Environmental Health Perspectives
Vol 90, pp. 209-213, 1991

Structural Equation Modeling in Environmental Risk Assessment

by Charles Ralph Buncher,* Paul A. Succop,* and Kim N. Dietrich*

Environmental epidemiology requires effective models that take individual observations of environmental factors and connect them into meaningful patterns. Single-factor relationships have given way to multivariable analyses; simple additive models have been augmented by multiplicative (logistic) models. Each of these steps has produced greater enlightenment and understanding. Models that allow for factors causing outputs that can affect later outputs with putative causation working at several different time points (e.g., linkage) are not commonly used in the environmental literature. Structural equation models are a class of covariance structure models that have been used extensively in economics/business and social science but are still little used in the realm of biostatistics. Path analysis in genetic studies is one simplified form of this class of models. We have been using these models in a study of the health and development of infants who have been exposed to lead *in utero* and in the postnatal home environment. These models require as input the directionality of the relationship and then produce fitted models for multiple inputs causing each factor and the opportunity to have outputs serve as input variables into the next phase of the simultaneously fitted model. Some examples of these models from our research are presented to increase familiarity with this class of models. Use of these models can provide insight into the effect of changing an environmental factor when assessing risk. The usual cautions concerning believing a model, believing causation has been proven, and the assumptions that are required for each model are operative.

Introduction

Studying the environment is more difficult than it used to be. In the past, one could observe one particular environmental contaminant and one particular health effect and report that they were related. As our knowledge of other effects on health grew, we had to consider more than one factor when studying a particular health effect, whether these additional factors were other causes, modifying factors, or confounding factors. Consideration of multiple factors led to the use of multivariate models, in which no longer was one agent at a time observed, but a multitude of risk factors were considered simultaneously. The actual methodology used was frequently a multiple linear regression with least-squares fit because the computer programs were available and the computers could readily handle such calculations.

For each study, there was a theoretical model available and a set of data. None of the models used in these situations was particularly realistic. For example, the authors of such reports knew that contaminants did not act one at a time, but the model was easy to apply in the real world, especially when one was looking at large effects. Likewise, few really believed that risk factors in the arbitrary scale used to measure them in multiple linear regression could be cumulated on a simple additive scale over the possible range of values in accord with the model for the regression. The addition of interaction terms in the equations helped but did not solve these problems, especially if the interacting variables were effect modifiers or were causally related to the outcome themselves. Still, these regression models provided a useful first step in the research.

A next step was the development of more realistic models. In particular, the multiplicative family of models that has been derived from the work of Cox on logistic regression has been very helpful (*1*). Many biologic phenomena seem to work proportionally, that is, there is a 2-fold increase regardless of the baseline value over a wide array of values. This is quite commonly observed, and in the minds of many, it is a more common situation than a two-unit (additive) increase regardless of the baseline value. The literature shows that the introduction of Cox models into epidemiologic and environmental research has produced many new insights into relationships. We believe that it is time for a new type of model to introduce a new round of research resulting in greater understanding.

Structural Equation Models

There are many newer models that are helpful in learning more about environmental problems. We take the view that models provide additional insight into a situation by showing the relative importance of various factors and by providing quantitative relations between variables and by providing testable predictions of relationships in new situations.

One can create nonlinear models that are usually more complex than the linear models, including those that deal with the

*Department of Environmental Health, University of Cincinnati Medical Center, Cincinnati, OH 45267-0183.

Address reprint requests to C. R. Buncher, Department of Environmental Health, University of Cincinnati Medical Center, Cincinnati, OH 45267-0183.

kinetics of the situation. In some situations these models can be very helpful. Our research problem concerned a less defined area involving both mathematically well-defined variables and variables whose specifications were not well known. We investigated and have now used for some 6 years structural equation models. These models have received wide use in the econometric literature and in the psychometric literature but little use to date in the environmetric literature.

We would like to point out the advantages of structural equation models in longitudinal environmental research because we feel they are of great value in researching those environmental problems in which measurements are taken over several time points. These models are still little known in the environmental/epidemiologic community, and they are not used as frequently as they should be at this time. We believe that increased use of these models will create more knowledge and interest in environmental problems, will be more informative than most current models, and will help us predict how best to control environmental hazards in the most efficient way.

The practical concerns are important. One technique that may be used for estimating a structural equation model involves the use of the computer program LISREL (linear structural relations) (*2*). The output is a maximum likelihood fit of the model to the data set showing the fitted regression coefficients in either standardized or unstandardized form. One of the results calculated by the LISREL program is a chi-square goodness-of-fit statistic to evaluate how closely the total model fits the total data set. This is useful when there are competing models because one can fit each model and then compare the chi-square values that emerge. Because the difference between two chi-square values also has a chi-square distribution, the models can be compared and tested. If the models are nested in a hierarchical manner, then the two models can be compared directly by using the difference in the chi-square values for the two models. The result is either that one model is significantly better than the other or that the data do not find a statistically significant difference in the fit of the two models and thus both are contenders for explaining the relationships among the variables. As in most modeling procedures, one prefers to use the most parsimonious explanation for the data.

Let us work through a research example to illustrate how these models are developed and how they can be more informative than regression models. The problem was to relate data from observations at several different time points into a coherent picture of the influence of an environmental pollutant. In those instances in which the data can be collected at time point one to predict or model the data to be obtained at time point two, one traditionally uses some type of regression analysis. A dependent variable to be measured at time point two is predicted by independent variables at time point one. The question facing our group was how to use data from at least four different time points to model results.

Suppose one wants to understand the deficits in learning ability or accomplishment at 3 years of age due to the effects of lead in that child's lifetime. One can use data from early life, say age 1, to predict results at age 3, but there are also birth data that can be used, and in addition data concerning the pregnancy and even information concerning the preconception characteristics of the prospective parents. Usual regression models tend to be flat over time and are not well adapted to using the information from stage one to predict stage two and in turn to predict stage three or four.

Structural equation modeling is one possible solution to this dilemma. These models make use of the pairwise correlations or covariances among the variables and a statement of presumed relations among the variables as input and then the system produces a best fit, in the maximum likelihood sense, to create a model of the total system.

Our group has been studying the effects of environmental lead on the development of infants living in the inner city of Cincinnati. This predominantly lower socioeconomic status sample of approximately 300 children resides in an area with a long history of cases of pediatric lead poisoning. Environmental studies have shown conclusively that in this cohort lead from paint, dust, and soil associated with poor housing stock is the major contributor to body burden (*3,4*).

We started with regression models for the relations among the environmental factors, the confounding factors, the other characteristics of the families, and the outcome variables. These models did not take into account the natural time relationships of the measurements. For example, there were factors measured during the index pregnancy including the mother's blood lead, there were factors measured at birth, there were factors measured at the various longitudinal follow-up times such as 3 months after birth, 6 months after birth, 1 year after birth, and so forth.

A factor measured at birth could be considered a cause or modifying factor of some outcome at 6 months, but at the same time, the factor measured at birth can be considered an outcome of prenatal factors. Regression models do not have any natural methodology for acknowledging these relationships in time, while structural equation models are exactly appropriate for this use of data (*5*).

Accordingly, we outline progress in the last 6 years of this study to help other researchers make use of this methodology. We believe that structural equation models will enhance our understanding of environmental hazards. If we are effective in convincing more people analyzing data to use these methods, there will be the inevitable abuses. At that time, we shall be happy to discuss further the possible misuses of the methodology.

An Example

Let us walk through the creation of a model which we have used previously to study the effects of prenatal lead exposure on sensorimotor development in early infancy (*6*). One possible model is that the blood lead of the child directly influenced the development of the child as measured with the Bayley Psychomotor Development Index (PDI) or the Mental Development Index (MDI), resulting in a linear regression model.

Most people think that the mother's blood lead (PbB) during pregnancy is predictive of the child's blood lead at birth (PbBB) or in later infancy, and more so the closer in time for the two measurements. In other words, the easily measurable mother's lead will reflect the *in utero* fetal exposure, which would be more difficult or impossible to measure. The fetal exposure in turn may influence development. Of course, the mother's smoking and alcohol use as well as measures of the environment such as type of housing and past history of exposure will influence the mother's blood lead. The child's birth weight and gestational age are also known covariates of early infant development.

Table 1. Reduced multiple regression model of log prenatal maternal blood lead on Bayley Mental Development Index at 6 months.[a]

Variable	Beta	SE	*t*	*p*
Prenatal PbB	− 5.95	2.61	−2.28	0.023
Sex of child	−17.45	8.05	−2.17	0.031
Prenatal PbB × sex	10.54	3.96	2.66	0.008
Birth weight, g	0.008	0.002	3.53	0.0005
Gestational age, weeks	1.83	0.63	2.93	0.004

[a] Beta is the regression slope; SE is the standard error of beta; PbB is blood lead (μg/dL); sex was coded 0 for males and 1 for females.

We started with a multivariate multiple regression strategy that tested the hypothesis of whether prenatal or early neonatal PbB had adverse direct effects on 6-month Bayley MDI (*6*). The final reduced model (*6*) indicated that child's sex, birth weight, and gestational age each made independent and statistically significant ($p < 0.05$) contributions to the child's sensorimotor development. The prenatal (maternal) blood Pb and a sex by prenatal blood Pb interaction were also significant. The interaction indicated that greater effects of Pb were observed in the male infants (Table 1).

Since a neonate's PbB before 6 months of age is predominantly determined by *in utero* exposure, a simple structural equation model was also employed, using only an indicator variable (0 versus 1) of tobacco and/or alcohol use (CITAC), fetal Pb as indicated by prenatal PbB, intrauterine development as indexed by the child's birth weight and gestational age; and neurobehavioral status at 6 months, as indicated by the child's MDI and PDI. A backward elimination of any nonsignificant structural equation paths was calculated, resulting in the model shown (Fig. 1). This model is the product of an interim analysis based on approximately half of the total subjects in the full longitudinal cohort (*7*).

One problem that must be faced in using all models is the role of modifying variables. Thus, prenatal PbB may influence birth weight, which in turn influences development at 6 months of age. The regression model usually uses both prenatal blood lead and birth weight as predictors of development at some later time. This has the advantage that the predictor is controlled or adjusted for the intervening variable. The problem is knowing when one has controlled enough. If both terms are in the model equally, i.e., both are independent variables with equal roles in the matrix, one runs the risk of overcontrolling. The model says that A and B jointly predict the effect E. In this situation, the influence of the variable of interest may be lost after controlling for one or more intervening variables even though the variable of interest has a true effect. This situation is a product of the intercorrelations of the variables with the variable of interest.

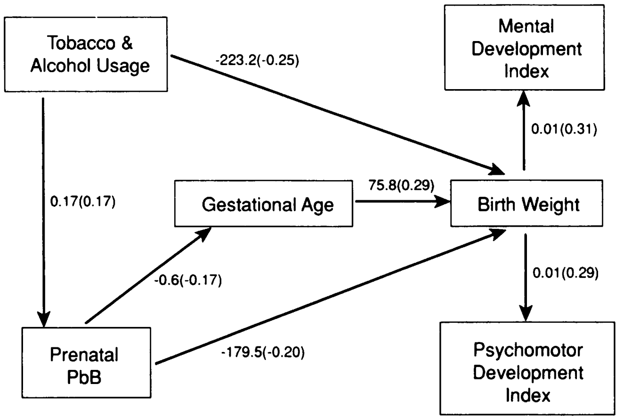

FIGURE 1. A simple structural equation model for MDI and PDI; adapted from Dietrich et al. (7).

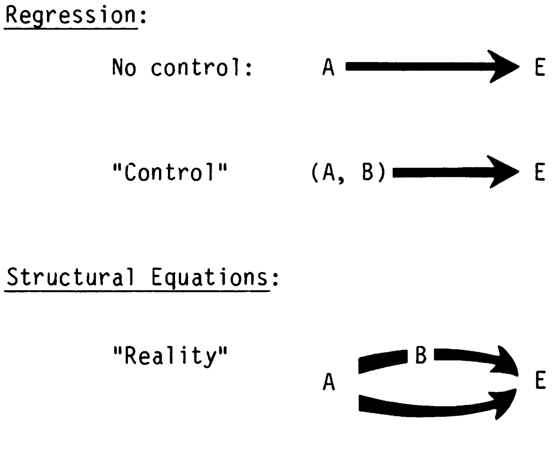

FIGURE 2. Regression models may control over other variables.

Structural equation modeling offers another possibility for the case in which the variable of interest and the intervening variable are measured at different time points. The model can be set up so that A predicts B which in turn predicts E, but at the same time A independently predicts E (Fig. 2). In this situation, A is allowed both of its roles, and the simultaneous model fitting procedure will search for the best subdivision of its influence into these two paths. More complicated relations are an easy byproduct of structural equation modeling.

One variable that appears in these models as a statistically significant predictor of both prenatal Pb and birth weight is the so-called composite index of tobacco and alcohol consumption (CITAC) variable. This is a dichotomous factor indicating the use of tobacco and/or alcohol during pregnancy or the non-use of either. To refine this rather global variable, the tobacco consumption (expressed as number of one-half packs of cigarettes smoked per day) and alcohol consumption (yes/no) were separated. In addition, self-reports of marijuana and narcotics use were included as two additional variables (each yes/no).

The use of this model allowed for the investigation of the variables with greater specification. It also allowed some theoretical investigations along the lines of "what if this variable reached some extreme value," that is, a sensitivity analysis. Moreover, one could consider the theoretical effect of extreme values in one variable on another variable.

To determine if this greater specification of substance use during pregnancy would change the results of previous analyses, the structural model of the 6-month developmental data was reanalyzed (*8*). In the refined models, cigarette smoking and alcohol use predicted prenatal PbB, although cigarette consump-

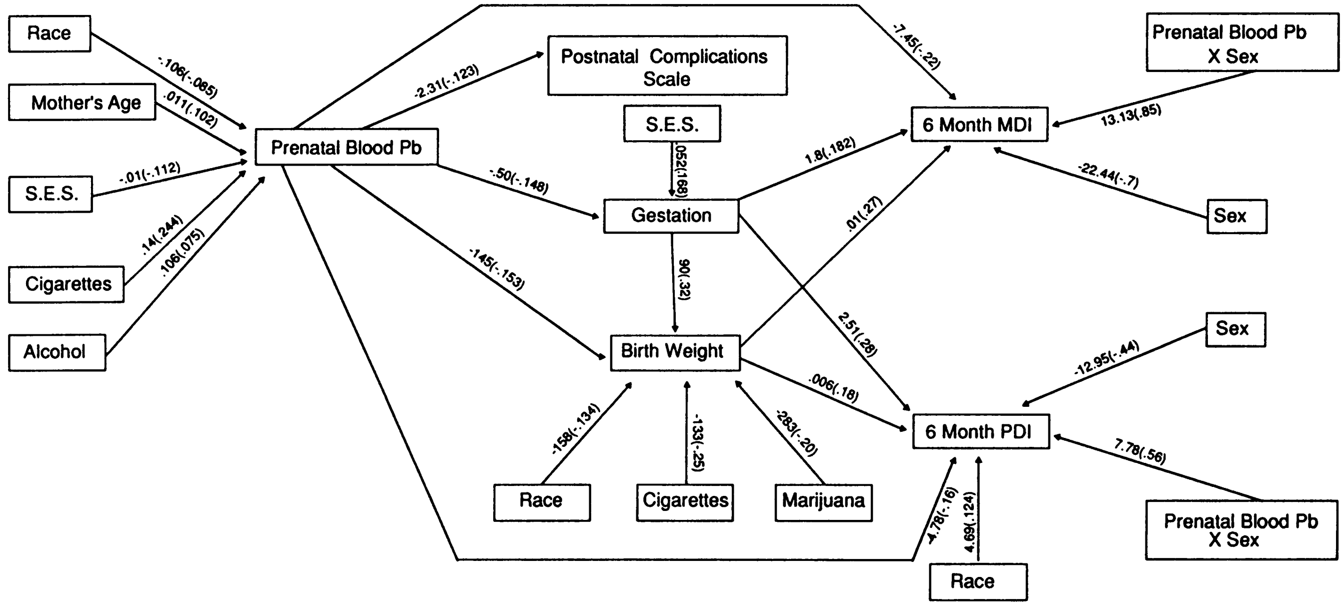

FIGURE 3. A refined structural equation model for 6-month MDI and 6-month PDI showing, among other factors, the effect of marijuana use on birth weight as well as the effect of prenatal blood lead; update of structural analysis of complete data set reported in Dietrich et al. (*6*).

tion predicted more of the variance in maternal PbB. As in the original models, prenatal PbB was significantly and inversely related to birth weight, with little difference between the parameter estimates in these models. An interesting finding in this refined model was that marijuana use was significantly and inversely related to birth weight. Figure 3 shows the resulting structural model.

Both direct and indirect effects of prenatal Pb were found in this larger cohort of children, with indirect effects once again seen through the intervening variables of gestation and birth weight, i.e., as Pb exposure *in utero* increased, a smaller baby was born somewhat earlier that would also be expected to have poorer neurobehaviorial status at 6 months. The direct effect was seen to be modified by a prenatal Pb by sex interaction, not seen in the smaller cohort or entertained in the simpler structural model. This static interaction (static at least so far as MDI and PDI are concerned) indicated that the direct effect of *in utero* Pb exposure (maternal PbB) was experienced only by male children.

One of the advantages of structural equation modeling is the path diagram that emerges as a byproduct of the analysis. Figure 3 shows the independent, intervening, and dependent variables. The independent variables, which are traditionally those which are first in time, are represented along the left-hand margin. The intervening variables, usually intermediate in time, are then displayed in the middle of the diagram showing their interrelationships. Finally the dependent, later outcome variables are displayed near the right-hand margin. Thus, time has the usual progression from left to right on these displays.

An endogenous variable is one that serves as a dependent variable in one or more of the structural equations. These are contrasted with exogenous variables that do not serve as a dependent variable in any of the structural equations.

Many of the less mathematically sophisticated consumers of the analysis can relate to the path diagram effectively. These same individuals may not be able to get much value from the series of regression equations since there is no pictoral presentation of results and these individuals may not relate well to mathematical expressions. It should be noted that the path diagram is an information-rich presentation. The path diagram shows the variables under consideration, their relative standing in time, their interconnections, and the numeric values for the relationships.

Characteristics

Some characteristics of the structural equation model are of importance. If there is proper specification of the model, the model always converges. Thus, the conceptual problems to getting answers that exist in other methodologies such as saddle points, singularities, and collinearity are not operative.

The analysis procedure does involve quite a lot of calculation. Thus, one needs a fairly sophisticated computer for the procedure and needs quite a bit of time on that computer. There are personal computer versions of structural equation programs, but they are still limited in scope. This means that currently, for most environmental problems of interest, a large mainframe is necessary but a supercomputer is not. We do not find that this computer-intensive characteristic is a critical factor with analysis. As machine computation continues to decline in cost, it is unlikely to be a major concern for any environmental study that has already expended many thousands or millions of dollars.

We would also like to contrast path analysis, which is well known to many statisticians, with structural equation modeling. Both use correlations as a starting point, but they are quite different. For example, path analysis is a least-squares procedure; structural equations use maximum likelihood fitting. Also in path analysis, all nodes must be connected with a correlation paths, while structural equations permit nodes to be omitted from the model.

Structural equation modeling is called by some causal modeling in recognition that the pathways may be interpreted as variable A causes variable B. We do not use that notation in this report to avoid raising issues that are not essential to the use of this procedure. Thus, in one view, structural equation models are just one more method of fitting a model to data. Just because one calls these models causal does not overcome all of the usual problems with discerning causation including variables not included in the model, inexact measurements such as measuring blood levels when tissue levels are pertinent, alternative models, and so forth. Others may choose to believe that they are working on the causal level. We do not think that these points are specific to structural equation modeling and therefore curtail this discussion at this point.

In summary, we think that the methods of structural equation modeling will be useful in studying environmental issues. We believe that studies that involve measurements at successive time points with complex interrelationships are better studied through structural equation models. The time dimension is better used in this type of model than in the more flat regression analysis. Like all other models, structural equation models do not solve all problems, but their greater use in the environmental literature will move this field ahead more rapidly.

Some of this work was supported by a program project grant from the National Institute of Environmental Health Sciences (P01-ES01566), Paul B.Hammond, Principal Investigator. The authors also thank the technicians and support staff at the University of Cincinnati College of Medicine and Babies Milk Fund Association Clinic who helped to make this research possible.

REFERENCES

1. Cox, D. R. Regression models and life-tables (with discussion). J. R. Stat. Soc. B 34: 187-220 (1972).
2. Joreskoq, K. G., and Sorbom, D. LISREL: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood User Guide. International Educational Resources, Chicago, 1981.
3. Bornschein, R. L., Succop, P. A., Dietrich, K. N., Clark, C. S., Que Hee, S., and Hammond, P. B. The influence of social and environmental factors on dust lead, hand lead, and blood lead levels in young children. Environ. Res. 38: 108-118 (1985).
4. Bornschein, R. L., Succop, P. A., Krafft, K. M., Clark, C. S., Peace, B., and Hammond P. B. Exterior surface dust lead, interior house dust lead and childhood lead exposure in an urban environment. In: A Symposium, Trace Substances in Environmental Health, II (D. D. Hamphill, Ed.), University of Missouri, Columbia, 1986, pp. 322-332.
5. Zellner, A. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. J. Am. Stat. Assoc. 57: 348-368 (1962).
6. Dietrich, K. N., Krafft, K. M., Bornschein, R. L., Hammond, P. B., Berger, O., Succop, P. A., and Beir, M. Low level fetal lead exposure effect on neurobehavioral development in early infancy. Pediatrics 80: 721-730 (1987).
7. Dietrich, K. N., Krafft, K. M., Shukla, R., Bornschein, R. L., and Succop, P. A. The neurobehavioral effects of early lead exposure. In: Toxic Substances and Mental Retardation: Neurobehavioral Toxicology and Teratology. AAMD Monographs 8: 71-95 (1987).
8. Dietrich, K. N., Succop, P. A., and Bornschein A. L. The effects of prenatal and postnatal lead exposure on fetal growth and child development. Paper presented at the Neurobehavioral Workshop on Lead sponsored by the United States Environmental Protection Agency and the International Lead Zinc Research Organization, Inc., March 20-22, 1989, Research Triangle Park, NC.