METHODS FOR ANALYZING PANEL STUDIES OF ACUTE
HEALTH EFFECTS OF AIR POLLUTION © 117

BY

EDWARD L. KORN

and

ALICE S. WHITEMORE

TECHNICAL REPORT NO. 25

MAY 1979

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

STUDY ON STATISTICS AND ENVIRONMENTAL
FACTORS IN HEALTH (SIMS)

PREPARED UNDER SUPPORT TO SIMS FROM

DEPARTMENT OF ENERGY (DOE)

ROCKEFELLER FOUNDATION

SLOAN FOUNDATION

ENVIRONMENTAL PROTECTION AGENCY (EPA)

NATIONAL SCIENCE FOUNDATION (NSF)

✓ DEPARTMENT OF STATISTICS
✓ STANFORD UNIVERSITY
STANFORD, CALIFORNIA

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

28

DISCLAIMER

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

DISCLAIMER

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

Methods for Analyzing Panel Studies of Acute
Health Effects of Air Pollution

EDWARD L. KORN

Department of Biomathematics, University of California,
Los Angeles, California 90024, U.S.A.

ALICE S. WHITEMORE

Department of Statistics, Stanford University,
Stanford, California 94305, U.S.A.

Summary

New methods are presented for analyzing repeated binary health measurements of individuals exposed to varying levels of air pollution. The methods involve a separate logistic regression of response against environmental covariates for each individual under study. Parameters reflecting individual susceptibility to pollutants and weather are estimated using the regression techniques developed by Cox (1970, 1972a). The parameters are then combined over the individuals in the study to obtain summary estimates of environmental effects. The approach does not require the independence of successive health measurements. It is illustrated with data on asthma and air pollution in the Los Angeles area. A special case of the approach can be adapted to continuous health measurements; the details of this procedure are outlined in an Appendix.

Key Words: Repeated binary responses, multiple logistic regression, Markov chain.

1. Introduction

1.1. Background

Panel studies, involving repeated observations of susceptible population subgroups, are frequently used to investigate the health effects of air pollution. Studies have been published of the influence of weather and daily levels of air pollution on bronchitis, cardiopulmonary symptoms, asthma attacks, irritation symptoms, and pulmonary function (see Stebbings and Hayes (1976) for related references). It is generally agreed (Hasselblad (1978), US Congressional Report (1976), Stebbings and Hayes (1976)), that the methodology for collecting and analyzing the data from these studies has not been adequately developed. The purpose of this paper is to describe a method for analyzing the relationship between a binary health response and air pollution which avoids many of the problems of existing approaches.

In order to illustrate the nature of a binary panel response study, we will describe the asthma studies conducted by the Environmental Protection Agency's Community Health and Environmental Surveillance System (EPA-CHESS). Panels of asthmatics are selected from each of several communities in a metropolitan area by consultations with physicians. Those panelists who meet certain screening requirements are asked for their age, sex, race, smoking status, educational level, as well as for characteristics associated with their asthma. They are then asked to record each day in a diary the presence or absence of an asthma attack and to return the diary at the end of each week for approximately eight months. Daily measurements of several air pollutants, and of

meteorological variables such as temperature and humidity, are obtained simultaneously for each community. Thus both the binary response of an individual (dependent variable) and the aerometric and meteorological covariates (independent variables) are time series.

1.2. Motivation

The statistical method most commonly used to analyze such panel study data is linear regression of the daily panel event rate against aerometric and meteorological measurements, and against other covariates thought to influence response. There are several problems with this approach. First, it requires that the event rates on successive days be statistically independent with constant variability about the expected value. It is evident (Roth, Viren, and Colucci (1977), Hasselblad (1978), US Congressional Report (1976)) that these two assumptions are not met. Hasselblad (1978) has shown, for example, that asthmatics are significantly more likely to have an attack on days following an attack day than on days after an attack-free day. Moreover, the variance of the panel event rate for a given day depends upon the true event probabilities for the panelists who report on that day.

A third problem with the usual linear regression approach is due to missing response data. Association or lack of association between a panel event rate and pollutant levels may be caused by "drop-ins" and "drop-outs". As an extreme example, on one day the panel event rate might be low because those reporting are relatively health individuals with low overall event rates. Another day might see a less healthy group of individuals reporting, and thus have a high event rate. This

situation can lead either to spurious association or loss of power, depending on the pollutant levels for the two types of days.

Yet another problem with linear regression of the panel event rate against environmental covariates is the constraints placed upon the regression parameters by the requirement that the rate lie between zero and one. The undesirability of representing proportions as linear functions is discussed by Cox (1970). Although some nonlinear (e.g., logistic or probit) regression models avoid this problem, it is not clear how these models could arise from any assumption about the functional form of the individual response probabilities. Thus no plausible assumption about the individual probabilities allows a meaningful interpretation of the parameters in such a nonlinear regression of the panel event rate.

1.3. Description of Proposed Analysis

We shall now describe a method of analysis that avoids each of the above problems. An essential feature of the method is the separate

regression of response data for each individual. Accordingly, each panelist is assigned a vector of regression coefficients which represents his or her susceptibility to the various environmental factors included in the regression. These vectors are then averaged over the panelists to obtain summary measures of the effects of the covariates upon the panel, each individual's vector being weighted according to the amount of data he has contributed. In the next section we present the details of this method and illustrate them with an application to CHES asthma data. To facilitate comparison with such data, we shall formulate the model in discrete time units of days, and not consider the analogous continuous time version.

2. Statistical Analysis

2.1. The Model

We regard the diary of a panelist as a set of observations on a discrete time stochastic process Z_t , $t = 1, 2, \dots, n$. The process has two states, $Z_t = 0, 1$ corresponding to the absence or presence of an event during day t . Let $p(t)$ denote the probability of an event on day t , given the environmental covariates from the start of the study until day t , and given the history of the process Z_t from the start of the study until day $t - 1$.

We assume for $p(t)$ the logistic form

$$p(t) = \exp(\alpha + \beta'x_t) / [1 + \exp(\alpha + \beta'x_t)] \quad (1)$$

Here $\tilde{x}_t$ is a vector of pollution levels, temperature, and other covariates and their interactions associated with day t , and $\tilde{\beta}$ is a vector of unknown regression coefficients. The function α , which depends upon one or more unknown parameters, may also depend on both t and the previous history of the process. The parameters in the model may vary from person to person. On the other hand, for many panel study applications, the covariates $\tilde{x}_t$ will vary only from day to day, and not between panelists. The model (1) and its special cases arise as consequences of a detailed theory for the role of pollutants in the aggravation of asthma (Whittemore and Keller (1979)).

The likelihood of a panelist's data $z_1, \dots, z_n$, conditional on the covariates $\tilde{x}_t$, $t = 1, \dots, n$, is

$$L = \prod_{t=1}^n p(t)^{z_t} [1 - p(t)]^{1-z_t}, \quad (2)$$

where $p(t)$ is given by (1). Our plan is to i) specify the function α ; ii) estimate the parameters in this function and the regression vector $\tilde{\beta}$ for each panelist by maximizing his likelihood (2); iii) combine the estimates of $\tilde{\beta}$ over the individuals in the panel.

We now consider some special cases of (1), corresponding to various assumptions about the function α . These special cases, and their interpretation for asthma, are also discussed by Whittemore and Keller (1979).

i) The simplest case occurs when $\alpha \equiv \text{constant}$. Then the state of the process on any day is independent of its state on any other day.

Hence this formulation is relevant only to those panel studies for which independence of successive responses can be assumed. The likelihood (2) can be written

$$L(\alpha, \beta) = \exp(\alpha z + \beta' \underline{x}) / \prod_{t=1}^n [1 + \exp(\alpha + \beta' \underline{x}_t)] , \quad (3)$$

where $z = \sum z_t$, $\underline{x} = \sum z_t \underline{x}_t$. Maximum likelihood estimates of α and β , and the asymptotic covariance matrix of these estimates, are given by Cox (1970) for L of the form (3). Because responses are independent in this version, days with missing data can simply be deleted from the study. Then n is the number of days without missing data, and equation (3) refers to the relabeled sequences $\{z_t\}$, $\{\underline{x}_t\}$.

An alternate method for estimating β that eliminates the nuisance parameter α is obtained by maximizing the likelihood conditional on z , the total number of events. This likelihood is

$$L_c(\beta) = \exp(\beta' \underline{x}) / \sum_A \exp(\beta' \sum_{t \in A} \underline{x}_t) , \quad (4)$$

the sum in the denominator being taken over all sets A of z days in the period. The use of (4) may be limited by the computational expense incurred in evaluating the denominator when the number of sets A is large.

ii) We next suppose that α assumes two values, α_0 or $\alpha_0 + \alpha_1$, depending on whether $z_{t-1} = 0$ or 1. This dependence can be expressed succinctly as $\alpha = \alpha_0 + z_{t-1} \alpha_1$. In this case the process Z_t is a two-state Markov chain, modulated by the covariates $\underline{x}_t$. This is the simplest example of (1) that yields serially correlated response.

In order to describe the likelihood of the data, we must define the distribution of Z_1 . This can be done by formally setting Z_0 equal to Z_n , making the time sequence circular. Then the likelihood (2) is

$$L(\alpha_0, \alpha_1, \beta) = \exp(\alpha_0 z + \alpha_1 y + \beta' \underline{x}) / \prod_{t=1}^n [1 + \exp(\alpha_0 + \alpha_1 z_{t-1} + \beta' \underline{x}_t)] , \quad (5)$$

where z , $\underline{x}$ are as in (3), and where $y = \sum z_t z_{t-1}$ is the number of pairs of 1's in the circular sequence. For each individual, maximum likelihood estimates of α_0 , α_1 and β can again be found in Cox (1970).

However, unlike case (i), the second derivative matrix of the negative of the logarithm of $L(\alpha_0, \alpha_1, \beta)$ depends upon the data through Z_t . The expected value of this matrix evaluated at $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\beta}$, i.e., the expected Fisher information, is difficult to compute. More simply, there is justification (Efron and Hinkley (1978)) for using the "observed" Fisher information, i.e., minus the second derivative of the log likelihood evaluated at $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\beta}$, to obtain the asymptotic variance-covariance matrix of the estimates.

This case also differs from the previous one in that the likelihood (5) contains no factor, analogous to (4), which is independent of α_0 and α_1 and which can be regarded as a conditional likelihood for the data. Thus the maximization of a conditional likelihood cannot be used to estimate β alone.

Days with missing data can again be deleted from the analysis. When the missing information for day $t - 1$ is the response rather than a covariate level however, the next day's event probability $p(t)$ cannot

be specified. If the number of days t for which z_{t-1} is missing is small, then these days can also be omitted from consideration. If the number of such days is not small, it may be desirable to assign a value to the missing z_{t-1} solely in order to use nonmissing data on day t . For example, z_{t-1} might be equated to the individual's overall event rate, i.e., to the average value of z_t taken over all of the days he reported. The effect of this substitution on estimates of β is likely to be small.

iii) As a third case, let $\alpha = \alpha[k(t)]$, where $k(t) \geq 1$ is the backward recurrence time to the previous event. Thus $k(t) = s$ if the most recent event occurred on day $t - s$; we take $k(t) = t$ if no event occurred prior to t . With this assumption, the process Z_t is a discrete time version of the modulated renewal process defined by Cox (1972a). The intervals between successive events are the interarrival times of the renewal process. The expression (1) represents the hazard function for an interarrival time ending on day t .

Estimation of β for continuous time modulated renewal processes has been discussed by Cox (1972a). He argues that no information about β can be provided except from the observed times at which events occur. The function α is thus eliminated from the inference process for β by arguing conditionally on the set of observed interarrival times. Accordingly, for each individual a "partial likelihood" is obtained by conditioning appropriately on the order statistics of his interarrival times. The method compares a score $\exp(\beta'x_t)$ for each event day t with the scores for all days having the same backward recurrence time to a preceding event. The reader is referred to Cox (1972a) for details.

The treatment of computational difficulties caused by ties among a panelist's interarrival times has been discussed by several authors (see for example, Cox (1972b)). A more serious problem with this example is the existence of days with missing data. A missing observation is a right censoring of the preceding interarrival time, and a left censoring of the following one. While right censoring is easily accommodated by Cox's method, left censoring is not. Several ad hoc solutions can be used; however, their effect on the resulting inferences about β is unknown. In addition, even with complete data, there has not been adequate study of the statistical properties of estimates obtained from Cox's partial likelihood as applied to time series.

For some data it may be reasonable to assume that the vector of parameters β does not vary from person to person, interpanelist variation occurring only among the parameters involved in α . In this case there is an alternative to the separate maximization of a likelihood for each person, namely maximization of the product of the individual likelihoods with respect to a single β . This alternative is most feasible

when the function α has been eliminated from the individual likelihoods, as in (4) of case (i) or in Cox's partial likelihood as applied to case (iii).

When separate estimates of β are obtained for each panelist, it may be desirable to combine these estimates over the panel in order to obtain summary measures of association between the covariates and the response. We will now describe two methods for doing this.

2.2. Combining Regression Coefficients Over the Individuals in a Panel

Using one of the procedures outlined in Section 2.1, we estimate a vector of regression coefficients corresponding to the covariates for each of the I panelists. For simplicity we assume first that there is one covariate. Thus the estimated regression coefficient $\hat{\beta}^{(i)}$ for the i th panelist is asymptotically normally distributed with mean $\beta^{(i)}$, the true coefficient for that panelist, and variance σ_i^2 . That is $\hat{\beta}^{(i)} \sim N(\beta^{(i)}, \sigma_i^2)$, where the $\hat{\beta}^{(i)}$ are independent for $i = 1, 2, \dots, I$. If the $\beta^{(i)}$ themselves are considered a random sample from an underlying population distribution, then the distribution of $\hat{\beta}^{(i)}$ is considered conditional on $\beta = \beta^{(i)}$ for the i th panelist (random effects model). The situation in which the $\beta^{(i)}$ are all considered to be equal will be called the fixed effects model.

In the fixed effects model, we wish to estimate the single β that is the true regression coefficient of all of the panelists. Assuming that normality holds for the given sample estimates $\hat{\beta}^{(i)}$, and that the σ_i^2 are known, the maximum likelihood estimate $\hat{\beta}$ of β is given by a weighted average of the $\hat{\beta}^{(i)}$: $\hat{\beta} = \sum w_i \hat{\beta}^{(i)}$. The weights are inversely

proportional to the variances and sum to one, i.e., $w_i = (\sigma_i^2)^{-1} \cdot (\sum (\sigma_j^2)^{-1})^{-1}$. The estimate β is unbiased, and its variance is given by $\text{Var}(\hat{\beta}) = (\sum (\sigma_j^2)^{-1})^{-1}$. This variance is the same as the asymptotic variance of the maximum likelihood estimate of β based upon the combined likelihood of the data from all of the panelists, as described at the end of Section 2.1. In practice, estimates of the variances σ_i^2 would be used to compute the weights. The fixed effects model is appropriate for testing the null hypothesis that the covariate is not associated with response for any of the panelists, i.e., that all of the $\beta^{(i)}$ are equal to 0.

For the random effects model, we assume that the true coefficients $\beta^{(i)}$ of the panelists are a random sample from a normal distribution with mean β and variance η^2 . The mean β represents the overall population association of the covariate with the response; the variance η^2 represents the variability of β from panelist to panelist. The unconditional distribution of the $\hat{\beta}^{(i)}$ is $\hat{\beta}^{(i)} \sim N(\beta, \sigma_i^2 + \eta^2)$, where the $\hat{\beta}^{(i)}$ are independent for $i = 1, 2, \dots, I$. Treating the σ_i^2 as known, the maximum likelihood estimates $\hat{\beta}$ and $\hat{\eta}^2$ of β and η^2 are given by the solution of the maximum likelihood equations:

$$\sum (\hat{\beta}^{(i)} - \hat{\beta}) / (\sigma_i^2 + \hat{\eta}^2) = 0 ,$$

$$\sum (\hat{\beta}^{(i)} - \hat{\beta})^2 (\sigma_i^2 + \hat{\eta}^2)^{-2} - \sum (\sigma_i^2 + \hat{\eta}^2)^{-1} = 0 ,$$

and $\hat{\eta}^2 \geq 0$. Thus $\hat{\beta}$ is again a weighted average of the $\hat{\beta}^{(i)}$:

$\hat{\beta} = \sum w_i \hat{\beta}^{(i)}$, where the weights are inversely proportional to the estimated variances: $w_i = (\sigma_i^2 + \hat{\eta}^2)^{-1} [\sum (\sigma_j^2 + \hat{\eta}^2)^{-1}]^{-1}$.

Comparison of the form and distribution of $\hat{\beta}$ in the random effects model with their counterparts in the fixed effects model shows that they differ only in the additional variance component which is due to variation between panelists. The estimate $\hat{\eta}^2$ can be calculated by solving numerically the maximum likelihood equations. By inverting the Fisher information matrix, we obtain the variances of $\hat{\beta}$ and $\hat{\eta}^2$: $\text{Var}(\hat{\beta}) = (\Sigma(\sigma_j^2 + \hat{\eta}^2)^{-1})^{-1}$, $\text{Var}(\hat{\eta}^2) = (\frac{1}{2} \Sigma(\sigma_j^2 + \hat{\eta}^2)^{-2})^{-1}$, and $\text{Cov}(\hat{\beta}, \hat{\eta}^2) = 0$. The random effects model is appropriate for testing the null hypothesis that the average effect of the covariate on the response of all of the panelists is zero. This model has the following advantage: if the variances σ_i^2 are underestimated by a constant amount, then the additional variability will be incorporated into $\hat{\eta}^2$, yielding an asymptotically unbiased estimate of the variance of $\hat{\beta}$.

If there is more than one covariate, then a vector of regression coefficients $\hat{\beta}^{(i)}$ is estimated for each panelist. The above techniques can be applied to each coordinate of the vectors $\hat{\beta}^{(i)}$ separately. Alternatively, for the fixed effects model, one can estimate β with a weighted average of the $\hat{\beta}^{(i)}$, where now each weight is a matrix. The weights are chosen to be proportional to the inverses of the estimated covariance matrices, and to sum to an identity matrix. In the random effects model, a covariance matrix T representing the underlying covariance of the β 's in the population is estimated by the solution of the appropriate maximum likelihood equations.

2.3. Example

Data from the EPA-CHESS Los Angeles asthma studies were analyzed using the Markov case (ii) of Section 2.1. While a full analysis of these data will be reported elsewhere, we illustrate this method by showing in Table 1 the results for a single panel of asthmatics in Garden Grove, California during the period November 17, 1974 to June 29, 1975.

For each of 41 panelists, daily asthma attack response was regressed against the following covariates: the constant 1, the indicator z_{t-1} for attack response on the previous day, daily average level of total suspended particulates (TSP), minimum daily temperature, humidity, "day of study", and the six weekdays Monday through Saturday. The covariate for day of study is an indicator assuming the value one during the first 46 days of the study and zero elsewhere. The six weekday covariates are indicators for the days Monday through Saturday, with all of these covariates assuming the level zero on Sunday. Only the regression coefficient corresponding to Saturday is shown in Table 1. The covariate for day of study was included to adjust for an increase in attacks at the start of the study which was unaccompanied by a similar change in pollutants or weather. This increase, noted in several panels which were observed over the same period in different years, may reflect a seasonal effect due to pollen levels. The weekday covariates were included to adjust for the possibility of a systematic weekly pattern of asthma propensity, such as increased attacks on Mondays or weekends.

Table 1

Average Regression Coefficients and Variance Components $\hat{\eta}^2$
 for
 EPA-CHESS Los Angeles Asthma Study, Garden Grove, 1974-75 (41 panelists)

	Markov covariate	total suspended particulates ($\cdot 10^3$)	minimum daily temperature ($\cdot 10^3$)	humidity ($\cdot 10^3$)	day of study	day of week
regression coefficient (using fixed effects model)	1.968	1.58	-8.91	-3.68	0.178	0.556
standard error	0.076	0.83	5.96	2.90	0.095	0.134
$\hat{\eta}$	0.296	0.00	18.57	0.00	0.338	0.159
regression coefficient (using random effects model)	2.131	1.58	-9.06	-3.68	0.176	0.577
standard error	0.093	0.83	6.79	2.90	0.099	0.147

The first two rows of Table 1 give the estimated fixed effects parameters and their standard errors as specified by the equations for the fixed effects model. The third row is the square root of $\hat{\eta}^2$, the estimated between-panelists variance component of the random effects model. The last two rows of the table correspond to the random effects parameters.

The highly significant values of the estimated Markov parameter $\hat{\alpha}_1$ for both fixed and random effects models provide strong evidence against the assumption in case (i) that successive responses are independent. Indeed, $\hat{\alpha}_1$ was nonzero at a level of significance less than .05 for all but 5 of the 41 panelists in the analysis. This indicates the existence of autocorrelation in the response variable that cannot be attributed to autocorrelation of the covariates.

The positive values of $\hat{\eta}$ for the Markov parameter $\hat{\alpha}_1$ and the coefficients for temperature, day of study and day of week indicate that the interpanelist variability of these estimates cannot be explained by their Fisher information variances. For the Markov parameter $\hat{\alpha}_1$, this is consistent with the large differences in attack frequency and pattern known to exist among asthmatics. By contrast, the estimated coefficients for TSP and humidity do not vary appreciably among the panelists relative to their variability according to the Fisher information. In this case the fixed effects and random effects estimates coincide.

While these data show a marginally significant increase in panel attack propensity due to increased levels of TSP, the largest effects are those due to the Markov covariate and to the day of week (i.e.,

Saturday vs. Sunday). This suggests that small pollutant effects, as well as interactions among pollutants and weather variables, cannot be detected without altering the study design.

In hopes of detecting susceptible subgroups of the asthmatic population, the regression coefficients for TSP, minimum temperature, humidity, day of study and Saturday were each classified according to sex, age (under or over age 16), and the reported presence or absence of hay fever, increased attacks during cold weather, and increased attacks during periods of emotional stress. Using the random effects model, summary estimates of the above coefficients were obtained for each level of these dichotomous variables. It was found that panelists who reported suffering from hay fever had a significantly lower coefficient corresponding to TSP, a significantly higher coefficient corresponding to Saturday, and were affected significantly more by colder temperatures than those panelists without hay fever. Panelists who reported increased attacks in cold weather were affected significantly more by colder temperatures and significantly less by drier weather than those who did not. No other significant differences were found among the above comparisons. A more detailed analysis would obtain summary estimates of the regression coefficients for the population stratified on these dichotomous variables simultaneously.

3. Discussion

Each of the methods described in Section 2.1 involves the separate regression of the response of individual panelists against environmental covariates. The approach has a number of attractive features. First, only the simplest of the four cases assumes that successive responses are uncorrelated, an assumption regarded as untenable for many health panel studies. Second, each person's periods of morbidity are compared only with his own periods in good health. Unlike methods which regress daily panel averages against aerometric measurements, this avoids possible bias due to the missing data of sickly or healthy individuals. Third, estimation of a regression vector for each individual enables one to assess the relationship between characteristics such as age, sex or smoking status and particular covariates of interest. Thus subgroups of the study population which are especially susceptible to a given environmental factor can be detected via their high regression coefficients.

When cases (i) and (ii) are appropriate, visual tests of model fit can be obtained by plotting standardized residuals versus time, versus deciles of response probability, or versus levels of particular covariates of interest. However, we know of no formal tests of goodness-of-fit that are applicable to these two cases. This is an area in need of further work.

Finally, all of the methods permit the study of interactive effects among two or more of the covariates. Such effects can be investigated by including joint non-linear functions of the relevant covariates in the regression.

Acknowledgments

This work was supported in part by grants to the SIAM Institute for Mathematics and Society from Department of Energy, Rockefeller Foundation, Sloan Foundation, Environmental Protection Agency, and National Science Foundation. The second author was supported by a Rockefeller Foundation Fellowship in Environmental Affairs.

The authors wish to thank Donald Pierce for helpful discussions related to this work.

References

- Anderson, T. W. (1971). The Statistical Analysis of Time Series. John Wiley and Sons, New York.
- Cox, D. R. (1970). Analysis of Binary Data. Methuen and Co. London, Ltd.
- Cox, D. R. (1972a). The statistical analyses of dependencies in point processes. In Stochastic Point Processes: Statistical Analysis, Theory and Applications. P. A. W. Lewis (ed.), John Wiley and Sons, New York.
- Cox, D. R. (1972b). Regression models and life tables. Journal of the Royal Statistical Society, Series B 34, 187-220.
- Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. Journal of the American Statistical Association, 72, 551-565.
- Efron, B. and Hinkley, D. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed vs. expected Fisher information. Biometrika, 65, 457-482.
- Hasselblad, V. (1978). Comparison of methods for the analysis of panel studies. Report of the United States Environmental Protection Agency, EPA-600/1-78-043, Research Triangle Park, North Carolina.
- Roth, H. D., Viren, J. R. and Colucci, A. V. (1977). Evaluation of CHESS: New York Asthma Data 1970-71, Vol. I. EPRI EA-50. Electric Power Research Institute, Palo Alto, California.

Stebbing, J. H. and Hayes, C. G. (1976). Panel studies of acute health effects of air pollution. Environmental Research, 11, 89-111.

Whittemore, A. S. and Keller, J. B. (1979). Asthma and air pollution: A quantitative theory. Proceedings of SIMS Conference on Energy and Health. N. E. Breslow and A. S. Whittemore (eds.), SIAM, Philadelphia, Pennsylvania.

United States Congressional Report: The Environmental Protection Agency's research program, with primary emphasis on CHESS: An investigative report. Prepared for the Committee on Science and Technology, United States House of Representatives (1976).

Appendix: Application to Continuous Response Variables

The Markov property used in Section 2.3 can be adapted for use with panel studies of continuous health measurements, such as impairment of lung function, severity of bronchitis, etc. Here we indicate briefly how classical time series methods can be applied to studies in which the measurement Z_t observed for an individual at time t is continuous. We assume that

$$Z_t = \alpha_0 + \alpha_1 Z_{t-1} + \beta' \tilde{x}_t + \epsilon_t, \quad t = 1, \dots, n,$$

where, as usual, $\tilde{x}_t$ is a p -dimensional vector of environmental covariates, $\alpha' = (\alpha_0, \alpha_1)$ and β are vectors of unknown parameters, and where the ϵ_t are independent, identically distributed normal variates $\epsilon_t \sim N(0, \sigma^2)$, $t = 1, \dots, n$. For convenience we set $Z_0 = Z_n$. The maximum likelihood estimates $\hat{\alpha}$, $\hat{\beta}$ can be shown (see for example Anderson (1971, Chapter 5)) to satisfy the normal equations

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} z \\ y \\ x \end{pmatrix},$$

where $z = \sum z_t$, $y = \sum z_t z_{t-1}$ and $x = \sum z_t \tilde{x}_t$, and where

$$A_{11} = \begin{pmatrix} n & z \\ z & \sum z_t^2 \end{pmatrix}, \quad A'_{12} = A_{21} = (\sum \tilde{x}_t \quad \sum z_{t-1} \tilde{x}_t),$$

and $A_{22} = \sum \tilde{x}_t \tilde{x}_t'$. The maximum likelihood estimate of σ^2 is $\hat{\sigma}^2 = \frac{1}{n} \sum [z_t - (\hat{\alpha}_0 + \hat{\alpha}_1 z_{t-1} + \hat{\beta}' \tilde{x}_t)]^2$. The estimates $\hat{\alpha}$, $\hat{\beta}$ are consistent and asymptotically normally distributed as n increases. The asymptotic covariance matrix for $\hat{\alpha}$ and $\hat{\beta}$ can be estimated by

$$\frac{n}{n-p-2} \hat{\sigma}^2 \begin{pmatrix} 1 & \frac{\hat{\alpha}_0}{1-\hat{\alpha}_1} & \frac{1}{n} \sum \tilde{x}_t' \\ \frac{\hat{\alpha}_0}{1-\hat{\alpha}_1} & \frac{1}{n} \sum \tilde{z}_t^2 & \frac{1}{n} \sum \tilde{z}_{t-1} \tilde{x}_t' \\ \frac{1}{n} \sum \tilde{x}_t & \frac{1}{n} \sum \tilde{z}_{t-1} \tilde{x}_t & \frac{1}{n} \sum \tilde{x}_t \tilde{x}_t' \end{pmatrix}^{-1}.$$

Missing variables can be treated as discussed in Section 2.1. As before, each individual has a vector $\hat{\beta}$ of estimated regression coefficients. The vectors of the individuals in a panel can be combined as discussed in Section 2.2.