefficients, many of which switch from positive to negative for some of the 18 pairs of regressions presented. In all 18 regressions, the coefficient for sulfates is markedly reduced when new variables are introduced, in most instances to less than half its previous value. The coefficients for particulates are reduced for all but three of the 18 comparisons, but the reductions are not generally as marked as for sulfates.

The results cited by the authors are similar for the other variables considered in this chapter. The heating and climate variables reduced the air pollution coefficients, and a further reduction to nonsignificant levels occurred when home-heating fuels were added. They believe these analyses do not mean "the previous association between air pollution and mortality is disproved, but rather that it is made more specific by directing the association to home-heating fuels, rather than to all sources of air pollution" (p. 89).

In Chapter 5, the authors are trying to determine causality. If both portions of this chapter are considered together, we find that introducing occupation mix removes much of the apparent effects of sulfates on total mortality and the authors believe that the inclusion of occupation casts doubt on their

analysis of cancer mortality rates in Chapter 4. The climate and home-heating variables cast doubt on the association of air pollution variables with disease-specific rates. We are left with only the infant mortality rates showing relatively stable association with air pollution, although the association is not as strong as for some of the adult regressions. These results demonstrate that unravelling cause and effect in the presence of many interrelated variables is very difficult. Yet it is indeed a necessary precursor to effective policy formulation. The authors seem to be rather too optimistic when they conclude, "We view the results of this chapter as giving a qualified endorsement to the hypothesis of causality" (p. 107).

Is Air Pollution a Surrogate?

The rationale behind Chapter 6 is quite different from that of preceding chapters. The authors chose suicide, venereal disease, and crime rates as social ills associated with urbanization but not caused by air pollution. They then relate these to the same air pollution and demographic variables as they used in their mortality analysis. The rationale behind this analysis is that if air pollution is found to be related to these variables, then it is acting as a surrogate for omitted factors. Such a relationship would cast doubt on the previous analyses where mortality is related to air pollution because air pollution might also be a surrogate for other factors in this analysis. The authors suggest that "personal habits" might be such a factor.

Of the three types of rates investigated in order to rule out the possibility that air pollution is acting as a surrogate, only crime rates are at all convincing. The venereal disease rates as analyzed show incidence rates by counties to be lower than by cities, in contradiction to the data; thus the validity of these analyses is in doubt. Some of the suicide rate analyses do indeed show a positive correlation with air quality. The authors repeated the analyses for 1961 data for suicide and venereal disease but make no mention of doing this for crime rates and do not explain why they did not carry out a parallel study. Our impression is that this chapter does not contribute a great deal to our understanding of the problem of the association between health and air pollution. Nevertheless, it is an example of the efforts that Lave and Seskin have taken to try all possible approaches and to counter all objections that might be raised.

Do the Results Hold up in Later Years?

The last chapter of this section, Chapter 7, is called "1969 Replication, further Verification and Summary." Lave and Seskin summarize as follows (p. 158): "The jackknife technique and the deletion of outliers was applied to the 1969 data, and it was found that the estimated air pollution effects were not very sensitive to particular data points or extreme observations, while the estimated socioeconomic effects exhibited less stability." They go on to explain that alternative specifications to the simple linear form were also investigated, and that the residuals were examined. They include, "There was some evidence that an important variable (or variables) had been omitted, since SMSAs with unexpectedly large (or small) mortality rates in 1960 had large (or small) mortality rates in 1969. Thus, some other factors in addition to those variables used in the analysis were significant in raising or lowering the mortality rates."

One of the most interesting findings of Chapter 7 is captured in Table 7.7, showing age-sex-race-specific mortality rates for 1969, which documents the startling difference in observed 1969 air pollution effects over age, sex, and race. The effects generally increase with age. However, so does the mortality rate. It is not clear whether we are observing an increased effect of pollutants with age, a cumulative effect, or a statistical artifact. This matter is troubling, because a difference in susceptibility of different age groups would have substantial impact on the use of these results for cost-benefit analysis.

In general this chapter, as well as some other chapters of this section, seems to be responding to methodology and not to the potential analytic problems. Rather than constructing a (null) hypothetical relationship and showing the data inconsistent with other alternatives, Lave and Seskin select tests, apply them and show that the implied alternatives are at least no better than the base model. While these criticisms do not invalidate their results, they do cast doubts on the strength of their conclusions. A systematic model-building approach would be more satisfying, however, we appreciate that this book is a report of the evolution of the authors' ideas over many years.

Comments on Section III

Annual and Daily Time-Series Analyses

In Chapters 8 and 9, the authors report analyses of air pollution and mortality data over time in various cities. The authors argue that cross-sectional comparisons of different areas are designed primarily to detect long-term effects of air pollution, while short-term daily analysis is most useful in identifying short-term effects. Analyzing annual data represents a mix of these two strategies, although the annual data cannot be sensitive to very short-term effects.

The results reported in Chapter 8 are based on annual data from 1960 to 1969 in 26 SMSAs. Chapter 9 reports results based on series of daily data of varying lengths in five cities.

Yearly Time Series, 1960-1969

The bulk of Chapter 8 is made up of regression analyses similar to those presented in earlier parts of the book with the addition of three different methods for incorporating the time variable into the analysis. In a preliminary analysis, the authors investigated regression equations for age-sex-race adjusted total mortality rates with dummy variables for SMSAs and years. They found that the SMSA dummy variables alone explained "more than 78.7 percent" of the variance of the rates, and that the area and time dummy variables explained "almost 91.1 percent" (p. 166). Results for unadjusted total mortality and race-adjusted infant mortality rates were somewhat higher. Under these circumstances, the socioeconomic and air pollution variables had little more to contribute. "Indeed, when the SMSA and time dummy variables were included, the estimated coefficients of the air pollution variables were statistically nonsignificant in explaining the three mortality rates (the coefficients were generally negative as well). As a consequence of these preliminary findings, we chose to exclude the SMSA dummy variables" (p. 166).

Daily Time Series

In Chapter 9, daily series were studied separately in five cities. Air pollution effects were found in the Chicago data but generally not in data from Denver, Philadelphia, Saint Louis, and Washington, D.C. The authors feel that the relatively smaller size of the last four cities and consequent greater random variation in daily mortality may be the reason for this result.

The finding that the SMSA dummy variables explained a high proportion of the variance in death rates implies that in the annual series analysis, the bulk of the findings arise from the differences between areas rather than from year-to-year variation. To the extent that this is so, the authors' observation that the cross-sectional time series analysis results are in general agreement with the previous cross-sectional analyses is more a matter of repetition than corroboration.

The examination of the daily series incorporates a set of climate variables because of their relation to air pollution and mortality, but the authors do not attempt to examine interaction effects of air pollution and climate. These regressions also included dummy variables for day of week to control for weekly cycles in activities and consequent effects on mortality. Socioeconomic variables were not included because the analyses were carried out within city and these variables were assumed to be relatively constant over time. Variability within a city was not considered.

The authors were unable in the daily time-series analysis to come to a clear-cut conclusion as to whether these short-term effects "hastened" death or had a more additive effect on overall mortality, but they assert that these time-series analyses support their earlier findings from cross-sectional analysis of an association between air pollution and mortality.

Comments on Section IV

Policy Implications

Section IV consists of two chapters, 10 and 11. In Chapter 10, Lave and Seskin present a framework for benefit-cost analysis of air pollution abatement. They give a lucid and comprehensive discourse on the application of benefit-cost analysis to air pollution control. Each point is well made, clearly discussed, and positioned in the analysis. An important point of the analysis is that, "there is an inherent bias in benefit-cost analysis toward underestimating the benefits and overestimating the costs" (p. 212). Thus, the EPA's estimated cost of stationary-source abatement for 1979 probably overestimates the actual cost of abatement.

Lave and Seskin also present a discussion of benefits with special emphasis on the difficulties associated with assessing benefits in three categories: (1) real property, including cleaning, maintenance, life of materials, and general value; (2) plants and animals; (3) human health. However, the authors justify abatement programs solely on the benefits to human health.

From their analysis they conclude, that a "conservative estimate of the effect of a 58 percent reduction in particulates and an 88 percent reduction in sulfur oxides (reductions corresponding to proposed control levels) would lead to a 7.0 percent decrease in total mortality . . . a national annual benefit of $16.6 billion in 1979 (1973 dollars). We have confidence that substantial abatement of air pollution from the major stationary-source categories of . . . is warranted" (p. 244).

Additionally, Lave and Seskin compare the National Academy of Sciences, 1974 report on controlling mobile-source emissions (on presently mandated standards) to their results relative to mobile-source emission on human health (Chapters 7 and 9). They conclude that, "the anticipated costs clearly exceed the expected benefits, . . . This conclusion

holds even for areas such as Los Angeles" (p. 244).

Thus, Lave and Seskin support substantial abatement of stationary-source air pollution and do not support current mandated controls on mobile-source emissions. Specifically, their analysis supports the current national ambient air standard for particulates, and suggests a more stringent standard for sulfur dioxide.

Two points are bothersome in this chapter; the lack of sensitivity analysis and the extrapolation of benefits (mortality rates). Generally, decision analyses, like that presented by Lave and Seskin, depend on rather "soft" data and debatable assumptions. Sensitivity analysis is simply a method of examining the effects that changes in the data or assumptions have on the results. In this case, Lave and Seskin could have varied the estimated effects by a standard error either way and listed the results.

The second point refers to the inherent problems with extrapolation. Clearly, extrapolation from any analysis is risky. In this case, it is suspect. Our review of Section II points out evidence of potential nonlinearity. Nonlinearity would imply that the effect of a change depends on the initial value; certainly a reasonable assumption when considering dose-response relationships.

In Perspective

Chapter 11 is presented as a summary of preceding chapters together with conclusions. The authors start by listing nine criteria formulated by Hill (*2*). These criteria have been "used by epidemiologists to judge whether causality was a reasonable interpretation of their findings" (p. 235). Lave and Seskin then summarize chapter by chapter the analyses carried out in the earlier part of the book, and follow this with a page and a half summary of the policy implications. The last part of the chapter deals with future research needs and improving the measurement of benefits and cost.

This chapter does more than summarize the analyses that Lave and Seskin have done, it puts them in perspective. Had they chosen to present this as the first chapter, instead of the last, it would have alerted the reader to the idea that this book is not viewed by its authors as a definitive study of a very complex problem. Rather, it is a pioneering step, that points up the need for further studies.

In their summary of the statistical analyses, the authors explain that alternative functional forms were tried but "While some had greater explanatory power than the linear model, we decided to continue relying on the linear form because of its simplicity and ease of interpretation" (p. 239). This explanation is understandable if their objective is to demonstrate the existence of an association between mortality and air pollution that needs further investigation. But if the main objective is predicting cost and benefits as accurately as possible, simplicity and ease of interpretation may not be good reasons for selecting a particular model. They elaborate on this further in Appendix B, where they discuss the present standards. They state, " . . . our results do not lend support to the threshold concept. When we examined alternative specifications, we found that generally a simple linear specification 'fit' the data as well as other functional forms" (p. 316). This remark is in contradiction to the argument given for a linear model (p. 239), that it was chosen because of its simplicity and ease of interpretation.

The authors' discussion of future research needs stresses the importance of obtaining better air quality data; and of studies of morbidity as well as mortality. When they discuss improving the measurement of benefits and cost, they write, "Two crucial areas in need of closer examination are the impacts of air pollution on esthetics and its related impact on the quality of life generally" (p. 248). The usual progression of epidemiologic studies is from observation of an association gleaned from existing data, to prospective studies where data are collected for the purpose of answering a specific question. Lave and Seskin have taken the first step and demonstrated the need for further steps to be taken.

Comments on Appendices

Rather more than a fifth of Lave and Seskin's book is devoted to appendices. Appendix A is a review of the literature relating air pollution to health. Appendix B presents the implications of the authors' findings on the national ambient air quality standards. Appendix C is a listing of the SMSAs used in each of the analyses, and Appendix D gives the means and standard deviations of every variable used in every analysis together with the sources of the data. Appendices E and F describe respectively the methodology for direct adjustment of rates and the adjustments used to calculate costs for the cost-benefit analysis. The last Appendix, G, is a cross-reference listing of the literature, grouped according to the pollutants studied.

The sheer volume of documentation reflects the care that Lave and Seskin have taken to explain very precisely exactly how they carried out their analyses. Appendices A and G are likely to be useful reference sources. Appendices C and D were useful to us when we carried out our robust analysis and will no doubt be useful to other researchers who want to investigate a different sort of analysis.

Reanalysis of the Lave and Seskin 1960 Total Mortality Data

Our goals in analyzing the 1960 total mortality data were limited to a duplication of Lave and Seskin's primary analysis of the 1960 total mortality data and a preliminary look at the sensitivity of their results. In this section, we duplicate Lave and Seskin's least-squares regression of 1960 total mortality on data reported to be Lave and Seskin's data set. While checking these data against the sources cited by Lave and Seskin we found discrepancies which we corrected; we present least-squares regression analysis of the corrected data. Next we identify outlying data points, SMSAs, by two methods and examine their influence on the reported results. The first method identifies SMSAs which have widely different explanatory variable vectors. The second detects large residuals from a reanalysis of the data.

Duplication and Correction

The first step in our duplication of Lave and Seskin's 1960 total mortality regression analysis was to establish that the data we had reconstructed were essentially the same data used by the authors (see Appendix I for details). Means and standard deviations were calculated for the variables used in the least-squares regression analyses of 1960 total mortality. These statistics agreed (± 2%) with statistics presented by Lave and Seskin (p. 321). Thus, we assume that our reconstructed data are essentially the same data analyzed by Lave and Seskin.

Using reconstructed data, we duplicated two total mortality regression analyses reported by Lave and Seskin (Table 3.1, p. 31). The difference in estimated regression coefficients between our analysis and that of Lave and Seskin were within the bounds of machine error.

Before proceeding with a more extensive reanalysis, we cross-checked the reported data with the sources cited by Lave and Seskin. A few discrepancies were noted and corrected (see Appendix I for details). Using the corrected data, we again duplicated the same two total mortality regression analyses. The results are reported in Table 2 under the columns labeled "117 SMSAs." Comparing the regressions on the corrected data to those reported by Lave and Seskin reveals no significant change in the explanatory power ($R^2 \approx 0.83$). However, examining regressions with 11 explanatory variables reveals substantial variability in the estimated effects of the pollution variables: min $S = -7\%$, mean $S = -69\%$, max $S = +154\%$; min $P = +51\%$, mean $P = +13\%$; max $P = -67\%$. The differences between estimated effects for the regression with seven explanatory variables are less pronounced: min $S = -10\%$; mean $P = +7\%$.

We used the corrected data for the remainder of our reanalysis.

Sensitivity

Outlying SMSAs. A principal concern of this reanalysis was the possibility that some unusual SMSAs unduly affected the results. Our technique of detecting such SMSAs was to treat the explanatory variables defining an SMSA as a single multivariate observation. After logarithmic transformation of the explanatory variables to achieve near distributional symmetry, we used a robust estimate of covariance to scale the distance between each SMSA and the median explanatory vector. SMSAs that were widely separated from the others were excluded from the data used in the reanalysis of the total mortality data (for details see Appendix II).

Six SMSAs were identified as having explanatory variable vectors which were widely separated from the median vector characterizing the remaining SMSAs. These outlying SMSAs are: Charleston, W.V.; Fresno, Calif.; Jersey City, N.J.; Las Vegas, Nev.; Macon, Ga.; Phoenix, Ariz. Table 3 provides some insight into the variables which contributed most to each of the six SMSAs selected as outlying. For example, Jersey City, N.J., is uncharacteristic in that its population per square mile (P/M^2) is 1357.2, which is very different from the median for the 117 SMSAs of 42.5. Also note the sulfate (S) levels for Jersey City, N.J.

Repeating two of Lave and Seskin's analyses of the 1960 total mortality data, without the six outlying SMSAs, resulted in the coefficients listed in Table 2 under the column headed, "111 SMSAs." Again, the most pronounced differences are noted between the regressions with 11 explanatory variables. Comparing the reanalyses with 117 and 111 SMSAs, we note that while the explained variability is quite stable ($R^2 \approx 0.83$) the estimated air pollution effects vary considerably; min $S = -42\%$, mean $S = +39\%$, max $S = +124\%$; min $P = +4\%$, mean $P = +18\%$, max $P = -83\%$. Again, variations of the estimated coefficients between the seven explanatory variable regressions is less pronounced; min $S = -8\%$, mean $P = +10\%$.

These observations are important not simply because the estimated coefficients vary, we expect them to vary, but because of the potential effect on the subsequent analysis. That is, Lave and Seskin use the estimated pollution effects from the 11 variable regressions to decide on which pollution vari-

ables to use in the further analyses. They state, "Since our interest was centered on the air pollution variables, we initially retained only those whose coefficients were positive and exceeded their standard errors, with the further constraint that at least one sulfate measure and one particulate measure were retained" (p. 32). The reanalysis of the 111 most characteristic SMSAs indicates that the Max S would be the selected sulfate variable instead of Min S using these criteria. Clearly such a choice is significant if it is not critically examined and if it carries over to other analyses. In this case, both criticisms apply. Lave and Seskin use these criteria throughout the subsequently reported analyses; the result is that min *S* and mean *P* are the pollution variable selected for the seven variable regressions. Variables

Table 2. 1960 Coefficients from regressions on total mortality rate: Lave and Seskin regression analyses for 117 SMSAs; and the reanalysis for 117 SMSAs, 111 SMSAs, and 108 SMSAs.[a]

	11 Explanatory Variables				7 Explanatory Variables			
	Lave and Seskin[b]	Reanalysis			Lave and Seskin[b]	Reanalysis		
	117 SMSAs	117 SMSAs	111 SMSAs	108 SMSAs	117 SMSAs	117 SMSAs	111 SMSAs	108 SMSAs
R^2	0.831	0.831	0.833	0.873	0.828	0.828	0.826	0.864
Constant	343.381	362.731	373.987	272.526	301.205	324.885	329.260	214.935
Air pollution variables ($\mu g/m^3$)								
Min *S* (10)	0.473	0.440	0.254	0.082	0.631	0.565	0.521	0.404
	(0.283)	(0.285)	(0.304)	(0.245)	(0.233)	(0.224)	(0.234)	(0.189)
Mean *S* (× 10)	0.173	0.054	0.075	0.261				
	(0.326)	(0.327)	(0.337)	(0.272)				
Max *S* (× 10)	0.028	0.071	0.159	0.080				
	(0.112)	(0.114)	(0.122)	(0.098)				
Min *P*	0.199	0.301	0.312	0.568				
	(0.622)	(0.611)	(0.623)	(0.501)				
Mean *P*	0.303	0.343	0.406	0.012	0.452	0.482	0.531	0.405
	(0.427)	(0.420)	(0.471)	(0.379)	(0.169)	(0.170)	(0.192)	(0.159)
Max *P*	−0.018	−0.030	−0.055	0.013				
	(0.095)	(0.099)	(0.134)	(0.109)				
Socioeconomic variables								
P/M^2 (× 0.1)	0.083	0.094	0.143	0.091	0.089	0.098	0.204	0.150
	(0.054)	(0.053)	(0.152)	(0.121)	(0.052)	(0.052)	(0.146)	(0.118)
≥65 (× 10)	6.880	6.916	6.820	7.364	7.028	7.058	7.064	7.675
	(0.414)	(0.410)	(0.422)	(0.382)	(0.389)	(0.384)	(0.396)	(0.355)
NW (× 10)	0.396	0.409	0.352	0.451	0.422	0.430	0.387	0.501
	(0.104)	(0.104)	(0.108)	(0.087)	(0.098)	(0.097)	(0.103)	(0.085)
Poor (× 10)	0.038	0.028	0.140	0.092	−0.002	−0.013	0.067	0.018
	(0.146)	(0.148)	(0.154)	(0.128)	(0.100)	(0.141)	(0.148)	(0.126)
Log Pop (× 100)	−0.276	−0.323	−0.380	−0.244	−0.212	−0.260	−0.303	−0.161
	(0.200)	(0.194)	(0.228)	(0.185)	(0.189)	(0.185)	(0.224)	(0.183)

[a]*Standard errors are in parentheses below the corresponding regression coefficients.*
[b]Source: Lave and Seskin, p. 31.

Table 3. Variables describing the six outlying SMSAs.

SMSA	Values of the raw data											Total morality rate
	Min *S*	Mean *S*	Max *S*	Min *P*	Mean *P*	Max *P*	P/M^2	≥65	NW	Poor	Log Pop	
Charleston, W.V.	15	283	940	55	225	958	27.9	70	58	214	5.403	780
Fresno, Calif.	18	34	198	45	119	304	6.1	81	75	215	5.563	850
Jersey City, N.J.	155	229	340	63	147	253	1357.	103	69	127	5.786	1200
Las Vegas, Nev.	31	79	201	50	145	389	1.6	45	95	113	5.104	730
Macon, Ga.	18	27	128	22	122	754	28.6	62	310	263	5.256	870
Phoenix, Ariz.	15	86	266	98	247	573	7.2	72	55	191	5.822	760
Overall statistics												
3rd Quartile	60	134	287	56	145	327	70.2	97	176	230	5.842	1000
Median	42	81	210	43	113	241	42.3	84	94	153	5.536	900
1st Quartile	27	62	147	30	83	174	24.3	70	45	128	5.338	800

selected may also have significance in policy decisions; e.g., do we lower min S ignoring max S, or do we strive to lower max S?

Residual Analysis. The next step in the reanalysis was to examine the residuals from the regressions. Figure 1*a* shows a plot of the expected normal order statistics versus residuals from the regressions with 11 variables on the corrected data (117 SMSAs). Figure 1*b* shows the same plot for the regressions with seven variables. Clearly, three residuals are widely separated from the others in each plot. In both cases the outlying residuals were for the SMSAs, Scranton, Pa., Tampa, Fla., and Wilkes-Barre, Pa. Surprisingly, these residuals were more than three standard errors away from zero in our analysis and so we assume in Lave and Seskin's analysis, and yet Lave and Seskin do not mention them. For Scranton and Wilkes-Barre, Pa., the predicted mortality rates were too low, while for Tampa, Fla., the predicted mortality rates were too high. The reasons for these large deviations are not entirely clear. Two potential explanations are, nonnormal errors, or model misspecification. Note, these SMSAs have outlying residuals, (more than three standard errors) in the 111 SMSA regressions also.

To gain insight into the effects of these outlying rates, we omitted these three SMSAs along with the previous six outlying SMSAs, and recalculated the regressions. The resulting coefficients are reported in Table 2 in the columns labeled "108 SMSAs." The ability of this regression to explain the observed variability is enhanced because we removed the three most deviant points ($R^2 \approx 0.87$), but the signifi-

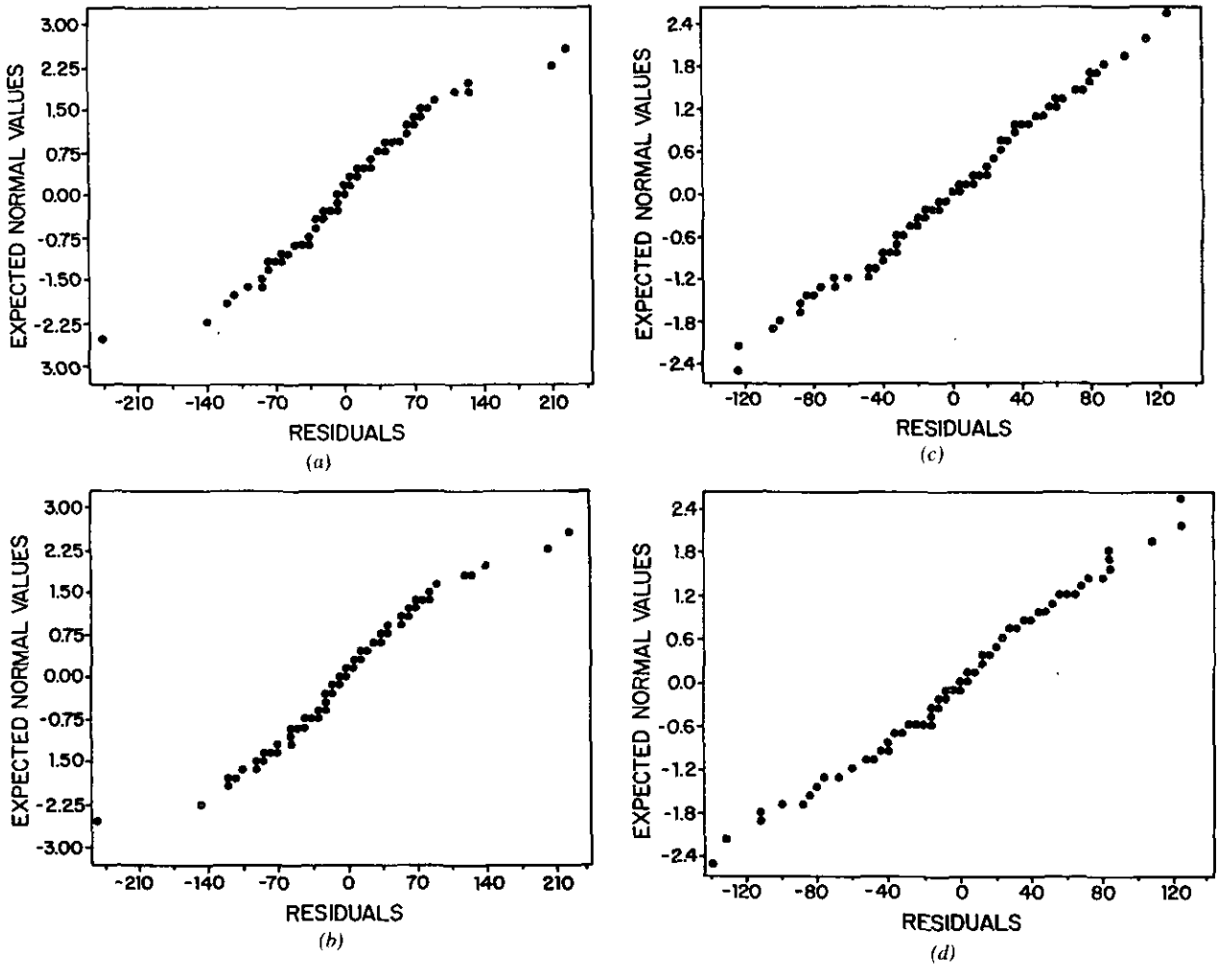

FIGURE 1. Normal probability plots of residuals (observed minus predicted mortality) from the least-squares regression of 11 or 7 variables on total mortality: (*a*) 11 variables, 117 SMSAs; (*b*) 7 variables, 117 SMSAs; (*c*) 11 variables, 108 SMSAs; (*d*) 7 variables, 108 SMSAs.

Table 4. Mortality rates (per 100,000), predicted and observed, and standardized residuals for the nine omitted outlying SMSAs.

SMSA	Observed	11 Variables			7 Variables		
		Predicted[a]	Standard error[b]	Residual[c]	Predicted[a]	Standard error[b]	Residual[c]
Charleston, W.V.	780	901.2	81.37	−1.489	799.5	55.93	−0.348
Fresno, Calif.	850	844.6	54.32	0.099	844.9	54.03	0.094
Jersey City, N.J.	1200	1196.7	162.31	0.020	1274.9	157.49	−0.476
Las Vegas, Nev.	730	607.2	55.49	2.213	599.3	56.29	2.322
Macon, Ga.	870	801.0	72.54	0.951	827.2	54.63	0.782
Phoenix, Ariz.	760	814.8	57.97	−0.945	812.0	56.42	−0.922
Scranton, Pa.	1400	1174.7	57.68	3.906	1187.8	56.37	3.764
Tampa, Fla.	1260	1552.5	63.78	−4.586	1566.8	62.57	−4.903
Wilkes-Barre, Pa.	1280	1051.8	54.96	4.152	1063.1	55.55	3.905

[a]Predicted from the regression without these SMSAs.
[b]Standard error of residual.
[c]Standardized by its standard error.

cant change is in the estimated coefficients. The variation between coefficients from the 117 SMSA reanalysis and the 108 SMSA reanalysis is, while not statistically significant, marked. Comparing the regressions with the 11 explanatory variables, the changes are: min $S = -81\%$, mean $S = +383\%$, max $S = +13\%$; min $P = 89\%$, mean $P = -96\%$, max $P = +143\%$. Note the max P coefficient is now positive. On comparing the seven variable regressions, the pollution coefficient changes are less marked. Min $S = -28\%$ and mean $P = -16\%$. Residual plots for these regressions are displayed in Figures 1*c* and 1*d*; no additional deviant points are evident.

Finally, using the models estimated from the 108 SMSAs, we computed predicted values and standard errors of prediction for each of the nine omitted SMSAs. These results are listed in Table 4. The differences are substantial and bothersome, especially with regard to their consequences in the estimated effects of pollution policies.

Conclusions from Reanalysis

We are confident that the reconstructed data are essentially the data used by Lave and Seskin. Finding such obviously deviant points is disconcerting. So we can only conclude that Lave and Seskin did not believe that these SMSAs were sufficiently deviant to comment on or to consider in detail.

We conclude from the differences found when deviant SMSAs are excluded that the regression coefficients are quite unstable and so must be used with care. Use of the coefficients for cost-benefit computations without extensive sensitivity analysis could, therefore, be misleading.

Computing

The following statistical computing packages were used to do the calculations necessary for this reanalysis: BMDP-77 (*4*), SAS, SNAP, and SPSS (*4-7*).

Overall Conclusions

The authors present us in Chapter 3 with the analysis which they finally use in their cost-benefit discussion. Thus the strength of their conclusions depends on three aspects of this analysis: was it based on a good choice of variables? was the model selected a good choice? was the methodology used for estimation of parameters adequate?

The authors do considerable investigation into the use of different sets of variables. At the end of each subsequent chapter, they give conclusions in terms of whether the new analysis supports the findings presented in Chapter 3. Their enthusiasm differs somewhat between chapters, but in each case they find that they are still convinced of the association between mortality and air pollutants. This conclusion is based on the use of existing data and Lave and Seskin are careful to point out many of the difficulties with the available data. We do not disagree with the conclusion of the existence of an association but have some reservations about their methods of estimating its magnitude.

We were particularly concerned that Lave and Seskin did not investigate as fully as we would wish how well their models fit their data. To determine whether their estimates might be severely distorted by outlying values, we undertook a robust analysis and re-estimated the regression given in Eq. 3.1-1,

after some data correction and removal of outliers. The estimates that we obtained for the air pollution coefficients are listed in Table 5.

Table 5. Air pollution coefficients

	Sulfates		Particulates	
	Lave & Seskin	Reanalysis	Lave & Seskin	Reanalysis
Minimum	0.47	0.08	0.20	0.57
Mean	0.17	0.26	0.30	0.01
Maximum	0.03	0.08	−0.02	0.01

They differ considerably from the values given in Eq. 3.1-1 by Lave and Seskin and used in their cost-benefit analysis.

Our other area of concern is that the linear model may not be the best-fitting model. Even if it is a good description of current interrelationships, it may not be suitable for predicting what would occur if the value of the air pollution variables were changed. There are several reasons for doubting the stability of the linear model. One reason is that a dose-response relationship that can be regarded as linear over a particular range of values may not be linear outside that range. Another reason is that no attention has been paid to the competing risks that cause mortality — the authors assume that all the other socioeconomic factors would remain constant while the air quality changed.

Our final conclusion is that Lave and Seskin have made a pioneering effort in showing an association between mortality rates and air pollution. The next steps are to assume a cause-and-effect relationship and to assess the relative costs and benefits of reducing air pollution. These steps cannot, in our opinion, be undertaken with any degree of confidence given the quality and nature of the available data. This conclusion seems to be very close to the authors' own views as expressed in their last chapter, Chapter 11. We believe readers would be well advised to read Chapter 11 before other chapters, because it not only summarizes the rest of the book but also places the work in perspective as a first step in the solution of a complex problem.

Appendix I: Data

Reconstructing the Data

The 1960 data on air pollution and certain other variables used by Lave and Seskin were given by Eugene Seskin to Theodore Thomas, who in turn listed the data in his Ph.D. thesis (*8*). The variables coded for 106 Standard Metropolitan Statistical Areas (SMSAs) and 10 State Economic Areas (SEAs) (Milwaukee, Wisc. was not included) and used in our regression analyses are listed in Table 6. Total population and total mortality rates for each SMSA or SEA were extracted by us from the same sources used by Lave and Seskin (*9, 10*). Data for Milwaukee were reconstructed by us from the original sources cited by the authors (*9-11*) and added to the Thomas data.

Means and standard deviations for the variables used in the regression analyses were calculated from the Thomas data. Since these statistics agree (±2%) with those presented by Lave and Seskin in Table D.1, we assumed the data used by Thomas were essentially the same data analyzed by the authors and by us.

Editing

This data set was checked against the original data sources cited by Lave and Seskin (*9-11*).

Values recorded for the air pollution variables (variables 1-6 in Table 6) were compared with published values (*11*). Lave and Seskin note (p. 30, footnote 2) that the data available for suspended particulate concentration were more complete than that for sulfate concentration. When 1960 data were not available, their rule was to use data from the closest year available. These procedures resulted in the breakdowns of SMSAs by year of data source shown in Table 7.

Table 6. Variables used in regression analyses.

Variable number	Variable name
1	Minimum observed particulate concentration (min *P*)
2	Maximum observed particulate concentration (max *P*)
3	Arithmetic average particulate concentration (mean *P*)
4	Minimum observed sulfate concentration (× 10) (min *S*)
5	Maximum observed sulfate concentration (× 10) (max *S*)
6	Arithmetic average sulfate concentration (× 10) (mean *S*)
7	Population per square mile (P/M^2)
8	Percent nonwhite (× 10) (NW)
9	Percent over age 65 (× 10) (≥ 65)
10	Percent family income less than $300 (× 10) (poor)
11	Total population (Pop)
12	Total mortality (rate per 100,000)

Table 7. Breakdowns of SMSAs by year of data source.

Year	Suspended particulate data	Sulfate data
1957	—	16
1958	—	64
1959	4	32
1960	80	3
1961	30	—
Unknown	3	2

Additionally, the PHS publication (*11*) reports the sulfate data for each year in one of three formats: frequency distribution for the individual year, quarterly composite, or quarterly heavy. It appeared to us that Lave and Seskin extracted the data from the frequency format, or if that was not available, obtained the minimum and maximum readings from the latter two tables combined together. Table 8 presents the breakdowns by year and format for the sulfate data.

The socioeconomic variables used in the regression analyses (variables 7-10 in Table 6), were compared to the authors' data source (*10*). The errors in coding are noted in the next section.

Table 8. Sources of sulfate concentration data.[a]

Year	Frequency distribution	Composite and heavy	Heavy only	Total
1957	1	14	1	16
1958	41	22	1	64
1959	10	18	4	32
1960	3	0	0	3
Total	55	54	6	115

[a]Two SMSAs had data of undetermined origin.

Data Correction

Coding errors were detected for two SMSAs: Atlanta, Ga., variables 4-6, 8-10 (see Table 6); and Bridgeport, Conn., variables 4-6. After correcting these errors, means and standard deviations of the corrected data were calculated. These did not markedly deviate from the ones reported by the authors.

The corrected data set was used in our reanalysis.

Appendix II: Outlying SMSAs (SEAs)

Our method of detecting outlying SMSAs is to look for SMSAs with explanatory variable vectors which are beyond the expected range. Specifically, we use the distance between the explanatory vector defining an SMSA and a vector of medians, scaled by a robust estimate of the covariance matrix. We now describe the methods and results for robust estimation of the covariance matrix. Following this, we consider each SMSA (SEA) separately and identify six outlying SMSAs(SEAs). A general discussion of obtaining robust estimates of covariance and identifying outliers has been given by Devlin, Gnanadesikan, and Kettenring (*3*).

Our first step was to examine the distributions of the explanatory variables. Histograms are mostly skewed and longtailed, for example see Fig. 2.1.

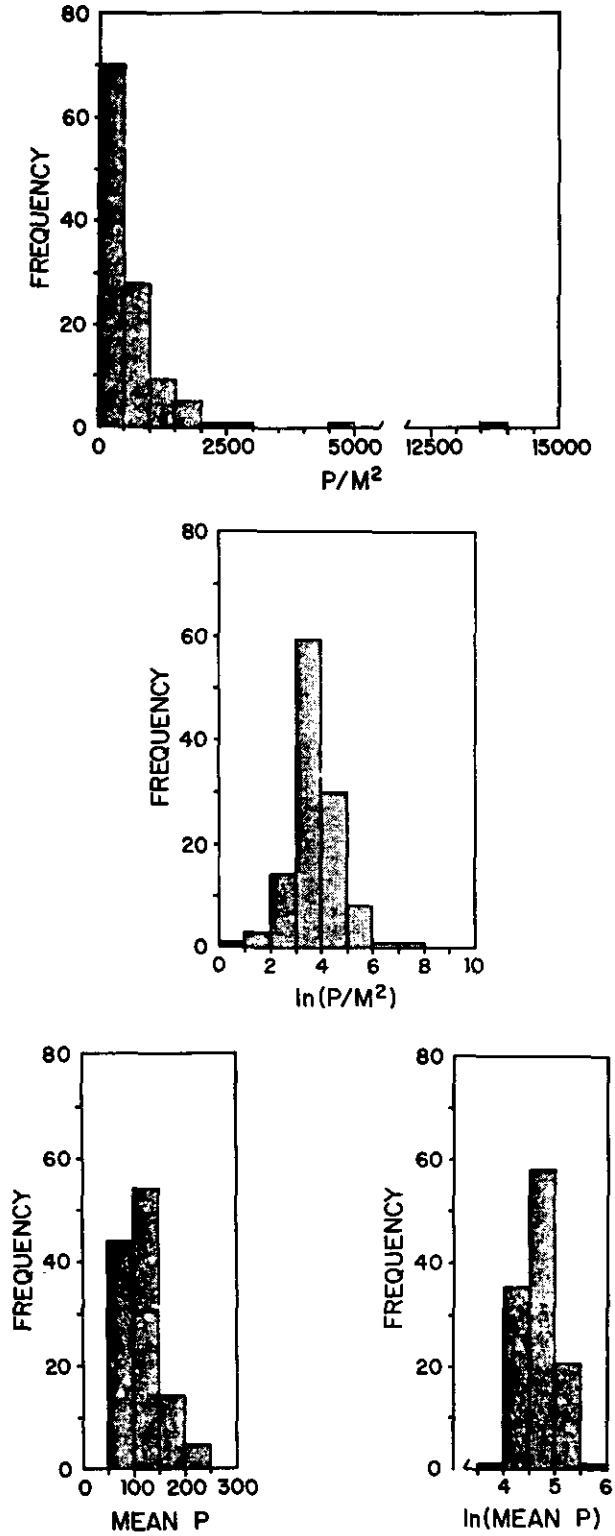

FIGURE 2. Distributions of population per square mile (P/M²) and annual mean particulate level (Mean P) for 117 SMSAs in original and logarithmic scale, illustrating the symmetrizing effect of logarithmic variable transformation.

Logarithmic scaled histograms of the variables suggested that ln-transformed data are more nearly symmetrically distributed than the data in the original scale. Natural logarithms of the raw data were used in the next step of our analysis, the robust determination of covariance.

The next step in developing an estimate of the covariance matrix was to obtain a robust estimate of the standard deviation ($\hat{\sigma}_i$) for each of the eleven explanatory variables (x_i). Using the interquartile ranges $I(x_i)$ ($i = 1, \ldots, 11$), the estimates $\hat{\sigma}_i$ were calculated as,

$$\hat{\sigma}_i = \frac{I(x_i)}{2(0.6745)} \qquad i = 1, \ldots, 11$$

Then for each possible unique pair of explanatory variables (x_i, x_j), interquartile ranges for the sum and difference of the standardized variables were calculated;

$$I_{ij}^{+} = I \left[\frac{x_i}{\hat{\sigma}_i} + \frac{x_j}{\hat{\sigma}_j} \right]$$

and

$$I_{ij}^{-} = I \left[\frac{x_i}{\hat{\sigma}_i} - \frac{x_j}{\hat{\sigma}_j} \right]$$

Robust estimates of correlation coefficients (r_{ij}) were then calculated by using:

$$r_{ij} = \frac{(I_{ij}^{+})^2 - (I_{ij}^{-})^2}{4[2(0.6745)]^2} \qquad i \neq j$$

These gave an initial estimate $\tilde{R}$ of the correlation matrix:

$$\tilde{R} = [r_{ij}], \; r_{ii} = 1$$

Eigenvalues and eigenvectors were calculated for $\tilde{R}$. Since some of the eigenvalues were negative, we proceed to apply an iterative technique recommended by Devlin, Gnanadesikan, and Kettenring (*3*) to obtain a positive definite estimate of the correlation matrix (R^*): Each r_{ij} ($i \neq W$) was multiplied by 0.99 and eigenvalues and eigenvectors were recalculated. If any eigenvalue was still negative, each r_{ij} ($i \neq j$) was again multiplied by 0.99 and the process repeated until all eigenvalues were positive. Convergence occurred after 33 iterations.

To obtain the robust estimate of the covariance matrix, a diagonal matrix of $\hat{\sigma}_i$ $i = 1, \ldots, 11$ was formed:

$$S = \begin{bmatrix} \hat{\sigma}_1 & & 0 \\ & \ddots & \\ 0 & & \hat{\sigma}_{11} \end{bmatrix}$$

Then the covariance matrix $\tilde{\Sigma}$ was estimated by:

$$\tilde{\Sigma} = S\, R^* S$$

The computed values for $\hat{\sigma}_i$, $i = 1, \ldots, 11$ and R^* are given in Tables 9 and 10.

Table 9. Medians and estimated standard deviations for the ln-transformed explanatory variables.

Index number	Variable[a]	Median	Standard deviation
1	ln(min $S \times 10$)	3.74	0.592
2	ln(mean $S \times 10$)	4.39	0.571
3	ln(max $S \times 10$)	5.35	0.496
4	ln(min P)	3.76	0.463
5	ln(mean P)	4.73	0.414
6	ln(max P)	5.48	0.468
7	ln(P/M^2)	3.75	0.786
8	ln(≥65×10)	4.43	0.241
9	ln(NW×10)	4.54	1.011
10	ln(Poor×10)	5.03	0.435
11	ln (Pop)	12.75	0.859

[a]Pollution variables S and P are in μg/m^3. Socioeconomic variables ≥65, NW, and Poor are in percent of total population.

Table 10. Robust estimate of the correlation matrix of the ln-transformed explanatory variables. Row and column numbers correspond to the index numbers in Table 9.

	1	2	3	4	5	6	7	8	9	10	11
1	1.0000	0.5982	0.6376	0.2514	0.4493	0.6115	0.4098	−0.1584	0.1680	−0.2825	0.3310
2		1.0000	0.8806	0.2520	0.3899	0.8982	0.7445	−0.3519	0.6399	−0.6941	0.2460
3			1.0000	0.2558	0.4801	0.7235	0.3938	0.0797	0.1824	−0.2147	0.3640
4				1.0000	0.6217	0.5538	0.3829	0.1698	0.3922	−0.0771	−0.0172
5					1.0000	0.8982	0.7445	−0.3519	0.6399	−0.4549	0.4060
6						1.0000	0.6081	0.1731	0.6430	−0.4826	0.5333
7							1.0000	0.1269	0.1628	−0.3650	0.6866
8								1.0000	−0.5348	0.6714	−0.1407
9									1.0000	−0.5528	0.0258
10										1.0000	−0.4580
11											1.0000

Table 11. Distances of each SMSA explanatory vector from the median vector with respect to the robustly estimated variance-covariance matrix.

SMSA	*Distance*	*SMSA*	*Distance*
Albuquerque, N.M.	98.3	Manchester, N.H.[a]	200.2
Allentown, Pa.	36.4	Memphis, Tenn.	55.6
Atlanta, Ga.	60.1	Miami, Fla.	106.7
Atlantic City, N.J.	185.9	Milwaukee, Wis.	55.3
Augusta, Ga.	70.3	Minneapolis, Minn.	14.9
Austin, Tex.	54.2	Mobile, Ala.	181.5
Baltimore, Md.	152.5	Montgomery, Ala.	9.9
Baton Rouge, La.	31.1	Nashville, Tenn.	19.8
Beaumont, Tex.	24.3	Newark, N.J.	56.8
Birmingham, Ala.	129.2	New Haven, Conn.[a]	47.2
Boston, Mass.[a]	69.7	New Orleans, La.	58.5
Bridgeport, Conn.[a]	49.8	New York City, N.Y.	36.8
Brockton, Mass.[a]	81.2	Norfolk, Va.	22.1
Canton, Ohio	24.2	Oklahoma City, Okla.	16.2
Charleston, S.C.	151.4	Omaha, Neb.	19.5
Charleston, W.V.	616.7	Orlando, Fla.	8.9
Charlotte, N.C.	5.8	Philadelphia, Pa.	68.7
Chattanooga, Tenn.	32.8	Phoenix, Ariz.	463.0
Chicago, Ill.	14.8	Pittsburgh, Pa.	81.8
Cincinnati, Ohio	39.6	Portland, Maine	127.0
Cleveland, Ohio	45.0	Portland, Ore.	15.6
Columbia, S.C.	104.5	Providence, R.I.[a]	79.8
Columbus, Ga.	27.8	Raleigh, N.C.	10.7
Columbus, Ohio	86.9	Reading, Pa.	41.1
Dallas, Tex.	44.5	Richmond, Va.	5.0
Dayton, Ohio	17.3	Roanoke, Va.	46.4
Denver, Col.	37.8	Rockford, Ill.	61.6
Des Moines, Iowa	24.7	Sacramento, Calif.	23.9
Detroit, Mich.	20.8	Saginaw, Mich.	13.7
Duluth, Minn.	52.1	Saint Louis, Mo.	8.7
El Paso, Tex.	107.2	Salt Lake City, Utah	10.3
Fall River, Mass.[a]	175.4	San Antonio, Tex.	57.4
Flint, Mich.	98.6	San Diego, Calif.	6.0
Fort Worth, Tex.	32.3	San Francisco, Calif.	15.6
Fresno, Calif.	388.3	San Jose, Calif.	9.1
Galveston, Tex.	75.4	Savannah, Ga.	7.5
Gary, Ind.	33.1	Scranton, Pa.	26.8
Greensboro, N.C.	177.2	Seattle, Wash.	56.3
Greenville, S.C.	84.0	Shreveport, La.	105.8
Hamilton, Ohio	39.3	Sioux Falls, S.D.	24.6
Harrisburg, Pa.	18.0	South Bend, Ind.	28.1
Hartford, Conn.[a]	32.2	Springfield, Mass.[a]	2.5
Houston, Tex.	129.5	Springfield, Ohio	18.1
Huntington, W.V.	12.6	Tacoma, Wash.	39.7
Indianapolis, Ind.	77.2	Tampa, Fla.	37.0
Jackson, Mich.	63.1	Terre Haute, Ind.	80.2
Jackson, Miss.	7.1	Toledo, Ohio	46.0
Jacksonville, Fla.	8.7	Topeka, Kans.	108.9
Jersey City, N.J.	565.6	Waco, Tex.	17.0
Johnstown, Pa.	81.0	Washington, D.C.	18.7
Kansas City, Mo.	27.9	Wheeling, W.V.	52.3
Knoxville, Tenn.	36.9	Wichita, Kans.	18.2
Lansing, Mich.	13.3	Wilkes Barre, Pa.	49.9
Las Vegas, Nev.	629.7	Wilmington, Del.	21.0
Little Rock, Ark.	6.6	Winston-Salem, N.C.	6.9
Lorain, Ohio	87.8	Worcester, Mass.[a]	59.8
Los Angeles, Calif.	24.2	York, Pa.	88.7
Macon, Ga.	562.3	Youngstown, Ohio	8.1
Madison, Wis.	7.6		

[a]State Economic Area (SEA).

Table 12. Outlying SMSAs.

SMSA	Distance
Charleston, W.V.	616.7
Fresno, Calif.	388.3
Jersey City, N.J.	565.6
Las Vegas, Nev.	629.7
Macon, Ga.	562.3
Phoenix, Ariz.	463.0

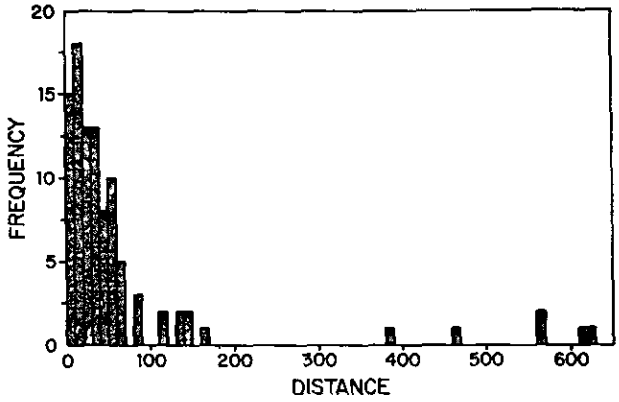

FIGURE 3. Distribution of distance between SMSA explanatory variable vector and the median exploratory vector, standardized by the robustly estimated variance-covariance matrix.

Identification of Outlying SMSAs

Outlying SMSAs were identified by calculating the distance between the observed and median explanatory vectors using the robust estimate of covariance as the norm. Let $M = (m_1, m_2, \ldots, m_{11})$ be the vector of medians and $X_j = (x_{1j}, x_{2j}, \ldots, x_{11j})$, $j = 1, \ldots, 117$, be the observed explanatory vector defining the j-th SMSA. Then for each SMSA, the distance between the observed vector X_j and the median vector M is calculated as

$$\text{Distance } (j) = (X_j - M)' \, \tilde{\Sigma}^{-1} \, (X_j - M)$$

where $\tilde{\Sigma}^{-1}$ is the inverse of the robust estimate of the covariance matrix, $\tilde{\Sigma}$.

The distances for each of the 117 SMSAs were computed (Table 11). Figure 3 is a histogram of the 117 distance measures, one for each SMSA. Several of the SMSAs have distance measures which are markedly different from the other SMSAs. The median distance is 41 and the third quartile is 81. Clearly the six largest distance measures are markedly different from the remaining 111. Specifically we identify the six SMSAs shown in Table 12 as outlying.

These SMSAs were subsequently omitted from the 1969 data set and the data reanalyzed. Among the remaining 111 SMSAs, the largest measure is 200.2, Manchester, N.H.

We would like to extend our appreciation to Professors F. Mosteller, F. Speizer, and J. Whittenberger for their helpful comments and advice as this work progressed. Typing support was provided by Ellen Collins and Nina Leech.

Our thanks go also to the Electric Power Research Institute, Palo Alto, California, for supporting this work.

REFERENCES

1. Lave, L. B., and Seskin, E. P. Air Pollution and Human Health. Johns Hopkins University Press, Baltimore, 1977.
2. Hill, A. B. The environment and diseases: associations and causation. Proc. Roy. Soc. Med. Sect. Occup. Med. 58: 272 (1965).
3. Devlin, S. J., Gnanadesikan, R., and Kettenring, J. R. Robust estimation and outlier detection with correlation coefficients. Biometrika 62 : 531(1975).
4. Dixon, W. J., and Brown, M. B. BMDP-77, Biomedical Computer Programs, P-Series. University of California Press, Berkeley, 1977.
5. Barr, A. J., Goodnight, J. H., Sall, J. P., and Helwig, J. P. A User's Guide to SAS 76. SAS Institute, Inc., Raleigh, N.C., 1976.
6. Hoaglin, D. C., and Welsch, R. E. MIT-SNAP, An Interactive Data Analysis System. Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Mass., 1977.
7. Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., and Bent, D. H. SPSS, Statistical Package for the Social Sciences, 2nd ed., McGraw-Hill, New York, 1975.
8. Thomas, T. J. An investigation into excess mortality and morbidity due to air pollution. Purdue University, Ph.D. Thesis, Environmental Sciences. University Microfilms, Ann Arbor, 1973.
9. U.S. Department of Health, Education and Welfare. Vital Statistics of the United States, 1960. GPO, Washington, 1963.
10. U.S. Bureau of the Census. County and City Data Book, 1962. GPO, Washington, 1962.
11. U.S. Department of Health, Education and Welfare. Air Pollution Measurements of the National Air Sampling Network, Analyses of Suspended Particulates, 1957-61. U.S. Public Health Service Publication 978, Cincinnati, 1962.

Comments on "Air Pollution and Human Health: A Reanalysis"

Lester B. Lave* and Eugene P. Seskin†: Ferreting out the relationship between various air pollutants and their effects on human health is as good an intellectual puzzle as one can find. But unlike world-class bridge, the results and the policies that follow can have major impacts on our lives, ranging from shorter life expectancies with more chronic disease to paralysis of the economy from needlessly restrictive regulations. While the work does not have the glamour of center court at Wimbledon, or even of fundamental cancer research, it does attempt to keep the government from giving all-or-nothing answers to complicated environmental questions.

Thibodeau, Reed, and Bishop (T-R-B) have contributed to the solution of this puzzle, in part by investigating areas in which we were less than complete, and in part by examining independently the same basic data. Before getting down to "hand-to-hand combat," we want to stress the conclusions that we share with T-R-B. (1) There is a close, statistically significant association between air pollution (as measured by sulfates and suspended particulates)

and mortality rates in the United States; this relationship is evident over time and across metropolitan areas. (2) The close association is relatively robust, as evidenced by analyses of various data bases, alternative functional forms, different statistical techniques, and other exploratory methods. (3) If the estimated relationship is a causal one, the estimated effect of air pollution on health is large, warranting stringent abatement of sulfur oxides and particulates (and perhaps other air pollutants).

At this point, agreement between T-R-B and ourselves on the principal issue of causality is not evident. We believe that the statistical evidence, in conjunction with knowledge gained from epidemiological studies of particular groups, laboratory experiments, physiology, and air chemistry are consistent with the conclusion that a causal relationship has been demonstrated. T-R-B do not commit themselves, although they express a lack of confidence in using the results to show cause-and-effect, given shortcomings of the underlying data. At the same time, T-R-B provide no evidence that the relationship is spurious.

While our investigation has only scratched the surface in trying to get answers to this intellectual puzzle, we think enough is now known both to sharpen the questions for new investigations and to reevaluate public policy. Clearly, the latter conclusion is not accepted by everyone. Both conclusions warrant a careful examination of possible flaws in our analysis, attempts by T-R-B to expose and correct them, and a further look at some of the difficulties of the problem itself. Thus, we turn to the civilized "thrust and parry" of disputation concerning epidemiological research.

The major areas of contention are those surrounding method, and perhaps the most controversial of these is whether to exclude "unusual" observations. We did not exclude observations with less than a pleasing appearance, although we did employ various methods of exploratory data analysis. In contrast, T-R-B chose to exclude outlying observations and much of the complexity and differences in their analysis can be attributed to the procedure they use to "drop" observations and the resulting estimates they obtain in the subsequent reanalysis.

Since least-squares regression analysis minimizes the sum of squared residuals, it accords a great deal of weight to unusual or outlying observations. Occasionally, the sign of an estimated coefficient can be reversed by dropping only a few observations. Given the considerable deficiencies in the underlying data, particularly the aerometric data, it is undesirable to rely on results that are sensitive to a few, perhaps erroneous, data points. Nevertheless, before scampering off to banish outliers, one must remember that nature is less than beneficent in performing experiments. Observational data exhibit an alarming appearance of homogeneity and seldom occur in the design of a Latin square. Instead natural experiments often "produce" data clustered around a multidimensional central tendency. Thus, to toss out unusual observations can amount to throwing out the baby with the bathwater, with the result being greater collinearity between the independent variables.

We approached the problem of "extreme" observations differently. First, we used so-called "jackknife" tests, which systematically exclude observations in order to investigate the sensitivity of parameter estimates to omitted data. For example, one method ordered the observations according to the level of particulate pollution and then dropped blocks of contiguous observations. From that analysis it was determined that the parameter estimates were no more sensitive to outlying data points than would be expected from the previously estimated standard errors. A second approach to the problem involved replicating the original analysis using data from different years, including a cross-sectional time-series investigation as well as a pure time-series study. The similarity of the parameter estimates across these different data sets lent confidence to our conclusions.

There are no simple solutions to the problems associated with the inherent collinearity among socioeconomic data. In our case, industrial activity will be correlated with pollution levels, as well as with such factors as income levels. Furthermore, these variables are not likely to exhibit great changes over a period of time as short as a decade. This problem becomes especially apparent when we add variables for occupation mix and home-heating characteristics. The resulting collinearity precludes estimating the parameters with precision or confidence. Nevertheless, we think the changes exhibited in the parameter estimates are explicable and that the results are in general corroboration with the hypothesis that air pollution causes mortality. At the same time, we recognize fully that our explanations are no more than further hypotheses which need further exploration.

T-R-B went through the excruciatingly difficult task of checking our data against original sources. Anyone engaged in such a labor of love of scientific accuracy is to be commended. The book notes that 1960 aerometric data were not universally available and so data from neighboring years occasionally were used; they documented the extent of substitution. However, they mention that two SMSAs could not be checked. We still do not agree with their data but, rather than quarrel about minutiae, we point out that the 1969 replication, which had no comparable data problems, confirmed the 1960 estimates.

One of the lessons we learned in our attempts to discern patterns in the parameter estimates is that one should not place too much emphasis on any single estimate. Thus, our preferences changed from relying on individual coefficients for the air pollution variables to considering the sum of elasticities of several coefficients (for example, the sum of the coefficients for the minimum, mean, and maximum sulfate levels) at a time. Unfortunately, T-R-B report only individual air pollution coefficients in their reanalysis instead of the sums of elasticities. Thus, while there appears to be considerable variation in their estimated parameters, some of this is spurious because specific coefficients increase while others decrease (for example, their coefficient for maximum sulfate rises while that for minimum sulfate falls).

T-R-B assert that we should have tailored an analysis to the infant mortality rates rather than using the same specification that we used for the total mortality rate. However, we would point out that our analyses encompass more than forty mortality rates and a number of different data sets. Clearly, we did not want to estimate an *ad hoc* specification for each mortality rate and each data set. Instead, we chose to preserve the basic 1960 specification throughout most of the succeeding analyses so comparisons could be made with genuine replications. In making this decision, it appeared to us that the same basic socioeconomic variables should be included in the various analyses. At the same time, we recognized that the estimated coefficients for certain variables would take on different interpretations in the different analyses. Be that as it may, there is merit in the T-R-B assertion, and we hope that they will turn their worthy talents to examining further the infant mortality rates.

To our surprise, T-R-B give short shrift to our analysis of suicides, venereal disease, and crime rates. We think that analysis provides a stringent test of the hypothesis that the set of socioeconomic and air pollution variables is masking some other variable (or variables) that is the true cause of the association. The statistical significance of these analyses certainly suggests that the data are far from random. T-R-B dismiss the venereal diseases analysis on grounds that make little sense. Both the Center for Disease Control (CDC) and previous analyses have indicated difficulties with these data. However, they are the standard data used in such work and they lead to plausible estimates. We know of no grounds for dismissal. In addition, T-R-B criticize our failure to replicate the crime rate analysis. Given the results from the 1960 data, we saw little to be gained by a replication. Furthermore, at the time of the analysis, no data were available for a replication. We invite T-R-B to collect and analyze 1970 data if they deem it worthwhile.

There are a host of other, more minor points that T-R-B raise;

these are handled best by private communication. However, we must plead innocent publicly to some of their charges. For example, T-R-B take us to task for failing to present an analysis of residuals for the 1960 data. Actually, that analysis is presented later in the book, together with the analysis of residuals for the 1969 data. Furthermore, contrary to T-R-B's assertion, we also explicitly discuss outliers such as Tampa, Florida and Scranton and Wilkes-Barre, Pennsylvania. T-R-B also note that linearity may not hold outside the range of data and therefore extrapolations may be in error. We specifically "flag" this issue in Chapter 10.

T-R-B have written a review that is almost as long as our book; we do not want to succumb to the temptation to write a rejoinder longer than the review. Having stressed our differences with T-R-B, we want to express now our high regard for their careful work and we want to reiterate the similarities in the results. There is a close association between the levels of specific air pollutants and mortality rates in the United States. Much has been learned from our two investigations. However, they represent only the first steps in an enormously difficult field of research.

*Carnegie-Mellon University and Brookings Institution, Washington, D.C. 20036.

†Bureau of Economic Analysis, U.S. Department of Commerce.