ESTIMATING DOSE/RESPONSE RELATIONS FROM FILTERED TIME SERIES

S. MAZUMDAR, H. SCHIMMEL, AND I. HIGGINS

TECHNICAL REPORT NO. 78
JUNE 1985
ESTIMATING DOSE/RESPONSE RELATIONS
FROM FILTERED TIME SERIES

by

S. MAZUMDAR, H. SCHIMMEL, and I. HIGGINS

TECHNICAL REPORT NO. 78
JUNE 1985

STUDY ON STATISTICS AND ENVIRONMENTAL FACTORS IN HEALTH (SIMS)

PREPARED UNDER A GRANT TO SIMS FROM
ENVIRONMENTAL PROTECTION AGENCY (EPA)
SLOAN FOUNDATION
NATIONAL SCIENCE FOUNDATION (NSF)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ESTIMATING DOSE/RESPONSE RELATIONS FROM FILTERED TIME SERIES
S. Mazumdar, H. Schimmel, and I. Higgins*

The procedure described here was developed during a study attempting to relate the premature mortality response to ambient air pollution. Time series were available for daily mortality, average daily air pollution (SO₂, Smoke) and a number of daily weather variates for winters (i.e. November - February) beginning November 1, 1965 and for full year data beginning April 1, 1965 and ending December 31, 1976. The problem is relatively difficult because of the confounding effects of seasonal trends and weather variates and we as well as other investigators have obtained a variety of answers according to procedures employed in the analysis. Our own approach emphasized the following points:

1) The analysis is directed toward acute effects and slow rhythms must be removed from the time series so as to remove the confounding effects usually described as due to long term and seasonal trends;

2) Although a number of deseasonalizing procedures have been employed we have found most of these faulty as they include the effects of year to year differences in level and we consider filtering of all slow rhythms essential;

3) Since results of the pollution/mortality relation are usually interpreted in dose/response terms we consider it necessary that the analysis specifically address this relation; and finally

4) Because daily variations in weather are influential with respect to both pollution and mortality we consider it essential to correct for this influence in a thorough fashion.

In defining the new procedure presented here and illustrating its use for the London study we will proceed by stages, first considering the elementary problem of defining a relation between y, a dependent variable, and z, an independent or explanatory variable, given N

* This report was delivered by H. Schimmel at a SIMS Seminar, March 6, 1985, Stanford, California. Address correspondence to Dr. Herbert Schimmel, 25 Usonia Road, Pleasantville, NY 10570, Phone (914) 769-0199, see addendum of June 27, 1985 for limitations on quoting illustrative data.
samples $x_n, y_n$. The problem is usually formulated by means of the following equation:

$$y_n = \eta(x_n) + \epsilon_n$$  \hspace{1cm} (1)

where $\eta(x)$ expresses the functional dependence of $y$ on $x$ and $\epsilon_n$ is considered error or unexplained residual.

$\eta(x)$ is usually derived by defining its form and estimating parameters for that form according to some criterion and attendant procedure, such as LLSR or maximum likelihood.

I first encountered such a problem almost 50 years ago when a friend, a fellow graduate student in physics, had as his thesis project the diffusion of one metal into another using the newly available radioactive isotopes. The isotope metal was plated at the end of a metal rod which was heated to promote diffusion. After cooling the rod was shaved along its length and the radioactive counts gave a measure, $y$, of the amount of diffused metal while $x$ was the distance of each slice from the end of the rod. A theoretical formula was available for the function $y(x)$ but the experimental curves proved far more ragged than my friend and his thesis advisor expected and three years work was about to be abandoned on the grounds of faulty experimental procedure. Although physicists knew that radioactive emission was a random process, very few were accustomed to practical statistical analyses. When it was finally recognized that $y$ was a random variable the results were found consistent with both physical diffusion theory and the expected Poissonian variability. The desired diffusion constants were readily estimated and the thesis accepted.

In the more general problem of estimating $\eta(x)$ from (1) a functional form is frequently assumed for $\eta(x)$ either based on some hypothesis regarding the $x, y$ relation or for descriptive purposes using for example a polynomial in $x$. Many investigators explore the $\eta(x)$ relation initially by creating scatterplots. When the sample number $N$ is large I usually explore the $x, y$ relation by ordering the data according to $x$ and quantize it into $K$ quantiles and compute the means of these quantiles which we shall denote here by $\eta_k$. This provides for a visual inspection more focused on the $x, y$ relation when $N$ is large; also the results may be smoothed and even used to estimate $\eta(x)$ without recourse to the large full data base.

We should note that these $\eta_k$ values can be derived by LLSR of a stepwise function fitted
to the quantiles. Having noted this, we note also that the following equation

\[ y_n = \eta(x_n) + z(u_n) + \varepsilon_n \]  

(2)

may be used to derive a stepwise function \( \eta(x) \) where the estimates \( \eta_k \) are corrected for the influence of a covariate (or covariates) \( u_n \).

The fact that the \( \eta_k \) are frequently corrected to give a piecewise linear plot of \( \eta \) rather than plotted as a step function suggests a procedure where we directly estimate such a form of \( \eta \) based on the quantile boundaries. Let us designate by \( B_k \), \( 0 \leq k \leq K \), estimates of \( \eta(x) \) at the quantile boundaries; also having ordered the \( z \) by magnitude let the indices \( m, k \) correspond to \( n \) where \( x_n \) occupies the \( m \)th position in the \( k \)th quantile interval and \( 1 \leq m \leq M \) with \( MK = N \). With \( \eta(x) \) being piecewise linear we have

\[ \eta(x_n) \equiv \eta(x_{m,k}) = \sum_{k=0}^{K} \beta_{n,k} B_k \]  

(3)

where all the \( \beta \) coefficients are zero except those for the boundary values of the \( k \)th quantile, namely \( B_{k-1} \) and \( B_k \).

Since \( \eta(x_n) \equiv \eta(x_{m,k}) \) is a linear function between \( B_{k-1} \) and \( B_k \) the values of these two \( \beta \) coefficients are given by

\[ \beta_{n,k-1} = \frac{z_{m,k} - z_{M,k-1}}{z_{M-k} - z_{M,k-1}} \]
\[ \beta_{n,k} = \frac{z_{M,k} - z_{m,k}}{z_{M,k} - z_{M,k-1}} \]  

(4)

To estimate the \( B_k \) we introduce the estimates \( \eta(x_n) \) into (1) or (2) as the case may be. Using LLSR we identify the \( B_k \) as the parameters to be estimated and identify the \( \beta_{n,k} \) as the independent variables.

Figure 1 is based on a hypothetical example where (1) has been used to estimate the \( \eta_k \) and \( B_k \) which are estimates respectively for the quantile values for a step function and for a piecewise linear function connecting estimates for the quantile boundaries. In this illustration \( \varepsilon_n \) values are assumed to be small and the errors of estimate for both \( \eta_k \) and \( B_k \) have been multiplied by \( 2M \). Note that the errors of estimate of \( \eta_k \) are much larger than those of the \( B_k \) reflecting the trend within quantiles.
Filtering. In the problem which stimulated the procedure described here it was necessary to filter the time series to remove the slow components. This was done by taking residuals about a 15 day centered moving average. It is one of a class of weighted moving average linear filters in the time domain. If we use the expression for \( \eta(z_n) \) in (3) for the ordered time series and apply a moving average we note that \( \eta'(z_n) \) where the accent denotes the filtered function, we get

\[
\eta'(z_n) = \sum_{0}^{k} \beta'_{n,k} B_k
\]  

(3')

where the \( \beta' \) represents the new coefficients arising from the moving average operation. This \( \eta' \) may be introduced into modified (1) or (2), to wit

\[
y'(z_n) = \eta'(z_n) + \epsilon'_n
\]  

(1')

\[
y'(z_n) = \eta'(z_n) + z'(u_n) + \epsilon'_n
\]  

(2')

The \( B_k \) are then estimated by LLISR with the confounding influence of long term trends and seasonal movements having been removed.

Illustrations from Exploratory Studies
of the London Data Set

We have carried out numerous exploratory studies on this data set. In the course of these we refined our procedures for including the dose/response function in the analysis to the following statement:

\[
M_n/E(M_n) + [1 + R(P_n)] + Z_n + \epsilon_n
\]  

(5)

where \( M_n \) represents mortality on the \( n \)th day, \( E(M_n) \) – an expected value of mortality, \( P_n \) – pollution level, \( Z_n \) a term representing confounding covariates which may and usually does include unidentified slowly varying influential factors and \( \epsilon_n \), a residual. \( R \) represents the mortality response.

For estimation purposes, and having observed that the only rapidly varying components
are acute effects of weather (5) has been progressively simplified for estimation purposes to

\[ M_n/E(M_n) = R'(P_n) + Z'_n + \epsilon_n \]

or taking logs

\[ \log' M_n = \log'(1 + R_n) + \epsilon' \]

\[ \cong R'_n + \epsilon'_n \]  \hspace{1cm} (5')

In the last equation, (5'), it is assumed that the influences of all slowly varying factors have been included in \( E(M_n) \) and that \( \log' E(M_n) \cong 0 \). Also because for most samples \( R_n \) is found to be smaller than \( \pm 0.1 \) it is assumed that \( \log(1 + R_n) \cong R_n \). Note \( \log' M_n \) and \( R'_n \) still have to be corrected for weather and DOW, i.e. day-of-week, which remain influential variables.

Our initial exploratory analyses showed weather variables influential as well as DOW. Using linear and quadratic expressions in \( P_n \) to express \( R_n \), we observed

1) In a partition of effects as between \( SO_2 \) and Smoke measure, the pollution effect, if present, was associated primarily with Smoke; and

2) As Smoke levels declined regression coefficients increased in the polynomial models. This reflected in effect a constant \( \bar{R}_n \), i.e. mean response. Our first explorations showed that after 1958–61 winters when Smoke averages 400 \( \mu g/m^3 \) we were obtaining \( \bar{R} \) values which were the same for the winters of 1961–64 when \( P \cong 200 \ \mu g/m^3 \) as for the summers of 1974–76 when \( P \cong 20 \ \mu g/m^3 \).

The procedure using \( B_k \) for the dose/response function described here was developed some 3 years ago but we were reluctant to apply it in the presence of the statistical artifact of constant \( \bar{R}_n \) in the presence of falling \( P \) over the years. We described this as surrogate behavior.

We accordingly concentrated on refining the weather corrections. The use of a complex, nonlinear, weather correction which included lag effects up to one week resulted in an \( \bar{R}_n \) which was effectively reduced to zero in the summer and was not significant during spring and fall. Surrogate behavior persisted however after 1962 during the winter months.

Recently we have examined the acute Mortality-Temperature relation expressing \( R_n \) in the form of a piecewise linear function defined by \( B_k \). The results are shown in Figure 2 which
are free hand drawings based on plots of the form shown in Figures 3 and 4 for February and August. Note Temperature is in °C and includes corrections for windchill in the winter and humidity and wind in the summer as tabulated by NOAA under the designation apparent temperature. Note Smoke and DOW have been included as covariates in the analyses on which Figures 2, 3, and 4 are based.

As a test we have also used the new procedure to express the Pollution/Mortality response based on filtered, weather corrected variables. The data show results for the first 9 winters and the last 9 winters separately. The levels in the first were approximately 3 times as high as in the second, yet response levels are about the same and \( \bar{R}_n \) would be about the same except for the presence of the very high pollution levels of the earliest winters. Thus as we expected this procedure brings out the surrogate behavior observed earlier with polynomial models.

Concluding Remarks.

On heuristic grounds we consider \( M \equiv 16 \) should be the minimum number of samples per quantile. If \( N \) is sufficiently large we can simultaneously estimate the \( B_k \) for same day and lag functions. Figure 2 suggests that we can pool 12 month data, since the filtered functions show approximately the same behavior over a connected total range. With 12 full years of data we could develop model free corrections of filtered mortality for weather using 5 sets of \( B_k \). For each set with \( N = 720 \) we could use \( M = 18, K = 40 \) and after initial analysis develop smoothed functions.

In the light of the results obtained in Figure 2 it would be useful to make a similar analysis for the relation of pollution or rather log pollution to temperature by month. If it is found that the data can be pooled for a full year than pollution like mortality can be corrected for its relation to temperature taking into account lags. This would seem to be a reasonable last ditch effort for removing the surrogate behavior of pollution in its relation to mortality. Previous explorations suggest it will fail but results should be as interesting as those shown in Figure 2.
FIGURE 1
Estimates of Quantile Means $\hat{\eta}_k - x$ and Estimates of Quantile Boundary Values $B_k - o$. Standard error of estimate $\hat{\tau}$ is multiplied by $2M$. (See text)
FIGURE 2. Relation of Mortality to Temperature - See Text
DOSE RESPONSE controlling for DOW, LP2
Data: Crude, Model: Piecewise Linear, Time Period: February 1966–76

FIGURE 3a
DOSE RESPONSE controlling for DOW, LP2
Data: Filtered, Model: Piecewise Linear, Time Period: February 1966-76

FIGURE 3b
DOSE RESPONSE controlling for DOW, LP2
Data: Filtered, Model: Piecewise Linear Lag 1, Period: February 1966–76

FIGURE 3c
DOSE RESPONSE controlling for DOW, LP2
Data: Filtered, Model: Piecewise Linear Lag 2, Period: February 1966-76

FIGURE 3d
DOSE RESPONSE controlling for DOW, LP2
Data: Crude, Model: Piecewise Linear, Time Period: August 1965–76

FIGURE 4a
DOSE RESPONSE controlling for DOW, LP2
Data: Filtered, Model: Piecewise Linear, Time Period: August 1965–76

FIGURE 4b
DOSE RESPONSE controlling for DOW, LP2
Data: Filtered, Model: Piecewise Linear Lag 1, Period: August 1965–76

FIGURE 4c
DOSE RESPONSE controlling for DOW, LP2
Data: Filtered, Model: Piecewise Linear Lag 2, Period: August 1965-76

FIGURE 4d
DOSE RESPONSE
Data: Corrected, Model: Piecewise Linear, Time Period: Nov-Feb 1967-76

Figure 5a
Ln of M1
ACP2 Quantile Limits
Following my SIMS talk I reviewed it with Dr. Mazumdar and also discussed our plans for further analyses. At first I was tempted to modify the "Concluding Remarks" of the talk as a result of our discussions. However, since we are dealing with work in progress I concluded it would be most useful if the following observations were added.

1. The exploratory material presented in Figures 2 through 5 may not be valid because: first the computations and plots have not been double checked; and second my conclusions based on visual inspection may not be valid because of variability in the results. Note this exploratory material was introduced to illustrate the procedure and was not intended to represent definitive results.

2. In my "Concluding Remarks" after inspection of Figure 2 I suggest pooling the monthly temperature/mortality results illustrated by Figures 3 and 4. After an analysis with pooled data statistical tests should be employed to check whether the monthly dta results are consistent with the pooled data results.

3. In my concluding remarks I state that if the procedure is employed to include the lag temperature effects then DK values must be estimated. This is consistent with our published work where D temperature values were used and no interaction was provided for. In the more general case where interaction is taken into account then $K^D$ values will have to be estimated. In our earlier analyses of the joint effects of S0 and Smoke on mortality we allowed for such interaction. See reference given in point #4 below. In applying the procedure to the two dimensional case we will be estimating $K^2$ values $B(k_1, k_2)$. If we wish to use rectangular coordinates, we would include a cross product term in the interpolation in addition to lienra terms in variables 1, 2. The equivalent of Equation (3) page 3 would remain linear in the $B$ values and could be filtered as in the one dimensional case.

4. It was our intention in using the procedures to obtain a smooth curve by applying a moving average. Figures 2-5 have been so smoothed. The variability in Figure 5 shows that we will lose a substantial part of the lower and upper ends of the range which are
of greatest interest if we average a sufficient number of points to obtain a moderately smooth curve.

The biological nature of the pollution/mortality response relation suggests that we should constrain the estimates so that \( B_k \geq B_{k-1} \). Barlow, Bartholomew, Bremmer and Brunk in the Wiley Series in Probability and Mathematical Statistics deal with this question and we are arranging for a computer program which has been developed to handle this constraint. Depending on the structure of the program obtained we will apply it to the full data base or to the \( B \) values set forth in Equations (3) and (3'). As to the more general case of regression on \( D \) independent variables our preliminary exploration has not revealed any computer programs available except for the cases \( D \) equal 1 or 2.

5. Finally, although the material presented here derives from the joint work of the authors who appear on the title page, I alone am responsible for any errors of fact or interpretation. For earlier published results see our article in Archives for Environmental Health, July/August 1982 [Vol. 37 (No. 4), pp. 213-220].