Working Paper

A SIMULATION STUDY OF THE CONDITIONAL GAUSSIAN DIFFUSION PROCESS MODEL OF SURVIVAL ANALYSIS

Fernando Rajulton
Anatoli Yashin

February 1986
WP-86-6

International Institute for Applied Systems Analysis
A-2361 Laxenburg, Austria

A SIMULATION STUDY OF THE CONDITIONAL GAUSSIAN DIFFUSION PROCESS MODEL OF SURVIVAL ANALYSIS

Fernando Rajulton
Anatoli Yashin

February 1986
WP-86-6

This research was conducted in conjunction with a summer research seminar on heterogeneity dynamics, under the direction of James W. Vaupel and Anatoli I. Yashin, in the Population Program at IIASA led by Nathan Keyfitz.

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
2361 Laxenburg, Austria

Foreword

A group of eleven Ph.D. candidates from seven countries--Robin Cowan, Andrew Foster, Nedka Gateva, William Hodges, Arno Kitts, Eva Lelievre, Fernando Rajulton, Lucky Tedrow, Marc Tremblay, John Wilmoth, and Zeng Yi--worked together at IIASA from June 17 through September 6, 1985, in a seminar on population heterogeneity. The seminar was led by the two of us with the help of Nathan Keyfitz, leader of the Population Program, and Bradley Gambill, Dianne Goodwin, and Alan Bernstein, researchers in the Population Program, as well as the occasional participation of guest scholars at IIASA, including Michael Stoto, Sergei Scherbov, Joel Cohen, Frans Willekens, Vladimir Crechuha, and Geert Ridder. Susanne Stock, our secretary, and Margaret Traber, managed the seminar superbly.

Each of the eleven students in the seminar succeeded in writing a report on the research they had done. With only one exception, the students evaluated the seminar as "very productive"; the exception thought it was "productive". The two of us agree: the quality of the research produced exceeded our expectations and made the summer a thoroughly enjoyable experience. We were particularly pleased by the interest and sparkle displayed in our daily, hour-long colloquium, and by the spirit of cooperation all the participants, both students and more senior researchers, displayed in generously sharing ideas and otherwise helping each other.

This paper by Fernando Rajulton deals with the simulation study for a stochastic process model of human mortality and aging. Using the Monte Carlo method, the author generated the survival data and developed the parameter estimation algorithms in the case of stochastically changing unobserved covariates. The efficacy of the approach is discussed. The results show that the model developed is applicable for many fields where hidden heterogeneity and selection are present.

James W. Vaupel
Anatoli I. Yashin

Abstract

The recently developed conditional Gaussian diffusion process model is a powerful tool of survival analysis. Its generality not only encompasses the survival models to date but also brings into focus the influence of unobserved variables related to "death" of individuals. Further, that the model makes feasible a unique estimation of the parameters of the underlying unobserved or partially observed process is shown in this paper through a set of simulated data on death times and an unobserved variable. Possibilities of extensive use of the model to areas other than mortality are pointed out.

Acknowledgments

The author is grateful to IIASA and to the staff of the Population Program for the opportunities provided during the summer program. In particular, thanks are due to Dr. Anatoli Yashin for fruitful discussions on the model and for the practical guidance in simulation.

Contents

A SIMULATION STUDY OF THE CONDITIONAL GAUSSIAN DIFFUSION PROCESS MODEL OF SURVIVAL ANALYSIS

Fernando Rajulton

Inter-University Program in Demography
Vrije Universiteit Brussel
2 Pleinlaan
B-1050 Brussels
Belgium

INTRODUCTION

No population is homogeneous and no one individual is like unto another. This self-evident truth defies any mathematical modelling of human behaviour, however sophisticated it may be. That no population is homogeneous has motivated demographers to take into account at least the differential behaviours of homogeneous subgroups of a heterogeneous population. But recent efforts directed at examining the heterogeneity of individuals under the motif of *heterogeneity dynamics*, though mostly restricted to mortality analysis to date, have thrown open new vistas for further research in areas other than mortality. Analytical experience learned from recent efforts can lead to developing powerful means of analysing human behaviour in general, as data at individual level become more and more available and as the mathematical apparatus being developed with a special orientation to analysing human mortality becomes more and more generalized in the light of practical applications in many fields.

This paper addresses one such application--of the Gaussian diffusion process model of physiological change and mortality basically derived in Woodbury and Manton (1977) and extended by Yashin, Manton and Stallard (1985) to incorporate variables, both unobserved and observed at fixed times. The main purpose of this paper is to examine whether the model works--in the sense that it leads to unique estimation of the parameters involved--through a simulation of data and to present certain practical guidelines for its application. First, as a sort of preparation, survival models in common use are briefly recalled in the order of their complexity and the conditional Gaussian model is introduced in its generality and with the basic notions involved. Second, data on death times and an unobserved variable are generated through Monte Carlo simulations with specific values of the parameters of the model, and then, conversely, the estimation of the parameters is carried out with the only information on death times. Finally, possible extensive application of the model to other demographic variables is pointed out.

SURVIVAL MODELS AT A GLANCE

Variables that heterogenize a population are many. Some are observed, others not. Some are observables, many are not. To deal with the observed variables, mathematical tools have long since been developed and refined to study the impact of the explanatory variables on the "death time" of individuals.

In the simplest case, when only data on death times of individuals are known, Kaplan-Meir product-limit estimation procedure is a common-sense tool. When some explanatory variables (mainly expressed as covariates) along with death times are known, proportional hazard model and Cox regression model are at one's disposal. More recently, mixed proportional hazard model has been forwarded (Ridder and Verbakel 1983) to take into account the omitted covariates as well under certain assumptions, the main one being the absence of any correlation between the included and the omitted covariates.

At a higher level, when data on death times and *longitudinal data* on some observed variables are available, the Gaussian diffusion process model (Woodbury and Manton 1977; Yashin 1984; Yashin, Manton and Stallard 1985a; Yashin 1985; Yashin, Manton and Vaupel 1985; Yashin, Manton and Stallard 1985b; Yashin and Manton 1984) can be gainfully employed not only for the usual survival analysis but also for predicting more precisely the hazard rate and the "health changes". Because, the model in its generality covers most of the other survival models and explicitly takes into account the role and impact of the unobserved variables on mortality.

What has been brought to light in recent times is the influence, though unknown and hidden, of the unobservables in any heterogeneity analysis (Heckman and Singer 1982; Tuma and Hannan 1984). However, in most of the studies which purport to examine the influence of the unobserved variables, an explicit consideration of the relationship between the realized hazard rate and the (parameters of the) underlying unobservable *process* has not been considered. This lack of consideration can lead not only to wrong inferences but also to simplistic attempts at using the observed data to make forecasts merely on the basis of the long-recognized age pattern of heterogeneity. On the other hand, the conditional Gaussian diffusion process model, drawing from the analytical experience in disciplines other than demography (for example, in communication theory, information theory and control) is built on the relevance of the relationship between the observed rate and the (parameters of the) underlying process both in the presence and absence of information. Using any little information available on the observed and the unobserved processes will surely enhance the precision of analysis and of forecasts.

The basic model as developed by Woodbury and Manton (1977) recognized the impact on mortality of the physiological variables such as serum cholesterol, blood pressure etc., which evolve over time in a manner that can be described by a Gaussian diffusion process--and hence its name. It helped in establishing the mathematical relationships between the observed mortality and the parameters of the process governing ***change in the means and covariances of the physiological variables related to mortality.*** The model was actually based on the Kolmogorov-Fokker-Planck Equation, the application of which required the assumption that the underlying process was Markovian. This assumption would imply that the individual's future profile of physiological values is a result of both a deterministic function of his current value and a stochastic term. Yashin et al., however, generalized this basic model to deal with non-Markovian processes (Yashin, Manton and Vaupel, 1985; Yashin, 1985) and with the combination of observed and unobserved variables.[1]

[1] The history of the development of the Gaussian diffusion process model goes back to the initial attempts by Cameron and Martin in 1940's at calculating the mathematical expectation of an exponent which is a functional of a Wiener process. Much later, in 1980, Myers developed the approach due to Novikov and found the formula for averaging the exponent, where instead of a Wiener process, there is a process satisfying a linear stochastic differential equation driven by a Wiener process. With the on-set of heterogeneity dynamics, the concept of averaging the exponent came into prominence; because any exponent can be considered as a conditional survival function and the observed rate (of the population without considering heterogeneity) is nothing else but the expectation of individual rates. And, Yashin shows that if the functional involved is of a quadratic form one can get another constructive way of averaging the exponent using the conditional gaussian property which avoids all the complications involved in the earlier approaches. For details and references herein, see Yashin (1984).

A BRIEF REVIEW OF THE MODEL

The Gaussian diffusion process can be described by (1) a linear autoregressive model of change in the physiological variables and (2) a quadratic function describing the relation between the hazard rate and the values of the physiological variables. In other words, *linear dynamics* and *quadratic dependency* are the key components of the model. These two components have been found to describe human physiological change and mortality in a number of epidemiological studies of chronic disease.

Suppose that the mortality rate for individual i in a population of N individuals depends on a process Y_t which evolves over time.[2] Assume that the process for each individual evolves independently from that of all other individuals. Further, as mentioned above, the mortality rate for an individual denoted by $\mu(t,Y_t)$ is assumed to be a quadratic function of the set of values Y_t:

$$\mu(t,Y_t) = \mu_{0t} + \mu_{1t}Y_t + \mu_{2t}Y_t^2 \quad . \tag{1}$$

Assume also that Y_t satisfies the linear diffusion type stochastic differential equations:[3]

$$dY_t = (a_{0t} + a_{1t}Y_t)dt + b_t dW_t \tag{2}$$

where a_{0t}, a_{1t} and b_t are well-bounded functions and W_t is a Wiener process which does not depend on the initial condition Y_0 which is gaussian distributed with *known* mean m_0 and variance γ_0.

The relation between $\bar{\mu}_t$—the observed hazard rate—and the conditional mortality rate $\mu(t,Y_t)$ described by the conditional gaussian process can be expressed in the form:

$$\bar{\mu}_t = E[\mu(t,Y_t)\,|\,T_i \geq t] \tag{3}$$

where E denotes the mathematical expectation and T_i is the death time of the i-th individual associated with the mortality rate $\mu(t,Y_t)$. When $Y_0 \sim N(m_0,\gamma_0)$ and $\mu(t,Y_t)$ is quadratic, then $\bar{\mu}_t$ is the age-specific mortality rate among *survivors* to time t with the following relation to the parameters of the distribution of Y_t.

[2] Note that Y_t can be generalized to a multidimensional case by considering appropriate vectors and matrices instead of scalars.

[3] This process can also be adapted to the notions of "frailty" introduced by Vaupel et al. (1979) and Vaupel and Yashin (1982) and suitable modifications can always be done to express the mathematical relationship.

$$\bar{\mu}_t = \mu_{0t} + \mu_{1t} m_t + \mu_{2t}(m_t^2 + \gamma_t) \tag{4}$$

where m_t and γ_t satisfy the non-linear differential equations:[4]

$$\frac{dm_t}{dt} = a_{0t} + a_{1t} m_t - \gamma_t \mu_{1t} - 2\mu_{2t}\gamma_t m_t \tag{5}$$

and

$$\frac{d\gamma_t}{dt} = 2a_{1t}\gamma_t + b_t^2 - 2\gamma_t^2 \mu_{2t} \tag{6}$$

with the initial conditions m_0 and γ_0.

Three points are in order.

(a) The relationship between the observed mortality and the underlying physiological process expressed in equation (4) can be used to develop the likelihood function for the estimation of the parameters of the process from the distribution of the observed death times of individuals in the population [see equation (21)].

(b) Yashin has shown that this likelihood function can be generalized to deal with the estimation of the parameters conditional upon the realized values of a process partially observed at fixed times. In this case, the underlying process would be a jump process and the same equations (1) through (6) developed for the continuous case hold good also. Only the likelihood function has to be modified with an additional term involving the conditionality [see below equation (24)].

(c) Further, the same can be extended to the cases where some additional variables (covariates) have been measured [see equation (28) for the likelihood].

This paper addresses these three points in the simple case when μ_{0t} and μ_{1t} in (1) are zero and μ_{2t}, a_{0t}, a_{1t} and b_t are constant over time; that is, $\mu_{2t} = \mu$, $a_{0t} = a_0$, $a_{1t} = a_1$ and $b_t = b$, whereby (1) is reduced to

$$\mu(t, Y_t) = \mu Y_t^2 \ . \tag{7}$$

[4] These two equations (5) and (6) are similar to the Kalman filter equations in communication theory used to estimate signals. Here, they have been generalized to include mortality selection. cf. for details Yashin, Manton and Vaupel (1985).

Before entering into the application of the model, the interpretation of the parameters α_0, α_1 and b of the linear dynamic process deserves one's attention. Following the interpretations given by Woodbury and Manton (1977) to the parameters of the Kolmogorov-Fokker-Planck equation, the "drift" denoted by α_0 is the *systematic change* in mean values, the "regression" effect denoted by α_1 is the *convergence* to mean values due perhaps to homeostatic tendencies and the "diffusion" denoted by b is the *divergence* due to random influences. These interpretations are comparable to the usual ones adduced to any linear regression model, as the dynamics of the physiological variables is assumed to be linear.

Application of the Model through Simulation

Case 1: *physiological variables Y_t are unobserved*

(a) Data Simulation

We have

$$dY_t = (\alpha_0 + \alpha_1 Y_t)dt + bdW_t \quad . \tag{2'}$$

The discrete approximation of this equation is

$$\begin{aligned} Y_{t+\delta} - Y_t &= (\alpha_o + \alpha_1 Y_t)\delta + b\,\delta W_t \\ &= (\alpha_0 + \alpha_1 Y_t)\delta + b\sqrt{\delta}\varepsilon_{t+\delta} \end{aligned} \tag{8}$$

that is,

$$Y_{t+\delta} = \alpha_0 + (1 + \alpha_1\delta)Y_t + b\sqrt{\delta}\varepsilon_{t+\delta} \quad . \tag{9}$$

When $\delta = 1$,

$$Y_{t+1} = \alpha_0 + (1 + \alpha_1\delta)Y_t + b\,\varepsilon_{t+1} \tag{10}$$

where $\varepsilon_t \sim N(0,1)$. Denoting by Y_0^t the process up to time t

$$\begin{aligned} F(t) &= Prob\,[T \le t \mid Y_0^t] \\ &= 1 - e^{-\int_0^t \mu(s,Y_s)ds} \quad , \end{aligned} \tag{11}$$

we generate a uniform random variate r such that

$$F(t) = Prob\,[r \le F(t)] = Prob[F^{-1}(r) \le t] \tag{12}$$

For details, see Rubinstein (1981). Once we have generated r, we know that

$$r \leq 1 - e^{-\int_0^t \mu(s,Y_s)ds} = 1 - e^{-\int_0^t \mu Y_s^2 ds} \tag{13}$$

that is,

$$-log(1 - r) \leq \int_0^t \mu Y_s^2 ds \tag{14}$$

If, therefore, τ is the death time of an individual, then

$$\tau = inf\,[t: \int_0^t \mu Y_s^2 ds \geq -\log(1 - r)] \quad . \tag{15}$$

Similarly, from (5) and (6) adapted for the constancy of parameters over time, the corresponding means and variances of the process m_t and γ_t will be given by (when $\delta = 1$):

$$m_{t+1} = a_0 + (1 + a_1)m_t - 2\gamma_t \mu m_t \tag{16}$$

and

$$\gamma_{t+1} = (1 + 2a_1)\gamma_t + b^2 - 2\gamma_t^2 \mu \quad . \tag{17}$$

Thus, the algorithm for simulation would be as follows: for each individual, with fixed a_0, a_1, μ, m_0, γ_0,

(a) generate r_t, a uniform random variate

(b) calculate $-\log(1 - r_t)$

(c) generate $\varepsilon_t \sim N(0,1)$

(d) generate $Y_0 \sim N(m_0, \gamma_0)$

(e) with the generated Y_0, calculate Y_t and $\sum_{s=0}^{t} \mu Y_s^2$. When this sum becomes $> -\log(1 - r_t)$ at a particular value of t, that value of t is the required τ.

With a_0=0.02, a_1=−0.1, b=0.001, μ=1.0, m_0=0.2 and γ_0=0.001, the generated death times for 100 individuals (with the computer program SIMUL1 given in the Appendix) are shown in Table 1.

Table 1. Simulated times to death of 100 individuals with specifications of $\alpha_0 = 0.02$, $\alpha_1 = -0.1$, $b = 0.001$, $\mu = 1.0$, $m_0 = 0.2$, and $\gamma_0 = 0.001$.

17	5	18	13	1	8	2	22	1	7
122	6	11	21	3	27	3	40	10	43
55	9	9	4	1	16	2	20	51	5
62	17	25	10	97	13	2	3	1	21
13	6	2	66	40	62	65	51	48	11
21	28	21	28	3	2	63	16	2	66
5	38	13	1	12	11	18	5	18	23
53	71	1	14	14	21	60	1	1	57
108	1	27	48	26	72	27	33	18	19
16	7	19	1	59	38	8	27	2	24

A practical guideline for fixing the parameter values would be as follows: *under stability conditions*, when the random disturbances would be negligible, (2') can be written as

$$\frac{dY_t}{dt} = \alpha_0 + \alpha_1 Y_t = 0 \qquad (18)$$

which yields the relation

$$Y_t = -\frac{\alpha_0}{\alpha_1} = Y \qquad (19)$$

and hence the mean death time $\bar{\tau}$ of individuals would be

$$\bar{\tau} = \frac{1}{Y^2} = \left[\frac{\alpha_1}{\alpha_0}\right]^2 . \qquad (20)$$

This helps in choosing proper values for the parameters α_0 and α_1 according to the situational requirements. Thus, for example, if we have information that mean death time is 25, then α_0 could be given a value of 0.02 and α_1 a value of -0.1 such that $Y = 0.2$ when the process becomes stable and hence $\bar{\tau} = 25$. The death times in

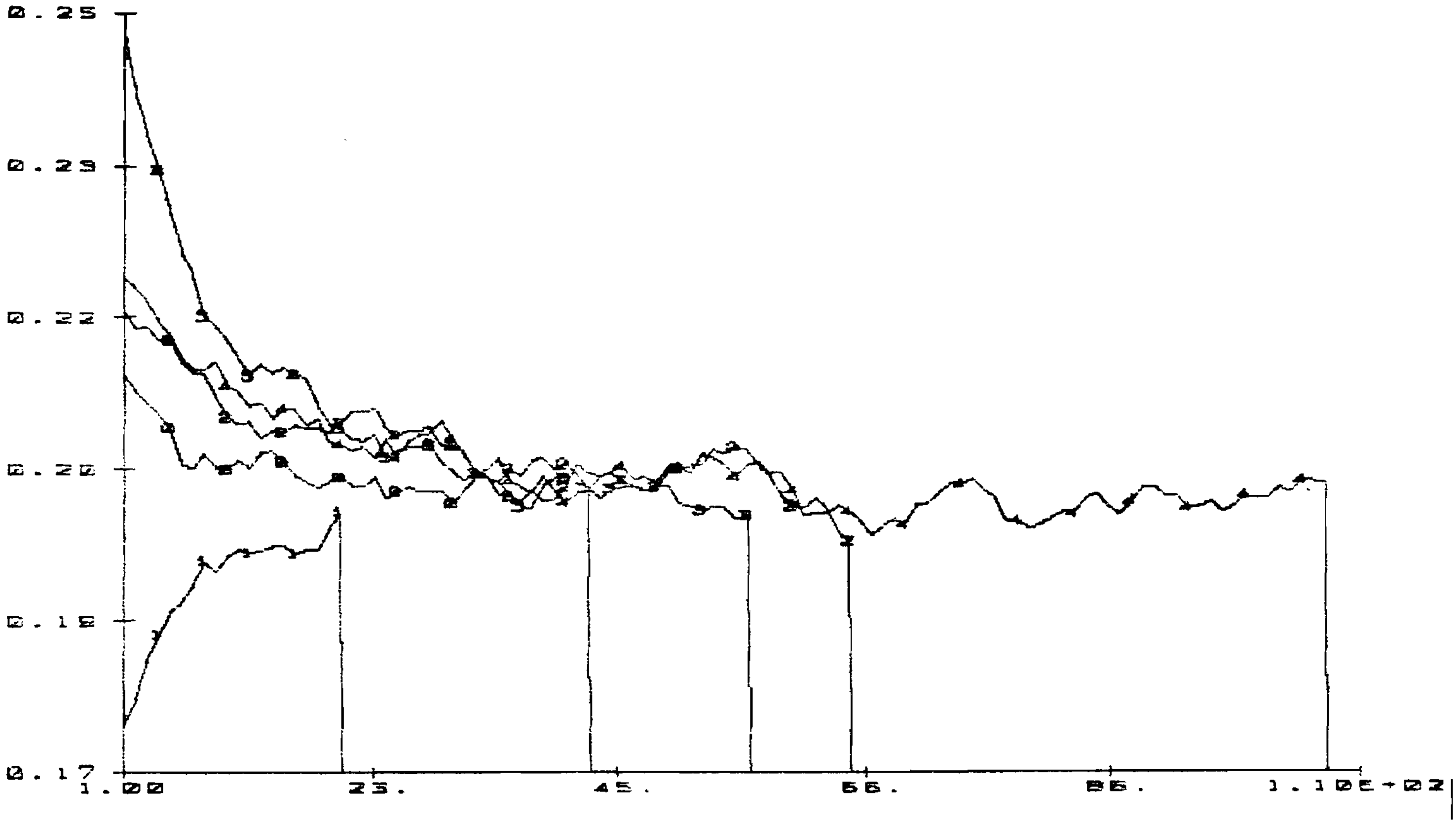


Figure 1. Trajectory of Y_t for chosen individuals whose death times are 21, 43, 57, 66 and 108 years.

We have generated the death times of 100 individuals with the specifications of the parameters as shown in Table 1. Now, if we estimate the parameters, given the death times, can we get the estimates to be as close as possible to those with which we generated the death times? Maximizing (or minimizing) the log likelihood (or the negative log likelihood) function can be executed with the available libraries. We shall here minimize the function with the IMSL library subroutine ZXLSF available at IIASA. We shall carry out the estimation procedure for each parameter while keeping others fixed, and for all four parameters together.

It is important to note that m_t and γ_t values are expressed in terms of these parameters. And hence, m_t and γ_t are also to be estimated every time the estimation iteration is carried out. (The computer program SIMUL2 given in the Appendix may be of some help.) Before carrying out the estimation procedures, it would be of great help to examine the behaviour of the $-logL$ for various values of the parameters. Figures 2 through 5 describe the $-logL$ function for various values

Table 1 have been generated with these values of α_0 and α_1, and hence have the mean and standard deviation values as 24.34 and 25.09 respectively. In general, as $Y_t = -\frac{\alpha_0}{\alpha_1}$, it would be better to give values less than zero for α_1, letting α_0 be positive (the drift coefficient) and b be rather small. Examining the generated death times may be of some help; it is desirable, for example, to get the death times mostly under 100!

Care should be exercised on the possibility that the γ_t can become negative for higher values of t, while m_t will be normally assured to be around $\frac{\alpha_0}{|\alpha_1|}$. It would be practical to make the γ_t equal to zero when the actual calculation becomes negative.

Interests in different applications could specify different values of the parameters and examine the trajectory of the physiological process Y_t and the death times obtained. Figure 1 presents the trajectory of Y_t for five individuals whose death times are 43, 21, 66, 57 and 108 years (these are the 20th, 40th, 60th, 80th and 81st individuals in Table 1) and whose starting Y_0 values are 0.214, 0.169, 0.227, 0.258 and 0.225, respectively (not given in Figure 1). In other words, their initial values, say "frailty", are different. The trajectory of these individuals can be seen to become somewhat stable by $t = 45$ or so.

(b) Estimation of the Parameters

The likelihood function for N individuals is given by

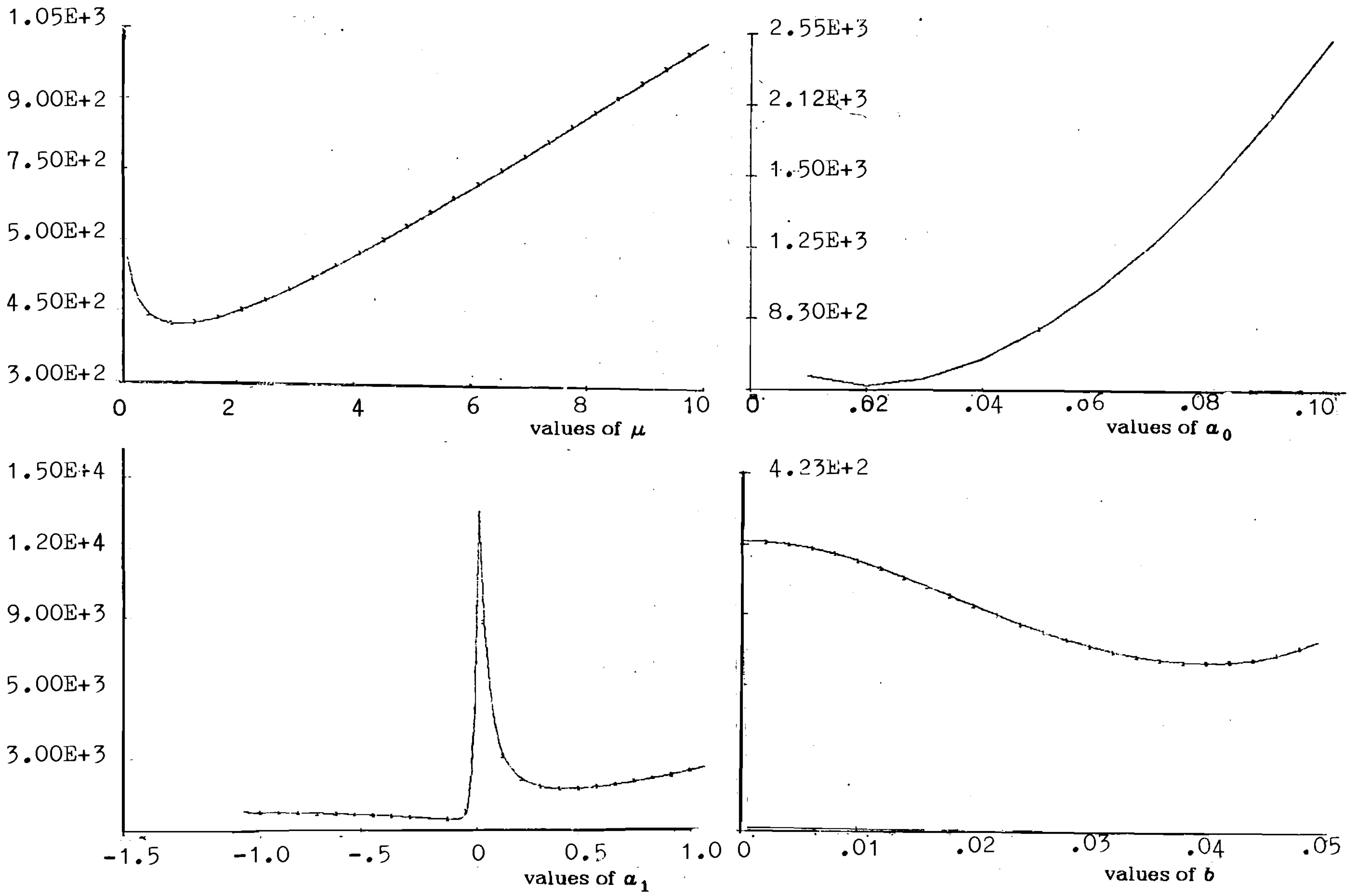
$$L = \prod_{i=1}^{N} \bar{\mu}_{\tau_i} e^{-\int_0^{\tau_i} \bar{\mu}_s ds} \tag{21}$$

where

$$\bar{\mu}_{\tau_i} = \mu(m_{\tau_i}^2 + \gamma_{\tau_i}) \quad . \tag{4'}$$

And hence,

$$\log L = \sum_{i=1}^{N} [\log \mu(m_{\tau_i}^2 + \gamma_{\tau_i}) - \sum_{s=0}^{\tau_i} \mu(m_s^2 + \gamma_s)] \quad . \tag{22}$$



Figures 2-5. Negative log likelihood function for various values of the four parameters, μ, a_0, a_1 and b.

of α_0, α_1, μ and b. In Figure 2 the likelihood function is seen to be a very smooth function for different values of μ, attaining minimum at $\mu = 1.0$ as expected. In Figure 3 also, the function is found to be smooth for different values of α_0, having the minimum at $\alpha_0 = 0.02$ as again expected. However, for various values of α_1 the function is no longer smooth in Figure 4; it attains minimum at two points, one at below zero and the other at above zero. This "anomaly" however is the result of using the α_1 in the denominator for the estimate of m_0; an arithmetic overflow could occur at $\alpha_1 = 0$ when α_1 is allowed to vary from negative to positive values. In fact, Figure 4 plots the likelihood function with $Y_t = 0$ when $\alpha_1 = 0$. It is possible also to obtain a smooth curve of the likelihood function by specifying proper values of Y_t when $\alpha_1 = 0$, such that it becomes unimodal.

Figure 5 plotting the likelihood function against various values of b is informative. Though we have generated the data on death times with $b = 0.001$, the $-logL$ attains minimum at about $b = 0.04$. The random influence of the process is clearly brought out here; the coefficient of "noise" or disturbance cannot be estimated properly. Without being aware of this, one could easily make wrong inference about the random influence with the estimate of b. What is important to note here, however, is that the other parameters of the dynamics (drift and regression) are *less* sensitive to the random influence.

With these preliminary observations at hand, the minimization program ZXLSF of the IMSL library is found to yield the estimates presented in Table 2. The minimization has been carried out for each parameter with the other three fixed, for various trial initial values and BOUNDs (required by the ZXLSF). The estimates are in general very good.[5]

It is worth emphasizing here that the parameter μ is the most important in this analysis, as it is the mortality parameter (or epidemiological coefficient), while α_0 and α_1 are physiological coefficients and b is the noise coefficient. From this experiment, we learn that if the BOUNDs are specified properly, μ is very well estimated; so too, α_0 and α_1. But b is always difficult to estimate uniquely, it has to

[5] The first few trials with the ZXLSF indicated that the estimation depends on the specification of the variable BOUND in the program. This BOUND has to be specified properly along with the initial entry for the parameter to be estimated, such that $x_0 - BOUND \leq X \leq x_0 + BOUND$.

Table 2. Estimates of the parameters a_0, a_1, b and μ for given initial entries and BOUNDs.

Variable parameter	Initial value specified	BOUND	Estimate	Comments
μ	0.1	5.0	1.02943	A proper specification of BOUND yields very good estimate of μ, whatever be the initial value
	0.5	5.0	1.02915	
	1.0	5.0	1.03135	
	1.6	5.0	1.0265	
	2.0	1.0	1.02991	
	3.0	2.0	1.03009	
	4.0	3.5	1.03113	
	5.0	4.5	1.02915	
	10.0	9.5	1.02937	
a_0	0.1	1.0	0.0204	The estimate of a_0 fluctuates around the true minimum point for different initial values and BOUNDs. Higher initial values and BOUNDs lead surprisingly to symmetrically negative estimates. This has to be examined.
	0.5	1.0	0.0178	
	1.0	1.0	0.0235	
	2.0	2.0	0.0172	
	5.0	6.0	-0.0186	
	5.0	7.0	-0.0224	
	0.001	1.0	-0.0171	
a_1	-1.5	1.495	-0.0968	a_1 is estimated well for different BOUNDs, provided initial values are negative. cf. Figure 4.
	-2.0	1.999	-0.1044	
	-2.0	2.5	-0.1017	
	-2.0	2.9	-0.1035	
	-1.0	2.9	-0.1025	
	-1.0	3.9	-0.1025	
	-1.0	5.0	-0.1025	
	-1.0	7.0	-0.1025	
b	0.1	1.0	0.0276	The estimate of b is very different from what Figure 5 indicates. Whatever be the initial values and BOUNDs, estimate of b seems to converge to 0.027.
	0.01	1.0	0.0272	
	0.01	0.1	0.0277	
	0.01	0.5	0.0277	
	0.01	10.0	0.0277	
	0.0001	0.1	0.0242	

be estimated through some exogenous studies, if they are available (see the comments in Table 2).

Case 2: *physiological variables partially observed at fixed times*

With the physiological variables observed at fixed times, the same formulations hold good except for a minor adjustment in the calculation of m_t and γ_t values and an additional conditional term in the likelihood function. Let us assume that the data on physiological variables have been observed at every fifth year.

$$\bar{\mu}(\tau_i, \hat{Y}_{\tau_i}) = \mu(m^2_{\tau_i} + \gamma_{\tau_i}) \tag{4''}$$

and the likelihood function becomes

$$L = \bar{\mu}(\tau_i, \hat{Y}_{\tau_i}) e^{-\int_0^{\tau_i} \bar{\mu}(s, \hat{Y}_s) ds} \prod_{j=2}^{\tau_i} f(Y_{t_j} | \hat{Y}_{t_{j-1}}) \tag{24}$$

where

$$f(Y_{t_j} | \hat{Y}_{t_{j-1}}) = e^{-\frac{(Y_{t_j} - m_{t_j-1})^2}{2\gamma_{t_j-1}}} \ . \tag{25}$$

Note the difference in the suffices t_{j-1} and $t_j - 1$. In other words, it is a jump process at every 5-th year, and f above has the mean and variance of the year previous to the observed one.

The estimation procedure can be carried out in the same way as before, for each parameter with the others fixed or for all four parameters together. For simplicity, only the estimates carried out for all 4 parameters together with the program ZXMIN of the IMSL library are given below. The minimization has been carried out with the initial values for a_0, a_1, b and μ equal to 0.05, -0.05, 0.05, 0.5 respectively and with parameter IOPT = 2 (causing ZXMIN to compute the diagonal values of the Hessian matrix). The estimates are:

(1) $a_0 = 0.01959$

(2) $a_1 = -0.0922$

(3) $b = 0.0483$

(4) $\mu = 0.9155$

The estimates are quite close to the values we expect to obtain. Note also that the estimate of b is, as in Figure 5, at about 0.04.

Case 3: *one covariate has been observed*

When some additional informations on covariates which affect mortality are available, it is easy to extend the above formulations. For convenience, let us consider one dichotomous covariate whose exponential parameter is β. To start with, death times have to be simulated with the inclusion of the covariate parameter, and thus the algorithm given for data simulation has to be modified to read $\sum_0^t \mu e^{\beta X} Y_s^2$ for finding the death time τ of an individual. With these data on death times, the estimation can be carried out as before. The likelihood function then will be as follows:

$$L = \prod_{i=1}^{N} \bar{\mu}(\tau_i, \beta) e^{-\int_0^{\tau_i} \bar{\mu}(s, \beta) ds} \tag{26}$$

where

$$\bar{\mu}(\tau_i, \beta) = \mu e^{\beta X_i} (m_{\tau_i}^2 + \gamma_{\tau_i}) \quad . \tag{27}$$

And hence,

$$L = \prod_{i=1}^{N} \mu e^{\beta X_i} (m_{\tau_i}^2 + \gamma_{\tau_i}) e^{-\int_0^{\tau_i} \mu e^{\beta X_i} (m_s^2 + \gamma_s) ds} \tag{28}$$

with

$$\frac{dm_t}{dt} = a_0 + a_1 m_t - 2\gamma_t \mu e^{\beta X} m_t \tag{29}$$

$$\frac{d\gamma_t}{dt} = 2a_1 \gamma_t + b^2 - 2\gamma_t^2 \mu e^{\beta X} \quad . \tag{30}$$

If we maximize (or minimize) the log$L$ (or $-logL$) with $\beta = 0$, we get the estimate for μ. If we maximize (or minimize) with $\beta \neq 0$, we would get the estimates for both μ and β.

DISCUSSION AND POINTERS TO FURTHER APPLICATIONS

In a rather simple (simulated) application of the Gaussian diffusion process model which takes into account the influences of both the observed and unobserved variables on mortality, the results are encouraging especially with respect to the unique estimation of the parameters of the underlying and yet unknown (possibly partially known) processes. It is a simple experiment in that the parameters of the processes have been held constant over time.

The next complicated application would be to let the parameters change over time t as described in equation (1). This would obviously require a more sophisticated minimization or maximum likelihood program than those used here. The author is aware of an efficient computer program for minimization - MINUIT from the CERN library. This program executes the minimization through sophisticated procedures (of Monte Carlo, Nelder and Mead and Fletcher). The program also offers an error analysis through covariance matrix and confidence intervals and a multiplicity of possibilities for fixing and varying the parameters at one's pleasure. If such a program is available, more complicated experiments on varying parameters over time and both X_t and Y_t evolving over time could be carried out. But what is presented here is basic to any experiment, whether simulated or not, on the Gaussian model. What has been achieved through this simple experiment shows that the model can be a powerful tool for survival analysis.

If other studies indicate a functional form for $\overline{\mu}_t$, that is, the nature of the functional dependence of population hazard rate on the means and variances of Y_t, this functional form could be better utilized for purposes of forecasting. For example, the functional form of $\overline{\mu}_t$ has often been confirmed to be a Gompertz function especially for older ages. If this knowledge could be exploited, it would undoubtedly make forecasts of health changes and mortality more precise.

Various other exogenous factors and their effect on mortality forecasts, such as environmental pollution, economic condition etc., may be included in the model. Very often these exogenous forces are observable. And Yashin, Manton and Stallard (1985) indicate the way this could be utilized in the model by extending the vector Y_t to include such factors.

Much more, the endogenous component Y_t could be modelled as dependent on the exogenous components. For example, consider the climatic factors influencing human mortality. These variables (sometimes referred to as "state variables") on climatic changes, economic conditions etc., would form the exogenous component in the vector Y_t. And changes in Y_t could be dependent on these exogenous components. For example, these factors could have greater impact on certain population groups, say, older people. In such a case, interaction between these factors and age would have to be included in the hazard function. This implies a more rapid selection of certain groups under certain changes in these factors. However, the (linear) dynamic equations describing the changes in these factors (these perhaps could be gathered from econometrics and other disciplines) would *not* be affected by selection so that no further modification of the time series (auto-regressive) equations for these factors would be necessary, as the added stochasticity is not due to diffusion process describing Y_t but to the stochasticity of these factors alone.

The Gaussian diffusion process model is so general that it can be applied to any type of survival analysis. Since Y_t follows Gaussian, modifications of Y_t could afford a variety of applications in many fields. Consider, for example, the nuptiality. Hernes (1972) in developing a (diffusion) model to describe the process of entry into marriage estimated the unobservable "marriageability" at the population level (he called it the "initial average marriageability"). Rajulton (1985) made use of this model in the context of estimating the parameters of the first passage probabilities of a semi-Markov model and extended the concept to divorceability and remarriageability as well. Now, the unobserved heterogeneity in marriageability of the single or the divorceability of the married could easily be incorporated into the Gaussian model, as much as these unobservables are particular cases of the Gaussian diffusion process. In particular, if Y_0 follows $N(m_0, \gamma0)$, then it is well known that Y^2 follows a Γ distribution (Kendall and Stuart 1977). This Γ distribution could be used profitably, as Vaupel and Yashin (1982) have shown, to study the influence of the varying marriageability or divorceability over time (in contrast to the constant decline in the Hernes model). Further, if certain covariates are found to influence this unobserved heterogeneity (for example, employment status or the number of children in the case of divorce), they could also be included in the model and the heterogeneous reality could be captured more precisely.

The Gaussian diffusion process model proves to be so useful a tool for further applications that it is certain that its extensive application in many fields would throw more light on the heterogeneity dynamics recently born.

References

Cox, D.R. and D. Oakes (1984) *Analysis of Survival Data*. New York: Chapman and Hall.

Heckman, J.J. and B. Singer (1982) Population heterogeneity in demographic models. Pages 567-599 in *Multidimensional Mathematical Demography*, edited by K. Land and A. Rogers. New York: Academic Press.

Hernes, G. (1972) The process of entry into first marriage. *American Sociological Review* 37:173-182.

Kendall, M. and A. Stuart (1977) *The Advanced Theory of Statistics, Vol.I, Distribution Theory*. London: Charles Griffin & Co. Ltd.

Miller, R.G. et al. (1981) *Survival Analysis*. New York: John Wiley and Sons.

Rajulton, F. (1985) Heterogeneous marital behaviour in Belgium, 1970 and 1977: An application of the semi-Markov model to period data. *Mathematical Biosciences* 73:197-225.

Ridder, G. and Verbakel, W. (1983) On the estimation of the proportional hazards model in the presence of unobserved heterogeneity. Unpublished paper, University of Amsterdam, Faculty of actuarial science and econometrics.

Rubinstein, R.Y. (1981) *Simulation and the Monte-Carlo Method*. New York: John Wiley and Sons.

Tuma, N.B. and M.T. Hannan (1984) *Social Dynamics, Models and Methods*. New York: Academic Press.

Vaupel, J.W., K.G. Manton, and E. Stallard (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 16:439-454.

Vaupel, J.W. and A.I. Yashin (1985) The deviant dynamics of death in heterogeneous populations. *Sociological Methodology 1985*, edited by Nancy Tuma. San Francisco: Jossey-Bass.

Woodbury, M.A. and K.G. Manton (1977) A random walk model of human mortality and aging. *Theoretical Population Biology* 11:37-48.

Yashin, A.I. (1984) Dynamics in survival analysis: conditional Gaussian property versus Cameron-Martin formula. WP-84-107. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Yashin, A.I. (1985) Survival analysis with partially observed covariates: discrete time. Unpublished paper.

Yashin, A.I. and K.G. Manton (1984) Evaluating the effects of observed and unobserved diffusion processes in survival analysis of longitudinal data. CP-84-58. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Yashin, A.I., K.G. Manton, and E. Stallard (1985a) Evaluating the effects of observed and unobserved diffusion processes in survival analysis of longitudinal data. Unpublished paper.

Yashin, A.I., K.G. Manton, and E. Stallard (1985b) *The Propagation of Uncertainty in Human Mortality Processes*. CP-85-36. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Yashin, A.I., K.G. Manton, and J.W. Vaupel (1985) Mortality and aging in a heterogeneous population: a stochastic process model with observed and unobserved variables. *Theoretical Population Biology* 27:154-175.

Appendix: Computer Program

```
      program simul1
c   -----------------------------------------------------
c    this is a program for simulating the times to death of
c    500 individuals and the unobserved Y(t) values, the
c    maximum death time being allowed is 150 years.
c
c    Remarks: n = no. of individuals considered
c             n1= value of maximum death time allowed
c
c      a maximum of 500 individuals and 150 years (death time) is given
c      in the program. Modify the dimension specification if necessary.
c     The parameter values are to be specified at the beginning
c     of the program on the terminal. The parameters are a(0), a(1),
c     b, mu, g(0), delta.
c
c     The seed for simulation is taken to be 151245. Anyother
c     seed, which is odd and less than 9 digits can be used.
c
c     The death times on output are stored in tape8 in the format
c     format(3x,10(i3,5x)). Tape7 contains the sample of individuals'
c     r, -log(1-r), y(0), eps, and t. Tape9 stores the data on trajectory
c     of 20,40,60,80,81st individuals for plotting if desired.
c
c   -----------------------------------------------------
c
      real  eps(150),y1(500,25),y(150),y2(150),y0(500),r(500),r1(500)
      real mu, var(150), m(150),m0
      integer t(500)
      call usearg
      write(6,11)
   11 format(42hgive parameter values, a0,a1,b,mu,g0,delta)
      read(5,*) n, n1, a0, a1, b, mu, g0, delta
      m0 = a0/(abs(a1))
      write(8,12) a0,a1,b,mu,m0,g0,delta
   12 format(23hwith the initial values,/23(1h-),/4ha0 =,1x,f8.4,2x,
     $ 4ha1 =,1x,f8.4,2x,3hb =,1x,f8.4,/4hmu =,1x,f8.4,2x,4hm0 =,1x,
     $ f8.4,2x,4hg0 =,1x,f8.4,/7hdelta =,1x,f8.4)
      write(8,66)
c
c   -------------------------------------------------------
c    calculate the mean and variance of the conditional distn
c    of y. cf. text
c   -------------------------------------------------------
c
      do 7 i=1,n1
      if(i.eq.1) m(i)=(a0*delta) +((1+a1*delta)*m0) -(2*g0*mu*delta*m0)
      if(i.eq.1) var(i)=((1+2*a1*delta)*g0) +((b**2)*delta)
     $   -(2*(g0**2)*mu*delta)
      if(i.ne.1) m(i)=(a0*delta) + ((1+a1*delta)*m(i-1))
     $     -(2*var(i-1)*mu*m(i-1)*delta)
      if(i.ne.1) var(i)= ((1+2*a1*delta)*var(i-1)) +((b**2)*delta)
     $    -(2*(var(i-1)**2)*mu*delta)
    7 continue
c   -----------------------------------------------------
      ix=151245
      do 1 i=1,n
c   -------------------------------------------------------
c    generate the uniform random variate r(i) and find r1(i)=
c    -alog(1-r(i)).
c
c   -------------------------------------------------------
c
      call randu(ix,iy,yf1)
      r(i)=yf1
      r1(i)=-alog(1-r(i))
c   --------------------------------------------------------
c     generate the initial random variate y0(i) ~N(m0,g0)
c   --------------------------------------------------------
```

```
c
      am = m0
      s = sqrt(g0)
      call gauss(ix,s,am,v)
      y0(i)=v
      yy=y0(i)**2
      do 2 k=1,n1
c ------------------------------------------------------------
c    generate the error random variate eps(i) ~ N(0,1)
c ------------------------------------------------------------
c
      am=0.
      s=1.
      call gauss(ix,s,am,v)
      eps(k)=v
c -----------------------------------------------------------
c    find the sequence of y variates given the initial y0
c    and the sum of squares of y to examine whether this sum
c    becomes greater than r1(i) for individual i at what time
c    to death
c -----------------------------------------------------------
c
      if(k.eq.1) y(k) =(a0*delta)+((1+a1*delta)*y0(i))+(b*delta*eps(k))
      if(k.ne.1) y(k) =(a0*delta)+((1+a1*delta)*y(k-1))+(b*delta*eps(k))
      w=k/5.0
      j=k/5
      if(w.eq.j) y1(i,j)=y(k)
      if(w.eq.j) write(10,65) i,k,j,y1(i,j)
      y2(k) = y(k)**2
      yy = yy + y2(k)
      if(yy.gt.r1(i)) t(i)=k
      if(k.eq.150.and.yy.lt.r1(i)) t(i)=k
      if(yy.gt.r1(i)) go to 3
    2 continue
    3 continue
    1 continue
      write(7,64) (t(i),i=1,n)
      do 99 i=1,n1
      w=i/5.0
      j=i/5
      if(w.eq.j) m(i)=y1(81,j)
      write(11,*) m(i)
   99 continue
   64 format(3x,10(i3,2x))
   66 format(1x)
   65 format(i3,2x,i3,2x,i3,2x,f10.6)
      stop
      end
c -------------------------------------------------------
c     subroutine gauss
c  purpose : to compute a numerically distributed
c            random number with a given mean and
c            standard deviation
c  usage   : call gauss(ix,s,am,v)
c  parameters:  ix - ix must contain an odd integer
c            number with 9 or less digits on the
c            first entry to gauss. thereafter it'll
c            contain a uniformly distributed integer
c            random number generated by the subrout
c            for use on the next entry to the subr.
c  remarks : this subr. uses randu which is machine
c            specific
c  method  : uses 12 uniform random numbers to compute
c            normal random numbers by central limit
c            theorem. the result is then adjusted to
c            match the given mean and s.d. the uniform
c            random numbers computed within the subr.
c            are found by the power residue method.
```

```
c-----------------------------------------------------
      subroutine gauss(ix,s,am,v)
      a=0.0
      do 50 i=1,12
      call randu(ix,iy,y)
      ix=iy
   50 a=a+y
      v=(a-6.0)*s+am
      return
      end

c-----------------------------------------------------
c         subroutine randu
      subroutine randu(ix,iy,yfl)
      iy=ix*65539
      if(iy)5,6,6
    5 iy=iy+2147483647+1
    6 yfl=iy
      yfl=yfl*.465613e-9
      return
      end
      program simui2
      external func
      common/data2/time(100),m0,a0,a1,b,g0,delta
      real m0
      integer maxfn,ier,time
      data step/0.1/,xacc/.01/,maxfn/100/,
     $     a0/0.02/,a1/-0.1/,b/0.001/,g0/0.001/,delta/1./
      call usearg
      m0 = a0/(abs(a1))
      write(6,11)
   11 format(41hgive the initial value of x and the bound)
      read(5,*) x, bound
      read(3,10) (time(i),i=1,100)
   10 format(10(i3,2x))
      call zxlsf(func,x,step,bound,xacc,maxfn,ier)
      write(6,*) ier, x
      stop
      end

      real function func(x)
      real m(125),m2(125),var(125),b1(100),sum(100)
      common/data2/time(100),m0,a0,a1,b,g0,delta
      real x,m,m2, m0
      integer time
      f=0.
      do 2 i=1,100
      n1=time(i)
c  -----------------------------------------------------------
c
      do 7 k=1,n1
      if(k.eq.1) m(k)=(a0*delta)+((1+a1*delta)*m0)-(2*g0*x*delta*m0)
      if(k.eq.1) var(k)=((1+2*a1*delta)*g0) +((b**2)*delta)
     $   -(2*(g0**2)*x*delta)
      if(k.ne.1) m(k)=(a0*delta) + ((1+a1*delta)*m(k-1))
     $    -(2*var(k-1)*x*m(k-1)*delta)
      if(k.ne.1) var(k)= ((1+2*a1*delta)*var(k-1)) +((b**2)*delta)
     $   -(2*(var(k-1)**2)*x*delta)
      m2(k) = m(k)**2
      if(var(k).lt.0) var(k)=0.
    7 continue
      sum(i)=m0**2 + g0
      do 4 k=1,n1
    4 sum(i)=sum(i) + x*(m2(k)+var(k))
      b1(i) =   -(alog(x * (m2(n1) + var(n1))) - sum(i))
      f = f + b1(i)
    2 continue
      func=f
      return
      end
```