EVALUATING THE EFFECTS OF OBSERVED AND UNOBSERVED DIFFUSION PROCESSES IN SURVIVAL ANALYSIS OF LONGITUDINAL DATA

Anatoli I. Yashin
Kenneth G. Manton

December 1984
CP-84-58

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

ABSTRACT

In analyses of human survival often explicit consideration of the dynamics of physiological processes governing survival is not made. Failure to consider the influence of such processes can lead to incorrect inferences about the operation of such processes and the inability to forecast future changes in survival. An explicit model of such processes has been presented by Woodbury and Manton (1977). Myers (1981) developed another approach based on the appropriate extension of the Cameron-Martin method. We show that estimation can be conducted using a conditional Gaussian strategy and that the conditional Gaussian approach offers several substantive and computational advantages.

THE AUTHORS

Dr. Anatoli Yashin is Senior Researcher at the Institute for Control Sciences, Academy of Sciences, Moscow, USSR. He is currently research scholar at IIASA with the Population Program.

Professor Kenneth Manton is Assistant Director of the Center for Demographic Studies at Duke University, Durham, North Carolina, USA. He has had frequent association with the Population Program at IIASA.

I. INTRODUCTION

In the analysis of mortality and morbidity in demographic and biostatistical studies the explicit relation of the realized event rate, say $\bar{\mu}(t)$, to the parameters of the underlying physiological process generating the health events, is often not considered. This can lead to a lack of precision in attempting to use the observed data to make forecasts about health changes or inaccurate statements about the effect on disease risk of altering risk factors in some specified way. In this paper we consider an approach for modeling the relation of the observed rates to the parameters of the underlying process both in the presence of auxilary information on the cross-temporal change of individual physiological values and when such information is not available. The procedure relies upon the fact that there will be information available, either from other studies or from theoretical models, that can be used to specify a reasonable structure for the process. Generally, for analyses of chronic disease such data will be available from prior epidemiological and clinical studies.

The basic approach utilized in this paper is derived from a multivariate Gaussian diffusion process model of human physiological change and mortality (Woodbury and Manton, 1977, 1983; Yashin et al., 1985). From that model we can establish the mathematical relation between observed mortality and the parameters of the processes governing change in the means and covariances of the physiological variables related to the risk of mortality. These relationships can be used to develop the likelihood function for the estimation of process parameters from the distribution of the observed time to death (failure) of persons in the population. The likelihood function can be generalized to deal with providing estimates conditional upon the realized values of an observed process measured at fixed times.

The conditional Gaussian approach to estimation presented in this paper can be contrasted with the more usual Cameron-Martin (1948) approach (and its extensions, Myers (1981) for determining the parameters of a stochastic process.

We will examine the advantage of the conditional Gaussian approach over Cameron-Martin based strategies.

II. PRELIMINARY SPECIFICATIONS

Suppose that the mortality rate for individual i in a population of I individuals depends on some ℓ dimensional process Z(t) which evolves over time. We will assume that the process for each individual evolves independently from that of all other individuals. We will also assume that the mortality rate is a quadratic function of the set of values Z(t), or

$$\mu(t, Z(t)) = Z'(t)\, Q(t)\, Z(t) + \mu_0(t) \tag{1}$$

where Q(t) is a non-negative definite symetric ℓ x ℓ matrix where $t \geq 0$. Assume that Z(t) satisfies the linear diffusion type, stochastic differential equation, defined for a probability space (Ω, H, P):

$$dZ(t) = (\alpha_0(t) + \alpha(t)\, Z(t))dt + b(t)d\, W_t \tag{2}$$

where $\alpha_0(t)$ is a ℓ-dimensional vector function of t with bounded elements for any $t \geq 0$; $\alpha(t)$ is a bounded ℓ x ℓ matrix for any $t \geq 0$; b(t) is a bounded ℓ x k matrix, and W_t is a k-dimensional Wiener process which does not depend on initial condition (Z(0)). The forms of (1) and (2) were selected on the basis that they have been found to be applicable in many biomedical applications (Manton and Woodbury, 1983, 1985). In general the form of (1) and (2) will be selected on the basis of prior relevant biostatistical studies or theoretical insights into the physiological processes of interest.

In purely demographic studies one often directly analyzes the realized age specific mortality rate $\bar{\mu}(t)$. More precise evaluation of the mortality process can be achieved by utilizing the relation between $\bar{\mu}(t)$ and $\mu(t,Z(t))$ for conditional Gaussian processes established in Woodbury and Manton (1977) (see further mathematical development in Yashin et al., 1984). This relationship was of the form

$$\bar{\mu}(t) = E(\mu(t, Z(t)) \,|\, T_i > t)$$

where T_i is the death time of the individual associated with the mortality rate

$\mu(t, Z(t))$.

Where the initial condition, $Z(0)$, has the multivariate normal distribution with the vector of means $m(0)$ and covariance matrix $\gamma(0)$, and $\mu(t, Z(t))$ is quadratic (see equation (1)), then $\bar{\mu}(t)$ is the age specific mortality rate among survivors to t with the following relation to the parameters of the distribution of Z:

$$\bar{\mu}(t) = m'(t)\ Q(t)\ m(t) + Tr(Q(t)\ \gamma(t)) + \mu_0(t) \tag{3}$$

where $m(t)$ and $\gamma(t)$ satisfy the nonlinear ordinary differential equations,

$$\frac{d\ m(t)}{dt} = \alpha_0 + \alpha(t)\ m(t) - 2\ m(t)\ Q(t)\ \gamma(t) \tag{4}$$

and

$$\frac{d\ \gamma(t)}{dt} = \alpha(t)\ \gamma(t) + \gamma(t)\alpha^*(t) + b(t)b^*(t) - 2\ \gamma(t)\ Q(t)\ \gamma(t) \tag{5}$$

with initial conditions $m(0)$ and $\gamma(0)$.

The relationships in (3), (4), and (5) have at least two important uses. First, they are convenient forms for including ancillary information from other studies or theoretical insights. For example, prior studies may indicate the functional form for $Q(t)$, i.e., the nature of the functional dependence of the population hazard rate on the means of the $Z(t)$ (e.g., Economos, 1982). For example, considerable evidence is available to suggest that the dependence of the $\bar{\mu}(t)$ on time, in the case of human adult mortality, could be Gompertz in form (Spiegelman, 1969). This would suggest that $Q(t)$ is a Gompertz function. Alternatively one may have information on the form of the process $(Z(t))$ influencing mortality. Clearly utilizing this information is very likely to increase the precision of forecasts of $\bar{\mu}(t)$ over naive procedures which ignore this information and simply project on the basis of the age pattern of $\bar{\mu}(t)$.

A second use of the form specified in (3) (and of the auxiliary equations in (4) and (5)) is to develop a likelihood estimation strategy to retrieve the parameters of the process from the trajectory of $\bar{\mu}(t)$. In the following we will consider how such strategies may be developed for two distinct observational plans. The first plan is for the continuous time monitoring of mortality, where mortality

is influenced by both observed and unobserved processes. Such a plan is seldom found in the type of longitudinal epidemiological studies in which we are most interested. Furthermore, since the continuous time formulation requires that we evaluate the parameters over the entire process from birth to death, it is computationally difficult to apply. As a consequence we present an approach for a second type of observational plan where measurements are made at fixed times. We will show that the same equations developed for the continuous time case for the conditional Gaussian model can be applied to the case of discrete time measurements if the correct initial conditions for the start of each interval are formulated.

III. A MODEL FOR ESTIMATING THE PARAMETERS OF A TWO-COMPONENT FAILURE PROCESS UNDER BOTH CONTINUOUS AND DISCRETE TIME OBSERVATIONAL PLANS

The first step in our development is to generalize the mortality process defined in equations (1) and (2) to the case where mortality is influenced by both an observed and unobserved process. Specifically, suppose that the duration of life for any individual in the cohort is a functional of the two component processes $Z(t) = X(t), Y(t)$. We may rewrite the quadratic form in (1) as

$$\mu(t, X(t), Y(t)) = (X'(t), Y'(t)) \begin{bmatrix} Q_{11}(t), & Q_{12}(t) \\ Q_{21}(t), & Q_{22}(t) \end{bmatrix} \begin{bmatrix} X(t) \\ Y(t) \end{bmatrix} + \mu_0(t) \qquad (6)$$

where $Q_{11}(t)$ and $Q_{22}(t)$ are positive-definite symetric matrices, and $Q'_{12}(t) = Q_{21}(t)$.

Furthermore, let us rewrite equation (2) as

$$d\begin{bmatrix} Y(t) \\ X(t) \end{bmatrix} = \left[\begin{bmatrix} \alpha_{01}(t) \\ \alpha_{02}(t) \end{bmatrix} + \begin{bmatrix} \alpha_{11}(t), & \alpha_{21}(t) \\ \alpha_{12}(t), & \alpha_{22}(t) \end{bmatrix} \begin{bmatrix} Y(t) \\ X(t) \end{bmatrix} \right] dt + \begin{bmatrix} b(t) \\ B(t) \end{bmatrix} d \begin{bmatrix} W_{1t} \\ W_{2t} \end{bmatrix} \qquad (7)$$

where W_{1t} and W_{2t} are vector valued Wiener processes, independent of initial values $X(0)$, $Y(0)$, and $b(t)$ and $B(t)$ are matrices with the appropriate dimensions. Thus, the processes $X(t)$ and $Y(t)$ are the solution of these linear stochastic differential equations. We may now consider the two different observational plans for multi-component processes of the type described by (6) and (7).

A. Continuous Observations

In Yashin et al. (1985) we considered the solution of these equations in the conditional Gaussian case by assuming that the distribution of the Y(t) was normal conditional on the observed process. We can demonstrate the validity of this observation by noting that one can always find a vector function F and a scalar G, such that the individual mortality rate, $\mu(t, X, Y)$, can be written

$$\mu(t, X, Y) = (Y - F)' Q_{22}(t) (Y - F) + G \tag{8}$$

where F and G are functions of t and X, i.e.,

$$F(t, X) = Q_{22}^{-1}(t) Q_{21}(t) X \tag{9}$$

and

$$G(t,X) = X'Q_{11}(t) X - X'Q_{12}(t) Q_{22}^{-1}(t) Q_{21} X + \mu_0(t). \tag{10}$$

The structure of (8) with respect to Y is similar to the hazard function considered by Myers (1981). However, it is difficult to use additional observations on the measured process X in his formulation. A more appropriate strategy seems to involve use of the conditional Gaussian approach developed in Yashin (1984) for a continuously observed process. This latter approach can be used for the evaluation of a process that is still under observation, e.g., to analyze data from the intermediate phases of a longitudinal study.

B. Fixed Time Observation

Let us now assume that the elements of X(t) are measured at a set of fixed times. Thus $X_i(t),\ldots, X_i(t_k)$ are the measurements on the ith individual. $Y_i(t)$ represents the variables that are not measured. We assume that both processes influence the mortality rate and that this dependence is as described by equation (6). Furthermore, we assume that the evolution of X(t) and Y(t) are described by (7). We wish to estimate the elements of Q(t) (in equation (6)) on the basis of data only on X, i.e., $X_i(t_1 \wedge T_i),\ldots, X_i(t_k \wedge T_i)$, $i = I$ where T_i are the observed death times. For simplicity we will suppress the index i and define $\hat{X}(t)$ as the matrix $X(t_1), X(t_2),\ldots, X(t_j(t))$ where

$$t_j(t) = \sup\{t_m: \ t_m < t\} \tag{11}$$

The survival function, conditional on the observed process X, say $S(t,\hat{X})$, may be defined

$$S(t, \hat{X}) = P(T > t \,|\, \hat{X}(t)) \tag{12}$$

so that

$$\bar{\mu}(t,\hat{X}(t)) = -\frac{\partial}{\partial t} \ln S(t,\hat{X}) \tag{13}$$

The problem of estimation is to find the appropriate relation of $\bar{\mu}(t, \hat{X}(t))$, the average mortality rate measured for fixed times, to the parameters of the underlying process and ultimately to the means ($m(t)$) and covariance ($\gamma(t)$) of the variables in both X and Y. The appropriate relations are presented in the following theorem:

Theorem: Suppose that we have a complex process defined by both measured and unmeasured variables with the structure presented in equation (7). Then the relation of the average mortality rate observed for survivors to a specific age to the underlying observed and unobserved processes is provided by

$$\bar{\mu}(t,\hat{X}(t)) = (m'(t)\ Q(t)\ m(t) + \mathrm{Tr}(Q(t)\ \gamma(t)) + \mu_0(t) \tag{14}$$

where $m(t) = \begin{pmatrix} m_1(t) \\ m_2(t) \end{pmatrix}$, $\gamma(t) = \begin{bmatrix} \gamma_{11}(t), & \gamma_{12}(t) \\ \gamma_{21}(t), & \gamma_{22}(t) \end{bmatrix}$ on the intervals $t_j \leq t < t_j+1$ satisfy the equations,

$$\frac{dm(t)}{dt} = \alpha_0(t) + \alpha(t)\ m(t) - 2m(t)\ Q(t)\ \gamma(t) \tag{15}$$

and

$$\frac{d\gamma(t)}{dt} = \alpha(t)\ \gamma(t) + \gamma(t)\alpha^*(t) + b(t)b^*(t) - 2\,\gamma(t)\ Q(t)\ \gamma(t)$$

where $\alpha_0(t) = \begin{bmatrix} \alpha_{01}(t) \\ \alpha_{02}(t) \end{bmatrix}$ and $\alpha(t) = \begin{bmatrix} \alpha_{11}(t), & \alpha_{12}(t) \\ \alpha_{21}(t), & \alpha_{22}(t) \end{bmatrix}$

The primary difference between these equations and those for the continuous time case is that, for each interval, a new set of initial conditions holds. Thus, we have defined a jump process in the observational plan where, at each time of measurement, there is a jump in information. Specifically, at time t_j, $j = 1,\ldots, K$ the initial values for the equation are

$$m_1(t_j) = m_1(t_j^-) + \gamma_{12}(t_j^-)\ \gamma_{22}^{-1}\ (t_j^-)\ (X(t_j) - m_2(t_j^-)) \tag{16}$$

$$m_2(t_j) = X(t_j) \tag{17}$$

$$\gamma_{11}(t_j) = \gamma_{11}(t_j^-) - \gamma_{12}(t_j^-)\ \gamma_{22}^{-1}(t_j^-)\ \gamma_{21}(t_j^-) \tag{18}$$

$$\gamma_{22}(t_j) = 0 \tag{19}$$

$$\gamma_{12}(t_j) = \gamma_{21}(t_j) = 0\ . \tag{20}$$

Thus the mean for X is equal to the observed value at the time of measurement (17) while the mean for Y is the mean conditional on X. The variance of the values of the observed variable is equal to 0 at the measurement time while for Y, we have the conditional variance. The initial conditions for each interval represents the jumps in information at these points. These results can be achieved through a two-stage proof. The first and most important step is to prove the conditional Gaussian property. This is done by examining the characteristic function conditional on the process X and the time to death. Once the conditional Gaussian properties are demonstrated (for details see Yashin et al. (1985)) we know that we only need the means and variances of the distribution of Y(t) to characterize the process. In the second step, we can specify the equations for the means and variances, again from the characteristic function.

IV. ESTIMATION

With the results above, estimation can be conducted quite simply. Specifically we may specify the likelihood function in terms of $\bar{\mu}(t,\hat{X}(t))$ as,

$$\mathcal{L} = \prod_{i=1}^{I} \bar{\mu}(t_i, \alpha)\ e^{-\int_0^{t_i} \bar{\mu}(u,\alpha)\,du}\ , \tag{21}$$

where $\bar{\mu}(t_i,\alpha)$ is given by equation (14), $m(t,\alpha)$ is given by (15) and $\gamma(t,\alpha)$ by (16). To evaluate (21) we need only specify $Q(t,\alpha)$ as some specific function (e.g., $\alpha e^{\beta t}$; the Gompertz) and write $\bar{\mu}(t_i,\alpha_1)$ in terms of $m(t,\alpha)$ and $\gamma(t,\alpha)$. Thus from the evaluation of $\bar{\mu}(t_i,\alpha)$ we can obtain the parameters of the underlying stochastic process. Since we cannot directly evaluate the forms in (21) we will have to use special numerical procedures (see Yashin, 1984).

V. A COMPARISON OF THE CAMERON-MARTIN AND CONDITIONAL GAUSSIAN APPROACHES

The Cameron-Martin approach (Yashin,1984) gives a way of calculating the mathematical expectation of an exponent which is the functional of a Wiener process. The exponent can be considered as a conditional survival function. Thus the approach has been suggested as a methodology for survival analysis where the stochastic process in the exponent is interpreted as covariates affecting the survival rate. Unfortunately, the Cameron-Martin approach has several significant limitations. To illustrate, it can be shown that for the linear diffusion process written in (7), the matrix of hazard coefficient, $Q(t)$, has the property

$$E \exp\left[-\int_0^t [Y(u),X(u)]' Q(u) [Y(u)X(u)]\, du\right] =$$

$$\exp[(Y(0),X(0))' \Gamma(0) (Y(0),X(0)) + \mathrm{Tr} \int_0^t [b(u),B(u)] [b(u),B(u)]' \Gamma(u)\, du \qquad (22)$$

where $\Gamma(u)$ is the solution of the matrix Ricatti equation

$$\frac{d\Gamma(u)}{du} = Q(u) - (\Gamma(u) + \Gamma'(u))\, a(u) - \tfrac{1}{2}(\Gamma(u) + \Gamma'(u))[b(u),B(u)] [b(u),B(u)]' (\Gamma(u) + \Gamma'(u)) \qquad (23)$$

with the terminal condition $\Gamma(t) = 0$.

The particular case of formula (22) corresponds to the well-known Cameron-Martin results (Yashin, 1984) specified for a Wiener process in the exponent of the form:

$$E \exp\left[-\int_0^t (W_u, Q(u)\, W_u)\, du\right] = \exp\left[\tfrac{1}{2} \int_0^t \mathrm{Tr}\, \Gamma(u)\, du\right] \qquad (24)$$

where $(W_u, Q(u), W_u)$ is the scalar product equal to the quadratic form, $W_u' Q(u)$,

W_u, and $\Gamma(u)$ is a symetric nonpositive definite matrix which is an unique solution of the matix Riccati equation

$$\frac{d\Gamma(u)}{du} = 2\ Q(u) - \Gamma^2(u) \tag{25}$$

and $\Gamma(t) = 0$ is a zero matrix.

To prove these relations one uses likelihood ratio principles applied to diffusion type processes (Novikov, 1972; Liptzer and Shirjaev, 1974). Using this approach, Myers (1981) found the formulas for averaging the exponent, when instead of a Wiener process, there is a process satisfying a linear stochastic differential equation driven by a Wiener process (i.e., 22) and (23)).

Unfortunately, the proof of the Cameron-Martin formula and its generalization (Myers, 1981) do not use the interpretation of the matrix Q as hazard coefficients and do not provide a direct physical interpretation of the variables $\Gamma(u)$ in (22) (or (24)). Furthermore the boundary condition on (23) (and (25)) makes it difficult to conduct the calculations either for subintervals, or when additional longitudinal measurements are made.

The methods described in this paper do not have these limitations. They involve the use of "Martingale" techniques to produce a general formula for averaging exponents which can be a more complex functional of a random process of a wider class (Yashin, 1984). In this paper we provide the specialization of these procedures to the case where the functional is a quadratic form for averaging the exponents. These procedures turn out to have a range of computationally important properties based upon the conditional Gaussian property.

VI. DISCUSSION

In this paper we present a procedure for evaluating the stochastic process underlying the observed population averaged survival rate. This procedure, using conditional Gaussian properties, leads to computationally powerful techniques for assessing human survival data. The conditional Gaussian approach can

be shown to have superior properties to the Cameron-Martin procedure. The procedure offers likelihood ratio techniques for estimating the basic parameters of the process.

The procedure has utility in several important areas. First, there has been much recent attention to the question of heterogeneity (unmeasured differentials in transition rates) and its effects on the analysis of human survival (Vaupel et al., 1979; Manton and Stallard, 1984; Heckman and Singer, 1982). Underlying this concern is the analytic problem of how systematic selection of persons by mortality affects the average force of transition among survivors. This involves examination of the effects of averaging of the exponent (and related functional) in the survival function. Past efforts have tended to resolve the problem by ignoring the particular effects of diffusion by using a deterministic trajectory for the temporal dependence of the individual hazard rate. This approach can only be an approximation and is problematic when one is attempting to infer the operation of the risk mechanism at the individual level. By explicitly including the diffusion process in the proposed model one can potentially greatly improve the precision of one's predictions and certainly has a much better procedure for determining the effects of intervention on the realization of risk.

A second major utility of the proposed approach is that it greatly facilitates the introduction of auxiliary information into one's analysis of the failure process. This is facilitated because one can directly examine the details of the process and thereby introduce information into the appropriate features of the model. This is a critically important property in analyzing human survival at advanced ages because the evolution of chronic diseases is a complex process operating over a lengthy time scale. Thus, though there is considerable empirical information on risk covariates and evolution of chronic disease from existing longitudinal studies, seldom have the dynamic properties of such data been completely exploited. For example, certain negative associations have been demonstrated between a risk factor (e.g., asbestos) and a specific disease outcome

(e.g., lung cancer) because of the systematic selection of the susceptible persons by a disease process (e.g., asbestosis) which had an earlier age assault pattern (Manton, 1985). Such dynamics and systematic selection require consideration of the basic dynamic process and the effects of selection on the average risk among survivors to unconfound such factors. Only by using auxiliary information and a model of the intrinsic processes can such public health questions be adequately resolved.

ACKNOWLEDGEMENTS

Dr. Manton's efforts in this research were supported by NIA Grant No. AG01159-08.

REFERENCES

Cameron, R.H., Martin, W.T: Transformation of Wiener Integrals by Nonlinear Transformation Transactions of American Mathematical Society 66:253-283, 1948.

Economos AC: Rate of aging, rate of dying and the mechanism of mortality. Arch Gerontol Geriatr 1:3-27, 1982.

Heckman JJ, Singer B: Population heterogeneity in demographic models. In Multidimensional Mathematical Demography (Land K, Rogers A, eds.). New York, Academic Press, 1982, pp. 567-599.

Liptzer, R.S., Shirjaev, A.N: Statistics of Random Processes, Nauka, 1974.

Manton KG: An evaluation of strategies for forecasting the implications of occupational exposure to asbestos. U.S. Library of Congress, Congressional Research Service, Government Division, 1985.

Manton KG, Stallard E: Heterogeneity and its effect on mortality measurement. Chapter 12 in Proceedings IUSSP Methodology and Data Collection in Mortality Studies Seminar, July 7-10, Dakar, Senegal, forthcoming in 1984.

Manton KG, Woodbury MA: A continuous-time multivariate Gaussian stochastic process model of change in discrete and continuous state variables. In Sociological Methodology, 1985 (Tuma N, ed.). Jossey-Bass, forthcoming in 1985.

Manton KG, Woodbury MA: A mathematical model of the physiological dynamics of aging and correlated selection: Part II-Application to the Duke Longitudinal Study. J Gerontol 38:406-413, 1983.

Myers L: Survival functions induced by stochastic covariate processes. J Applied Probability 18:523-529, 1981.

Novikov AA: On Parameters Estimation of Diffusion Processes: Studia Science Mathematics 7:201-209, 1972.

Spiegelman M: Introduction To Demography. Cambridge, Mass., Harvard University Press, 1969.

Vaupel JW, Manton KG, Stallard E: The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16:439-454, 1979.

Woodbury MA, Manton KG: A mathematical model of the physiological dynamics of aging and correlated mortality selection: Part I-Theoretical development and critiques. J Gerontol 38:398-405, 1983.

Woodbury MA, Manton KG: A random walk model of human mortality and aging. Theor Popul Biol 11:37-48, 1977.

Yashin AI: Dynamics in survival analysis: Conditional Gaussian property versus Cameron-Martin formula. WP-84, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1984.

Yashin AI, Manton KG, Vaupel JW: Mortality and aging in a heterogeneous population: A stochastic process model with observed and unobserved variables. Theor Popul Biol, forthcoming in 1985.