A FRAMEWORK THAT INCORPORATES REPEATED MEASUREMENTS INTO THE HAZARD

by

Jeffrey Joseph Gaynor

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1436

May 1983

A FRAMEWORK THAT INCORPORATES REPEATED MEASUREMENTS INTO THE HAZARD

by

Jeffrey Joseph Gaynor

A Dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill

1983

Approved by:

M. J. Symons
Advisor

Kenny L. Lee
Reader

James E. Wright
Reader

JEFFREY JOSEPH GAYNOR. A Framework that Incorporates Repeated Measurements into the Hazard (Under the direction of Michael J. Symons).

By allowing covariates to remain constant between consecutive measurements over time, the hazard is able to incorporate repeated measurements in addition to intervening event and baseline information. Within this context, proportionality receives a broader definition. Although an individual's hazard at time t may depend on the values of his most recent measurements, it may also depend on a particular piece of his covariate information that was measured earlier. A newly defined covariate that describes this "past history" information may then be included into the hazard. Examples are provided.

A general parametric representation of the hazard is proposed in which an underlying failure time distribution exists, and the model parameters are expressed as simple functions of the covariates. The multiplicative model is a reduced form of this more general model, and nonproportionality is given interpretability. Nonparametric estimation of the cumulative hazard as a function of covariate strata is used in determining an appropriate parametric form for the underlying hazard. Specifically, Breslow's (1972) piecewise exponential hazard is fitted to each covariate stratum, and the focus is upon the slopes of these cumulative hazard curves.

The proposed framework for incorporating repeated measurements into the hazard was applied to two data sets: a prospective follow-up study of heart diseased patients at Duke University and a

retrospective follow-up study of chrysotile asbestos textile workers. Measurements were repeated on an individual's congestive heart failure status in the first study, and the workers' annual exposures to chrysotile asbestos comprised the repeated measurements in the second study. The dose-response effect of a worker's cumulative exposure to chrysotile asbestos on his hazard of death from lung cancer is of particular interest. The results of the data analyses are presented.

ACKNOWLEDGEMENTS

I first want to thank my family for their steadfast support, faith, and love: my parents Milt and Lenore, my siblings, Dave, Rob, and Steph, and my grandfather Pop-Pop. I also want to thank my friends - Roger, Parke, and others - who provided much needed booster shots of morale for me from time to time.

I want to thank my committee members for their diligent time and effort that was invested into the molding of my dissertation: Professors Mike Symons, Regina Elandt-Johnson, Kerry Lee, Jim Grizzle, and Gerardo Heiss. I especially want to thank Mike, my advisor. His patience, persistent critiques, and expressed enthusiasm in my work were almost always well timed and well received. I also want to thank Professors Harvey Checkoway and John Dement for their assistance.

I want to thank the people at the Computation Center for their help over the years: Bob, Betty, Jim, Molly, Barb, two Waynes, and others. I also want to thank my typist, Ernestine Bland, for a finished product that is very pleasing to the eye. In addition, I want to thank Ruth Bahr for putting on some finishing touches.

Finally, I want to thank my therapist Dr. Laura Lowenbergh, as well as Dr. Grizzle; both have been my mentors. In the four years that it has taken me to complete this dissertation, I have come to understand more deeply the value of self-introspection.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION AND A REVIEW OF THE LITERATURE

1.1 Introduction

Epidemiological follow-up studies and clinical trials are commonly used by environmental and medical researchers. In these studies, each individual is observed until he fails from some cause, has become lost to follow-up, or has survived through the duration of the study. Age at death, age at disease onset, the length of time since entry into the study, the length of time since initial employment, and the length of time in remission are examples of the time-to-failure random variables that are of interest to these investigators.

One of the main purposes of an epidemiological follow-up study is to draw biologically sound conclusions about the effects of certain covariate information on individuals' patterns of failure. Only one type of failure may receive attention, for example, death from all causes. However, the covariates' effects on more than one cause-specific failure pattern may be determined.

In assessing the effect of an individual's treatment on survival, one may need to control for the confounding of that effect with extraneous factors. For example, if the individual's age is a predictor of one's failure time, then an adjustment for different age distributions

in the treatment groups may be required in order to validate any comparison made between the treatments. Even if the age distributions are identical among the treatment groups, but age is related to the outcome, incorporating it into the analysis should improve the precision of the treatment comparisons.

The use of survival models to aid in the interpretation of epidemiological follow-up data is advantageous for three reasons:

i) The estimated time-to-failure distribution describes the pattern of failure more precisely than an estimated probability of failure within one selected time period.

ii) A completely specified survival model has a descriptive advantage over nonparametric methods if it describes the data reasonably well using a minimum number of parameters, where each parameter in the model has a distinct interpretation.

iii) The stratification of continuous covariates is not a prerequisite, which is the case in a standardized mortality ratio analysis.

Until recently, survival analysis has been limited to the use of baseline covariates. Repeated measurements on individuals over time may yield additional information about their states of health during follow-up. The purpose of this dissertation is to investigate a methodology that incorporates repeated measurements into a survival analysis.

1.2 The Hazard Function

Let T be a nonnegative random variable that represents failure time. The probability that an individual survives past time t is called the survival function, and it is denoted by

$$S(t) = Pr(T > t), \quad t \geq 0.$$

The cumulative distribution function of T is then

$$F(t) = 1 - S(t).$$

The density function and the hazard function (the instantaneous rate of failure) for an individual at time t are represented by f(t) and $\lambda(t)$, respectively. They are defined by

$$f(t) = \lim_{\Delta t \to 0} \left\{ \frac{Pr(t \leq T \leq t + \Delta t)}{\Delta t} \right\} = \frac{-\partial S(t)}{\partial t}$$

and

$$\lambda(t) = \lim_{\Delta t \to 0} \left\{ \frac{Pr(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \right\} = \frac{-\partial \log S(t)}{\partial t} . \quad (1.1)$$

The functional form of S(t), f(t), or $\lambda(t)$ determines the other two functions. From (1.1) one obtains

$$S(t) = \exp \left\{ - \int_0^t \lambda(u) du \right\} , \quad (1.2)$$

and the integral of the hazard,

$$\Lambda(t) = \int_0^t \lambda(u) du, \quad (1.3)$$

is called the cumulative hazard function.

One fits a distribution to T via the hazard or cumulative hazard

because they are usually expressed by simpler mathematical forms than either the density or survival function. In addition, the concept of "the failure rate within a short time interval" is easy to comprehend.

The exponential, Gamma, Gompertz, Rayleigh, and Weibull distributions are useful in survival analysis, and their descriptions appear in Gross and Clark (1975).

Throughout this dissertation the terms location, scale, and shape parameters will appear. The standard definitions of these parameters, given by Elandt-Johnson and Johnson (1980), denote scale to be a parameter that multiplies the stochastic variable T, location to be a parameter that shifts T, and shape to be any parameter that is not a scale or location parameter. However, in this dissertation a scale parameter will be defined as one that multiplies the hazard and a shape parameter will be defined as one that gives functional form to the hazard. These alternative definitions allow for a more clear description of the hazard as a function of time. For example, consider the following four parameter hazard,

$$\lambda(t) = \alpha + \beta\, e^{\gamma(t-\delta)}.$$

Here, α, β, and γ are the location, scale, and shape parameters for $\lambda(t)$, and δ is the location parameter for T.

1.3 The Hazard as a Function of Baseline Covariates

Assume that s baseline covariates, i.e., concomitant information determined at the beginning of follow-up, are available for each individual. Denote z_u as the value of the u^{th} baseline covariate,

$u=1,\ldots,s$, and $\underset{\sim}{z} = (z_1,\ldots,z_s)'$. The hazard function, $\lambda(t;\underset{\sim}{z})$ is then defined conditionally on $\underset{\sim}{z}$. Specifically,

$$\lambda(t;\underset{\sim}{z}) = \lim_{\Delta t \to 0} \left\{ \frac{\Pr(t \leq T \leq t + \Delta t \mid T \geq t, \underset{\sim}{z})}{\Delta t} \right\}. \tag{1.4}$$

Two basic mathematical expressions for $\lambda(t;\underset{\sim}{z})$ have appeared in the literature: the additive model and the multiplicative model. The additive model is defined by

$$\lambda(t;\underset{\sim}{z}) = \lambda_0(t) + \sum_{u=1}^{s} h_u(t)\, g_u(z_u;\beta_u), \tag{1.5}$$

where $\lambda_0(t)$ and $h_u(t)$, $u = 1,\ldots,s$, are functions of t but not of $\underset{\sim}{z}$, and $g_u(z_u;\beta_u)$, $u = 1,\ldots,s$, do not depend on t. The multiplicative model is defined by

$$\lambda(t;\underset{\sim}{z}) = \lambda_0(t)\, g(\underset{\sim}{z};\underset{\sim}{\beta}), \tag{1.6}$$

where $\lambda_0(t)$ does not depend on $\underset{\sim}{z}$ and $g(\underset{\sim}{z};\underset{\sim}{\beta})$ does not depend on t. In (1.5) and (1.6), $\underset{\sim}{\beta} = (\beta_1,\ldots,\beta_s)'$, and the parameter β_u denotes the effect of the u^{th} covariate on the hazard, $u = 1,\ldots,s$. The expressions for $\lambda(t;\underset{\sim}{z})$ are defined in such a way that $\lambda(t;\underset{\sim}{z}=\underset{\sim}{z}_0) = \lambda_0(t)$ at some reference vector $\underset{\sim}{z}_0$. As a consequence, $\lambda_0(t)$ is known as the underlying hazard.

Elandt-Johnson (1980) depicts the additive form as being consistent with the standard competing risks theory since the overall hazard is expressed as a sum of cause-specific hazards. An increase in a cause-specific hazard causes an additive increment in the overall hazard. Similarly, under model (1.5), if a particular covariate is

positively associated with the hazard, then an increase in its value will cause an additive increment in the hazard.

The multiplicative model infers proportionality, i.e., the ratio of any two individuals' hazards with baseline values $\underset{\sim}{z}_i$ and $\underset{\sim}{z}_{i'}$,

$$\frac{\lambda(t;\underset{\sim}{z}_i)}{\lambda(t;\underset{\sim}{z}_{i'})} ,$$

will not depend on time t. Under the additive model, individuals will have proportional hazards only when $h_u(t) = \alpha_u \lambda_0(t)$, $u = 1,\ldots,s$, α_u being a proportionality constant.

Simple examples of the additive and multiplicative models were initially given by Feigl and Zelen (1965). They used an exponential underlying hazard, $\lambda_0(t) = \lambda$, and a multiplicative model with one covariate using $g(z;\beta) = (1+\beta z)^{-1}$. Based on this hazard, the expected time to failure is a simple linear function of z. Feigl and Zelen (1965) had also suggested the use of the additive model $\lambda(t;z) = \lambda+\beta z$, and the multiplicative model $\lambda(t;z) = \lambda \exp(\beta z)$, which Glasser (1967) extended by using two covariates. Cox (1972) then generalized the multiplicative model to s covariates by

$$\lambda(t;\underset{\sim}{z}) = \lambda_0(t) \exp\{\underset{\sim}{\beta}'\underset{\sim}{z}\} . \qquad (1.7)$$

The form $g(\underset{\sim}{z};\underset{\sim}{\beta}) = \exp\{\underset{\sim}{\beta}'\underset{\sim}{z}\}$ is appealing in that the hazard remains nonnegative for any choice of $\underset{\sim}{\beta}$.

The paper by Cox (1972) has stimulated a large amount of research in survival analysis. Cox assumed the multiplicative model in (1.7), but $\lambda_0(t)$ was not specified parametrically. The estimation of $\underset{\sim}{\beta}$ without knowledge of $\lambda_0(t)$ was performed by maximizing a portion of

the likelihood function which did not contain $\lambda_0(t)$. Cox (1975) defined this portion as a partial likelihood, and its description appears in Section 1.6.

Cox's (1972) partial likelihood technique of estimating $\underset{\sim}{\beta}$ in (1.7) has four attractive features:

i) The estimation of the effects of the covariate vector $\underset{\sim}{z}$ is made without having to completely specify the parametric form of the hazard.

ii) The effects of multiple covariates on an individual's hazard can be addressed in a computationally straightforward manner.

iii) Cox (1975) showed that the usual large sample properties of the maximum likelihood procedure apply to the partial likelihood, even when time dependent covariates of the form $z(t) = h(t)z$ are used ($h(t)$ is a specified function of t).

iv) Given that individuals' hazard rates are proportional, good efficiency has been demonstrated for Cox's model (1.7), where $\lambda_0(t)$ is left unspecified. Kalbfleisch (1974) assumed exponential and Weibull distributions for $\lambda_0(t)$ in (1.7), using a zero-one covariate for z. A more general form for $\lambda_0(t)$ that included these two cases was considered by Efron (1977). In those examples, the asymptotic relative efficiency of Cox's partial likelihood estimate of β, in comparison with the maximum likelihood estimate of β

obtained from the complete likelihood function, remained close to one for "practical" values of β, i.e., for $|\beta| < 1.15$.

Breslow (1972,1974) proposed a piecewise exponential underlying hazard in model (1.7). Denoting $0 < t_{(1)} < t_{(2)} < \ldots < t_{(r)}$ to be the r distinct failure times, he defined

$$\lambda_0(t) = \begin{cases} \lambda_j, & t \in (t_{(j-1)}, t_{(j)}], \quad j=1,\ldots,r \\ 0, & t > t_{(r)}, \end{cases} \tag{1.8}$$

where $t_{(o)} = 0$. This saturated representation of the underlying hazard is nonparametric in that r parameters are used to describe r distinct failure times. Many functional forms for $\lambda_0(t)$ are closely approximated by (1.8) when the lengths of the time intervals $(t_{(j-1)}, t_{(j)}]$, $j=1,\ldots,r$, are short. This will occur as the number of distinct failure times in the cohort becomes large. Using models (1.7) and (1.8), Breslow (1972,1974) obtained a likelihood expression for $\underset{\sim}{\beta}$ similar to that of Cox (1972). In fact, Breslow's incomplete likelihood, described in Section 1.6, is often used as an approximation to Cox's partial likelihood because of its simpler formulation in the presence of tied failure times. Without tied failure times, the two expressions become equivalent.

Kalbfleisch and Prentice (1973) demonstrated that a marginal likelihood of the failure times' ranks is equivalent to Cox's partial likelihood, even in the presence of censoring. If there are tied failure times, then the two approaches yield different likelihoods. Since only Cox's partial likelihood allows the inclusion of time

dependent covariates, the marginal likelihood approach of Kalbfleisch and Prentice will receive no further attention in this dissertation.

One drawback to Cox's assumption of a multiplicative model is that it may be too restrictive in some cases. If individuals' hazards, as functions of baseline covariate information, show a distinct departure from proportionality, then a hazard function that can explain the existing type of nonproportionality is required. To cover the nonproportional case, Cox's model (1.7) generalizes to

$$\lambda(t;\underset{\sim}{z}) = \lambda_0(t)\ \exp\{\underset{\sim}{\beta}_1'\underset{\sim}{z} + (\underset{\sim}{\beta}_2'\underset{\sim}{z})h(t)\}\ , \qquad (1.9)$$

where $h(t)$ is some monotonic function of t. This hazard reduces to the multiplicative model when $\underset{\sim}{\beta}_2 = \underset{\sim}{0}$. Kalbfleisch and Prentice (1980) generalized $(\underset{\sim}{\beta}_2'\underset{\sim}{z})h(t)$ in (1.9) to become

$$\sum_{u=1}^{s} \beta_{2u} z_u\ h_u(t),$$

thereby assigning a distinct nonproportionality function, $h_u(t)$, to the covariate z_u, $u=1,\ldots,s$.

In an example using one covariate, a zero-one variable designating the type of treatment for leukemia patients, Cox (1972) chose $h(t) = t-10$, where 10 was a convenient value in the center of the range of the survival times. Kalbfleisch and Prentice (1980), in a similar example using cancer data and a zero-one covariate, chose $h(t) = \log(t)-c$, where c denoted the observed mean log survival time. In both examples, $\beta_2 > 0$ ($\beta_2 < 0$) implies that the ratio of the hazards for sample 1 versus sample 0 is increasing (decreasing) with time. Note that if the underlying hazard is Gompertz (Weibull), then

the test of H_0: $\underset{\sim}{\beta}_2 = 0$ in model (1.9) when $h(t) = t$ ($h(t) = \log(t)$) would test specifically if that underlying hazard's shape parameter depends on $\underset{\sim}{z}$. Finally, in a study of multiple drug chemotherapy on mice injected with leukemia cells, Stablein, Carter, and Wampler (1980) replaced $(\underset{\sim}{\beta}_2'\underset{\sim}{z})h(t)$ in (1.9) with $(\underset{\sim}{\beta}_2'\underset{\sim}{z})t + (\underset{\sim}{\beta}_3'\underset{\sim}{z})t^2$, where $\underset{\sim}{\beta}_1'\underset{\sim}{z}$ contained quadratic functions of dose. Their model allowed for the possibility that low dose therapy prolongs survival, while aggressive treatment may increase the chance for longer survival or cure, but at a much greater risk of early toxic death.

If the multiplicative model (1.7) is only satisfied within particular strata of individuals, then the hazard for individuals belonging to the g^{th} stratum may be defined by

$$\lambda_g(t;\underset{\sim}{z}) = \lambda_{og}(t)\exp\{\underset{\sim}{\beta}_g'\underset{\sim}{z}\}, \quad g=1,\ldots,h,$$

where h denotes the number of strata. If it is found that $\underset{\sim}{\beta}_1 = \underset{\sim}{\beta}_2 = \ldots = \underset{\sim}{\beta}_h$, then the covariate vector's effects are consistent across the h strata.

Taulbee (1977,1979) proposed a different hazard for representing nonproportionality. He incorporated baseline covariate information into a generalized Rayleigh hazard function by

$$\lambda(t;\underset{\sim}{z}) = \sum_{k=0}^{m} \lambda_k \, h(\underset{\sim}{z};\underset{\sim}{\beta}_k)t^k, \qquad (1.10)$$

where m is some nonnegative integer. In his examples, Taulbee used $h(\underset{\sim}{z};\underset{\sim}{\beta}_k) = \exp\{\underset{\sim}{\beta}_k'\underset{\sim}{z}\}$. This hazard is a completely specified parametric form but with flexibility in that the maximum polynomial degree, m, is suitably determined from the data analysis. If $\underset{\sim}{\beta}_0 = \underset{\sim}{\beta}_1 = \ldots = \underset{\sim}{\beta}_m$,

then $\lambda(t;\underset{\sim}{z})$ becomes a proportional hazards model. Taulbee (1977) remarked about his model, "The assumption of a polynomial hazard is arbitrary and is made in the hope that the true but unknown hazard may be satisfactorily approximated by a polynomial of reasonably low degree."

In an analysis of kidney graft survival, Bailey, Homer, and Summe (1977) used the nonproportional hazard function

$$\lambda(t;\underset{\sim}{z}) = \alpha^*(\underset{\sim}{z}) + \beta^*(\underset{\sim}{z}) \exp\{-\gamma^*(\underset{\sim}{z})t\}, \qquad (1.11)$$

where t denotes the time survived beyond surgery, $\alpha^*(\underset{\sim}{z}) = \alpha_0 \exp\{\underset{\sim}{\alpha}'\underset{\sim}{z}\}$, $\beta^*(\underset{\sim}{z}) = \beta_0 \exp\{\underset{\sim}{\beta}'\underset{\sim}{z}\}$, and $\gamma^*(\underset{\sim}{z}) = \gamma_0 \exp\{\underset{\sim}{\gamma}'\underset{\sim}{z}\}$. An underlying hazard exists at $\underset{\sim}{z} = \underset{\sim}{0}$, and it is defined by

$$\lambda_0(t) = \alpha_0 + \beta_0 e^{-\gamma_0 t}.$$

This distribution was chosen in order to reflect a maximum death rate immediately after kidney transplant surgery which diminishes with time toward a lower bound. The multiplicative model is attained from (1.11) if $\underset{\sim}{\alpha} = \underset{\sim}{\beta}$ and $\underset{\sim}{\gamma} = \underset{\sim}{0}$.

In a recent paper by Hazelrig, Turner, and Blackstone (1982) a generic family of survival functions was defined by a common differential equation. An underlying hazard was then selected from this family, and the model parameters were expressed as functions of the covariates.

These examples describe a general approach for incorporating covariates into the hazard. Namely, an underlying hazard exists, and each parameter in that distribution may be expressed as a simple

function of the covariates. This approach includes many of the models discussed in this section, and it establishes a framework for testing the hypothesis of proportionality. Furthermore, if an appropriate underlying failure time distribution has been selected, then nonproportionality of the hazard may be represented in a most informative manner. For instance, if $\alpha_0 \neq 0$ and $\underset{\sim}{\alpha}=\underset{\sim}{0}$ in model (1.11), then the effects of $\underset{\sim}{z}$ on the hazard are greatest immediately following surgery and diminish with time until they no longer have an effect.

A more detailed discussion of this framework for incorporating covariates into the hazard, along with a nonparametric procedure for determining an appropriate underlying failure time distribution, will be presented in Chapter 2.

1.4 The Incorporation of Repeated Measurements and Intervening Events into the Hazard

Certain covariates change with time, for example, an individual's cholesterol level. When it is of interest to assess the effect of this type of covariate on an individual's hazard, the use of repeated measurements may yield a more powerful analysis. Here, the individual's hazard at time t would depend on his most recent measurement rather than his baseline value.

One may also be interested in the effect of an intervening event's status on the hazard. Any event that occurs for certain individuals after their follow-up has begun is an intervening event. For example, consider the occurrence of a nonfatal myocardial infarction during follow-up. It is possible to include the individual's

infarct history at time t as a covariate in the hazard.

In order that an individual's hazard is defined throughout his follow-up period, his covariates must be defined over that period. To accomplish this with repeated measurements, each covariate is assumed to remain constant between consecutive measurements over time. That is to say, time dependent covariates are defined by the repeated measurements as step functions of time. Specifically, denote τ_0, τ_1, ..., τ_q (τ_0=0) to be the follow-up times that measurements are performed. Assume that individual i has measurements recorded at times $\tau_0,\tau_1,\ldots,\tau_{q_i}$ where τ_{q_i} is the last measurement time prior to his observed follow-up time, t_i. Let z_{ium} be the value of the u^{th} covariate that was recorded for individual i at time τ_m. Then one obtains

$$z_{iu}(t) = z_{ium}, \quad t \in [\tau_m,\tau_{m+1}), \quad m = 0,\ldots,q_i, \tag{1.12}$$

where $t_i = \tau_{q_i+1}$. With this definition of $z_{iu}(t)$, one might consider z_{ium} to approximate the mean value of the i^{th} individual's u^{th} covariate during the interval $[\tau_m,\tau_{m+1})$.

The status of a dichotomous intervening event is simply expressed as

$$z(t) = \begin{cases} 0 & \text{if } t < w \\ 1 & \text{if } t \geq w, \end{cases} \tag{1.13}$$

where w represents the waiting time to the occurrence of that event.

In Farewell's (1979) study of the time to infection following bone marrow transplantation, two time dependent covariates were incorporated into his analysis. One was defined as a function of repeated measurements and the other was defined as the status of an intervening event.

Specifically, an individual's granulocyte level was measured daily. Denoting $z_{1\tau}$ as his value at the beginning of his $(\tau+1)^{st}$ day of follow-up, $z_1(t) = z_{1\tau}$ during that day, i.e., for $t \in [\tau,\tau+1)$. Here one may consider $z_{1\tau}$ to approximate the individual's mean granulocyte level during his $(\tau+1)^{st}$ day of follow-up. The other covariate was defined as $z_2(t) = 1$ if GVHD (graft versus host disease) had occurred for the individual prior to t days of follow-up, 0 otherwise.

The u^{th} covariate for an individual at time t, $z_u(t)$, is defined in this dissertation as a step function of time by (1.12) or (1.13), u=1,...,s. A baseline covariate is defined simply by $z_u(t) = z_{uo}$, for all t. Denote $\underset{\sim}{z}(t) = (z_1(t),...,z_s(t))'$. Then an individual's hazard at time t, conditional on $\underset{\sim}{z}(t)$, is defined by

$$\lambda(t;\underset{\sim}{z}(t)) = \lim_{\Delta t \to o} \left\{ \frac{\Pr[t \leq T \leq t + \Delta t | T \geq t, \underset{\sim}{z}(t)]}{\Delta t} \right\} . \qquad (1.14)$$

One may consider modelling the i^{th} individual's covariate values $z_{iu0}, z_{iu1}, \ldots, z_{iuq_i}$ as a parametric function of time and z_{iu0} prior to their incorporation into the hazard (Elandt-Johnson, 1981). However, in this dissertation only the question of the hazard's dependency on the observed covariate values will be addressed.

Although an individual's hazard at time t may depend on the values of his most recent measurements, it may also depend on a particular piece of his covariate information that was measured earlier. A newly defined covariate that describes this "past history" information may then be included into the hazard. For example, suppose that in a follow-up study of heart disease patients, chest pain status is

measured on individuals once a month. Denoting CP_m to be an individual's chest pain status that is measured at the beginning of his $(m+1)^{st}$ month of follow-up, $CP(t)=CP_m$ for t contained in that month. This covariate represents an individual's mean chest pain status within one month of time t. Now, define an individual's mean chest pain progression status within four months of time t by

$$PR(t) = (CP_m + CP_{m-1} - CP_{m-2} - CP_{m-3})/2,$$

where t is contained in his $(m+1)^{st}$ month of follow-up. This covariate represents the difference between an individual's mean chest pain status within two months of t and his mean chest pain status between two and four months of t. If one presumes that the hazard depends on CP(t) and PR(t), then one might fit a model to $\lambda(t;CP(t),PR(t))$. The covariate PR(t) represents past history information on chest pain status, and testing for its effect on the hazard is analogous to testing for the Markov assumption in a stochastic process.

As another example, consider the paper by Gail (1981) in which several aspects of an individual's CEA (carcinoembryonic antigen) cancer marker history were investigated for their effects on the hazard of death and the hazard of colorectal cancer recurrence. Patients were followed from the time of colorectal cancer resection, and CEA levels were measured periodically throughout each patient's follow-up. Denoting $z_1(t)$ to be the CEA level at t days of follow-up, Gail used two types of "past history" covariates:

i) $z_2(t) = z_1(t-w)$, a covariate that identified the CEA level w days prior to time t, and

ii) $z_3(t) = \begin{cases} 1 & \text{if } \sup z_1(u) \geq x,\ 0 \leq u \leq t, \\ 0 & \text{otherwise,} \end{cases}$

a covariate that identified whether or not an individual's CEA level surpassed the value x prior to time t.

The hazard models in Section 1.3 may be generalized to include covariates that are defined by repeated measurements or intervening events. In particular, Cox's (1972) multiplicative model, where $\underset{\sim}{z}(t)$ is defined as a step function of time by (1.12) or (1.13), generalizes to

$$\lambda(t;\underset{\sim}{z}(t)) = \lambda_0(t)\ \exp\{\underset{\sim}{\beta}'\underset{\sim}{z}(t)\}. \qquad (1.15)$$

Model (1.15) is, in fact, a proportional hazards model since the hazard curves that formulate this model are proportional (see Section 2.1). Both Farewell (1979) and Gail (1981) had used this model in conjunction with Cox's partial likelihood procedure. This approach was also taken by Crowley and Hu (1977) and Kalbfleisch and Prentice (1980) in their testing for the effectiveness of a heart transplant in prolonging life.

The data from the Stanford Heart Transplant Study has been useful to statisticians concerned with the incorporation of intervening event information into the hazard. Patients, once accepted into the program, would wait until a suitable heart donor was found. Denoting t as the follow-up time since acceptance, a patient's heart transplant status is defined by

$$z_1(t) = \begin{cases} 1 & \text{if } w \leq t \\ 0 & \text{otherwise,} \end{cases}$$

where w denotes the patient's waiting time to heart transplant. The test of $z_1(t)$'s effect on the hazard compares the death rates between patients that have and have not yet received a heart transplant. The patient's waiting times to transplant comprise the person-times for the pretransplant group. Once a patient has received a heart transplant he immediately begins contributing person-time to the posttransplant group.

Crowley and Hu (1977) also considered a "waiting time to heart transplant" effect on posttransplant survival by including

$$z_2(t) = \begin{cases} w & \text{if } w \leq t \\ 0 & \text{otherwise,} \end{cases}$$

as a covariate in model (1.15).

Mantel and Byar (1974) tested for the effectiveness of a heart transplant in prolonging life by performing the nonparametric Mantel-Haenszel test (Mantel 1966, Breslow 1975). Their approach was similar to that of Crowley and Hu (1977) and Kalbfleisch and Prentice (1980) in that it allowed for the possibility of patients leaving the pretransplant group and entering the posttransplant group during follow-up. The Mantel-Haenszel test assumes under the alternative hypothesis that one group's hazard is consistently greater than the other group's hazard over time. This includes the case where transplant status acts multiplicatively on the hazard.

Turnbull, Brown, and Hu (1974) used the multiplicative model with an underlying exponential hazard,

$$\lambda(t;z_1(t)) = \lambda_0 \exp\{\beta_1 z_1(t)\},$$

in their analysis of the heart transplant effect. No allowance was made for extraneous covariates. Under this model, the fitted probabilities of certain pretransplant individuals surviving as long as they did were suspiciously small. The authors then adopted an empirical Bayes approach in which patients had different but unknown exponential hazards, allowing for "hardier" patients to have smaller hazards than "regular" patients. The exponential parameter among pretransplant patients, θ, became $\tau\theta$ once a patient received a transplant. By assuming a gamma (γ,ρ) prior distribution for $\Theta = \{\theta: \theta > 0\}$, Turnbull, Brown, and Hu (1974) arrived at the following marginal distribution for T, given $z_1(t)$:

$$\lambda(t;z_1(t)) = \rho[\gamma\tau^{-z_1(t)} + t]^{-1}. \qquad (1.16)$$

Evidence was provided to demonstrate a better fit of this model over the multiplicative model with an underlying exponential distribution.

The interesting feature about model (1.16) is that it is not a particular form of the multiplicative model. This hazard represents the Pareto distribution, where ρ is the scale parameter and $\gamma^*(z_1(t)) = \gamma\tau^{-z_1(t)}$ is the location parameter. The hazard depends on the covariate $z_1(t)$ only through its location parameter.

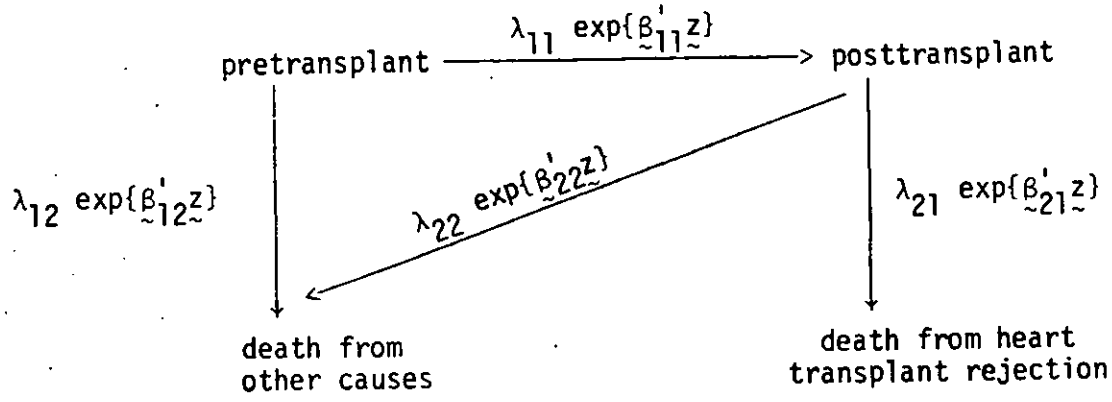
Finally, one might additionally be interested in the effects of certain covariate information on the waiting time distribution of an intervening event. Under these circumstances, a stochastic representation of the intervening event's waiting time is required. Lagakos (1976) proposed such a model, where an exponential hazard was assumed for the intervening event's waiting time distribution. Two distinct exponential hazards of death were assumed: one for individuals in

which the intervening event had not yet occurred and the other for individuals in which the event had occurred.

Beck (1979) incorporated more than one cause of death into this model. Using the Stanford Heart Transplant Study's data as an example, he assumed distinct exponential hazards for two causes of posttransplant death: heart transplant rejection and other causes. Figure 1.1 displays the stochastic components of Beck's model, with the inclusion of multiplicative effects for the baseline covariate vector $\underset{\sim}{z}$. Here, λ_{11} is the assumed underlying hazard for an individual's waiting time to heart transplant, λ_{12} is the assumed pretransplant underlying hazard of death, λ_{21} is the assumed posttransplant underlying hazard of death due to heart transplant rejection, and λ_{22} is the assumed posttransplant underlying hazard of death due to causes other than heart transplant rejection. The parameters $\underset{\sim}{\beta}_{11}$, $\underset{\sim}{\beta}_{12}$, $\underset{\sim}{\beta}_{21}$, and $\underset{\sim}{\beta}_{22}$ allow the four hazards to depend differently on $\underset{\sim}{z}$.

FIGURE 1.1

The Stochastic Components of Beck's (1979) Model



In a similar application involving a follow-up study of U.S. rubber industry workers, Higgins (1978) fitted underlying Weibull

hazards. In this study, the intervening event of interest was disability retirement. Higgins also tested the assumption of time homogeneity, that is, an individual's hazard for a future transition depends only on the length of time spent in his current state, and not on the time at which he entered that state. This assumption was not met since there was a significant inverse relationship between the post-disability retirement hazard of death and the individual's waiting time to disability retirement.

The inclusion of an intervening event's waiting time distribution into a survival model is part of what is known as a "crude hazard rate analysis", and its description appears in the next section.

1.5 The Application of Competing Risks Theory to Epidemiological Follow-Up Studies

In many epidemiological follow-up studies the investigators are interested in the effects of covariate information on different types of failure. The term "competing risks" derives its name from the notion that different causes of death "compete" for the life of an individual. However, failure need not refer only to a cause of death. In the paper by Beck (1979), the occurrence of an intervening event represented one type of failure, and deaths from two causes represented the other types of failure. When nonfatal causes of failure are of interest to the investigator, individuals may fail more than once during their follow-up. Each individual then contributes multivariate failure time data to the study. Such data define a stochastic process in which time is continuous, and the transition probabilities are replaced by cause-specific hazard functions. The

modelling of multivariate failure time data has received attention in Kalbfleisch and Prentice (1980) and Prentice, Williams, and Peterson (1981). However, for the purpose of brevity, the following discussion of competing risks will assume that the modes of failure are deaths from different causes.

Much of the competing risks literature to date has assumed the existence of a "potential" time to death for each cause of death (Chiang 1968, Hoel 1972, Gail 1975, Tsiatis 1975, Elandt-Johnson 1976, David and Moeschberger 1978). That is to say, an individual has a death time for each cause of death, but only the earliest occurring death is observed. Assuming that individuals die from only one of c underlying causes, let X_k be the random variable which denotes an individual's "potential" time-to-death from cause k. Denote $\underset{\sim}{X} = (X_1, X_2, \ldots, X_c)'$. Then the joint survival function of $\underset{\sim}{X}$ is expressed by

$$S_{\underset{\sim}{X}}(t_1, t_2, \ldots, t_c) = \Pr(X_1 > t_1, X_2 > t_2, \ldots, X_c > t_c). \quad (1.17)$$

Two distinct types of hazards may be derived from the joint survival function, the net hazards and the crude hazards. The net hazard rate, $\mu_k(t)$, supposedly represents the time-to-death distribution for cause k when individuals are only able to fail from that cause. These functions are defined by

$$\mu_k(t) = \frac{-\partial \log S_k(t)}{\partial t}, \quad k=1,\ldots,c,$$

where $S_k(t)$ is the marginal survival function of X_k. The crude hazard rate, $\lambda_k(t)$, known also as the k^{th} cause-specific hazard rate, is used

for representing the time-to-death distribution for cause k when individuals are at risk of death from any one of the c causes. These functions are derived from (1.17) by

$$\lambda_k(t) = \left[\frac{-\partial \log S_{\underset{\sim}{X}}(t_1,\ldots,t_c)}{\partial t_k}\right]_{t_1=\ldots=t_c=t}, \quad k=1,\ldots,c.$$

The crude hazards are identifiable and estimable from survival data since an individual's time to death and cause of death may be observed. However, knowledge of the appropriate parametric forms for $\lambda_k(t)$, k=1,...,c, is not enough information to identify the parametric form for the joint distribution of $\underset{\sim}{X}$ (Tsiatis 1975, Birnbaum 1979). Elandt-Johnson and Johnson (1980) have shown that one unique independent risks model and an infinite number of dependent risks models for the joint distribution of $\underset{\sim}{X}$ in (1.17) may give rise to the same crude hazards. Since the joint survival function of $\underset{\sim}{X}$ designates the net hazards, they are also not identifiable from survival data. If one assumes independence of the X_k's, k=1,...,c, then the joint distribution of $\underset{\sim}{X}$ becomes identifiable. However, if one assumes a particular distributional form for $\underset{\sim}{X}$ in (1.17) without the independence assumption, then the parameters contained in that model may still not be identified by the crude hazards (Elandt-Johnson and Johnson 1980). This would preclude a test for independence of the X_k's, k=1,...,c. Finally, the joint survival function of the "potential" times to death for c-1 causes may be severely affected by the biological changes that occur when individuals are no longer able to die from a c^{th} cause. Consequently, it is not necessarily rational to

assume that each net hazard represents the time-to-death distribution for a specific cause when individuals may no longer die from other causes (Prentice, Kalbfleisch, et al. 1978, Kalbfleisch and Prentice 1980). These problems are known as the nonidentifiability aspects of competing risks.

The focus of this dissertation is on the estimation of the crude hazards. They may be defined without any reference to a joint distribution of "potential" cause-specific failure times. Specifically,

$$\lambda_k(t) = \lim_{\Delta t \to 0} \left\{ \frac{\Pr[t \le T \le t + \Delta t, K=k | T \ge t]}{\Delta t} \right\}, \quad k=1,\ldots,c, \qquad (1.18)$$

where T denotes the failure time of an individual, and K is the random variable that identifies an individual's cause of failure. The function $\lambda_k(t)dt$ intuitively represents the probability of death from cause k in the small interval [t,t+dt), conditional on survival to t.

Under the assumption that an individual may die from only one underlying cause, the overall hazard is expressed as the sum of the c crude hazards,

$$\lambda(t) = \sum_{k=1}^{c} \lambda_k(t).$$

The overall survival function is then defined by

$$S(t) = \exp\left\{ -\int_0^t \left(\sum_{k=1}^{c} \lambda_k(u) \right) du \right\}.$$

It should be noted that if one takes the "potential" times to death approach, then $\lambda(t)$ represents the distribution of T = minimum (X_1, ...,X_c).

Covariate information is incorporated into the crude hazards by

$$\lambda_k(t;\underset{\sim}{z}(t)) = \lim_{\Delta t \to o} \left\{ \frac{\Pr[t \le T \le t+\Delta t, K=k \mid T \ge t, \underset{\sim}{z}(t)]}{\Delta t} \right\}, \quad k=1,\ldots,c, \quad (1.19)$$

where $\underset{\sim}{z}(t)$ is defined as a step function of time by (1.12) or (1.13). The individual's overall hazard becomes

$$\lambda(t;\underset{\sim}{z}(t)) = \sum_{k=1}^{c} \lambda_k(t;\underset{\sim}{z}(t)),$$

and his overall survival function is defined by

$$S(t;\underset{\sim}{Z}(t)) = \exp\left\{ -\int_o^t \left[\sum_{k=1}^{c} \lambda_k(u;\underset{\sim}{z}(u)) \right] du \right\}, \quad (1.20)$$

where $\underset{\sim}{Z}(t) = \{\underset{\sim}{z}(u): u \le t\}$ denotes the individual's accumulated covariate information by time t.

In many epidemiological follow-up studies, the estimation of the relationship between covariates and cause-specific patterns of failure is important. Usually, there are c causes of failure present in the data, but only $c^* < c$ of them are of interest to the investigator. Model (1.19) is the tool by which such data are analyzed, hence the name crude hazard rate analysis.

1.6 Construction of the Likelihood

Introduction

In this section a thorough discussion of the likelihood, as a function of cause-specific hazards, is presented. Four topics will be covered: the random censorship model, complete and incomplete likelihoods, Cox's partial likelihood, and Breslow's approximation to Cox's partial likelihood in the presence of tied

failure times. Holt (1978) and Prentice, Kalbfleisch, et al. (1978) defined Cox's partial likelihood for the situation where more than one cause of failure is of interest. Within this context, they defined the partial likelihood for two models: i) the covariates act multiplicatively on each cause-specific underlying hazard, and ii) both the cause of failure and the covariates act multiplicatively on an underlying hazard. Model (ii) is a more restricted version of model (i), and if its assumptions are met, then its use should improve the efficiency of the covariates' estimated effects (Prentice, Kalbfleisch, et al. 1978). Tied failure times were not considered in these 1978 papers, and the partial likelihood derivations were omitted. In Kalbfleisch and Prentice's (1980) book, they suggest the use of Breslow's approximation to Cox's partial likelihood if tied failure times are present, but the derivations for cases (i) and (ii) were not presented. Since Cox's partial likelihood approach is used in this dissertation, the derivations of these likelihoods are presented in this section.

Finally, it is assumed in this section that the time dependent covariates within the covariate vector $\underset{\sim}{z}(t)$ are defined only by (1.12) or (1.13), that is, as step functions of time based on the repeated measurements. Although one could define the u^{th} covariate by $z_u(t) = h(t)z_{uo}$ in order to allow for a nonmultiplicative effect of the baseline covariate z_{uo} on the hazard, this will not be assumed here.

Random Censorship

In many follow-up studies there are individuals who have not failed by the time an analysis of the data is desired. Also, there may be participants that either withdraw early from the study or are lost to follow-up. These individuals are considered to have been censored from the right. Left censoring occurs if individuals enter the study at different times.

It is usually assumed that the pattern of censoring has no influence on the estimation of the time-to-failure distribution. Specifically, independent "potential" time-to-failure (T) and time to censoring (Y) distributions are assumed such that

$$S_{T,Y}(t,y) = S_T(t)\ S_Y(y).$$

This is called the random censorship model. Under this assumption, the inclusion of a time-to-censoring distribution into the likelihood adds no information about the parameters in the time-to-failure distribution, $S_T(t)$. Therefore, one may estimate the hazard rate of failure without having to fit a time-to-censoring distribution, and the likelihood is created by treating the censoring times as fixed observations.

Williams and Lagakos (1977) and Lagakos (1979) defined noninformative censoring as the condition in which the inclusion of a time-to-censoring distribution into the likelihood adds no information about the parameters in the failure time distribution, $S_T(t)$. They demonstrated that the random censorship model is but a subset of a well defined class of noninformative models for $S_{T,Y}(t,y)$. The reader is referred to these papers for more detail.

Random censorship assumes that an individual will not be censored due to a change in health related to the failure mechanism under study. For example, in a heart disease study, it is assumed that patients will not withdraw early from the study because they no longer have chest pain. Clearly, this type of censoring may affect the estimation of the time-to-death from heart disease distribution. Any knowledge that "informative" censoring has occurred in the data should be recorded and accounted for in the interpretation of the results.

The random censorship assumption will be made throughout this dissertation, and the censoring times will be treated as fixed observations.

Complete and Incomplete Likelihoods

The complete likelihood function is written as a product of components. If some of these terms are omitted from the likelihood, then the resulting function will be called an incomplete likelihood. Such likelihoods are used in situations where the completely specified survival model contains parameters that are not of primary concern.

Assume that c causes of death are present in the data, and that the investigator is interested in only $c^* < c$ of them. Denote $\lambda_k(t;\underset{\sim}{\theta}_k,\underset{\sim}{z}_i(t))$ to be the i^{th} individual's crude hazard for cause k, comprised of the parameters $\underset{\sim}{\theta}_k$, k=1,...,c. Let t_i represent the i^{th} individual's observed follow-up time, i=1,...,n, and define

$$\delta_{ki} = \begin{cases} 1 & \text{if individual i dies at } t_i \text{ from cause k} \\ 0 & \text{otherwise.} \end{cases}$$

Then the complete likelihood is written as a product of two components,

$$L_C(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_c) = L_{I1}(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_{c^*}) \cdot L_{I2}(\underset{\sim}{\theta}_{c^*+1},\ldots,\underset{\sim}{\theta}_c),$$

where

$$L_{I1}(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_{c^*}) = \prod_{k=1}^{c^*} \prod_{i=1}^{n} \left[\lambda_k^{\delta_{ki}}(t_i;\underset{\sim}{\theta}_k,\underset{\sim}{z}_i(t)) \cdot \exp\left\{-\int_0^{t_i} \lambda_k(u;\underset{\sim}{\theta}_k,\underset{\sim}{z}_i(u))du\right\}\right] . \tag{1.21}$$

Here, $L_{I1}(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_{c^*})$ contains the likelihood information on the c* cause-specific hazards of interest. By assuming that the parameters $(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_{c^*})$ are functionally independent of $(\underset{\sim}{\theta}_{c^*+1},\ldots,\underset{\sim}{\theta}_c)$, the maximum likelihood estimates of $(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_{c^*})$ obtained from either $\log L_C(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_c)$ or $\log L_{I1}(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_{c^*})$ are equivalent. Inclusion of the nuisance hazards, $\lambda_k(t;\underset{\sim}{\theta}_k,\underset{\sim}{z}_i(t))$, k=c*+1,...,c, into the likelihood is noninformative. Therefore, in constructing the likelihood for c* cause-specific hazards of interest, one may use the incomplete likelihood in (1.21).

Likelihood (1.21) treats the deaths from causes c*+1,...,c as censored observations. Consequently, the use of L_{I1} instead of L_C precludes the fitting of an overall survival function.

Cox's Partial Likelihood

Again, assume that c* causes of death are of interest, and consider the deaths from causes c*+1,...,c as censored observations. Define $0 < t_{(1)} < t_{(2)} < \ldots < t_{(r)}$ to be the r distinct follow-up times to death among the c* causes. Denote $I_j = (t_{(j-1)},t_{(j)}]$, j=1,...,r, where $t_{(o)} = 0$. Define R_j' as the group of individuals that survived past $t_{(j-1)}$, j=1,...,r+1. Let $I_{j\ell}$ be the portion of

the interval I_j that individual ℓ was observed to survive through, $j=1,\ldots,r$, and $I_{(r+1)\ell}$ to be the length of time survived by individual ℓ past $t_{(r)}$. Define $D_{j(k)}$ as the group of $d_{j(k)}$ individuals that died at $t_{(j)}$ from cause k. Then the incomplete likelihood in (1.21) is rewritten as

$$L_{I1}(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_{c^*}) = \left[\prod_{j=1}^{r}\prod_{k=1}^{c^*}\prod_{i\in D_{j(k)}} \lambda_k(t_{(j)};\underset{\sim}{\theta}_k,\underset{\sim}{z}_i(t_{(j)}))\right] \times$$

$$\left[\prod_{j=1}^{r+1} \exp\left\{-\sum_{\ell\in R_j}\int_{I_{j\ell}} \sum_{k=1}^{c^*} \lambda_k(t;\underset{\sim}{\theta}_k,\underset{\sim}{z}_\ell(t))dt\right\}\right]. \qquad (1.22)$$

Define the sequence of events $\{U_1,\ldots,U_{r+1}\}$, $\{V_{j(k)},j=1,\ldots,r, k=1,\ldots,c^*\}$ and $\{S_{j(k)},j=1,\ldots,r,k=1,\ldots,c^*\}$ by

U_j = the observed survival experience for individuals within the interval I_j, $j=1,\ldots,r$,

U_{r+1} = the observed survival experience for individuals beyond $t_{(r)}$,

$V_{j(k)}$ = the event that $d_{j(k)}$ deaths from cause k occur at $t_{(j)}$,

and

$S_{j(k)}$ = the event that $d_{j(k)}$ particular individuals die from cause k at $t_{(j)}$.

Denoting $U^{(j)} = (U_1,\ldots,U_j)$ and $S^{(j)} = \{S_{\ell(k)}, \ell=1,\ldots,j, k=1,\ldots,c^*\}$, the likelihood of the events $\{S_{j(k)}|U^{(j)}, S^{(j-1)}\}$, $j=1,\ldots,r$, $k=1,\ldots,c^*$, comprises the first product of terms in (1.22), and the likelihood of the events $\{U_j|U^{(j-1)}, S^{(j-1)}\}$, $j=1,\ldots,r+1$, comprises the second product of terms in (1.22).

The sequence of events $\{V_{j(k)}, j=1,\ldots,r, k=1,\ldots,c^*\}$ is used for

constructing Cox's partial likelihood. Denote $V^{(j)} = \{V_{\ell(k)}, \ell=1,\ldots, j, k=1,\ldots,c^*\}$. The incomplete likelihood in (1.22) may be re-expressed by the joint likelihood of the events

$$\left\{S_{j(k)}\middle|U^{(j)},V^{(j)},S^{(j-1)}\right\}, \left\{V_{j(k)}\middle|U^{(j)},V^{(j-1)},S^{(j-1)}\right\}, \begin{matrix} j=1,\ldots,r, \\ k=1,\ldots,c^*, \end{matrix}$$

and

$$\left\{U_j\middle|U^{(j-1)},V^{(j-1)},S^{(j-1)}\right\}, \quad j=1,\ldots,r+1.$$

Denoting the contribution to the likelihood of the event

$$\left\{S_{j(k)}\middle|U^{(j)},V^{(j)},S^{(j-1)}\right\} \text{ by } L\left\{S_{j(k)}\middle|U^{(j)},V^{(j)},S^{(j-1)}\right\},$$

Cox (1975) defined the partial likelihood for the sequence of events $\{S_{j(k)}, j=1,\ldots,r, k=1,\ldots,c^*\}$ as

$$L_p = \prod_{j=1}^{r}\prod_{k=1}^{c^*} L\left\{S_{j(k)}\middle|U^{(j)},V^{(j)},S^{(j-1)}\right\}.$$

For each cause of death, $k=1,\ldots,c^*$, this partial likelihood is expressed by

$$L_p(\underset{\sim}{\theta}_k) = \prod_{j=1}^{r}\left\{\frac{\prod_{i\in D_{j(k)}} \lambda_k[t_{(j)};\underset{\sim}{\theta}_k,\underset{\sim}{z}_i(t_{(j)})]}{\sum_{L\in R_j(d_{j(k)})}\prod_{\ell\in L}\lambda_k[t_{(j)};\underset{\sim}{\theta}_k,\underset{\sim}{z}_\ell(t_{(j)})]}\right\}, \quad (1.23)$$

where $R_j(d_{j(k)})$ denotes the set of all subsets of $d_{j(k)}$ individuals chosen from the risk set R_j, the group of individuals at risk of failure at $t_{(j)}$, and ℓ denotes one of the $d_{j(k)}$ individuals from the particular subset L. The j^{th} component in the partial likelihood $L_p(\underset{\sim}{\theta}_k)$ represents the probability that $d_{j(k)}$ particular individuals died from cause k at time $t_{(j)}$ conditional on the information that

$d_{j(k)}$ deaths from cause k occurred at $t_{(j)}$ and that R_j individuals were at risk of death at $t_{(j)}$. If $d_{j(k)}=0$, then the j^{th} component of $L_p(\underset{\sim}{\theta}_k)$ is equal to one. If $d_{j(k)}=1$, then $R_j(d_{j(k)})=R_j$ and the denominator for the j^{th} component of $L_p(\underset{\sim}{\theta}_k)$ becomes

$$\sum_{\ell \epsilon R_j} \lambda_k(t_{(j)};\underset{\sim}{\theta}_k,\underset{\sim}{z}_\ell(t_{(j)})).$$

Marginal, conditional, and partial likelihoods, as defined by Cox (1975), are each examples of an incomplete likelihood. Only under certain conditions will two of these three types be equivalent. The reader is referred to Cox's (1975) paper for elaboration.

Now assume Cox's multiplicative model for each cause-specific hazard,

$$\lambda_k(t;\underset{\sim}{z}_i(t)) = \lambda_{ko}(t)\ \exp\{\underset{\sim}{\beta}_k'\ \underset{\sim}{z}_i(t)\},\quad k=1,\ldots,c^*. \tag{1.24}$$

The partial likelihood in (1.23) becomes a function of $\underset{\sim}{\beta}_k$,

$$L_p(\underset{\sim}{\beta}_k) = \prod_{j=1}^{r} \left\{ \frac{\prod_{i \epsilon D_{j(k)}} \exp\{\underset{\sim}{\beta}_k'\ \underset{\sim}{z}_i(t_{(j)})\}}{\sum_{L \epsilon R_j(d_{j(k)})} \prod_{\ell \epsilon L} \exp\{\underset{\sim}{\beta}_k'\ \underset{\sim}{z}_\ell(t_{(j)})\}} \right\},\ k=1,\ldots,c^*. \tag{1.25}$$

The parameters $\underset{\sim}{\beta}_k$, k=1,...,c*, may be estimated from $\prod_{k=1}^{c^*} L_p(\underset{\sim}{\beta}_k)$ without any knowledge of the cause-specific underlying hazards, $\lambda_{ko}(t)$, k=1,...,c*.

Finally, assume that the cause of failure acts multiplicatively on the hazard in (1.24) such that

$$\lambda_k(t;\underset{\sim}{z}_i(t)) = \lambda_o(t)\ \exp\{\alpha_k+\underset{\sim}{\beta}_k'\ \underset{\sim}{z}_i(t)\},\ k=1,\ldots,c^*, \tag{1.26}$$

where $\alpha_1=0$. Conditioning on the events V_j, that $d_j = \sum_{k=1}^{c^*} d_{j(k)}$

failures occur at $t_{(j)}$, $j=1,\ldots,r$, it can be shown that the partial likelihood for $\underset{\sim}{\alpha} = (\alpha_2,\ldots,\alpha_{c^*})'$ and $\underset{\sim}{\beta} = (\underset{\sim}{\beta}_1',\ldots,\underset{\sim}{\beta}_{c^*}')'$ becomes

$$L_p(\underset{\sim}{\alpha},\underset{\sim}{\beta}) = \prod_{j=1}^{r} \left\{ \frac{\prod_{k=1}^{c^*} \prod_{i\in D_{j(k)}} \exp\{\alpha_k + \underset{\sim}{\beta}_k' \underset{\sim}{z}_i(t_{(j)})\}}{\sum_{L\in R_j(d_j)} \prod_{\ell\in L} \sum_{k=1}^{c^*} \exp\{\alpha_k + \underset{\sim}{\beta}_k' \underset{\sim}{z}_\ell(t_{(j)})\}} \right\} \qquad (1.27)$$

where $R_j(d_j)$ denotes the set of all subsets of d_j individuals chosen from the risk set R_j.

Computation of the maximum partial likelihood estimates from (1.25) or (1.27) becomes difficult if there are a large number of tied failure times. For instance, if $R_j=100$ and $d_{j(k)}=d_j=2$ for some j, then the j^{th} component of the denominators in (1.25) and (1.27) would consist of $\binom{100}{2}=4,950$ summations. Breslow (1972) suggested an approximation to Cox's partial likelihood in which the denominator became more manageable. Its description is now given.

Breslow's Approximation to Cox's Partial Likelihood

Denote $\{R_j' - R_j\}$ to be the set of individuals that are censored within $I_j = (t_{(j-1)},t_{(j)}]$. Let $I_{j\ell}^*$ represent the period of time in I_j spent by individual $\ell\in R_j$ such that $\underset{\sim}{z}_\ell(t) \neq \underset{\sim}{z}_\ell(t_{(j)})$. Then the incomplete likelihood in (1.22) becomes

$$L_{I1}(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_{c^*}) = L_I(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_{c^*})$$

$$\cdot \left\{ \prod_{j=1}^{r} \exp\left[- \sum_{\ell\in\{R_j'-R_j\}} \int_{I_{j\ell}} \sum_{k=1}^{c^*} \lambda_k(t;\underset{\sim}{\theta}_k,\underset{\sim}{z}_\ell(t))dt \right] \right\} \qquad (F_1)$$

$$\cdot \left\{ \exp\left[- \sum_{\ell\in R_{r+1}} \int_{(r+1)\ell} \sum_{k=1}^{c^*} \lambda_k(t;\underset{\sim}{\theta}_k,\underset{\sim}{z}_\ell(t))dt \right] \right\} \qquad (F_2)$$

$$\cdot \left\{ \prod_{j=1}^{r} \exp\left[- \sum_{\ell \in R_j} \int_{I^*_{j\ell}} \sum_{k=1}^{c^*} [\lambda_k(t;\underset{\sim}{\theta}_k,\underset{\sim}{z}_\ell(t)) - \lambda_k(t;\underset{\sim}{\theta}_k,\underset{\sim}{z}_\ell(t_{(j)}))]dt \right] \right\}, \quad (F_3)$$

where

$$L_I(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_{c^*}) = \left\{ \prod_{j=1}^{r} \prod_{k=1}^{c^*} \prod_{i \in D_{j(k)}} \lambda_k(t_{(j)};\underset{\sim}{\theta}_k,\underset{\sim}{z}_i(t_{(j)})) \right\} \cdot$$

$$\left\{ \prod_{j=1}^{r} \exp\left[- \sum_{i \in R_j} \int_{t_{(j-1)}}^{t_{(j)}} \sum_{k=1}^{c^*} \lambda_k[t;\underset{\sim}{\theta}_k,\underset{\sim}{z}_i(t_{(j)})]dt \right] \right\} \qquad (1.28)$$

is Breslow's (1972,1974) incomplete likelihood. The two factors in (1.28) are the death time information from the c* causes and the individuals' survival experiences within the intervals I_j, j=1,...,r, excluding their censoring and covariate information within these intervals. The three factors of $L_{I1}(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_{c^*})$ that are deleted in obtaining Breslow's incomplete likelihood include F_1, the censoring information within the intervals I_j, j=1,...,r, F_2, the individuals' survival information beyond the last observed death time $t_{(r)}$, and F_3, the individuals' covariate information within interval I_j among those not censored in I_j, j=1,...,r. If there is much censoring, or if the time dependent covariates on an individual vary within the time intervals $(t_{(j-1)},t_{(j)}]$, j=1,...,r, then a loss of efficiency by using $L_I(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_{c^*})$ over $L_{I1}(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_{c^*})$ may occur. However, this problem is alleviated as the number of distinct times to death among the c* causes, r, becomes large, and the lengths between those times become small.

With the multiplicative model in (1.26), assume

$$\lambda_o(t) = \begin{cases} \lambda_j, & t \in (t_{(j-1)}, t_{(j)}], \quad j=1,\ldots,r, \\ 0, & t > t_{(r)}, \end{cases}$$

where $t_{(o)} = 0$. This underlying hazard is piecewise exponential, and the intervals are defined by the exact times to death. The likelihood in (1.28) becomes

$$L_I(\underset{\sim}{\lambda},\underset{\sim}{\alpha},\underset{\sim}{\beta}) = \left\{ \prod_{j=1}^{r} \left[\lambda_j^{d_j} \prod_{k=1}^{c^*} \prod_{i \in D_{j(k)}} \exp\{\alpha_k + \underset{\sim}{\beta}_k' \underset{\sim}{z}_i(t_{(j)})\} \right] \right\} \cdot$$

$$\left\{ \exp\left[- \sum_{j=1}^{r} \lambda_j h_j \sum_{\ell \in R_j} \sum_{k=1}^{c^*} \exp\{\alpha_k + \underset{\sim}{\beta}_k' \underset{\sim}{z}_\ell(t_{(j)})\} \right] \right\}, \qquad (1.29)$$

where $d_j = \sum_{k=1}^{c^*} d_{j(k)}$, $h_j = t_{(j)} - t_{(j-1)}$, $\underset{\sim}{\lambda} = (\lambda_1,\ldots,\lambda_r)'$, $\underset{\sim}{\alpha} = (\alpha_2,\ldots,\alpha_{c^*})'$, and $\underset{\sim}{\beta} = \underset{\sim}{\beta}_1',\ldots,\underset{\sim}{\beta}_{c^*}')'$. Maximizing $L_I(\underset{\sim}{\lambda},\underset{\sim}{\alpha},\underset{\sim}{\beta})$ with respect to $\underset{\sim}{\lambda}$, one obtains

$$\hat{\lambda}_j(\underset{\sim}{\alpha},\underset{\sim}{\beta}) = d_j \left[h_j \sum_{\ell \in R_j} \sum_{k=1}^{c^*} \exp\{\alpha_k + \underset{\sim}{\beta}_k' \underset{\sim}{z}_\ell(t_{(j)})\} \right]^{-1}, \quad j=1,\ldots,r. \qquad (1.30)$$

Inserting $\hat{\lambda}_j(\underset{\sim}{\alpha},\underset{\sim}{\beta})$, $j=1,\ldots,r$, back into (1.29), and omitting constant terms, Breslow's incomplete likelihood for $(\underset{\sim}{\alpha},\underset{\sim}{\beta})$ becomes

$$L_I(\underset{\sim}{\alpha},\underset{\sim}{\beta}) = \prod_{j=1}^{r} \left\{ \frac{\prod_{k=1}^{c^*} \prod_{i \in D_{j(k)}} \exp\{\alpha_k + \underset{\sim}{\beta}_k' \underset{\sim}{z}_i(t_{(j)})\}}{\left[\sum_{\ell \in R_j} \sum_{k=1}^{c^*} \exp\{\alpha_k + \underset{\sim}{\beta}_k' \underset{\sim}{z}_\ell(t_{(j)})\} \right]^{d_j}} \right\}. \qquad (1.31)$$

Since the individuals' uncensored survival experiences are maintained within the intervals $(t_{(j-1)}, t_{(j)}]$, $j=1,\ldots,r$, in (1.28), Breslow's incomplete likelihood, $L_I(\underset{\sim}{\alpha},\underset{\sim}{\beta})$ in (1.31), is not a partial

likelihood. But note its strong similarity with Cox's partial likelihood, $L_p(\underset{\sim}{\alpha},\underset{\sim}{\beta})$ in (1.27). If there are no tied times to death among the c* causes, i.e., if $\sum_{k=1}^{c^*} d_{j(k)}=1$ for j=1,...,r, then $L_p(\underset{\sim}{\alpha},\underset{\sim}{\beta})$ and $L_I(\underset{\sim}{\alpha},\underset{\sim}{\beta})$ are equivalent.

The incomplete likelihood $L_{I1}(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_{c^*})$ in (1.22) may be rewritten in the following way. Denote $0 < t_{(k1)} < t_{(k2)} < \ldots < t_{(kr_k)}$ to be the r_k distinct follow-up times to death from cause k, k=1,...,c*, and let $I_{kj} = (t_{(kj-1)}, t_{(kj)}]$, where $t_{(ko)} = 0$. Let R'_{kj} be the group of individuals that survived through $t_{(kj-1)}$, and denote $I_{kj\ell}$ to be the portion of interval I_{kj} that individual ℓ was observed to survive through. Denote $R'_{k(r_k+1)}$ as the group of individuals that survived past $t_{(kr_k)}$ and $I_{k(r_k+1)\ell}$ to be the length of time survived by individual ℓ past $t_{(kr_k)}$. Let D_{kj} be the group of d_{kj} individuals who fail at $t_{(kj)}$ from cause k, j=1,...,r_k, k=1,...,c*. One then obtains

$$L_{I1}(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_{c^*}) = \left\{ \prod_{k=1}^{c^*} \prod_{j=1}^{r_k} \prod_{i\in D_{kj}} \lambda_k\left(t_{(kj)};\underset{\sim}{\theta}_k,\underset{\sim}{z}_i(t_{(kj)})\right) \right\} \cdot \left\{ \prod_{k=1}^{c^*} \prod_{j=1}^{r_k+1} \exp\left[-\sum_{\ell\in R'_{kj}} \int_{I_{kj\ell}} \lambda_k\left(t;\underset{\sim}{\theta}_k,\underset{\sim}{z}_\ell(t)\right) dt \right] \right\}. \tag{1.32}$$

Now, assume model (1.24) and define

$$\lambda_{ko}(t) = \begin{cases} \lambda_{kj}, & t \in (t_{(kj-1)}, t_{(kj)}], \quad j=1,\ldots,r_k, \\ 0, & t > t_{(kr_k)}, \end{cases} \tag{1.33}$$

for k=1,...,c*. Here, each cause-specific underlying hazard is assumed to be piecewise exponential, and the intervals are defined

by the exact times to death for each cause. Omit from $L_{I1}(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_{c^*})$ the individuals' censoring and covariate information within the intervals I_{kj}, $j=1,\ldots,r_k$, $k=1,\ldots,c^*$, and the individuals' survival information from the k^{th} cause beyond $t_{(kr_k)}$, $k=1,\ldots,c^*$. One can verify that the maximum likelihood estimate of λ_{kj} in (1.33) becomes

$$\hat{\lambda}_{kj}(\underset{\sim}{\beta}_k) = d_{kj}\left[h_{kj} \sum_{\ell \in R_{kj}} \exp\{\underset{\sim}{\beta}_k' \underset{\sim}{z}_\ell(t_{(kj)})\}\right]^{-1}, \quad \begin{matrix} j=1,\ldots,r_k, \\ k=1,\ldots,c^*, \end{matrix} \qquad (1.34)$$

where R_{kj} denotes the group of individuals at risk of death at $t_{(kj)}$ and $h_{kj} = t_{(kj)} - t_{(kj-1)}$. Breslow's incomplete likelihood for $\underset{\sim}{\beta}_k$ becomes

$$L_I(\underset{\sim}{\beta}_k) = \prod_{j=1}^{r_k} \left\{ \frac{\prod_{i \in D_{kj}} \exp\{\underset{\sim}{\beta}_k' \underset{\sim}{z}_i(t_{(kj)})\}}{\left[\sum_{\ell \in R_{kj}} \exp\{\underset{\sim}{\beta}_k' \underset{\sim}{z}_\ell(t_{(kj)})\}\right]^{d_{kj}}} \right\}, \quad k=1,\ldots,c^*. \qquad (1.35)$$

Note the similarity between $L_I(\underset{\sim}{\beta}_k)$ and Cox's partial likelihood, $L_P(\underset{\sim}{\beta}_k)$ in (1.25). In fact, if there are no tied times to death from cause k, i.e., if $d_{kj}=1$ for $j=1,\ldots,r_k$, then $L_I(\underset{\sim}{\beta}_k) = L_P(\underset{\sim}{\beta}_k)$.

To summarize, Cox's model has two forms: i) model (1.24) in which the covariates act multiplicatively on each cause-specific underlying hazard, and ii) model (1.26) in which both the cause of death and the covariates act multiplicatively on an underlying hazard. For simplicity of computation, Breslow's approximation to Cox's partial likelihood is used: $L_I(\underset{\sim}{\beta}_k)$, $k=1,\ldots,c^*$, in (1.35) for case i) and $L_I(\underset{\sim}{\alpha},\underset{\sim}{\beta})$ in (1.31) for case ii).

1.7 Nonparametric Methods for Determining an Appropriate Parametric Form of the Hazard

Informal but meaningful interpretations can be obtained by a graphical display of the data. In survival analysis, nonparametric estimation of the hazard can yield valuable information about its parametric form. Actually, the estimated cumulative hazard is graphed against time because:

i) it is usually represented by a simpler mathematical expression than the survival function is, and

ii) it should yield a smoother curve than the estimated hazard.

One then considers the slope of the curve in determining the shape of the hazard.

Two nonparametric procedures in survival analysis are common:

i) the discrete hazard model (Kaplan and Meier 1958, Nelson 1972), and

ii) the piecewise exponential hazard model (Breslow 1972).

Both methods are nonparametric in that r exact times to failure are described by r parameters, a saturated parametric model. These methods can utilize either exact or grouped failure time data, and they can be generalized to multiple causes of failure.

The discrete hazard model assumes the hazard to be a distinct positive value at each exact failure time and to be zero within the intervals that separate those times. This contrasts with the piecewise exponential hazard model's assumption of a distinct constant hazard within each interval that separates the exact times to failure.

Under the discrete hazard model, it can be shown that the maximum likelihood estimate of the survival function is the Kaplan-Meier (1958) product limit estimate, and the maximum likelihood estimate of the cumulative hazard is the estimate proposed by Nelson (1972). The estimated variance of the Kaplan-Meier product limit estimate, obtained from the inverse of the observed Fisher information matrix and a Taylor series linear approximation, turns out to be Greenwood's formula (referenced in Kalbfleisch and Prentice 1980).

The discrete and piecewise exponential hazard models can be generalized to handle stratified covariate information. However, the estimates of Kaplan-Meier and Nelson do not employ the censoring and covariate information that exist within the intervals between the exact failure times. This information is precisely what Breslow (1972) omitted from the likelihood in order to obtain an approximation to Cox's partial likelihood. In a recent paper by Chen, Hollander, and Langberg (1982), it was demonstrated that the Kaplan-Meier product limit estimate underestimates the true survival function in small samples with censored data, and that the amount of bias increases with the amount of censoring. The piecewise exponential hazard model, on the other hand, uses all of the individuals' survival information between time zero and the last observed failure time. Consequently, it is anticipated that this model would yield cumulative hazard and survival function estimates with smaller mean squared errors in the presence of much censoring. In addition, the piecewise exponential hazard model should yield a more precise estimate of the cumulative hazard, as a function of covariate strata, when the exact times to

failure are spread out relative to the times at which covariate measurements are repeated. For these reasons, all nonparametric estimation of the cumulative hazard in this dissertation will be obtained by the piecewise exponential hazard model.

With $c^* < c$ pertinent causes of death, denote $0 < t_{(k1)} < t_{(k2)} < \ldots < t_{(kr_k)}$ to be the r_k distinct follow-up times to death from the k^{th} cause, $k=1,\ldots,c^*$. The piecewise exponential hazard model was defined in (1.33) without covariates. Here, $\lambda_k(t)$ replaces $\lambda_{ko}(t)$ as the crude hazard for cause k. The incomplete likelihood for model (1.33) without covariates is written as

$$L_{PE} = \prod_{k=1}^{c^*} \prod_{j=1}^{r_k} \lambda_{kj}^{d_{kj}} \exp\left\{-\lambda_{kj} \sum_{\ell \in R'_{kj}} h_{kj\ell}\right\}, \qquad (1.36)$$

where d_{kj} is the number of deaths from cause k at $t_{(kj)}$, $h_{kj\ell}$ is the length of time spent by individual ℓ in the interval $I_{kj} = (t_{(kj-1)}, t_{(kj)}]$, and R'_{kj} denotes the group of individuals that survived past $t_{(kj-1)}$. This likelihood is incomplete since there is no modelling of the hazards for the extraneous causes of death, $c^*+1,\ldots,c$. The term $L_{kj} = \sum_{\ell \in R'_{kj}} h_{kj\ell}$ represents the total amount of person-time spent by individuals in I_{kj}. Maximizing L_{PE} with respect to λ_{kj}, one obtains

$$\hat{\lambda}_{kj} = \frac{d_{kj}}{L_{kj}} , \qquad (1.37)$$

the observed death rate from cause k in the interval I_{kj}, $j=1,\ldots,r_k$, $k=1,\ldots,c^*$. The k^{th} cause-specific cumulative hazard at $t_{(kj)}$ is estimated by

$$\hat{\Lambda}_{PE,k}(t_{(kj)}) = \sum_{i=1}^{j} h_{ki} \frac{d_{ki}}{L_{ki}} , \qquad (1.38)$$

where $h_{ki} = t_{(ki)} - t_{(ki-1)}$. Its variance is estimated by

$$\hat{var}(\hat{\Lambda}_{PE,k}(t_{(kj)})) = \sum_{i=1}^{j} h_{ki}^2 \frac{d_{ki}}{L_{ki}^2}, \qquad (1.39)$$

using the inverse of the observed Fisher information matrix for the parameters λ_{kj}, $j=1,\ldots,r_k$, $k=1,\ldots,c^*$.

In Chapter 2, a method for incorporating stratified covariates, as functions of repeated measurements, into the piecewise exponential hazard model will be proposed. Cause-specific cumulative hazard estimates will be used for selecting appropriate parametric forms of the cause-specific underlying hazards. A graphical display of the covariates' effects on these hazards will also be obtained. It is of interest to determine if a nonmultiplicative model is suggested by the data. In order to do this, the multiplicative model assumption of Breslow (1972) and Holford (1976) will be discarded. However, as a consequence, the proposed method requires a categorization of the covariates into strata.

1.8 Outline of Subsequent Chapters

In Chapter 2, the hazard as a function of repeated measurements is discussed, and a broader definition of the proportional hazards model is presented. A general model for the hazard as a function of covariates is presented. It is based on two properties: an underlying failure time distribution exists, and the model parameters are expressed as simple functions of the covariates. This framework encompasses a wide variety of survival models, and each model defines

a distinct type of nonproportionality. Except for the exponential distribution, the multiplicative model is a reduced form of a more general model. Examples will be provided.

Also included in Chapter 2 is a method for nonparametric estimation of cause-specific cumulative hazards as functions of covariate strata. The piecewise exponential hazard model will be used, and its primary focus will be on the selection of appropriate parametric forms for the cause-specific underlying hazards. An application of this method to the Stanford Heart Transplant Study's data, using the zero-one covariate heart transplant status, will be presented.

In the last section of Chapter 2, the likelihood will be created as a function of minimal sufficient statistics. Computational efficiency is achieved by using the minimal sufficient statistics in place of the raw data.

The proposed framework for incorporating repeated measurements into the hazard was applied to a set of heart disease data from Duke University, and the results appear in Chapter 3. It was of interest to assess the effect of the individual's most recently measured congestive heart failure (CHF) status, a discrete covariate with five levels, on the hazards of i) death due to congestive heart failure and ii) death due to other heart disease causes. The data analysis was performed initially with one covariate. Two important baseline covariates were then included. After adjusting for the individual's most recently measured CHF status, the effects of prior CHF history on these hazards were addressed. This prior CHF history was defined by other covariates in the model. The appropriateness of the

multiplicative model for describing the data was investigated, and explicit likelihood ratio tests of this assumption were performed. As a validity check for the appropriateness of the assumed underlying hazard, the results were compared with those obtained by Cox's model. Finally, a determination was made of the information gained by using the repeated measurements on CHF status over just using its baseline value.

The analysis of a retrospective follow-up study of chrysotile asbestos textile workers is presented in Chapter 4 as a final example of incorporating repeated measurements into the hazard. Dement (1980, 1982) documented the workers' histories of asbestos exposure for those employed at the plant between 1940 and 1975. This information was synthesized from each individual's job history at the plant and from periodic measurements of asbestos levels in the air at the different job settings within the plant over the years. It was of interest to determine the effect of an individual's cumulative exposure to chrysotile asbestos on the hazard of death from lung cancer. Age at death from lung cancer was used as the outcome variable, and the workers' annual exposures to chrysotile asbestos, by age, comprised their repeated measurements. Three covariates were determined for each individual: his cumulative exposure to chrysotile asbestos at age t, the number of years elapsed between his initial employment and age t, and the calendar time (year) at age t. Using the methods described in Chapter 2, these covariates were analyzed separately and jointly for their effects. A discussion of the methodology's performance will appear at the end of Chapter 4.

In Chapter 5, topics for future research will be enumerated.

CHAPTER 2

A FRAMEWORK THAT INCORPORATES REPEATED MEASUREMENTS INTO THE HAZARD

2.1 The Hazard as a Function of Repeated Measurements: Proportionality Receives a Broader Definition

In Section 1.4, repeated measurements are incorporated into the hazard by defining the covariates to remain constant between successive sets of measurements over time. Specifically, if the covariate measurements are performed on each individual at the follow-up times τ_m, $m=0,\ldots,q$, where $\underset{\sim}{z}_m$ denotes the individual's measured value at τ_m, and τ_0 and τ_{q+1} denote the beginning and end of the study, then

$$\underset{\sim}{z}(t) = \underset{\sim}{z}_m \quad \text{for} \quad t \in [\tau_m, \tau_{m+1}), \quad m=0,\ldots,q.$$

If an individual's follow-up time occurs prior to the end of the study, then his covariate vector $\underset{\sim}{z}(t)$ is defined up to that time.

Now, assume there are p distinct values, possibly many, for $\underset{\sim}{z}(t)$ in the data, $\underset{\sim}{\zeta}_1,\ldots,\underset{\sim}{\zeta}_p$. Then a distinct hazard curve, $\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_\ell)$, is defined by some mathematical function for each value, $\underset{\sim}{\zeta}_\ell$, $\ell=1,\ldots,p$. The i^{th} individual's covariate vector, $\underset{\sim}{z}_i(t)$, may oscillate among these p covariate values over time. Therefore, at any given time during follow-up, the i^{th} individual's hazard lies on one of the p hazard curves, and it will follow that curve until his covariate

vector changes in value.

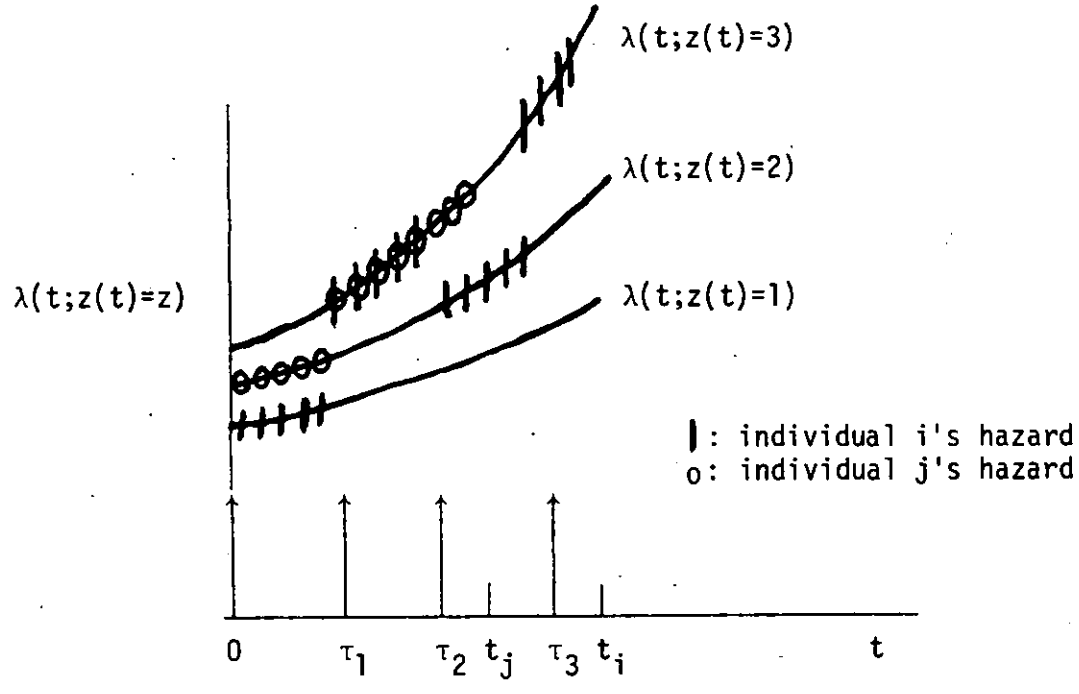
As an example, consider the three hazard curves in Figure 2.1. Using a single covariate, assume that $z(t)$ takes on the value 1,2, or 3. Assume that individual i maintains $z_i(t)=1$ during $0 \leq t < \tau_1$, $z_i(t)=3$ during $\tau_1 \leq t < \tau_2$, $z_i(t)=2$ during $\tau_2 \leq t < \tau_3$, and $z_i(t)=3$ during $\tau_3 \leq t \leq t_i$, where he fails at t_i. Assume that individual j maintains $z_j(t)=2$ during $0 \leq t < \tau_1$, and $z_j(t)=3$ during $\tau_1 \leq t \leq t_j$, where he fails at t_j. Then the vertical bars and circles along the hazard curves in Figure 2.1 denote the i^{th} and j^{th} individuals' hazards, $\lambda(t;z_i(t))$ and $\lambda(t;z_j(t))$, respectively.

Individuals' hazards, as functions of repeated measurements, may not be proportional over time. For instance, in Figure 2.1 the hazard for individual j is greater than the hazard for individual i during the intervals $[0,\tau_1)$ and $[\tau_2,t_j]$, but their hazards are equal during the interval $[\tau_1,\tau_2)$. However, proportionality may exist among the hazard curves that formulate the model, and it is this condition that defines multiplicative model. Therefore, any description of the hazard, as a function of repeated measurements, should focus on the hazard curves that define the model, $\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_\ell)$, $\ell=1,\ldots,p$, rather than the individuals' hazards. If the multiplicative model is not met, then these hazard curves are not proportional over time. This discussion is summarized as follows:

Property 2.1. Consider the hazard as a function of repeated measurements, and assume there are p distinct values for $\underset{\sim}{z}(t)$ in the data, $\underset{\sim}{\zeta}_1,\ldots,\underset{\sim}{\zeta}_p$. Then the multiplicative model,

FIGURE 2.1

AN EXAMPLE OF TWO INDIVIDUALS' HAZARDS,
AS FUNCTIONS OF REPEATED MEASUREMENTS



$$\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_{\ell}) = \lambda_0(t)\ g(\underset{\sim}{\zeta}_{\ell};\underset{\sim}{\beta}),\quad \ell=1,\ldots,p,$$

is satisfied if and only if for each pair of values, $\underset{\sim}{\zeta}_{\ell}$ and $\underset{\sim}{\zeta}_{\ell'}$, the ratio of their hazard curves

$$\frac{\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_{\ell})}{\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_{\ell'})},$$

does not depend on time. That is to say, the hazard curves which formulate the model are proportional rather than the ratio of hazards for two individuals.

An alternative characterization of the multiplicative model is obtained by considering each hazard curve separately. Define the relative change in the hazard curve $\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_{\ell})$ by

$$r.c.(t;\underset{\sim}{\zeta}_{\ell}) = \frac{\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_{\ell}) - \lambda(0;z(0)=\underset{\sim}{\zeta}_{\ell})}{\lambda(0;\underset{\sim}{z}(0)=\underset{\sim}{\zeta}_{\ell})},\quad \ell=1,\ldots,p.$$

Under the multiplicative model,

$$r.c.(t;\underset{\sim}{\zeta}_{\ell}) = \frac{\lambda_0(t)\ g(\underset{\sim}{\zeta}_{\ell};\beta) - \lambda_0(0)\ g(\underset{\sim}{\zeta}_{\ell};\beta)}{\lambda_0(0)\ g(\underset{\sim}{\zeta}_{\ell};\beta)}$$

does not depend on $\underset{\sim}{\zeta}_{\ell}$. Conversely, if $r.c.(t;\underset{\sim}{\zeta}_{\ell})$ does not depend on $\underset{\sim}{\zeta}_{\ell}$, then the ratio

$$\frac{\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_{\ell})}{\lambda(0;\underset{\sim}{z}(0)=\underset{\sim}{\zeta}_{\ell})}$$

does not depend on $\underset{\sim}{\zeta}_{\ell}$ for any t. This infers that $\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_{\ell})$ must factor into a product of two components, one a function of $\underset{\sim}{\zeta}_{\ell}$ and the

other a function of t. Thus the multiplicative model is satisfied, producing the following result:

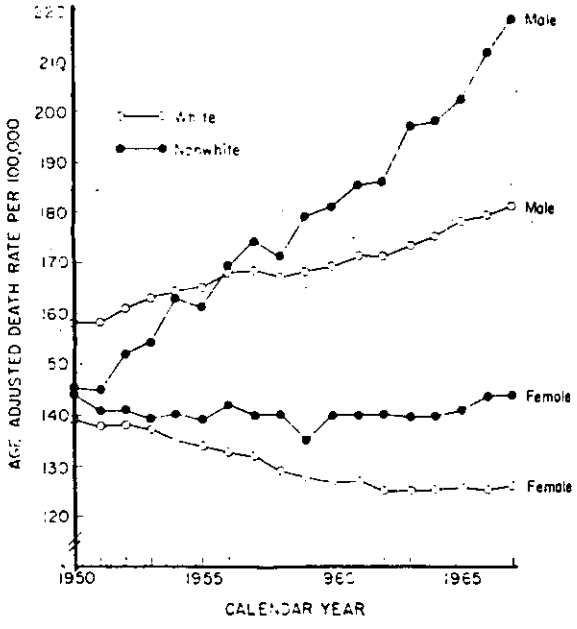
Property 2.2. The hazard as a function of repeated measurements, $\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_{\ell})$, $\ell=1,\ldots,p$, satisfies the multiplicative model if and only if for each ℓ, the relative change in the hazard curve, $r.c.(t;\underset{\sim}{\zeta}_{\ell})$, does not depend on $\underset{\sim}{\zeta}_{\ell}$, $\ell=1,\ldots,p$.

As an example, Figure 2.2 (Burbank and Fraumeni 1972) presents the annual U.S. black and white male age adjusted cancer mortality rates between 1950 and 1967. The cancer mortality rate ratio between black and white males was 0.93 in 1950, and had increased to 1.21 by 1967. Since these ratios are not equal, a lack of proportionality exists. Now consider the relative change in each race-specific cancer mortality rate curve. Between 1950 and 1967 the black male cancer mortality rate increased 50%, in contrast to only a 15% rise for white males. This information establishes that the black:white male cancer mortality rate ratio in 1967 is 1.5/1.15 times greater than in 1950, a result that also implies nonproportionality.

To conclude, the hazard ratio and the relative change in the hazard curve yield distinct as well as overlapping information. Either function may be used for describing nonproportionality. However, for the purpose of being concise, only one of these approaches will be used in this dissertation. The hazard ratio is chosen because it is analogous to the notion of relative risk.

FIGURE 2.2

TRENDS IN ANNUAL AGE ADJUSTED DEATH RATES FOR ALL SITES OF CANCER (INTERNATIONAL LIST NOS. 140-205 IN THE U.S., BY RACE AND SEX, 1950-67)



2.2 A Framework for Parametric Modelling of the Hazard

In Section 1.4 the hazard was defined as a function of covariates that describe baseline, repeated measurement, and intervening event information. Examples of such covariates were provided. In this section, a general parametric representation for the hazard curves $\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_\ell)$, $\ell=1,\ldots,p$, is proposed. Again, it is assumed that there are p distinct values for $\underset{\sim}{z}(t)$ in the data, $\underset{\sim}{\zeta}_1,\ldots,\underset{\sim}{\zeta}_p$.

The proposed model entails two features:

i) an underlying failure time distribution exists, and

ii) the model parameters are expressed as functions of the covariates.

Specifically, the hazard is defined by

$$\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_\ell) = \lambda_o(t;\underset{\sim}{\theta}^*(\underset{\sim}{\zeta}_\ell)), \quad \ell = 1,\ldots,p, \tag{2.1}$$

where $\lambda_o(t;\underset{\sim}{\theta}^*(\underset{\sim}{0})=\underset{\sim}{\theta}_o)$ denotes the underlying hazard and $\underset{\sim}{\theta}_o$ is comprised of w parameters. The vector $\underset{\sim}{\theta}^*(\underset{\sim}{\zeta}_\ell) = (\theta_1^*(\underset{\sim}{\zeta}_\ell),\ldots,\theta_w^*(\underset{\sim}{\zeta}_\ell))'$ is defined such that each $\theta_v^*(\underset{\sim}{\zeta}_\ell)$ is a particular parametric function of $\underset{\sim}{\zeta}_\ell$, $\ell=1,\ldots,p$, $v=1,\ldots,w$. Model (2.1) has the flexible quality of expressing any location, scale, or shape parameter that appears in the underlying hazard as a parametric function of the covariates. For instance, if one assumes a Gompertz underlying hazard, then one may define

$$\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_\ell) = \theta_1^*(\underset{\sim}{\zeta}_\ell)\, e^{\theta_2^*(\underset{\sim}{\zeta}_\ell)t}, \quad \ell=1,\ldots,p,$$

where $\theta_1^*(\underset{\sim}{\zeta}_\ell)$ and $\theta_2^*(\underset{\sim}{\zeta}_\ell)$ are greater than zero for each ℓ. In this model, the scale and shape parameters are expressed as functions of covariate information. Choose $\theta_v^*(\underset{\sim}{\zeta}_\ell) = \theta_{vo}\exp\{\underset{\sim}{\theta}_v'\underset{\sim}{\zeta}_\ell\}$, v=1,2. Then

the multiplicative model is satisfied if $\underset{\sim}{\theta}_2 = \underset{\sim}{0}$.

Model (2.1) defines a family of parametric representations for the hazard. Many of the forms presented in Chapter 1 are members of this family. Four examples are given below.

Example 2.1. The three models proposed by Feigl and Zelen (1965) are

$$\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_\ell) = \theta_1^*(\underset{\sim}{\zeta}_\ell) = \lambda(1+\underset{\sim}{\beta}'\underset{\sim}{\zeta}_\ell)^{-1},$$

$$\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_\ell) = \theta_1^*(\underset{\sim}{\zeta}_\ell) = \lambda + \underset{\sim}{\beta}'\underset{\sim}{\zeta}_\ell \ , \quad \text{and}$$

$$\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_\ell) = \theta_1^*(\underset{\sim}{\zeta}_\ell) = \lambda \exp\{\underset{\sim}{\beta}'\underset{\sim}{\zeta}_\ell\},$$

for $\ell=1,\ldots,p$. Here, the multiplicative model is satisfied and $\theta_1^*(\underset{\sim}{0})=\lambda$ denotes the exponential underlying hazard.

Example 2.2. The model proposed by Bailey, Homer, and Summe (1977) is

$$\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_\ell) = \theta_1^*(\underset{\sim}{\zeta}_\ell) + \theta_2^*(\underset{\sim}{\zeta}_\ell) \exp\{-\theta_3^*(\underset{\sim}{\zeta}_\ell)t\}, \quad \ell=1,\ldots,p, \quad (2.2)$$

where $\theta_v^*(\underset{\sim}{\zeta}_\ell) = \theta_{vo} \exp\left\{\sum_{u=1}^{s} \theta_{vu}\zeta_{u\ell}\right\}$, $v=1,2,3$, $z_u(t) = \zeta_{u\ell}$ is the value of the u^{th} covariate at time t, and s is the number of covariates. Here, $\theta_1^*(\underset{\sim}{\zeta}_\ell)$, $\theta_2^*(\underset{\sim}{\zeta}_\ell)$, and $\theta_3^*(\underset{\sim}{\zeta}_\ell)$ designate the location, scale, and shape parameters for the hazard. The underlying failure time distribution is defined by

$$\lambda_o(t;\underset{\sim}{\theta}_o) = \theta_{1o} + \theta_{2o}\, e^{-\theta_{3o}t},$$

where $\underset{\sim}{\theta}_o = (\theta_{1o},\theta_{2o},\theta_{3o})'$. If $\beta_u = \theta_{1u} = \theta_{2u}$ and $\theta_{3u} = 0$, $u=1,\ldots,s$, then the multiplicative model,

$$\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_\ell) = (\theta_{1o} + \theta_{2o}\, e^{-\theta_{3o}t}) \exp\{\underset{\sim}{\beta}'\underset{\sim}{\zeta}_\ell\}, \quad \ell=1,\ldots,p,$$

is attained.

Example 2.3. The Pareto model used by Turnbull, Brown, and Hu (1974) is generalized as

$$\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_\ell) = \theta_1^*(\underset{\sim}{\zeta}_\ell)\,[\theta_2^*(\underset{\sim}{\zeta}_\ell)+t]^{-1}, \quad \ell=1,\ldots,p,$$

where $\theta_1^*(\underset{\sim}{\zeta}_\ell)$ and $\theta_2^*(\underset{\sim}{\zeta}_\ell)$ designate the scale and location parameters for the hazard. Denoting $\theta_v^*(\underset{\sim}{\zeta}_\ell) = \theta_{vo} \exp\left\{\sum_{u=1}^{s} \theta_{vu}\zeta_{u\ell}\right\}$, v=1,2, and $\underset{\sim}{\theta}_o = (\theta_{1o},\theta_{2o})'$, the underlying hazard becomes

$$\lambda_o(t;\underset{\sim}{\theta}_o) = \theta_{1o}[\theta_{2o}+t]^{-1}.$$

If $\theta_{2u}=0$, u=1,...,s, then the multiplicative model is satisfied.

Example 2.4. The generalized Rayleigh hazard proposed by Taulbee (1977,1979), is

$$\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_\ell) = \sum_{v=0}^{w} \theta_v^*(\underset{\sim}{\zeta}_\ell)t^v, \quad \ell=1,\ldots,p,$$

where $\theta_v^*(\underset{\sim}{\zeta}_\ell) = \theta_{vo} \exp\left\{\sum_{u=1}^{s} \theta_{vu}\zeta_{u\ell}\right\}$, v=0,...,w. Writing $\underset{\sim}{\theta}_o = (\theta_{0o},\theta_{1o},\ldots,\theta_{wo})'$,

$$\lambda_o(t;\underset{\sim}{\theta}_o) = \sum_{v=0}^{w} \theta_{vo}t^v$$

denotes the underlying hazard. If $\beta_u=\theta_{0u}=\theta_{1u}=\ldots=\theta_{wu}$, u=1,...,s, then the multiplicative model,

$$\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_\ell) = \left\{\sum_{v=0}^{w} \theta_{vo}t^v\right\} \exp\{\underset{\sim}{\beta}'\underset{\sim}{\zeta}_\ell\}, \quad \ell=1,\ldots,p,$$

is attained.

The nonmultiplicative extension of Cox's model,

$$\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_{\ell}) = \lambda_0(t)\ \exp\{\underset{\sim}{\beta}_1'\underset{\sim}{\zeta}_{\ell} + (\underset{\sim}{\beta}_2'\underset{\sim}{\zeta}_{\ell})h(t)\},\ \ell=1,\ldots,p, \quad (2.3)$$

does not generalize the family of models in (2.1). For example, no representation of $\lambda_0(t)$ and $h(t)$ in (2.3) would yield the models in Examples 2.2-2.4.

To summarize, model (2.1) provides a framework for modelling nonproportionality among hazard curves. Examples 2.2-2.4 demonstrate that the multiplicative model is a reduced form of this more general model. Specifically, an underlying hazard exists, and each parameter in that hazard may be expressed as a function of the covariates.

Interpretation of Nonproportionality by the Hazard Ratio: Two Examples

Consider the underlying hazard that was proposed by Bailey, et al. (1977),

$$\lambda_0(t) = e^{\alpha_0}\left[1 + \beta_0\ e^{-\gamma_0 t}\right].$$

Incorporate the covariates $\underset{\sim}{z}(t)$ into this hazard by

$$\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_{\ell}) = \exp\{\alpha_0+\underset{\sim}{\alpha}'\underset{\sim}{\zeta}_{\ell}\}[1+\beta_0\ \exp\{-(\gamma_0+\underset{\sim}{\gamma}'\underset{\sim}{\zeta}_{\ell})t\}],\ \ell=1,\ldots,p. \quad (2.4)$$

The resulting hazard in (2.4) depends on $\underset{\sim}{\zeta}_{\ell}$ through the scale parameters $\underset{\sim}{\alpha}$ and the shape parameters $\underset{\sim}{\gamma}$. Each covariate vector $\underset{\sim}{\zeta}_{\ell}$ generates a distinct hazard curve $\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_{\ell})$ that decreases exponentially from time zero towards a lower bound. Clearly, the hazard curves that are generated by model (2.4) are not proportional unless $\underset{\sim}{\gamma} = \underset{\sim}{0}$.

The hazard ratio for model (2.4),

$$\frac{\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_{\ell})}{\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_{\ell'})} = e^{\underset{\sim}{\alpha}'(\underset{\sim}{\zeta}_{\ell}-\underset{\sim}{\zeta}_{\ell'})} \left\{\frac{1 + \beta_0 e^{-(\gamma_0+\underset{\sim}{\gamma}'\underset{\sim}{\zeta}_{\ell})t}}{1 + \beta_0 e^{-(\gamma_0+\underset{\sim}{\gamma}'\underset{\sim}{\zeta}_{\ell'})t}}\right\}, \qquad (2.5)$$

is a function of time t through $\underset{\sim}{\gamma}$. With only one covariate $z_1(t)$, assume $\gamma_1 < 0$ and $\zeta_{1\ell} > \zeta_{1\ell'}$. Then the hazard ratio in (2.5) equals $e^{\alpha_1(\zeta_{1\ell}-\zeta_{1\ell'})}$ at time 0, increases to a maximum value of

$$\left\{\frac{\gamma_0 + \gamma_1\zeta_{1\ell}}{\gamma_0 + \gamma_1\zeta_{1\ell'}}\right\} \exp\left\{(\alpha_1 - \gamma_1 t^*)(\zeta_{1\ell} - \zeta_{1\ell'})\right\}$$

at time t*, and then decreases back to $e^{\alpha_1(\zeta_{1\ell}-\zeta_{1\ell'})}$ by time ∞. The solution for t* is obtained by iteration.

In Chapter 3, model (2.4) is fitted to two cause-specific hazards of death in a follow-up study of heart diseased patients: the hazard of death due to congestive heart failure and the hazard of death due to other heart disease causes. The hazard ratio in (2.5) will be used for describing the observed lack of proportionality among the hazard curves.

As a second example, consider the following Gompertz hazard:

$$\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_{\ell}) = \begin{cases} \exp\{(\beta_0+\underset{\sim}{\beta}'\underset{\sim}{\zeta}_{\ell}) + (\gamma_0+\underset{\sim}{\gamma}'\underset{\sim}{\zeta}_{\ell})(t-\delta_0)\}, & t \geq \delta_0, \\ 0, \quad \text{otherwise.} \end{cases} \qquad (2.6)$$

for ℓ=1,...p. This hazard is a log linear function of the covariates, depending on $\underset{\sim}{\zeta}_{\ell}$ through the scale parameters $\underset{\sim}{\beta}$ and the shape parameters $\underset{\sim}{\gamma}$. Each covariate vector $\underset{\sim}{\zeta}_{\ell}$ generates a distinct hazard curve $\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{\zeta}_{\ell})$ that increases exponentially from time δ_0. Clearly, the

hazard curves that are generated by model (2.6) are not proportional unless $\underset{\sim}{\gamma} = \underset{\sim}{0}$.

The hazard in (2.6) assumes that the location parameter for T, δ_0, does not depend on $\underset{\sim}{z}_\ell$. That is to say, the time at which individuals are first at risk of failure is unaltered by their covariate levels.

The hazard ratio for model (2.6) is

$$\frac{\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{z}_\ell)}{\lambda(t;\underset{\sim}{z}(t)=\underset{\sim}{z}_{\ell'})} = \exp\left\{\underset{\sim}{\beta}'(\underset{\sim}{z}_\ell-\underset{\sim}{z}_{\ell'}) + \underset{\sim}{\gamma}'(\underset{\sim}{z}_\ell-\underset{\sim}{z}_{\ell'})(t-\delta_0)\right\}, \qquad (2.7)$$

and its log hazard ratio is a linear function of $(t-\delta_0)$. With one covariate $z_1(t)$, if $z_{1\ell} > z_{1\ell'}$ and $\gamma_1 < 0$, then the log hazard ratio decreases linearly with time, having a maximum value of $\beta_1(z_{1\ell}-z_{1\ell'})$ at the initial time $t=\delta_0$ and a slope of $\gamma_1(z_{1\ell}-z_{1\ell'})$.

In Chapter 4, model (2.6) is used for describing the dose-response effect of a textile worker's cumulative exposure to chrysotile asbestos on his hazard of death from lung cancer. The data had suggested that the age at which individuals began dying from lung cancer, δ_0, was not a decreasing function of exposure.

These examples demonstrate that each particular functional form for model (2.1) defines a distinct type of nonproportionality among hazard curves. For instance, the Gompertz and Weibull models cannot describe a hazard ratio that increases with time, achieves a maximum value, and then decreases with time. Even if it is known that the hazard increases monotonically with time, some other failure time distribution would be more appropriate.

2.3 Nonparametric Estimation of Cause-Specific Cumulative Hazards as Functions of Repeated Measurements

Nonparametric estimation of cause-specific cumulative hazards, as functions of repeated measurements, will be used primarily for selecting appropriate parametric forms of the cause-specific underlying hazards of interest. The multiplicative model is not assumed here, therefore, the proposed method requires a categorization of the covariates into strata. By assuming the piecewise exponential hazard model (1.33) within each covariate stratum, nonparametric cause-specific cumulative hazard curves are obtained. An appropriate parametric model is then selected by focusing on the slopes of these curves. In addition, these estimates will yield a graphical display of the covariate strata's effects on the cause-specific hazards of interest.

Assume that s covariates are available for each individual in a study. Categorize the u^{th} covariate into p_u values, u=1,...,s. Then there are a possible $p=p_1xp_2x...xp_s$ covariate strata in which an individual may belong at time t. Assuming that the covariates are defined by repeated measurements over time, an individual may belong to different covariate strata during different portions of his follow-up period. Clearly, an individual belongs to only one covariate stratum at any given time.

Denote C_ℓ to be the ℓ^{th} covariate stratum, ℓ=1,...,p. Assume that c causes of death are present in the data but only $c^* < c$ of them are of interest to the investigator. Let $0 < t_{(\ell k1)} < t_{(\ell k2)} < \ldots < t_{(\ell k r_{\ell k})}$ be the $r_{\ell k}$ distinct death times from cause k among

individuals that belonged to C_ℓ when they died, k=1,...,c*, ℓ=1,...,p. Then the following piecewise exponential cause-specific hazard is assumed for the k^{th} cause and the ℓ^{th} covariate stratum:

$$\lambda_k(t;\underset{\sim}{z}(t)\epsilon C_\ell) = \begin{cases} \lambda_{\ell kj}, t_{(\ell kj-1)} < t \leq t_{(\ell kj)}, \; j=1,\ldots,r_{\ell k}, \\ 0 \quad , t > t_{\ell k r_{\ell k}}, \end{cases} \tag{2.8}$$

where $t_{(\ell ko)}$=0, k=1,...,c*, ℓ=1,...,p. This model will generate a set of c* cause-specific cumulative hazard estimates for each of the p covariate strata, where $\sum_{\ell=1}^{p} \sum_{k=1}^{c^*} r_{\ell k}$ exact times to death are described by the same number of parameters.

The incomplete likelihood for model (2.8) - the survival information about the c-c* extraneous causes of death is omitted - is written as

$$L_{PE} = \prod_{\ell=1}^{p} \prod_{k=1}^{c^*} \prod_{j=1}^{r_{\ell k}} \lambda_{\ell kj}^{d_{\ell kj}} \exp\left\{-\lambda_{\ell kj} \sum_{i \epsilon R'_{\ell kj}} h_{\ell kji}\right\}. \tag{2.9}$$

Here, $d_{\ell kj}$ denotes the number of deaths from cause k at time $t_{(\ell kj)}$ among individuals that belonged to C_ℓ at $t_{(\ell kj)}$, $R'_{\ell kj}$ denotes the group of individuals that survived past $t_{(\ell kj-1)}$, and $h_{\ell kji}$ is the length of time spent by individual i in C_ℓ during the interval $(t_{(\ell kj-1)}, t_{(\ell kj)}]$. The term $L_{\ell kj} = \sum_{i \epsilon R'_{\ell kj}} h_{\ell kji}$ denotes the total amount of person-time spent by individuals in the covariate stratum C_ℓ during $(t_{(\ell kj-1)}, t_{(\ell kj)}]$. Maximizing L_{PE} with respect to $\lambda_{\ell kj}$, one obtains

$$\hat{\lambda}_{\ell kj} = \frac{d_{\ell kj}}{L_{\ell kj}}, \tag{2.10}$$

the observed death rate from cause k in $(t_{(\ell kj-1)}, t_{(\ell kj)}]$ among individuals that belonged to C_ℓ during that interval, $j=1,\ldots,r_{\ell k}$, $k=1,\ldots,c^*$, $\ell=1,\ldots,p$.

The k^{th} cause-specific cumulative hazard for the ℓ^{th} covariate stratum is defined by

$$\Lambda_k(t;\underset{\sim}{z}(u)\epsilon C_\ell,\ u\leq t) = \int_0^t \lambda_k(u;\underset{\sim}{z}(u)\epsilon C_\ell)du. \qquad (2.11)$$

Its maximum likelihood estimates at the observed death times are obtained from model (2.8) as

$$\hat{\Lambda}_{PE,k}(t_{(\ell kj)};\underset{\sim}{z}(u)\epsilon C_\ell,\ u\leq t) = \sum_{i=1}^{j} h_{\ell ki}\,\frac{d_{\ell ki}}{L_{\ell ki}}, \qquad (2.12)$$

where $h_{\ell ki} = t_{(\ell ki)} - t_{(\ell ki-1)}$, $j=1,\ldots,r_{\ell k}$, $k=1,\ldots,c^*$, $\ell=1,\ldots,p$. The estimated variances of these estimates are obtained from the inverse of the observed Fisher information matrix for $\lambda_{\ell kj}$, $j=1\ldots,r_{\ell k}$, $k=1,\ldots,c^*$, $\ell=1,\ldots,p$. Specifically,

$$\hat{var}\left[\hat{\Lambda}_{PE,k}(t_{(\ell kj)};\underset{\sim}{z}(u)\epsilon C_\ell,\ u\leq t)\right] = \sum_{i=1}^{j} h_{\ell ki}^2\,\frac{d_{\ell ki}}{L_{\ell ki}^2}. \qquad (2.13)$$

Graphs of $t_{(\ell kj)}$ versus $\hat{\Lambda}_{PE,k}(t_{(\ell kj)};\underset{\sim}{z}(u)\epsilon C_\ell,\ u\leq t)$, $k=1,\ldots,c^*$, are used for selecting appropriate parametric forms of the cause-specific underlying hazards. One evaluates the shape of $\lambda_k(t;\underset{\sim}{z}(t)\epsilon C_\ell)$ by focusing on the slope of the estimates for $\Lambda_k(t;\underset{\sim}{z}(u)\epsilon C_\ell,\ u\leq t)$.

2.4 An Example with a Zero-One Covariate: Heart Transplant Status

Consider the Stanford Heart Transplant data, given in Appendix I of Kalbfleisch and Prentice (1980). As defined in Section 1.4, let

$$z_1(t) = \begin{cases} 1 & \text{if the patient received a heart transplant within t days of follow-up,} \\ 0 & \text{otherwise.} \end{cases}$$

Then $\lambda(t;z_1(t)=0)$ and $\lambda(t;z_1(t)=1)$ denote a patient's pre-and post-transplant hazards of death from all causes.

Figure 2.3 presents the estimated pre-and post- transplant cumulative hazard curves,

$$\hat{\Lambda}_{PE}(t;z_1(u)=0,\ u\leq t) \quad \text{and} \quad \hat{\Lambda}_{PE}(t;z_1(u)=1,\ u\leq t),$$

obtained by the piecewise exponential formula (2.12). The slopes in both curves appear to decline steadily from time zero, and then achieve a lower bound by one year of follow-up. This suggests the underlying failure time distribution that Bailey, Homer, and Summe (1977) used for investigating survival among kidney transplant patients,

$$\lambda_0(t) = \alpha_0 + \beta_0 e^{-\gamma_0 t}. \tag{2.14}$$

Since the cumulative hazard's slopes for the two groups appear to remain equal during the first three months of rapid decline, the parameters β_0 and γ_0 may not be affected by transplant status. The three cumulative hazard points for the pretransplant group after 150 days of follow-up suggest that the hazard's asymptotic lower bound, α_0, may be greater for this group than for the posttransplant group.

The following four parameter model was fitted to the data by the method of maximum likelihood, using the MAXLIK program (Kaplan and

Elston 1972):

$$\lambda(t;z_1(t)) = \alpha_0\alpha_1^{z_1(t)} + \beta_0 e^{-\gamma_0 t}. \qquad (2.15)$$

The maximum likelihood estimate of α_1 was 0.945 (s.e.=0.88) and the likelihood ratio test of H_0: $\alpha_1=1$ yielded $\chi_1^2 = 0.02$, $p = .89$. If $\alpha_1 < 1$, then a patient's prognosis for survival would noticeably improve with a heart transplant only beyond six months of follow-up. Unfortunately, with only two pretransplant deaths after six months of follow-up, no powerful comparison between the two groups' hazards is possible beyond that time.

The maximum likelihood estimates of the parameters in model (2.14) were $\hat{\alpha}_0 = 0.00067$ (s.e.=0.00022), $\hat{\beta}_0 = 0.0102$ (s.e.=0.0022), and $\hat{\gamma}_0 = 0.0123$ (s.e.=0.0035). Its fitted cumulative hazard curve appears in Figure 2.3, and a close fit is achieved.

The most generalized version of model (2.14) was also fitted to the data:

$$\lambda(t;z_1(t)) = \alpha_0\alpha_1^{z_1(t)} + \beta_0\beta_1^{z_1(t)} e^{-\gamma_0\gamma_1^{z_1(t)}t}.$$

The log likelihood for this six parameter model, -490.92, was negligibly larger than that obtained by model (2.14), -490.98.

For comparison purposes, Cox's multiplicative model,

$$\lambda(t;z_1(t)) = \lambda_0(t)\, e^{\beta z_1(t)},$$

was fitted to the data, using Breslow's approximation to Cox's partial likelihood, (1.35). The heart transplant effect on survival was estimated as $\hat{\beta} = 0.106$ (s.e.=0.3). By assuming the full parametric model

(2.14) for $\lambda_0(t)$ above, the maximum likelihood estimate of β was $\hat{\beta} = 0.0454$ (s.e.=0.28). These results were consistent in that the test of H_0: $\beta=0$ was clearly accepted.

Finally, an exponential hazard model was fitted to the data,

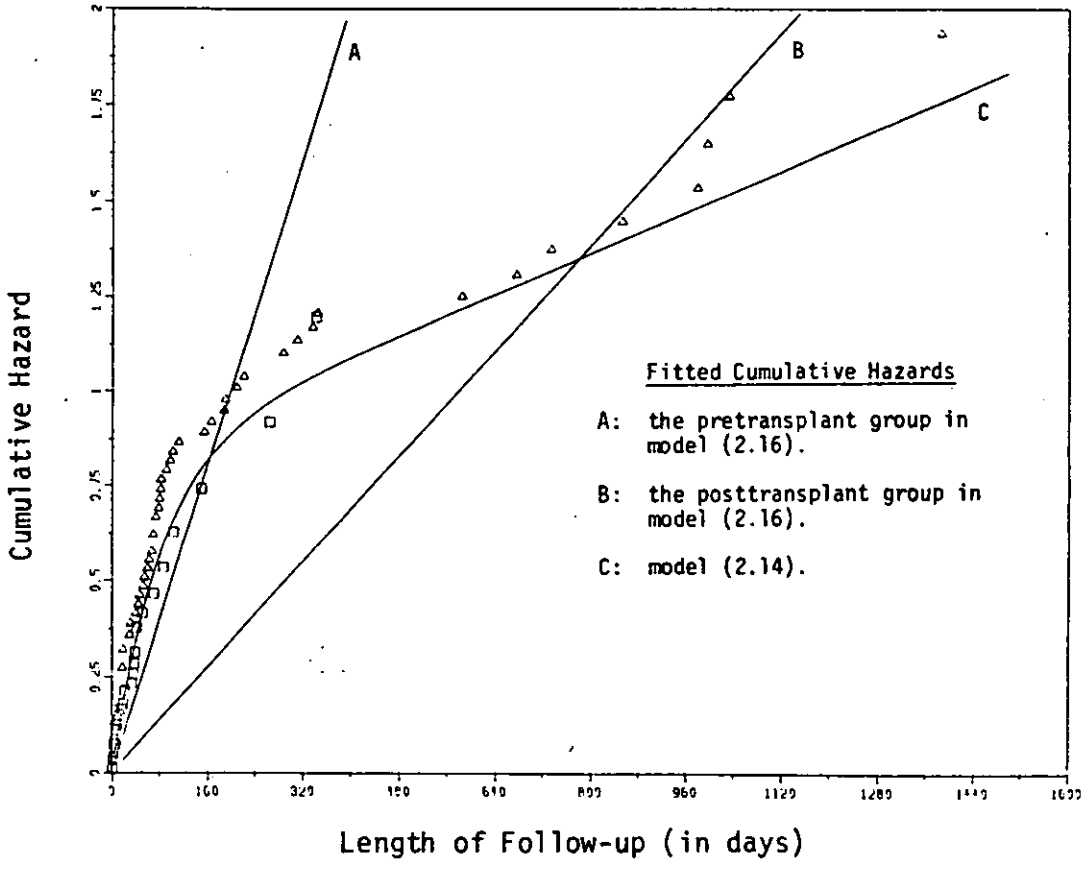
$$\lambda(t;z_1(t)) = \alpha_0\alpha_1^{z_1(t)}. \qquad (2.16)$$

The maximum likelihood estimates were $\hat{\alpha}_0 = 0.0051$ (s.e.=0.0009) and $\hat{\alpha}_1 = 0.3424$ (s.e.=0.0807). The likelihood ratio test of H_0: $\alpha_1=1$ was rejected ($\chi_1^2 = 18.5$, $p = .00002$), inferring that a heart transplant significantly improves one's prognosis for survival. This result contradicts that obtained by model (2.15). The maximized log likelihood for the exponential hazard model (2.16), -519.8, is substantially smaller than that achieved by the no covariate model (2.14). The poorer fit of model (2.16) becomes apparent in Figure 2.3, where the fitted cumulative hazards for this model, the two straight lines, deviate substantially from the nonparametric estimates. In summary, these results demonstrate that if a completely specified parametric hazard is to be used, then a nonparametric assessment of the appropriate underlying failure time distribution becomes essential. A visual contrast of the fitted curves with the nonparametric estimates can be very informative.

The hazard $\lambda(t;z_1(t))$ compares the death rates among patients that have and have not yet received a new heart by follow-up time t. This model does not consider the length of time that a patient has maintained a new heart. To allow for this information, a second covariate,

FIGURE 2.3

Cumulative Hazard Plots for the Stanford Heart Transplant Data. The squares and triangles denote the piecewise exponential cumulative hazard estimates for the pretransplant and posttransplant groups, respectively. Note a square between two triangles at 340 days of follow-up.



$$z_2(t) = \begin{cases} t-w, & t > w, \\ 0, & \text{otherwise,} \end{cases}$$

where w is the patient's waiting time to heart transplant, may be included into the hazard. One would then fit $\lambda(t;z_1(t),z_2(t))$ to the data.

A second approach is to define $\lambda_0(t)$ as the hazard for the pretransplant group and $\lambda_1(t^*)$ as the hazard for the posttransplant group, where t* denotes the length of follow-up since transplant. These two hazards at time $t^* = t$ compare the death rates for pretransplant individuals followed to time t with posttransplant individuals maintaining a new heart for length of time t. Since the pretransplant group's hazard was found to decrease exponentially during the first six months of follow-up, it may be important to adjust the posttransplant group's hazard for the individual's waiting time to heart transplant. Specifically, $\lambda_1(t^*;w)$ would be used.

2.5 Creating the Likelihood as a Function of the Minimal Sufficient Statistics: Computational Efficiency is Achieved

Assume that c causes of death are present in the data but only $c^* < c$ of them are of interest to the investigator. As discussed in Section 1.6, an incomplete likelihood is created in which deaths from the c-c* extraneous causes are treated as censored observations. Assume that repeated measurements are incorporated into the c* cause-specific hazards by (1.12), where $\tau_0,\tau_1,\tau_2,\ldots,\tau_q$ denote the q+1 follow-up times that measurements are taken (τ_0 denotes the beginning of the study). Further assume that there are

p distinct values for $\underset{\sim}{z}(t)$ in the data, $\underset{\sim}{\zeta}_1,\ldots,\underset{\sim}{\zeta}_p$. Let $t_{(m\ell k1)} < t_{(m\ell k2)} < \ldots < t_{(m\ell k r_{m\ell k})}$ denote the $r_{m\ell k}$ exact times to death from cause k in the interval $[\tau_m,\tau_{m+1})$ among individuals with covariate vector $\underset{\sim}{\zeta}_\ell$ at their deaths, k=1,...,c*, ℓ=1,...,p, m=0,...,q (τ_{q+1} denotes the time at which the study ends). Let $d_{m\ell kj}$ be the number of individuals that died from cause k at $t_{(m\ell kj)}$ with covariate value $\underset{\sim}{\zeta}_\ell$ when they died. Finally, assume that censoring occurs only at the follow-up times $\tau_0,\tau_1,\tau_2,\ldots,\tau_{q+1}$ - this assumption was met for the data in Chapters 3 and 4. Then the incomplete likelihood in (1.21) is rewritten as

$$L_{II}(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_{c^*}) = \left\{ \prod_{m=0}^{q} \prod_{\ell=1}^{p} \prod_{k=1}^{c^*} \prod_{j=1}^{r_{m\ell k}} \lambda_k(t_{(m\ell kj)};\underset{\sim}{\theta}_k,\underset{\sim}{\zeta}_\ell)^{d_{m\ell kj}} \right\} \cdot$$

$$\left\{ \prod_{m=0}^{q} \prod_{\ell=1}^{p} \exp\left[-\sum_{k=1}^{c^*} \sum_{j=1}^{r_{m\ell k}} d_{m\ell kj} \int_{\tau_m}^{t_{(m\ell kj)}} \left(\sum_{k=1}^{c^*} \lambda_k(t;\underset{\sim}{\theta}_k,\underset{\sim}{\zeta}_\ell) \right) dt \right. \right.$$

$$\left. \left. - n_{m\ell} \int_{\tau_m}^{\tau_{m+1}} \left(\sum_{k=1}^{c^*} \lambda_k(t;\underset{\sim}{\theta}_k,\underset{\sim}{\zeta}_\ell) \right) dt \right] \right\}, \quad (2.17)$$

where $n_{m\ell}$ denotes the number of individuals who survived the interval $[\tau_m,\tau_{m+1})$ and had the covariate value $\underset{\sim}{\zeta}_\ell$ during that time. If censoring had occurred between the times at which measurements were performed, then an additional term would be included into the exponent in (2.17).

The minimal sufficient statistics for the likelihood in (2.17) are $\{\underset{\sim}{\zeta}_\ell, \ell=1,\ldots,p\}$, $\{(t_{(m\ell kj)},d_{m\ell kj})$, j=1,...,$r_{m\ell k}$, k=1,...,c*, ℓ=1,...,p, m=0,...,q$\}$, and $\{n_{m\ell}$, ℓ=1,...,p, m=0,...,q$\}$. There are

a total of

$$p + \sum_{m=0}^{q} \sum_{\ell=1}^{p} \sum_{k=1}^{c^*} \sum_{j=1}^{r_{m\ell k}} \text{minimum}(2, d_{m\ell kj}) + \sum_{m=0}^{q} \sum_{\ell=1}^{p} \text{minimum}(1, n_{m\ell})$$

minimal sufficient statistics, in comparison with

$$p + \sum_{m=0}^{q} \sum_{\ell=1}^{p} \sum_{k=1}^{c^*} \sum_{j=1}^{r_{m\ell k}} d_{m\ell kj} + \sum_{m=0}^{q} \sum_{\ell=1}^{p} n_{m\ell}$$

bits of raw data.

By using the minimal sufficient statistics to create the likelihood, a large amount of data reduction may be accomplished. In addition, the computer time required for maximizing the likelihood may be substantially reduced. For example, in fitting the four covariate model (3.7) in Chapter 3, there were 7054 bits of raw data in comparison with 1191 minimal sufficient statistics.

Likelihood (2.17) was used for the model fitting in Chapter 3, where $c^* = 2$. A simplified version of this likelihood was used in Chapter 4, where $c^* = 1$ and all of the deaths within $[\tau_m, \tau_{m+1})$ were assumed to occur at τ_m (see Section 4.2). Maximization of these likelihoods were performed by the MAXLIK program (Kaplan and Elston, 1972).

CHAPTER 3

AN EXAMPLE WITH REPEATED MEASUREMENTS: HEART DISEASE DATA FROM DUKE UNIVERSITY

3.1 Introduction

The Duke University Heart Disease Study consists of patients that received a cardiac catheterization at Duke Hospital any time between November 1969 and the present. A patient's follow-up began on the day of his catheterization, time zero. The subsequent data analysis is based on 1,952 patients' follow-up through 1979, a ten-year period. These patients had chest pain prior to catheterization, significant coronary artery disease defined by at least 75% obstruction in one or more major coronary arteries, and received medical treatment as opposed to coronary bypass surgery.

In a Cox's model analysis of the effects of eighty-one baseline measurements on the hazard of death from all heart disease causes, Harris et al. (1979,1981) found the following three covariates to be among the most important prognostic indicators:

i) congestive heart failure (CHF) status, after being dichotomized into most severe (level 4) and less severe (levels 0,1,2, or 3) groups,

ii) the zero-one covariate "significant" left main stenosis (LMSIG), and

iii) left ventricular ejection fraction (LVEF).

The five levels of CHF status are defined by the New York Heart Association: 0 represents no symptoms, 1,2, and 3 represent progressively more severe levels, and 4 represents the most severe level. The covariate LMSIG equals one if the individual's left main coronary artery had at least 75% occlusion, zero otherwise. The covariates LMSIG and LVEF were determined from the results of the individual's catheterization.

Measurements were only repeated on CHF status, at the follow-up times 6 months, 1 year, and yearly thereafter. Using the notation in Section 1.4, denote CHF_m as the individual's measured CHF value at time τ_m, m=0,1,...,10, where τ_0=0, τ_1=6 months, τ_2=1 year, ..., τ_{10}=9 years. Then

$$CHF(t) = CHF_m,\ t \in [\tau_m,\tau_{m+1}),\ m=0,1,\ldots,10,$$

where τ_{11}=10 years, the maximum follow-up period. If the observed follow-up time on an individual is less than 10 years, then the covariate CHF(t) is defined for him up to that time. By allowing CHF_m to represent the individual's mean congestive heart failure status during the interval $[\tau_m,\tau_{m+1})$, the covariate CHF(t) approximates the individual's mean congestive heart failure status within one year of time t.

Although the individuals' measurements on CHF status were repeated over time, the investigators at Duke University did not intend to incorporate this information into the hazard. The data are being presented only as an application of the methodology. Clearly, an

optimal design would have required shorter time intervals between these measurements.

In this chapter, the effects of the covariates LMSIG, LVEF, and CHF(t) on the hazards of death due to congestive heart failure and other heart disease causes were investigated. In order to reduce the number of minimal sufficient statistics in the likelihood, thereby reducing computation costs, the covariate LVEF was subtracted by its observed mean, divided by its standard deviation, and then categorized into seven values: -3 (at least 2.5 standard deviations less than the mean), -2 (between 1.5 and 2.5 standard deviations less than the mean), and similarly -1,0,1,2, and 3. The five possible values for CHF(t) were maintained. The data included 44 observed deaths due to congestive heart failure and 235 observed deaths due to other heart disease causes.

Given that the individual's most recent CHF status, CHF(t), is included into the model as a covariate, it was of interest to determine if any prior knowledge of CHF status is also important. Four "past" CHF history covariates were used. The covariate CHF2(t) is defined by the individual's second most recent measurement on CHF status,

$$\text{CHF2}(t) = \begin{cases} \text{undefined}, & t < 6 \text{ months} \\ \text{CHF}_{m-1}, & t \in [\tau_m, \tau_{m+1}),\ m=2,\ldots,10. \end{cases}$$

If the observed follow-up time on an individual is less than ten years, then CHF2(t) is defined for him up to that time. This covariate approximates the individual's mean CHF status within two

years but beyond six months of time t. Since no measurements on CHF status were performed prior to time zero, CHF2(t) is undefined during the first six months of follow-up. Therefore, the effects of this covariate on the two cause-specific hazards of interest are assessed from patients' survival experiences beyond six months of follow-up.

The covariates CHF3(t), CHFP3(t), and CHFL3(t) synthesize the patient's CHF history within three years of time t:

$$\text{CHF3}(t) = \begin{cases} \text{undefined, } t < 2 \text{ years} \\ (.5\text{CHF}_0 + .5\text{CHF}_1 + \text{CHF}_2 + \text{CHF}_3)/3, \; t\varepsilon[2 \text{ years}, 3 \text{ years}), \\ (\text{CHF}_{m-2} + \text{CHF}_{m-1} + \text{CHF}_m)/3, \; t\varepsilon[\tau_m, \tau_{m+1}), \; m=4,\ldots,10; \end{cases}$$

$$\text{CHFP3}(t) = \begin{cases} \text{undefined, } t < 2 \text{ years} \\ 1 \quad \text{if} \quad \text{CHF}_0 \le \text{CHF}_1 \le \text{CHF}_2 \le \text{CHF}_3 \quad \text{and} \\ \qquad \text{CHF}_0 < \text{CHF}_3, \; t\varepsilon[2 \text{ years}, 3 \text{ years}), \\ 1 \quad \text{if} \quad \text{CHF}_{m-2} \le \text{CHF}_{m-1} \le \text{CHF}_m \quad \text{and} \\ \qquad \text{CHF}_{m-2} < \text{CHF}_m, \; t\varepsilon[\tau_m, \tau_{m+1}), \; m=4,\ldots,10, \\ 0 \quad \text{otherwise;} \end{cases}$$

and

$$\text{CHFL3}(t) = \begin{cases} \text{undefined, } t < 2 \text{ years} \\ .5\delta_0 + .5\delta_1 + \delta_2 + \delta_3, \; t\varepsilon[2 \text{ years}, 3 \text{ years}) \\ \delta_{m-2} + \delta_{m-1} + \delta_m, \; t\varepsilon[\tau_m, \tau_{m+1}), \; m=4,\ldots,10, \end{cases}$$

where

$$\delta_m = \begin{cases} 1 & \text{if } CHF_m = 4, \quad m=0,1,\ldots,10, \\ 0 & \text{otherwise.} \end{cases}$$

If the observed follow-up time on an individual is less than ten years, then CHF3(t), CHFP3(t), and CHFL3(t) are defined for him up to that time. The covariate CHF3(t) denotes the individual's mean CHF status within three years of time t. The covariate CHFP3(t) denotes the individual's CHF progression status within three years of time t. It equals one if the individual's CHF status became steadily worse during the three years prior to time t, zero otherwise. The covariate CHFL3(t) denotes the length of time within three years of t that an individual maintained the most severe CHF status, level 4. Since no information was available on the patients' CHF status prior to time 0, the covariates CHF3(t), CHFP3(t), and CHFL3(t) are undefined for $t < 2$ years. Their effects on the cause-specific hazards of interest are assessed from the individual's survival experiences beyond two years of follow-up.

If the baseline covariates LMSIG and LVEF appear in a model, and either LMSIG or LVEF is missing on an individual, then his survival experience is omitted from the likelihood. With repeated measurements, however, an individual's covariate information may be complete in some time intervals but incomplete in others. In only those time intervals where the individual's covariate values are missing will his survival experience be omitted from the likelihood. For example, if CHF(t) is included in a model and CHF_m was not recorded at time τ_m for an individual, then his covariate CHF(t) is missing in $[\tau_m, \tau_{m+1})$ and his survival experience during that interval

would be omitted from the likelihood. This would include the omission of the individual's hazard at time of death t if $t \in [\tau_m, \tau_{m+1})$.

Section 3.2 presents the analysis for one covariate, CHF(t). The cause-specific hazards for death due to congestive heart failure and death due to other heart disease causes are denoted by $\lambda_1(t;CHF(t))$ and $\lambda_2(t;CHF(t))$, respectively. The results of including the covariates LMSIG, LVEF, CHF2(t), CHF3(t), CHFP3(t), and CHFL3(t) into the analysis are presented in Section 3.3. To check the appropriateness of the assumed underlying hazard, Section 3.4 compares the results obtained by the parametric model in Section 3.3 with those obtained by Cox's (1972) model. Finally, in Section 3.5, the information gained by using the repeated measurements on CHF status instead of the baseline value is determined.

3.2 Analysis with One Covariate: Congestive Heart Failure (CHF) Status

Nonparametric maximum likelihood estimates of $\Lambda_k(t;CHF(u)=\ell, u \leq t)$, for $\ell=0,1,2,3,4$ and $k=1,2$ were computed by the piecewise exponential formula in (2.12). These cause-specific cumulative hazard estimates for death due to congestive heart failure and death due to other heart disease causes appear as the plotted symbols in Figures 3.1 and 3.2, respectively. For each cause of death and each level of the covariate, the hazard (the slope of the curve) appears to decrease from time zero towards a lower bound, which is obtained between three months and two years of follow-up. This result is consistent with the Stanford Heart Transplant Data in Section 2.4 and with the model used by Bailey et al. (1977) for investigating survival among kidney

transplant patients. The observed decreases in these hazards may be due to a large number of patients that were acutely ill at the time of their catheterizations (time zero) and then died shortly thereafter. In addition, patients' conditions may have become stabilized by the medical treatment, thereby producing a constant hazard with time by two years of follow-up.

The following parametric model was investigated for this data:

$$\lambda_k(t;\text{CHF}(t)=\ell) = \alpha_{k\ell} + \beta_{k\ell}e^{-\gamma_{k\ell}t}, \qquad (3.1)$$

where $\ell=0,1,2,3,4$, $k=1,2$, and t denotes the number of days of follow-up. This is a generalization of the model used by Bailey et al. (1977) to time dependent covariates and more than one cause of death. Model (3.1) assumes a distinct hazard with three parameters for each cause of death and covariate level, 30 parameters in all.

Since there were only two deaths due to congestive heart failure among patients that had CHF(t)=2 at the time of their deaths, the three parameters α_{12}, β_{12}, and γ_{12} in the hazard $\lambda_1(t;\text{CHF}(t)=2)$ could not be estimated from the data. To avoid this dilemma, α_{12}, β_{12}, and γ_{12} were set equal to α_{13}, β_{13}, and γ_{13}, respectively. The maximum likelihood estimates of the parameters in model (3.1) suggested the following 21 parameter, reduced model:

$$\lambda_k(t;\text{CHF}(t)=\ell) = \alpha_{k\ell}(1 + \beta_o e^{-\gamma_{k\ell}t}),\ k=1,2,\ \ell=0,\ldots,4. \qquad (3.2)$$

This model assumes the parameter β_o to be constant across cause of death and covariate level. Model (3.2) was fitted to the data with $\alpha_{12} = \alpha_{13}$ and $\gamma_{12} = \gamma_{13}$. The likelihood ratio test of H_o: $\beta_{k\ell}=\beta_o\alpha_{k\ell}$,

FIGURE 3.1

FITTED CUMULATIVE HAZARDS FOR MODEL (3.3):
DEATH DUE TO CONGESTIVE HEART FAILURE

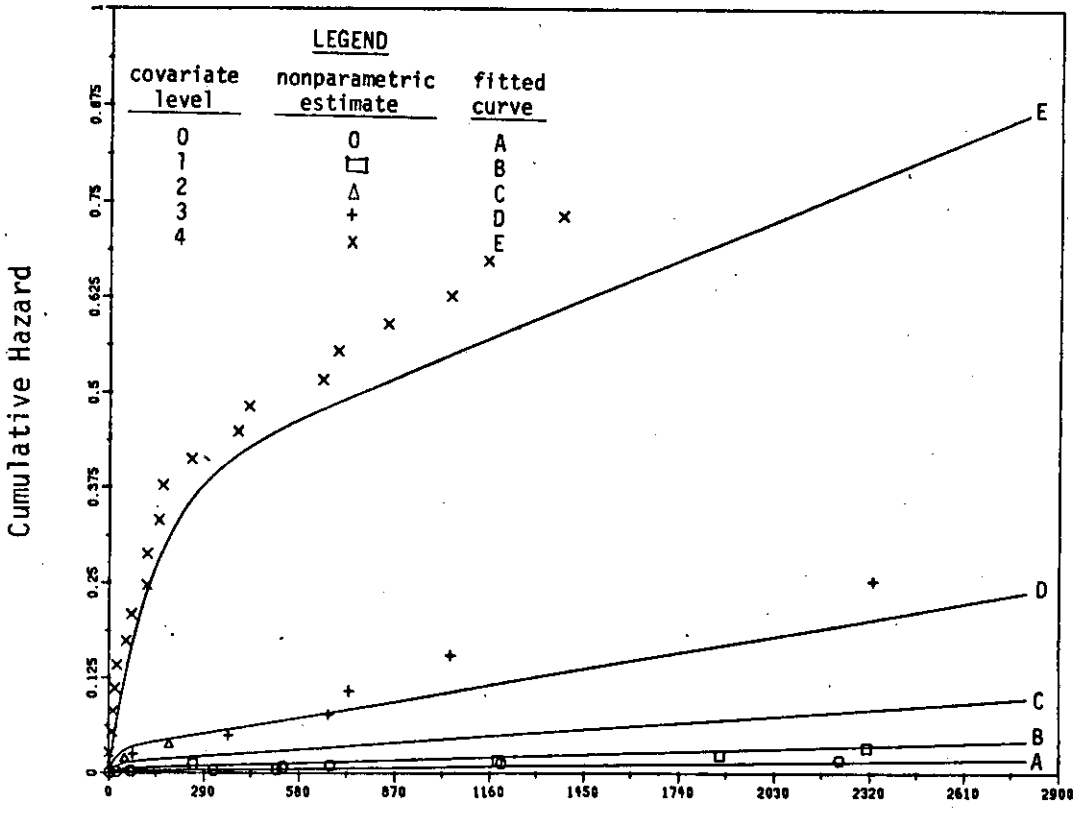
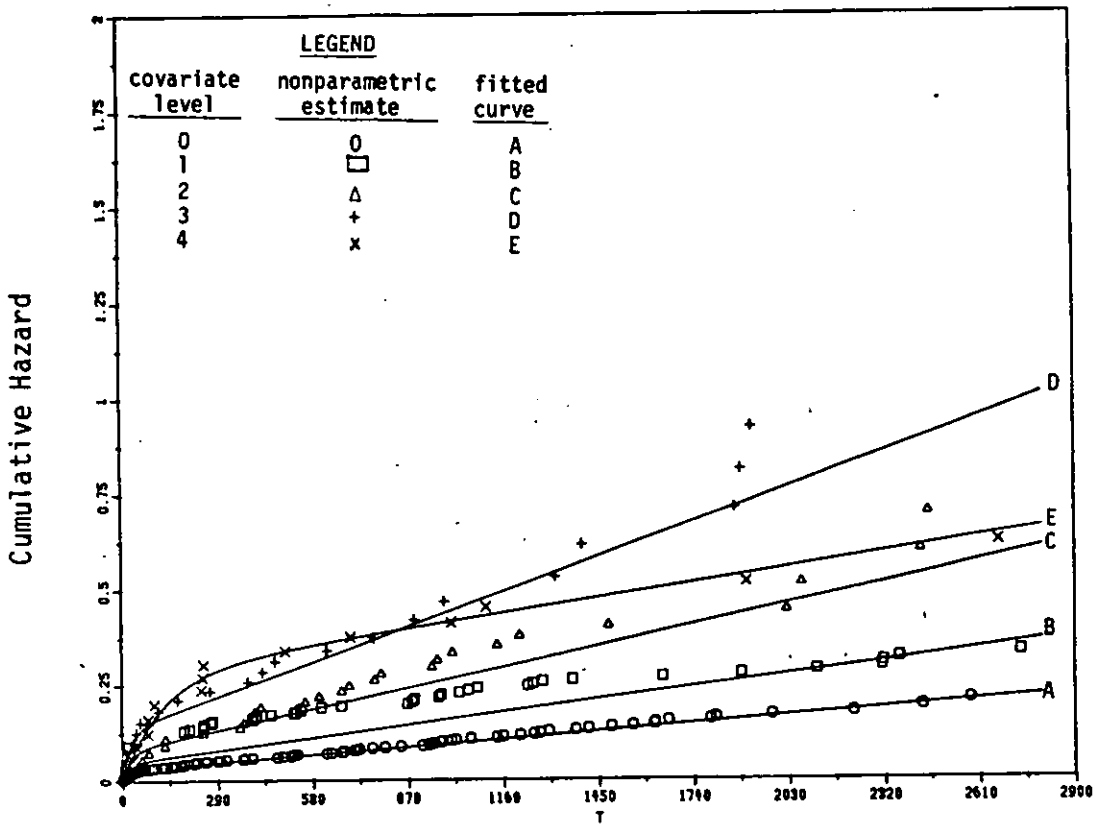


FIGURE 3.2

FITTED CUMULATIVE HAZARDS FOR MODEL (3.3):
DEATH DUE TO OTHER HEART DISEASE CAUSES



Length of Follow-up (in days)

for each k and ℓ, was nonsignificant, yielding $\chi^2_8 = 2.16$, $p = .98$. Model (3.2) was then refitted to the data without any restrictions on the parameters α_{12} and γ_{12}.

Further reduction of model (3.1) produced the eight parameter model

$$\lambda_k(t;CHF(t)) = \alpha_k(CHF(t))(1 + \beta_0 e^{-\gamma(CHF(t))t}), \qquad (3.3)$$

where

$$\alpha_1(CHF(t)) = \exp\{\alpha_{10} + \alpha_{11}CHF(t)\} ,$$

$$\alpha_2(CHF(t)) = \exp\{\alpha_{20} + \alpha_{21}CHF(t) + \alpha_{22}h(CHF(t))\} ,$$

$$\gamma(CHF(t)) = \gamma_0 + \gamma_1 h(CHF(t)), \text{ and}$$

$$h(CHF(t)) = \begin{cases} 1 & \text{if } CHF(t) = 4 \\ 0 & \text{otherwise.} \end{cases}$$

The scale parameters α_{10}, α_{11}, α_{20}, α_{21}, and α_{22} determine the hazard curves' asymptotic lower bounds in model (3.3). The underlying scale parameters for the two causes of death are $\exp(\alpha_{10})$ and $\exp(\alpha_{20})$. The parameters α_{11} and α_{21} represent convex, multiplicative effects of CHF(t) on the two hazards of death. The multiplicative term for the hazard of death due to other heart disease causes, $\alpha_2(CHF(t))$, was found to be smaller at CHF(t)=4 than at CHF(t)=3. This result is reflected in Figure 3.2 by a greater slope in the + curve than in the x curve beyond two years of follow-up. To account for this lack of monotonicity, the parameter α_{22} is included in model (3.3). In addition to the parameter β_0, the shape parameters γ_0 and γ_1 are consistent across the two causes of death. For $0 \leq CHF(t) \leq 3$, the shape

parameter is γ_0. At the most severe CHF level, CHF(t)=4, the shape parameter was found to be significantly smaller, with γ_1 denoting this effect.

The maximum likelihood estimates of the parameters in model (3.3), along with their standard errors, appear in Table 3.1. Graphical display of this model's fitted cause-specific cumulative hazard curves,

$$\hat{\Lambda}_k(t;\text{CHF}(u)=\ell,\ u\le t),\ k=1,2,\ \ell=0,\ldots,4,$$

appear in Figures (3.1) and (3.2). Comparison of these fitted curves with their corresponding nonparametric estimates suggest a good fit.

The cause of death acts multiplicatively on the hazard in model (3.3) since the ratio of the two cause-specific hazard curves at each covariate level,

$$\frac{\lambda_1(t;\text{CHF}(t)=\ell)}{\lambda_2(t;\text{CHF}(t)=\ell)} = \exp\left\{(\alpha_{10}-\alpha_{20}) + (\alpha_{11}-\alpha_{21})\ell-\alpha_{22}h(\ell)\right\},\ \ell=0,\ldots,4,$$

where $$h(\ell) = \begin{cases} 1 & \text{if } \ell=4, \\ 0 & \text{otherwise,} \end{cases}$$

does not depend on time t. The log ratio of the two underlying hazards, $(\alpha_{20}-\alpha_{10})$, was estimated to be 2.607 ± .297, demonstrating a substantially larger underlying hazard for death due to other heart disease causes. Although the likelihood ratio test of H_0: $\alpha_{22}=0$ yielded $\chi^2_1 = 8.4$, p = .0038, it will be shown in Section 3.3 that this lack of monotonicity in the CHF(t) effect is no longer significant after other covariates are included into the model. The maximum likelihood estimates $\hat{\alpha}_{11} = .91 \pm .11$ and $\hat{\alpha}_{21} = .50 \pm .07$ demonstrate that CHF

status is a stronger prognostic indicator for death due to congestive heart failure than for death due to other heart disease causes.

The covariate CHF(t) does not act multiplicatively on the cause-specific hazards in model (3.3) since the hazard ratio

$$\frac{\lambda_k(t;CHF(t)=4)}{\lambda_k(t;CHF(t)=\ell,\ell\neq 4)} = e^{\{\alpha_{k1}(4-\ell)+\alpha_{22}(k-1)\}} \cdot \frac{(1+\beta_0 e^{-(\gamma_0+\gamma_1)t})}{(1+\beta_0 e^{-\gamma_0 t})}, k=1,2, \tag{3.4}$$

is a function of time t through the shape parameter γ_1. The likelihood ratio test of H_0: $\gamma_1=0$, i.e., the test for the multiplicative model, was highly significant, yielding $\chi_1^2 = 14.1$, $p = .0002$. The fitted time dependent factor for the hazard ratio in (3.4),

$$\frac{(1 + \hat{\beta}_0 e^{-(\hat{\gamma}_0+\hat{\gamma}_1)t})}{(1 + \hat{\beta}_0 e^{-\hat{\gamma}_0 t})}, \tag{3.5}$$

appears in Figure 3.3. It demonstrates that the fitted hazard ratio in (3.4) increases from time zero to a maximum of approximately six times its initial value by four months of follow-up, and then descends back to its initial value. This result may reflect the severity of a patient's condition once he achieves the most severe level of CHF status. Since the estimate of γ_1 is negative, the relative decrease in either cause-specific hazard curve (see Section 2.1) is slower when CHF(t)=4.

The use of the nonparametric cause-specific cumulative hazards in (2.12) for determining goodness-of-fit was observed to be consistent with the likelihood ratio test results. For instance, the fitted

cause-specific cumulative hazard curves, based on the multiplicative version of model (3.3) where $\gamma_1 = 0$, appear in Figure 3.4 for death due to congestive heart failure and in Figure 3.5 for death due to other heart disease causes. The multiplicative model's fit to the piecewise exponential cumulative hazard estimates at the most severe CHF status, level 4, was distinctly poorer than that obtained by model (3.3) in Figures 3.1 and 3.2.

TABLE 3.1

MAXIMUM LIKELIHOOD ESTIMATES OF THE PARAMETERS IN MODEL (3.3)

Parameter	Estimate	Standard Error
α_{1o}	-12.246	.287
α_{2o}	- 9.639	.103
α_{11}	.9104	.1058
α_{21}	.5006	.0695
α_{22}	-1.0423	.3719
β_o	14.833	3.060
γ_o	.03795	.00850
γ_1	-.02997	.00856

FIGURE 3.3

THE FITTED TIME DEPENDENT FACTOR
FOR THE HAZARD RATIO IN (3.4)

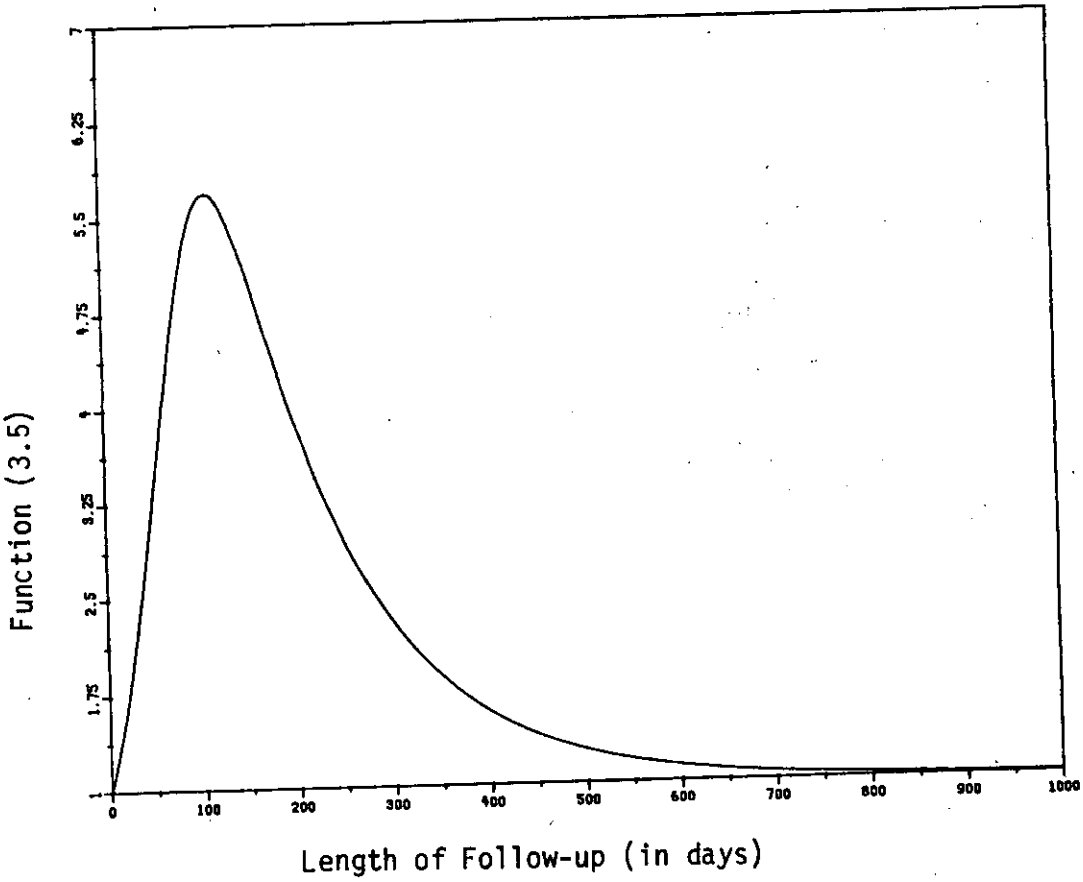


FIGURE 3.4

FITTED CUMULATIVE HAZARDS FOR THE MULTIPLICATIVE VERSION OF MODEL (3.3): DEATH DUE TO CONGESTIVE HEART FAILURE

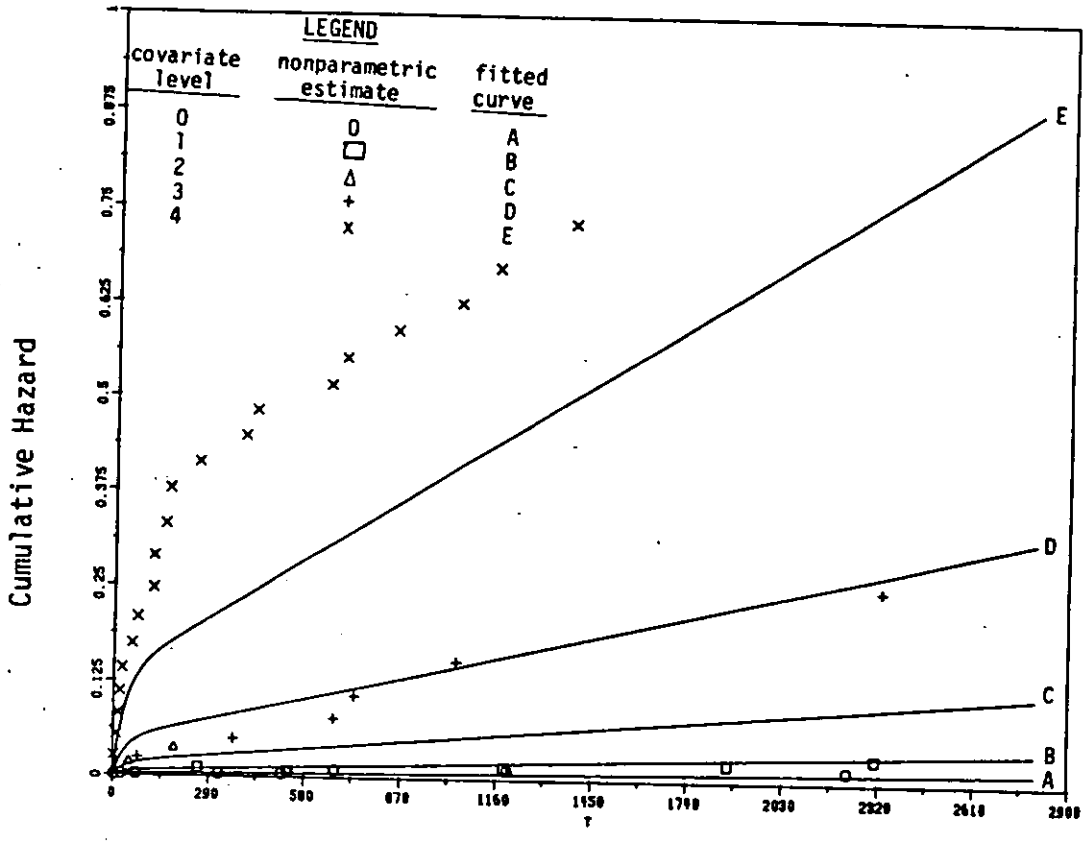
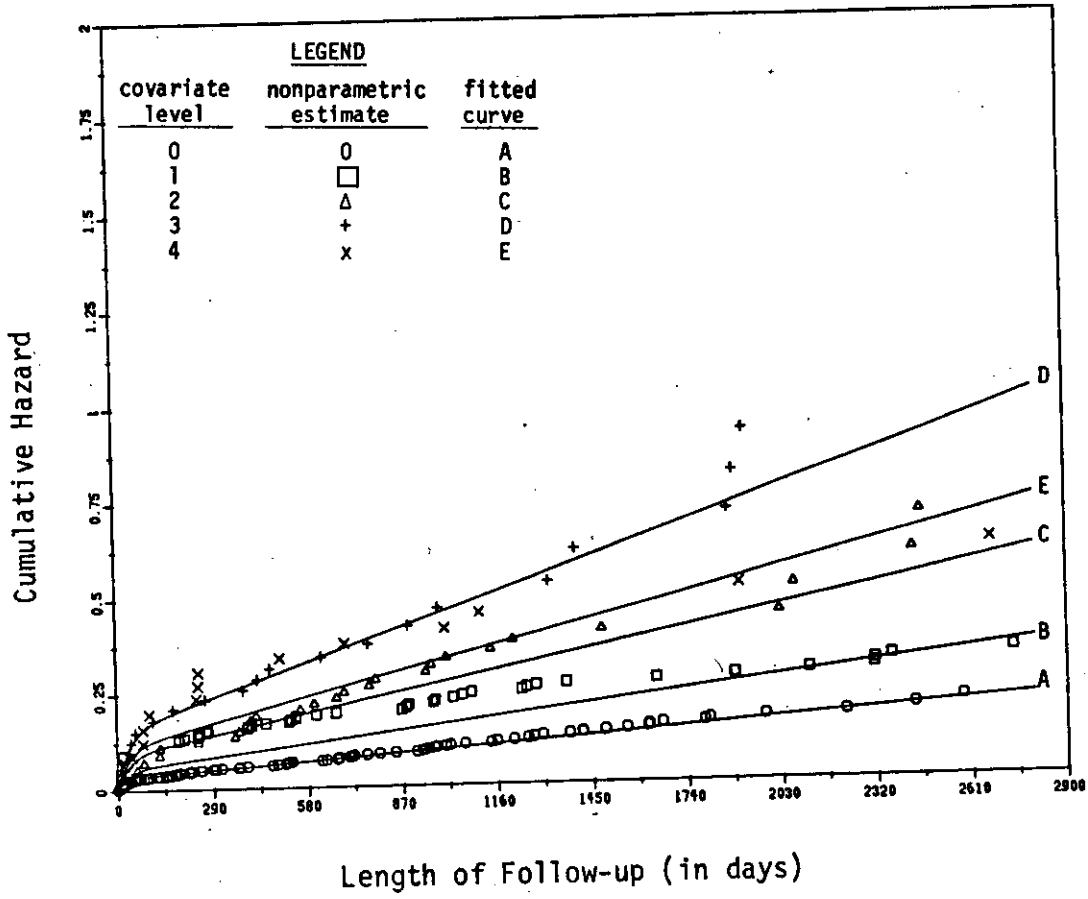


FIGURE 3.5

FITTED CUMULATIVE HAZARDS FOR THE MULTIPLICATIVE VERSION OF MODEL (3.3): DEATH DUE TO OTHER HEART DISEASE CAUSES



3.3 Extension to Seven Covariates

Preliminary investigation had produced the following results:

i) After adjusting for the effect of CHF(t) on the hazard of death due to congestive heart failure, separate effects of the covariates CHF2(t) and CHF3(t) were found on that hazard, but the covariates CHFP3(t) and CHFL3(t) had no influence.

ii) After adjusting for the effect of CHF(t) on the hazard of death due to other heart disease causes, none of the four "prior" CHF history covariates had an effect on that hazard.

iii) The baseline covariates LMSIG and LVEF act multiplicatively on the two cause-specific hazards.

iv) A LMSIG·CHF(t) antagonistic effect on the two cause-specific hazards was found. That is to say, the multiplicative effects of CHF(t) and LMSIG overestimate an individual's hazard if he has severe levels of both covariates.

Given these results, the following 23 parameter model was fitted to the data:

$$\lambda_k(t;\text{LMSIG},\text{LVEF},\text{CHF}(t),\text{CHF2}(t),\text{CHF3}(t)) =$$

$$\exp\left\{\begin{array}{l}\alpha_{ko}+\alpha_{k1}\text{LMSIG}+\alpha_{k2}\text{LVEF}+\alpha_{k3}\text{CHF}(t)+\alpha_{k4}\text{LMSIG}\cdot\text{CHF}(t)\\ +\ \alpha_{25}h(\text{CHF}(t))(1-f(k))+[\alpha_{16}+\alpha_{17}\text{CHF}(t)+\alpha_{18}\text{CHF2}(t)]g_1(t)f(k)\\ +\ [\alpha_{19}+\alpha_{1,10}\text{CHF}(t)+\alpha_{1,11}\text{CHF3}(t)]g_2(t)f(k)\end{array}\right\}$$

$$\cdot \left\{1 + \beta_{ko} e^{-[\gamma_{ko}+\gamma_{k1} h(CHF(t))]t}\right\}, \qquad (3.6)$$

where $f(k) = \begin{cases} 1 & \text{if } k=1 \\ 0 & \text{if } k=2, \end{cases}$ $g_1(t) = \begin{cases} 1 & \text{if } t \geq 6 \text{ months} \\ 0 & \text{otherwise,} \end{cases}$

$g_2(t) = \begin{cases} 1 & \text{if } t \geq 2 \text{ years} \\ 0 & \text{otherwise, and} \end{cases}$ $h(CHF(t)) = \begin{cases} 1 & \text{if } CHF(t)=4 \\ 0 & \text{otherwise.} \end{cases}$

The cause-specific underlying parameters in this model are α_{ko}, β_{ko}, and γ_{ko}, k=1,2. The parameters α_{k1}, α_{k2}, α_{k3}, and α_{k4} denote the cause-specific multiplicative effects of the covariates LMSIG, LVEF, CHF(t), and LMSIG·CHF(t). The parameter α_{25} denotes the lack of monotonicity in the multiplicative effect of CHF(t) on the hazard of death from other heart disease causes that was discovered by model (3.3). The parameters α_{18} and $\alpha_{1,11}$ denote the multiplicative effects of the "past" CHF history covariates CHF2(t) and CHF3(t) on the hazard of death due to congestive heart failure. The parameters α_{16} and α_{19} allow an adjustment in the underlying scale parameter for the hazard of death due to congestive heart failure, α_{1o}, when CHF2(t) and CHF3(t) are included into the model at six months and two years of follow-up, respectively. The parameters α_{17} and $\alpha_{1,10}$ allow an adjustment in the multiplicative effect of CHF(t) on the hazard of death due to congestive heart failure when CHF2(t) and CHF3(t) are included into the model at six months and two years of follow-up, respectively. Finally, the shape parameters γ_{k1}, k=1,2, denote nonmultiplicative effects of h(CHF(t)) on the two cause-specific hazards (similar in form to model (3.3)).

Table 3.2 summarizes the reductions in model (3.6) that were obtained by likelihood ratio tests. Each test was performed conditional on previously accepted null hypotheses. The underlying parameters β_{ko} and γ_{ko} were found to be constant across cause of death ($p = .94$). The adjusted underlying scale parameter and multiplicative CHF(t) effects at two years of follow-up on the hazard of death due to congestive heart failure, i.e., the parameters α_{19} and $\alpha_{1,10}$, were not significantly different from zero ($p = .34$), and the multiplicative effect of CHF3(t) on the hazard of death due to congestive heart failure was nonsignificant ($p = .26$). The parameters α_{16} and α_{17} were not significantly different from zero ($p = .11$). This result infers that the hazard of death due to congestive heart failure needed no adjustment of its underlying scale parameter and multiplicative CHF(t) effect at six months of follow-up, when the covariate CHF2(t) was included into the model. The multiplicative effect of CHF2(t) on the hazard of death due to congestive heart failure was significant ($p = .024$), and this effect was retained in the model. Therefore, since CHF3(t) has no effect after CHF(t) and CHF2(t) have been adjusted for, and with CHF(t) and CHF2(t) of importance, the hazard of death due to congestive heart failure at time t depends on an individual's CHF history only within two years of time t.

Constant multiplicative effects of the covariates LVEF and LMSIG·CHF(t) across cause of death were found ($p = .29$). The lack of monotonicity in the multiplicative CHF(t) effect on the hazard of death due to other heart disease causes was no longer significant ($p = .21$), and the h(CHF(t)) effect on the shape parameter was no

TABLE 3.2

REDUCTION OF MODEL (3.6) VIA LIKELIHOOD RATIO TESTS

Reduced Model	Number of Model Parameters	H_o	Maximized Log Likelihood	Likelihood Ratio Test of H_o, Conditional on Previously Accepted Null Hypotheses
(3.6)	23	-	-2626.65	-
(1)	21	$\beta_{ko}=\beta_o$, $\gamma_{ko}=\gamma_o$, for k=1,2	-2626.71	$\chi^2_2=0.12$, p=.94
(2)	19	$\alpha_{19}=\alpha_{1,10}=0$	-2627.78	$\chi^2_2=2.14$, p=.34
(3)	18	$\alpha_{1,11}=0$	-2628.41	$\chi^2_1=1.26$, p=.26
(4)	16	$\alpha_{16}=\alpha_{17}=0$	-2630.64	$\chi^2_2=4.46$, p=.11
(5)	15	$\alpha_{18}=0$	-2633.20	$\chi^2_1=5.12$, p=.024
(6)	14	$\alpha_{k2}=\alpha_{.2}$, $\alpha_{k4}=\alpha_{.4}$, for k=1,2	-2631.97	$\chi^2_2=2.46$, p=.29
(7)	13	$\alpha_{25}=0$	-2632.76	$\chi^2_1=1.58$, p=.21
(8)	12	$\gamma_{k1}=\gamma_{.1}$, for k=1,2	-2634.68	$\chi^2_1=3.84$, p=.05
(9)	12	$\gamma_{21}=0$	-2632.88	$\chi^2_1=0.24$, p=.62
(10)	11	$\gamma_{11}=0$	-2637.92	$\chi^2_1=10.08$, p=.0015

longer constant across cause of death ($p = .05$) - results obtained by the one covariate model (3.3). Here, the h(CHF(t)) effect on the shape parameter was highly significant for the hazard of death due to congestive heart failure ($p = .0015$) but nonsignificant for the hazard of death due to other heart disease causes ($p = .62$).

The reductions in model (3.6) are summarized by the following 12 parameter model:

$$\lambda_k(t;\text{LMSIG},\text{LVEF},\ \text{CHF}(t),\text{CHF2}(t)) =$$

$$\exp\left\{\alpha_{ko} + \alpha_{k1}\text{LMSIG} + \alpha_{.2}\text{LVEF} + \alpha_{k3}\text{CHF}(t) + \alpha_{.4}\text{LMSIG}\cdot\text{CHF}(t) + \alpha_{15}\text{CHF2}(t)g_1(t)f(k)\right\}$$

$$\cdot\left\{1 + \beta_o e^{-[\gamma_o+\gamma_{11}h(\text{CHF}(t))f(k)]t}\right\}. \qquad (3.7)$$

The maximum likelihood estimates of the parameters in this model appear in Table 3.3. The log ratio of the two underlying hazards in model (3.7), $(\alpha_{2o}-\alpha_{1o})$, was estimated to be $3.041 \pm .350$, demonstrating a much larger underlying hazard for death due to other heart disease causes. The log ratio of the cause-specific multiplicative LMSIG and CHF(t) effects, $(\alpha_{11}-\alpha_{21})$ and $(\alpha_{13}-\alpha_{23})$, were estimated to be $1.064 \pm .411$ and $.4217 \pm .1398$, respectively, demonstrating stronger LMSIG and CHF(t) effects on the hazard of death due to congestive heart failure. With an antagonistic LMSIG·CHF(t) effect that is constant across cause of death, $-.40 \pm .12$, the multiplicative effects of LMSIG and CHF(t) on either cause-specific hazard are diminished if the individual has severe levels of both covariates. The cause-specific

TABLE 3.3

MAXIMUM LIKELIHOOD ESTIMATES OF THE PARAMETERS IN MODEL (3.7)

Parameter	Estimate	Standard Error
α_{1o}	-12.747	.347
α_{2o}	- 9.706	.102
α_{11}	2.800	.431
α_{21}	1.736	.210
$\alpha_{.2}$	-.5773	.0521
α_{13}	.7117	.1364
α_{23}	.2900	.0549
$\alpha_{.4}$	-.3957	.1169
α_{15}	.2692	.1096
β_o	11.192	2.562
γ_o	.03751	.00915
γ_{11}	-.03190	.00944

hazards increase with diminishing LVEF since its multiplicative effect, constant across cause of death, was estimated to be $-.58 \pm .05$. The log ratio between the multiplicative effects of CHF(t) and CHF2(t) on the hazard of death due to congestive heart failure, $(\alpha_{13} - \alpha_{15})$, was estimated to be $.4425 \pm .1888$, demonstrating a stronger effect of the mean CHF status within one year of time t (CHF(t)) than that of the mean CHF status within two years but beyond six months of t (CHF2(t)). This result was anticipated.

The covariate CHF(t) acts multiplicatively on the hazard of death due to other heart disease causes in model (3.7), and the shape parameter for that cause-specific hazard is γ_0. However, CHF(t) does not act multiplicatively on the hazard of death due to congestive heart failure since its shape term, $\gamma_0 + \gamma_{11} h(\text{CHF}(t))$, is a function of CHF(t). The negative estimate of γ_{11} in Table 3.3 infers that the relative decrease in the hazard of death due to congestive heart failure (see Section 2.1) is slower when CHF(t)=4. The hazard ratio

$$\frac{\lambda_1(t;\text{LMSIG},\text{LVEF},\text{CHF}(t)=4,\text{CHF2}(t))}{\lambda_1(t;\text{LMSIG},\text{LVEF},\text{CHF}(t)=\ell,\ell\neq 4,\text{CHF2}(t))} = e^{\alpha_{13}(4-\ell)} \frac{(1+\beta_0 e^{-(\gamma_0+\gamma_{11})t})}{(1+\beta_0 e^{-\gamma_0 t})} \tag{3.8}$$

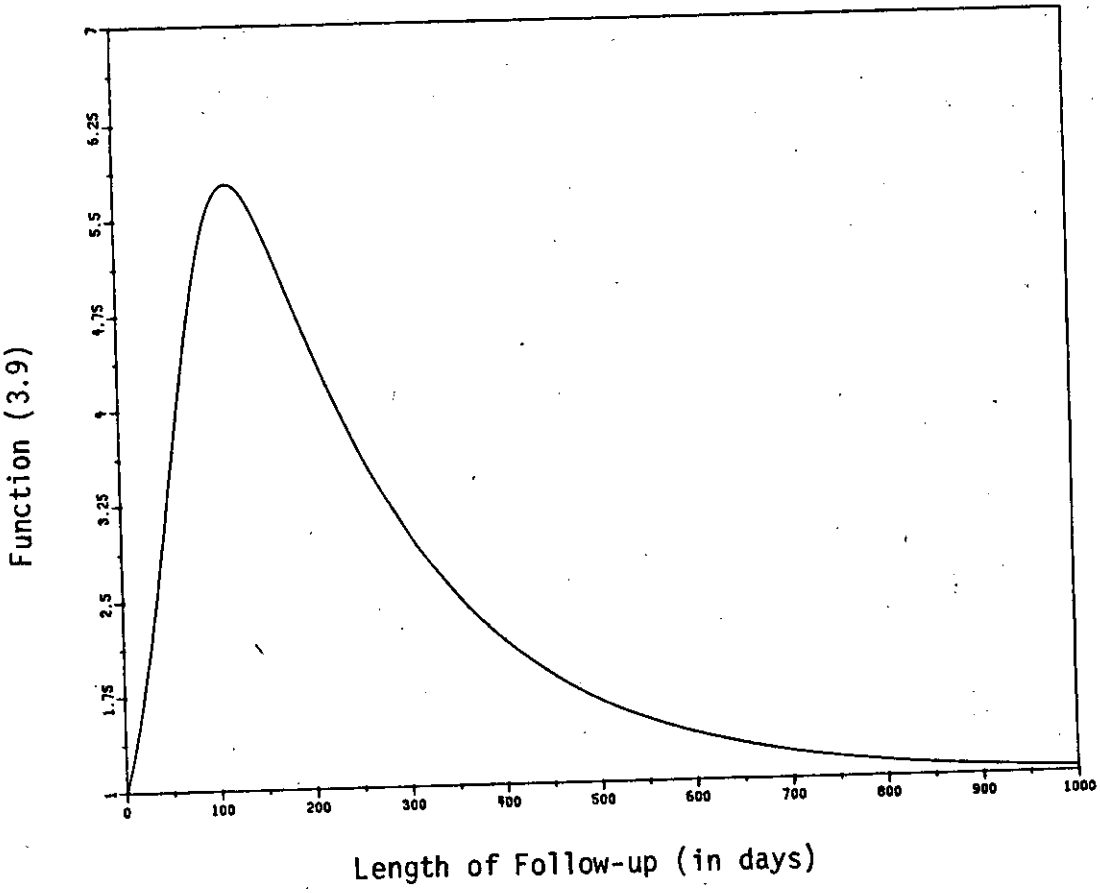
is a function of time t through the shape parameter γ_{11}. The fitted time dependent factor for the hazard ratio in (3.8),

$$\frac{(1+\hat{\beta}_0 e^{-(\hat{\gamma}_0+\hat{\gamma}_{11})t})}{(1+\hat{\beta}_0 e^{-\hat{\gamma}_0 t})}, \tag{3.9}$$

appears in Figure 3.6, and it is very similar to the curve in Figure

FIGURE 3.6

THE FITTED TIME DEPENDENT FACTOR FOR THE HAZARD RATIO IN (3.8)



3.3. It demonstrates that the hazard ratio in (3.8) increases from time zero to a maximum of approximately six times its initial value by four months of follow-up, and then descends back to its initial value.

Nonparametric maximum likelihood estimates of the cause-specific cumulative hazards were computed by the piecewise exponential formula in (2.12). Eight covariate strata were defined by dichotomous levels of CHF(t), LMSIG, and LVEF: CHF(t) was dichotomized into most severe (4) and less severe ($\leq$ 3) levels, and LVEF was dichotomized into below average ($<$ 0) and above average ($\geq$ 0) performances. For simplicity, the data were not stratified by CHF2(t). The eight stratum-specific piecewise exponential cumulative hazard estimates for death due to congestive heart failure and for death due to other heart disease causes appear as the plotted symbols in Figure 3.7 and 3.8, respectively. As expected, the slope of each covariate stratum's curve decreases sharply from time zero towards a lower bound that is achieved within two years of follow-up. These results validate the choice of the cause-specific underlying hazards in model (3.6) for describing the Duke University Heart Disease Data, i.e.,

$$\lambda_{ko}(t) = e^{\alpha_{ko}} (1 + \beta_{ko} e^{-\gamma_{ko} t}), \quad k=1,2.$$

Cumulative hazard curves were then fitted to these nonparametric estimates by the reduced version of model (3.7) in which the covariate CHF2(t) is excluded ($\alpha_{15}=0$). The maximum likelihood estimates and standard errors for this eleven parameter model appear in Table 3.6. In order to fit this model to the piecewise exponential

cumulative hazard estimates for the eight covariate strata, values for CHF(t) and LVEF had to be chosen. Stratum-specific CHF(t) and LVEF means at the death times among those who died, by cause, were used. Table 3.4 presents these stratum-specific CHF(t) and LVEF means, along with the number of deaths from each cause within strata. Figures 3.7 and 3.8 present the fitted cause-specific cumulative hazard curves for this reduced version of model (3.7), and reasonable fits to the nonparametric estimates were achieved. Curves were only fitted to strata that contained three or more distinct times to death.

Some of the fitted cumulative hazard curves in Figures 3.7 and 3.8 do not match up to their nonparametric estimates as closely as one might hope for. Therefore, an explanation of what is meant by "reasonable fits" is necessary. For example, compare stratum C's fitted curve in Figure 3.8 with its nonparametric estimates, the '+' symbols. The slope in the fitted curve C is too large prior to six months of follow-up. A steeper bend in that curve would have produced a better fit. However, if the six curves in Figure 3.8 are matched with the nonparametric symbols that they fit best, a correct matching is achieved. In addition, the importance of each stratum's fit must be weighted by the observed number of deaths in that stratum. Since 77.4% of the deaths from other heart disease causes had occurred in strata A and B, the model's fit to those strata is most important. In fact, the fitted curves A and B in Figure 3.8 fit the squares and triangles quite well. Now, recall that the ordinal values for CHF status, 0,1,...,4, were defined by the New York Heart Association's

TABLE 3.4

STRATUM-SPECIFIC COVARIATE MEANS AT THE DEATH TIMES AMONG THOSE WHO DIED, BY CAUSE OF DEATH

Strata			Deaths Due to Congestive Heart Failure				Deaths Due to Other Heart Disease Causes			
			Observed Number	Covariate Means			Observed Number	Covariate Means		
CHF(t)	LMSIG	LVEF		CHF(t)	LMSIG	LVEF		CHF(t)	LMSIG	LVEF
≤3	0	<0	8	2.000	0	-2.125	89	1.011	0	-1.697
≤3	0	≥0	8	0.750	0	0.250	93	0.634	0	0.387
≤3	1	<0	6	0.333	1	-1.500	19	0.737	1	-1.474
≤3	1	≥0	1	0.000	1	0.000	16	0.813	1	0.438
4	0	<0	12	4	0	-1.917	8	4	0	-1.875
4	0	≥0	2	4	0	0.000	8	4	0	0.250
4	1	<0	6	4	1	-2.000	2	4	1	-1.000
4	1	≥0	1	4	1	0.000	0	4	1	--

FIGURE 3.7

FITTED CUMULATIVE HAZARDS FOR THE REDUCED VERSION OF MODEL (3.7) IN WHICH CHF2(t) IS EXCLUDED: DEATH DUE TO CONGESTIVE HEART FAILURE

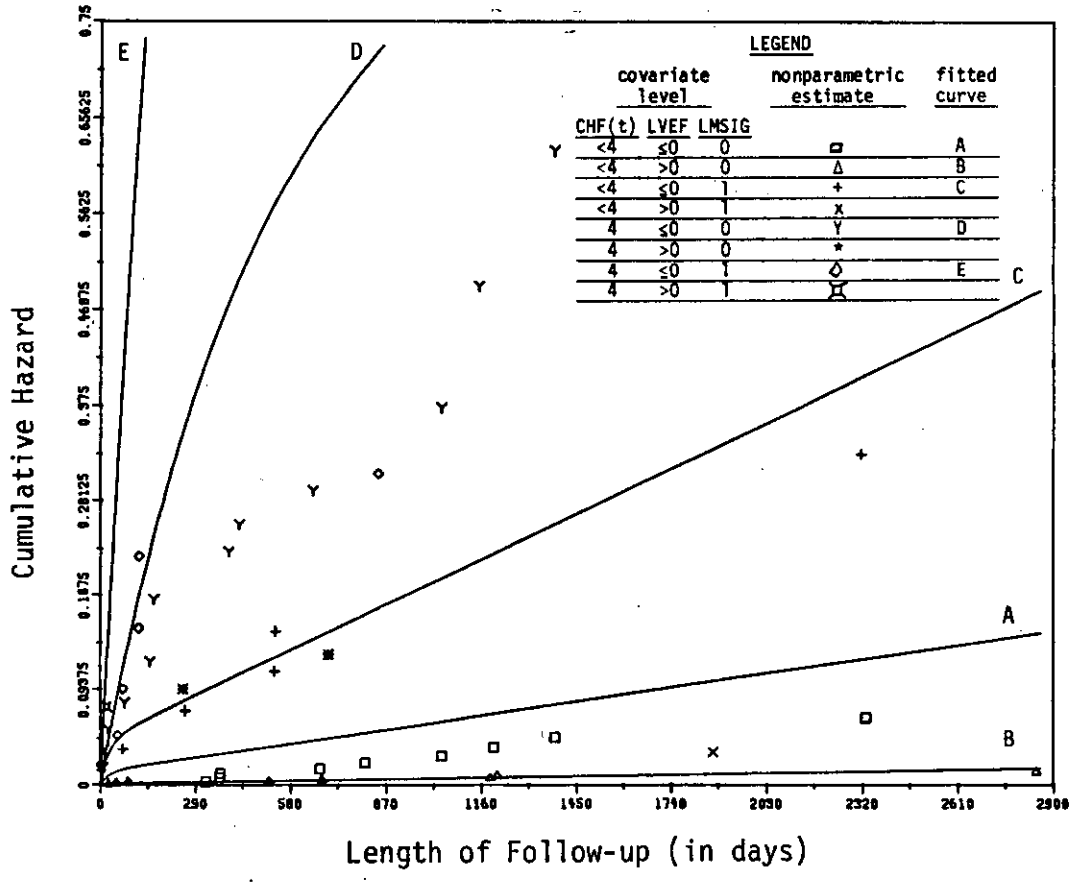
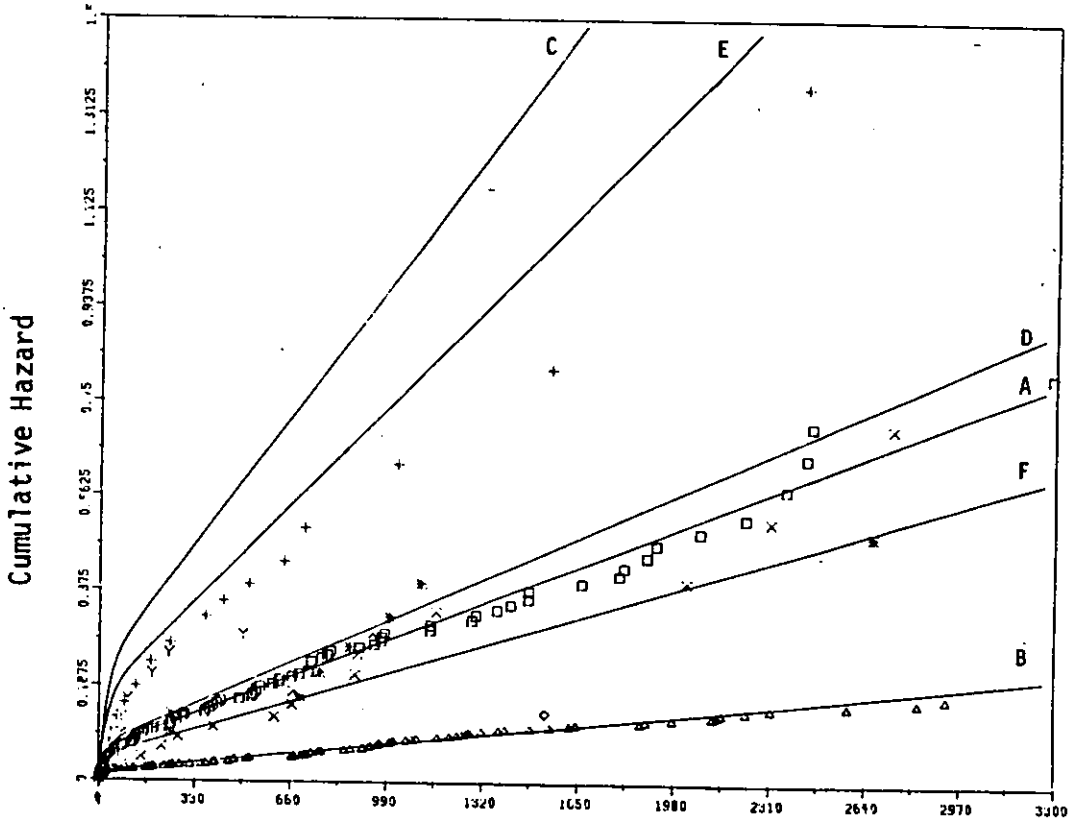


FIGURE 3.8

FITTED CUMULATIVE HAZARDS FOR THE REDUCED VERSION OF MODEL (3.7) IN WHICH CHF2(t) IS EXCLUDED: DEATH DUE TO OTHER HEART DISEASE CAUSES



LEGEND

covariate level			nonparametric estimate	fitted curve
CHF(t)	LVEF	LMSIG		
<4	<0	0	□	A
<4	≥0	0	△	B
<4	<0	1	+	C
<4	≥0	1	×	D
4	<0	0	▽	E
4	≥0	0	*	F
4	<0	1	◇	
4	≥0	1	⧖	

classifications. Therefore, any lack of the model's fit might suggest an alternative scale for representing CHF status. It might also suggest a transformation of some other covariate in the model. Finally, there is more than one choice of stratum-specific covariate values to use in fitting the model to the nonparametric estimates. Intuitively, the stratum-specific covariate means at the death times among those who died best represent the strata. Although these values were used in this dissertation, one could have also used the stratum-specific covariate means that are weighted by the person-years of follow-up attached to each distinct covariate value.

It should be noted that the estimated CHF(t) effect on the hazard of death due to congestive heart failure was stable with regard to the model's inclusion of the past CHF history covariate CHF2(t). The maximum likelihood estimate of the CHF(t) effect in model (3.7) was .712 ± .136 (Table 3.3). When CHF2(t) was excluded from this model, the estimated CHF(t) effect became .674 ± .144 (Table 3.6). The parameter estimate and its standard error remained nearly the same; therefore, there was no significant multicollinearity between CHF(t) and CHF2(t).

3.4 Comparison of the Maximum Likelihood Estimates Obtained by the Multiplicative Version of Model (3.7) With Those Obtained by Cox's Model

Cox's multiplicative model was fitted to the data,

$$\lambda_k(t;\text{LMSIG},\text{LVEF},\text{CHF}(t),\text{CHF2}(t)) = \lambda_o(t)\exp\left\{\begin{array}{l}[(\alpha_{2o}-\alpha_{1o})(1-f(k))]+\alpha_{k1}\text{LMSIG}+\alpha_{.2}\text{LVEF}+\alpha_{k3}\text{CHF}(t)+ \\ \alpha_{.4}\text{LMSIG}\cdot\text{CHF}(t)+\alpha_{15}\text{CHF2}(t)g_1(t)f(k)\end{array}\right\}, \quad (3.10)$$

where $g_1(t)$ and $f(k)$ are defined in (3.6), k=1,2. Maximum partial likelihood estimates of the parameters in model (3.10) were obtained from Breslow's approximation to Cox's partial likelihood, (1.31). These estimates, along with their standard errors, appear in Table 3.5.

The multiplicative version of model (3.7), i.e., $\gamma_{11}=0$ is assumed, was then fitted. In Table 3.5 the maximum likelihood estimate and standard error for the log ratio of the two cause-specific underlying hazards, $(\alpha_{20}-\alpha_{10})$, is reported, along with the maximum likelihood estimates and standard errors of the covariates' multiplicative effects.

The parameter estimates of the two models were remarkably alike. The standard errors obtained by the complete parametric model were negligibly smaller, on the average. Since there were only eleven repeated measurements on CHF status during the ten-year follow-up period, individuals' CHF(t) and CHF2(t) values changed seldom within the intervals that separate the exact times to death. With 238 exact times to death due to heart disease causes, the partial likelihood approach lost little information. In addition, since $\lambda_0(t)$ is left unspecified in Cox's model, the similarity in the results of the two approaches validates the choice of "decreasing exponential functions with lower bounds greater than zero" as the cause-specific underlying hazards in model (3.7).

3.5 Determination of the Information Gained by Using the Repeated Measurements on CHF Status

In order to determine the information gained by using the repeated measurements on CHF status, model (3.7) was extended to include

TABLE 3.5

COMPARISON OF THE MAXIMUM LIKELIHOOD ESTIMATES OBTAINED BY THE MULTIPLICATIVE VERSION OF MODEL (3.7) WITH THOSE OBTAINED BY COX'S MODEL

Parameter	Maximum Likelihood Estimate	Standard Error	Cox's Maximum Partial Likelihood Estimate	Standard Error
$\alpha_{2o}-\alpha_{1o}$	3.156	.328	3.148	.344
α_{11}	2.813	.423	2.795	.431
α_{21}	1.741	.210	1.724	.210
$\alpha_{.2}$	-.5814	.0517	-.5732	.0524
α_{13}	.9046	.1128	.8995	.1155
α_{23}	.2917	.0548	.2942	.0547
$\alpha_{.4}$	-.4023	.1167	-.4003	.1176
α_{15}	.1832	.1110	.1905	.1113

the baseline CHF status, CHF_0, as a covariate:

$$\lambda_k(t;LMSIG,LVEF,CHF(t),CHF2(t),CHF_0) =$$

$$\exp\left\{\begin{array}{l}\alpha_{ko}+\alpha_{k1}LMSIG + \alpha_{.2}LVEF + \alpha_{k3}CHF(t) + \alpha_{.4}LMSIG\cdot CHF(t) + \\ \alpha_{15}CHF2(t)\ g_1(t)f(k) + \alpha_{k6}CHF_0 + \alpha_{.7}LMSIG\cdot CHF_0\end{array}\right\}$$

$$\cdot\left\{1 + \beta_0 e^{-[\gamma_o+\gamma_{11}h(CHF(t))f(k) + \gamma_{12}h(CHF_0)f(k)]t}\right\}, \qquad (3.11)$$

where

$$h(CHF_0) = \begin{cases}1 & \text{if } CHF_0 = 4\\ 0 & \text{otherwise.}\end{cases}$$

Identical types of effects for CHF_0 and CHF(t) are assumed in this model, where the parameters α_{k6}, k=1,2, $\alpha_{.7}$, and γ_{12} correspond with α_{k3}, k=1,2, $\alpha_{.4}$, and γ_{11}, respectively. This sixteen parameter model enables one to test for the joint effects of the mean CHF status within one year of t (CHF(t)) and the mean CHF status within two years but beyond six months of t (CHF2(t)) after allowing for the effects of the baseline CHF status (CHF_0). Similarly, one can test for the effects of CHF_0 after allowing for the joint effects of CHF(t) and CHF2(t).

Model (3.11) was fitted to the data, and the likelihood ratio test for the joint effects of CHF(t) and CHF2(t) was highly significant, yielding $\chi^2_5 = 29.0$, p = .00002. The likelihood ratio test for the effects of CHF_0 was nonsignificant, yielding $\chi^2_4 = 5.4$, p = .25. Therefore, the covariates CHF(t) and CHF2(t), defined by the repeated measurements on CHF status, have a significant effect on the

cause-specific hazards even with the inclusion of the baseline covariate, CHF_0. On the other hand, the CHF_0 effects are no longer significant after one allows for the covariates CHF(t) and CHF2(t) in the model.

Finally, it was of interest to compare the estimated CHF(t) effects with those obtained when CHF(t) is replaced by the baseline covariate CHF_0. Two 11 parameter reduced versions of model (3.7) were fitted in which the covariate CHF2(t) was excluded:

i) CHF(t) is used, and

ii) CHF(t) is replaced by CHF_0.

Since it would not make practical sense to include $CHF_2(t)$ in a model with CHF_0, the covariate $CHF_2(t)$ was also excluded from the model when CHF(t) was used. The maximized log likelihood of this model using CHF(t), -2635.15, was substantially larger than the value obtained when CHF(t) was replaced by CHF_0, -2644.43. The maximum likelihood estimates of these two fitted models appear in Table 3.6. The multiplicative effects of CHF status, α_{k3}, k=1,2, were negligibly larger in the model using CHF(t). The absolute values of the other covariates' multiplicative effects were also slightly larger in the model using CHF(t). In fact, the only noticeable effect of replacing CHF(t) by CHF_0 were the changes in underlying scale parameter estimates: $e^{\alpha_{1o}}$ and $e^{\alpha_{2o}}$ were estimated to be 63% and 18% greater in the model using CHF_0. Since the cause-specific underlying hazards are obtained when the covariate values are zero, these results demonstrate that

$$\lambda_k(t;CHF_0=0) > \lambda_k(t;CHF(t)=0),\ k=1,2.$$

That is, an individual with a baseline CHF status of zero has a greater chance of attaining a more severe level by time t than if his most recent measurement on CHF status was zero. Therefore, his hazard of death from heart disease causes is estimated to be greater.

TABLE 3.6

MAXIMUM LIKELIHOOD ESTIMATES OF THE PARAMETERS IN TWO REDUCED VERSIONS OF MODEL (3.7), WHERE THE COVARIATE CHF2(t) IS EXCLUDED

Parameter	Estimates Obtained using CHF(t) (Standard Error)	Estimates Obtained using CHF_0 (Standard Error)
α_{1o}	-12.529 (.321)	-12.039 (.262)
α_{2o}	- 9.694 (.096)	- 9.526 (.089)
α_{11}	2.796 (.420)	2.480 (.398)
α_{21}	1.743 (.212)	1.604 (.197)
$\alpha_{.2}$	-.5804 (.0527)	-.5637 (.0546)
α_{13}	.6744 (.1442)	.6471 (.1186)
α_{23}	.2885 (.0537)	.2741 (.0592)
$\alpha_{.4}$	-.3963 (.1115)	-.3640 (.1116)
β_o	10.055 (2.270)	9.219 (2.119)
γ_o	.03583 (.00861)	.04027 (.00938)
γ_{11}	-.03336 (.00881)	-.03552 (.00964)

CHAPTER 4

AN ANALYSIS OF THE EFFECT OF CUMULATIVE EXPOSURE TO CHRYSOTILE ASBESTOS ON THE HAZARD OF DEATH FROM LUNG CANCER

4.1 Introduction

In a retrospective follow-up study of chrysotile asbestos textile workers, Dement (1980,1982) investigated a dose-response effect of cumulative exposure to chrysotile asbestos on the rate of death due to lung cancer. Asbestos exposure was determined from each individual's job history at the plant and from periodic measurements of asbestos levels in the air at the different job settings within the plant over the years. Specifically, an individual's length of employment at a given job was multiplied by the estimated asbestos level in the air at that site during the appropriate calendar time period. In this fashion, asbestos exposure in fiber-years/cc was accumulated on each worker.

The cohort consisted of 768 white males that were employed for at least six months in textile production operations. In addition, these men were employed for a minimum of one month between January 1, 1940 and December 31, 1965, ensuring a minimum time since initial employment of ten years for those alive at the study date end, December 31, 1975.

In Dement's study, cumulative exposure to chrysotile asbestos was categorized into four dose groups (see Table 4.1). Each worker contributed person-years of follow-up to one of these groups until his cumulative exposure became large enough to move him into a higher dose group. He then contributed person-years to the higher dose group. Therefore, the four dose groups do not represent strata of individuals; rather, they represent a stratification of each worker's person-years of follow-up into four cumulative exposure categories. Individuals' deaths were categorized into the four dose groups according to their cumulative exposures when they died.

Using the U.S. white male population as the standard for comparison, Dement performed a standardized mortality ratio (SMR) analysis on the deaths from lung cancer. For each dose group, the person-years of follow-up were distributed by five year age and calendar time periods, and the SMR was adjusted for these variables. In order to ensure a minimum "latency" of fifteen years, individuals' person-years of follow-up and deaths from lung cancer prior to fifteen years since initial employment were excluded from the analysis.

The results of Dement's (1980,1982) analysis on lung cancer mortality appear in Table 4.1. A significance test of each SMR was performed by assuming that the observed number of deaths follow a Poisson distribution. The SMR in each dose category was statistically significant, and their increasing values with cumulative exposure suggested a strong dose-response relationship. The SMR's may be compared directly if the expected lung cancer mortality rate for the j^{th} dose category and the k^{th} level of the extraneous variables (age and calendar time),

ρ_{jk}, factors into the product of two components: θ_j, an effect of the j^{th} dose category, and $\gamma\rho_{sk}$, a proportionality constant times the standard population rate for the k^{th} level of the extraneous variables (Freeman and Holford 1980). Under this multiplicative model assumption, the expected SMR in the j^{th} dose category becomes equal to $\gamma\theta_j$, and the dose-response effect between the j^{th} and j'^{th} dose categories, $\theta_j/\theta_{j'}$, is obtained by the ratio of their expected SMR's. Table 4.1 presents the SMR ratios between each of the higher dose categories and the lowest dose category.

Even if this multiplicative model assumption is met, the four SMR values in Table 4.1 do not allow for an adequate assessment of dose-response. Since there were only 26 deaths due to lung cancer, any further stratification by cumulative exposure may have sharply reduced the precision of each SMR estimate.

In this chapter, a survival analysis of lung cancer mortality is performed with the purpose of formalizing the suggested dose-response relationship in Table 4.1. Here, the dose-response effect is determined within the study cohort without the use of a standard population. Age at death from lung cancer is used as the outcome variable, and the workers' annual exposures to chrysotile asbestos, by age, comprise their repeated measurements. Three covariates are used: the individual's cumulative exposure to chrysotile asbestos at age t (CUMEX(t)), the number of years at age t since the individual's initial employment (LATENCY(t)), and the calendar year at age t (YEAR(t)). These covariates are defined by (1.12) as step functions of time, remaining constant until the completion of another year of age. In contrast to the

TABLE 4.1

RESULTS OF DEMENT'S (1980,1982) SMR ANALYSIS
ON DEATHS FROM LUNG CANCER

Cumulative Exposure to Chrysotile Asbestos (fiber-years/cc)	LUNG CANCER		SMR×100	SMR Ratio (lowest dose group is the standard)
	Observed Number Of Deaths	Expected Number Of Deaths		
< 27.4	8	3.59	223*	1.00
27.4 - 109.6	7	1.96	357**	1.60
109.6 - 274	9	0.92	978**	4.39
≥ 274	2	0.13	1553*	6.96
TOTAL	26	6.51	399**	

*p < .05

**p < .01

stratified analysis performed by Dement, the person-years of follow-up prior to fifteen years since initial employment are not deleted. Rather, the extraneous effects of "time since initial employment" and "calendar time" are controlled by the inclusion of the covariates LATENCY(t) and YEAR(t) into the survival analysis. Using the methods outlined in Chapter 2, the effect of an individual's cumulative exposure to chrysotile asbestos (CUMEX(t)) on the hazard of death from lung cancer is determined.

An updated version of Dement's (1980,1982) original data is used in this analysis. Three individuals whose causes of death were unknown at the time of Dement's study were later discovered to have died from lung cancer. Therefore, the subsequent analysis is based on 29 lung cancer deaths.

4.2 Covariate Definitions and Likelihood Construction

Denote EX_m as the individual's chrysotile asbestos exposure during the year following his m^{th} birthday, and

$$CUMEX_m = \sum_{\{i:i\leq m\}} EX_i$$

as the individual's cumulative exposure to chrysotile asbestos through age m. For simplicity of computation, an individual's age at death during the year following his m^{th} birthday was rounded off to the integer m. On the continuous time scale, m was chosen to represent the end of the year following the individual's m^{th} birthday, and the interval (m-1,m] denoted the time period during that year. This choice was made in order to include the individual's total cumulative exposure

at the time of his death into the likelihood. The cumulative exposure covariate CUMEX(t), measured in fiber-years/cc, was defined as

$$CUMEX(t) = CUMEX_m,$$

for $t \in (m-1,m]$. Therefore, an individual's cumulative exposure during the year following his m^{th} birthday was approximated by a constant value, that being his cumulative exposure at the end of that year. The other two covariates were also defined to remain constant during the year following the individual's m^{th} birthday:

$$YEAR(t) = YEAR_m,$$

and

$$LATENCY(t) = m - FEMPAGE,$$

where $t \in (m-1,m]$, $YEAR_m$ denotes the calendar year at age m, and m - FEMPAGE denotes the number of complete years between the individual's age at first employment and age m.

In order to simplify the computational construction of the likelihood, the covariates were rounded off. By reducing the number of distinct covariate values in the data, the total number of minimal sufficient statistics in the likelihood was reduced (see Section 2.5), and computer storage space was saved. Using the ROUND function in SAS - ROUND(x,i) rounds the variable x to the nearest integer i - the covariates CUMEX(t) and YEAR(t) were approximated by

$$CUMEX(t) = \begin{cases} \text{ROUND } (CUMEX(t),1) & \text{for } CUMEX(t) \leq 10, \\ \text{ROUND } (CUMEX(t),2) & \text{for } 10 < CUMEX(t) \leq 30, \\ \text{ROUND } (CUMEX(t),5) & \text{for } 30 < CUMEX(t) \leq 60, \\ \text{ROUND } (CUMEX(t),10) & \text{for } 60 < CUMEX(t) \leq 100, \\ \text{ROUND } (CUMEX(t),25) & \text{for } CUMEX(t) > 100, \end{cases}$$

and

$$YEAR(t) = ROUND\ (YEAR(t),5).$$

The covariate YEAR(t) was then standardized by

$$YEAR(t) = (YEAR(t) - 1925)/5.$$

The rounded off CUMEX(t) and YEAR(t) were used only for constructing and maximizing the likelihood. Their original values were used in estimating the nonparametric cumulative hazards. Theoretically, such rounding off is not required. However, the computer time required for maximizing the likelihood was substantial, using MAXLIK (Kaplan and Elston 1972) at UNC (University of North Carolina Computation Center). Since priority 9 had to be used (CPU time is free), the availability of computer storage space was severely limited. For future reference, TUCC (Triangle Universities Computation Center) just recently began allowing priority 9 jobs; its computer is extremely more time efficient than the one at UNC, and there is no restriction on storage space.

Table 4.2 presents the ages at death and covariate values at those ages for the 29 individuals who died of lung cancer. The first death occurred at age 46 and the last death occurred at age 73. The cumulative exposures for the individuals who died from lung cancer at the initial ages 46,47,47,51, and 51 years were 84.3, 3.1, 24.6, 230.4, and 111.4 fiber-years/cc, respectively. Since the maximum likelihood estimate of an individual's age at first risk of death is closely related to the first order statistic in the exponential, Gompertz, and Weibull distributions (see Appendix I), it does not appear that an individual's age at first risk of death from lung cancer decreases

with cumulative exposure. In fact, the modelling of such a relationship into the hazard was not significant (see the preliminary results in Section 4.3). Therefore, denoting $\lambda(t;\underset{\sim}{\theta},\underset{\sim}{z}(t))$ as the hazard of death due to lung cancer, where $\underset{\sim}{z}(t) = (\text{YEAR}(t),\text{LATENCY}(t),\text{CUMEX}(t))'$, a positive location parameter δ_o, independent of CUMEX(t), was included into the hazard $(\delta_o \epsilon \underset{\sim}{\theta})$ such that

$$\lambda(t;\underset{\sim}{\theta},\underset{\sim}{z}(t)) = 0 \quad \text{for} \quad t < \delta_o.$$

The location parameter δ_o estimates the individual's age at first risk of death from lung cancer.

The likelihood, as a function of $\underset{\sim}{\theta}$, is written as

$$L_{II}(\underset{\sim}{\theta}) = h(m_{(1)}-\delta_o) \cdot \prod_{\{m:m\geq\delta_o\}} \prod_{\ell=1}^{p} \left\{ \lambda(m;\underset{\sim}{\theta},\underset{\sim}{z}_\ell)^{d_{m\ell}} \cdot \exp\left[-n_{m\ell} \int_{\max(m-1,\delta_o)}^{m} \lambda(t;\underset{\sim}{\theta},\underset{\sim}{z}_\ell)\,dt \right] \right\}, \tag{4.1}$$

where $m_{(1)}$ denotes the earliest observed age at death from lung cancer,

$$h(m_{(1)}-\delta_o) = \begin{cases} 1 & \text{if } \delta_o \leq m_{(1)} \\ 0 & \text{otherwise,} \end{cases}$$

p denotes the number of distinct covariate values in the data, $\max(m-1,\delta_o)$ denotes the maximum of m-1 and δ_o, $d_{m\ell}$ denotes the number of individuals that died from lung cancer at age m with covariate value $\underset{\sim}{z}_\ell$ at that age, and $n_{m\ell}$ denotes the number of individuals that survived to age m with covariate value $\underset{\sim}{z}_\ell$ at that age. Likelihood (4.1) is a simplified version of likelihood (2.17), and it is a reexpression

of the incomplete likelihood in (1.21).

TABLE 4.2

THE AGES AT DEATH AND COVARIATE VALUES AT THOSE AGES FOR THE 29 INDIVIDUALS WHO DIED OF LUNG CANCER

OBS	AGE	YEAR	CUMEX	LATENCY
1	46	1949	84.309	31
2	47	1966	3.118	31
3	47	1969	24.621	29
4	51	1955	230.430	23
5	51	1963	111.383	30
6	52	1957	53.643	19
7	52	1959	141.640	36
8	52	1966	357.273	34
9	53	1947	4.356	10
10	54	1956	18.500	18
11	54	1952	65.081	20
12	55	1971	126.855	31
13	56	1958	254.820	26
14	56	1959	262.949	33
15	56	1968	68.441	25
16	57	1956	78.304	27
17	59	1963	109.170	31
18	60	1968	120.785	36
19	60	1969	158.364	29
20	61	1973	23.968	40
21	61	1974	197.270	42
22	63	1958	117.331	26
23	65	1975	155.671	33
24	65	1974	12.667	35
25	67	1974	7.081	34
26	68	1974	4.488	32
27	69	1964	280.314	31
28	71	1964	4.386	31
29	73	1968	5.164	24

4.3 Data Analysis - Results Obtained by the Gompertz Model

Nonparametric cumulative hazard estimates for death due to lung cancer were computed by the piecewise exponential formula in (2.12). Person-years of follow-up were stratified by CUMEX(t) into three dose groups, each with approximately ten lung cancer deaths:

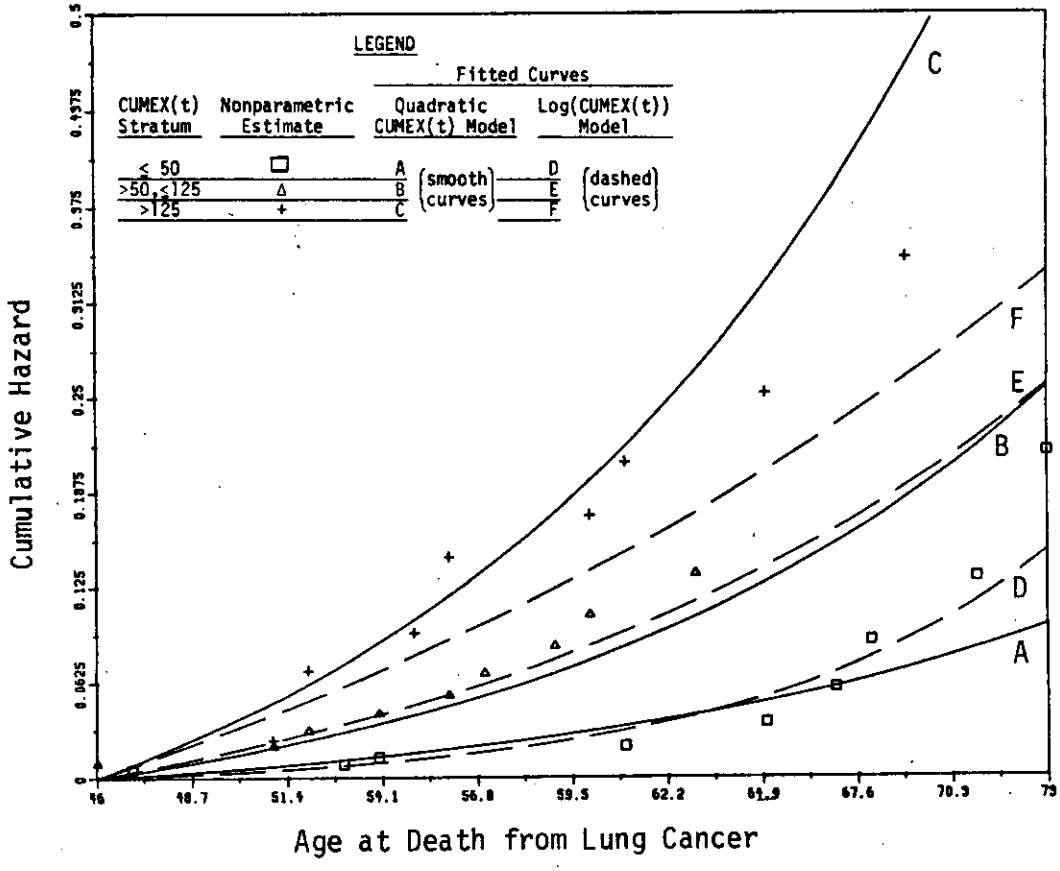
$\leq$ 50 fiber-years/cc,
> 50 but $\leq$ 125 fiber-years/cc, and
> 125 fiber-years/cc.

Since the earliest death from lung cancer occurred at age 46, person-years of follow-up prior to that age were not used. The three strata's cumulative hazard estimates appear as the plotted symbols in Figure 4.1, and the slope of each nonparametric curve increased monotonically with age. Cumulative hazard estimates were also obtained after stratifying the person-years of follow-up by LATENCY (t), and then again by YEAR(t). These nonparametric estimates (not presented here) similarly suggested an underlying hazard that increases with age. Due to the small number of lung cancer deaths, joint stratification of the person-years of follow-up by CUMEX(t), YEAR(t), and LATENCY(t) was not performed.

A deviation from the proportional hazards model was suggested by the three nonparametric cumulative hazard curves in Figure 4.1. Prior to age 60, the slopes of these curves increased strongly with CUMEX(t); however, by age 65 the slope of the "least CUMEX(t)" stratum had nearly caught up with that of the "most CUMEX(t)" stratum. With so few deaths beyond age 65, the validity of this observed lack of

FIGURE 4.1

FITTED CUMULATIVE HAZARDS FOR THE GOMPERTZ QUADRATIC CUMEX(t) AND LOG(CUMEX(t)) MODELS (Note a '+' directly above a 'Δ' at age 51 years)



proportionality is questioned. However, it suggests that a strong dose-response relationship exists at age 46, when individuals began dying from lung cancer, and that the strength of this relationship diminishes with age.

The results obtained by the nonparametric cumulative hazard estimates suggest a hazard that increases with age and a hazard ratio that decreases with age. Both the Gompertz and Weibull distributions meet these criteria. An underlying Gompertz hazard is used in this section,

$$\lambda(t) = \beta_0 e^{\gamma_0(t-\delta_0)}, \quad t \geq \delta_0,$$

and in Section 4.4, the results obtained by the Gompertz and Weibull models are compared.

Preliminary analysis via likelihood ratio tests suggested:

i) There were no significant interaction effects among the three covariates CUMEX(t), LATENCY(t), and YEAR(t) on the hazard of death from lung cancer ($p > .2$).

ii) The covariate log(LATENCY(t)) (a value of 0.0 for LATENCY(t) was replaced by 0.5) performed better than the covariate LATENCY(t), yielding a larger maximized log likelihood. Even though a quadratic expression for LATENCY(t), i.e., β_1LATENCY(t) + β_2(LATENCY(t))^2, yielded a larger maximized log likelihood than that obtained by log(LATENCY(t)), the difference was marginal. The quadratic effect of LATENCY(t) on the scale parameter β_0 was concave (β_1 is positive and β_2 is negative), and this was similar in shape to the log(LATENCY(t)) effect. Therefore, since the inclusion of a quadratic LATENCY(t)

effect would require two parameters, log(LATENCY(t)) was used in the analysis.

iii) The effects of log(LATENCY(t)) and YEAR(t) on the shape parameter γ_0 were nonsignificant ($p > .2$).

iv) The covariate log(CUMEX(t)) (a value of 0.0 for CUMEX(t) was replaced by 0.25) performed better than the covariate CUMEX(t), yielding a larger maximized log likelihood. On the other hand, a quadratic expression for CUMEX(t), i.e., $\beta_1 CUMEX(t) + \beta_2 (CUMEX(t))^2$, yielded a slightly larger maximized log likelihood than that obtained by log(CUMEX(t)). The log(CUMEX(t)) effect on the shape parameter γ_0 was significant, but the linear and quadratic effects of CUMEX(t) on the shape parameter γ were not significant ($p > .2$). Therefore, the results of two approaches are presented for comparative purposes, one using log(CUMEX(t)) and the other using a quadratic expression for CUMEX(t).

v) Cumulative exposure effects on the location parameter δ_0 were modelled into the hazard by two forms: $(\delta_0 + \delta_1 CUMEX(t))$ and $(\delta_0 + \delta_1 \log(CUMEX(t)))$. In both cases, the maximum likelihood estimate of δ_1 was zero. These results verify the assumption made in Section 4.2 that the individual's age at first risk of death from lung cancer is not a decreasing function of cumulative exposure.

The seven parametric "quadratic CUMEX(t)" model is defined by

$$\lambda(t; YEAR(t), \log(LATENCY(t)), CUMEX(t)) =$$

$$\exp\{\beta_0+\beta_1 YEAR(t) + \beta_2 \log(LATENCY(t)) + \beta_3 CUMEX(t) + \beta_4 (CUMEX(t))^2\} \cdot \exp\{\gamma_0 (t-\delta_0)\}, \quad t \geq \delta_0 , \tag{4.2}$$

where $\gamma_0 > 0$. The seven parameter "log(CUMEX(t))" model is defined by

$$\lambda(t;YEAR(t), \log(LATENCY(t)), \log(CUMEX(t))) =$$

$$\exp\{\beta_0+\beta_1 YEAR(t) + \beta_2 \log(LATENCY(t)) + \beta_3 \log(CUMEX(t))\} \cdot$$

$$\exp\{[\gamma_0 + \gamma_1 \log(CUMEX(t))] (t-\delta_0)\}, \quad t \geq \delta_0 , \tag{4.3}$$

where $[\gamma_0 + \gamma_1 \log(CUMEX(t))] > 0$ for all observed values of CUMEX(t). Since the expression $[\gamma_0 + \gamma_1 \log(CUMEX(t))]$ designates a shape parameter in the Gompertz distribution, it is restricted to be greater than zero.

The maximized log likelihood of model (4.2), -170.24, was slightly larger than that obtained by model (4.3), -171.18. The inclusion of a quadratic CUMEX(t) effect on the shape parameter γ_0 in (4.2), i.e., $\gamma_0 + \gamma_1 CUMEX(t) + \gamma_2 (CUMEX(t))^2$ replaces γ_0, was nonsignificant, yielding a likelihood ratio test statistic of $\chi^2_2 = 0.54$, $p = .763$. The likelihood ratio test of H_0: $\gamma_1 = 0$ in model (4.3) was significant, yielding $\chi^2_1 = 4.5$, $p = .034$. Therefore, the two models produced different interpretations of the data. Model (4.2) depicts a multiplicative quadratic effect of CUMEX(t), whereas model (4.3) includes a nonmultiplicative effect of log(CUMEX(t)). A confounding of these two effects may be due to the small number of observed lung cancer deaths.

The top half of Table 4.3 contains the maximum likelihood estimates of the parameters in models (4.2) and (4.3), along with their standard errors. Notice the strong effects of YEAR(t) and LATENCY(t) in both models. The likelihood ratio tests of the YEAR(t) effect were suggestive of

TABLE 4.3

MAXIMUM LIKELIHOOD ESTIMATES OF THE PARAMETERS IN MODELS (4.2), (4.3), (4.6), AND (4.7)

Quadratic CUMEX(t) Model (4.2): Gompertz Distribution

Parameter	Estimate	Standard Error
β_0	-9.95631	2.08674
β_1	-0.32047	0.16612
β_2	1.80331	0.88618
β_3	1.12034×10^{-2}	0.42263×10^{-2}
β_4	-0.19991×10^{-4}	0.09039×10^{-4}
γ_0	0.06726	0.02509
δ_0	46.0	[43.833, 46.0]*

*95% Confidence Interval

Log(CUMEX(t)) Model (4.3): Gompertz Distribution

Parameter	Estimate	Standard Error
β_0	-11.40608	2.09534
β_1	- 0.28754	0.16775
β_2	1.73503	0.89988
β_3	0.53823	0.19830
γ_0	0.16410	0.05032
γ_1	- 0.02505	0.00768
δ_0	46.0	[43.946, 46.0]*

Quadratic CUMEX(t) Model (4.6): Weibull Distribution

Parameter	Estimate	Standard Error
β_0	-10.90430	2.42505
β_1	- 0.31513	0.16509
β_2	1.76287	0.90022
β_3	1.08871×10^{-2}	0.42450×10^{-2}
β_4	-0.19160×10^{-4}	0.09110×10^{-4}
α_0	0.76671	0.50252
δ_0	44.63590	2.08308

Log(CUMEX(t)) Model (4.7): Weibull Distribution

Parameter	Estimate	Standard Error
β_0	-13.56041	3.40426
β_1	- 0.29099	0.16850
β_2	1.73278	0.88172
β_3	0.77971	0.55501
α_0	1.74053	1.08935
α_1	- 0.23445	0.20697
δ_0	44.17695	1.82897

statistical significance ($\chi^2_1 = 3.60$, $p = .058$ in (4.2); $\chi^2_1 = 3.60$, $p = .058$ in (4.3)), and the likelihood ratio tests of the log(LATENCY(t)) effect were significant ($\chi^2_1 = 4.64$, $p = .031$ in (4.2); $\chi^2_1 = 4.68$, $p = .031$ in (4.3)). The likelihood ratio test of the CUMEX(t) effect in model (4.2) yielded $\chi^2_2 = 7.88$, $p = .019$. Although not as strong, the log(CUMEX(t)) effect in model (4.3) was still significant, yielding $\chi^2_2 = 6.0$, $p = .050$.

The importance of controlling for the extraneous effects of calendar year and time since initial employment at age t is now demonstrated. When the covariates YEAR(t) and log(LATENCY(t)) are excluded from models (4.2) and (4.3), the significance of the CUMEX(t) and log(CUMEX(t)) effects increase substantially ($\chi^2_2 = 14.06$, $p = .001$ in (4.2); $\chi^2_2 = 12.70$, $p = .002$ in (4.3)). The maximum likelihood estimates of the parameters in models (4.2) and (4.3), when the covariates YEAR(t) and log(LATENCY(t)) are excluded, appear in Table 4.4. Notice how the inclusion of the covariates YEAR(t) and log(LATENCY(t)) into models (4.2) and (4.3) reduces the magnitudes of the estimated CUMEX(t) and log(CUMEX(t)) effects. The estimated linear component of the CUMEX(t) effect in model (4.2), $\hat{\beta}_3$, decreases from 0.0143 in Table 4.4 to 0.0112 in Table 4.3. The estimated multiplicative effect of log(CUMEX(t)) in model (4.3), $\hat{\beta}_3$, decreases from 0.73 in Table 4.4 to 0.54 in Table 4.3. In both cases, the amount of decrease is approximately one standard error. These results show that there is some confounding among the effects of cumulative exposure, calendar year, and the number of years since initial employment.

The first order statistic, age 46, turned out to be the maximum

TABLE 4.4

MAXIMUM LIKELIHOOD ESTIMATES OF THE PARAMETERS IN MODELS (4.2), (4.3), (4.6) AND (4.7), WHEN THE YEAR(t) AND LOG(LATENCY(t)) EFFECTS ARE EXCLUDED

Quadratic CUMEX(t) Model (4.2): Gompertz Distribution

Parameter	Estimate	Standard Error
β_0	-6.82218	0.42615
β_3	1.42685×10^{-2}	0.39899×10^{-2}
β_4	-0.22620×10^{-4}	0.09356×10^{-4}
γ_0	0.06870	0.02420
δ_0	46.0	-

Log(CUMEX(t)) Model (4.3): Gompertz Distribution

Parameter	Estimate	Standard Error
β_0	-8.77694	0.89576
β_3	0.73227	0.17743
γ_0	0.17753	0.05041
γ_1	-0.02710	0.00769
δ_0	46.0	-

Quadratic CUMEX(t) Model (4.6): Weibull Distribution

Parameter	Estimate	Standard Error
β_0	-8.48697	2.24444
β_3	1.38655×10^{-2}	0.40062×10^{-2}
β_4	-0.21653×10^{-4}	0.09353×10^{-4}
α_0	0.99225	0.74819
δ_0	43.82867	3.63581

Log(CUMEX(t)) Model (4.7): Weibull Distribution

Parameter	Estimate	Standard Error
β_0	-11.44036	3.29067
β_3	1.04646	0.60034
α_0	1.98189	1.17581
α_1	- 0.27156	0.22234
δ_0	43.98313	2.25832

likelihood estimate of the location parameter δ_0 in the Gompertz models 4.2 and 4.3. Such is not always the case in the Gompertz distribution. In fact, as the times to death become clustered - as opposed to being spread out - the maximum likelihood estimate of δ_0 would eventually become smaller than the first order statistic (see Appendix I). Since the likelihood is not right continuous at $\hat{\delta}_0 = 46.0$, MAXLIK was unable to compute a standard error of this estimate. Instead, a 95% confidence interval for δ_0 appears in Table 4.3. The 95% confidence interval for δ_0 is defined as the set of values $\{\delta_0^*: \delta_0^* \leq 46\}$ in which the likelihood ratio test of $H_0: \delta_0 = \delta_0^*$ is accepted at the .05 significance level. This was obtained by fitting models (4.2) and (4.3) for δ_0 fixed at each of seven values: 43.0, 43.5, 44.0, 44.5, 45.0, 45.5, and 46.0. Graphs of δ_0 versus the maximized log likelihood (δ_0) were then constructed (see Figure 4.2), and a linear interpolation of the 95% critical value for δ_0 was obtained.

Estimating Dose-Response Via the Hazard Ratio

Estimates of the hazard ratio, as a function of age, t, and CUMEX(t), were obtained for the quadratic CUMEX(t) model (4.2) and the log(CUMEX(t)) model (4.3). Using 1 fiber-year/cc as the standard exposure for comparison, the hazard ratio at age t is defined by models (4.2) and (4.3) as

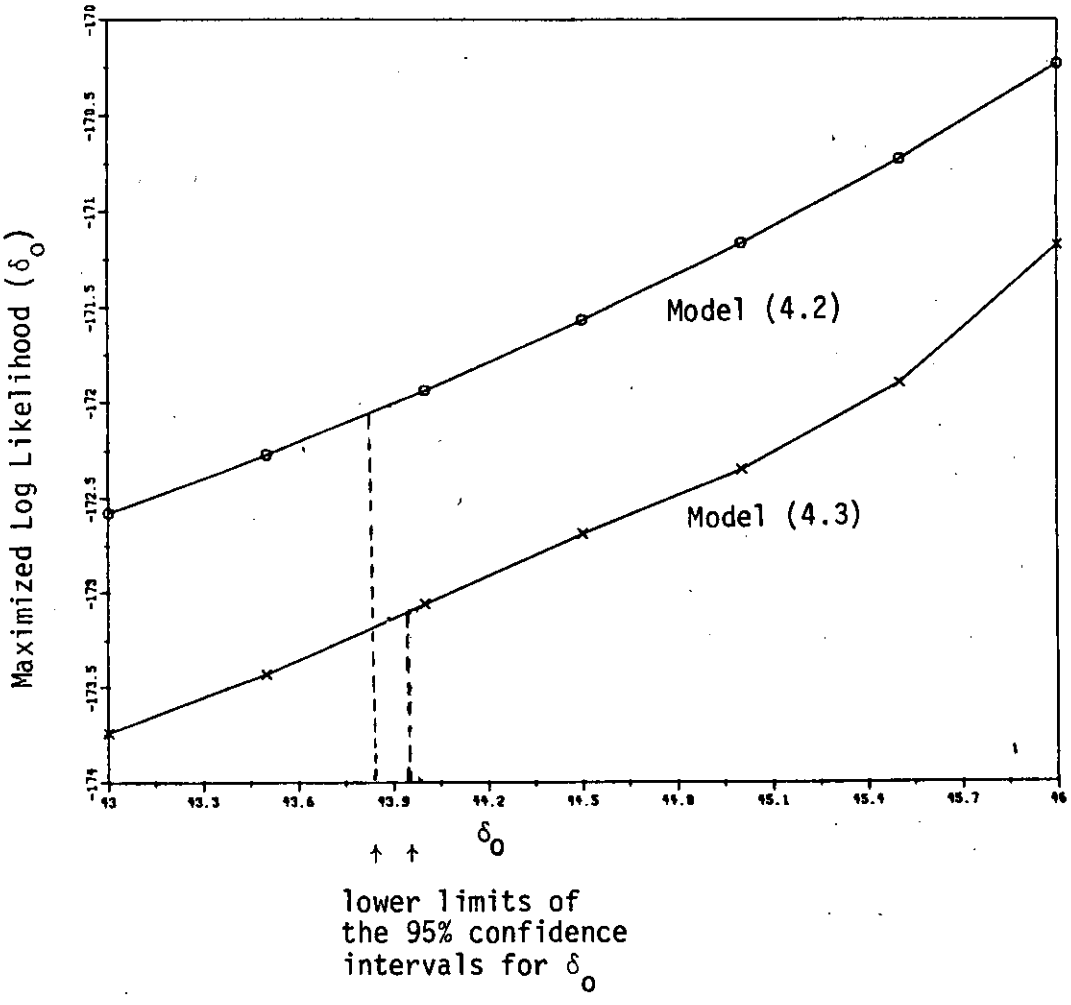
$$\text{h.r.}(t,\text{CUMEX}(t)) = \exp\{\beta_1(\text{CUMEX}(t)-1) + \beta_2((\text{CUMEX}(t))^2-1)\} \quad (4.4)$$

and

$$\text{h.r.}(t,\text{CUMEX}(t)) = \exp\{\beta_1 \log(\text{CUMEX}(t)) + \gamma_1 \log(\text{CUMEX}(t))(t-\delta_0)\}$$
$$= \text{CUMEX}(t)^{\{\beta_1+\gamma_1(t-\delta_0)\}}, \quad (4.5)$$

FIGURE 4.2

MAXIMIZED LOG LIKELIHOODS AS A FUNCTION OF δ_o, FOR MODELS (4.2) AND (4.3). THE LOCATION PARAMETER δ_o REPRESENTS THE INDIVIDUAL'S AGE AT FIRST RISK OF DEATH FROM LUNG CANCER



respectively. For a fixed value of CUMEX(t), the hazard ratio in (4.4) is constant with age, t, whereas in (4.5) it is a log linear function of age. Using the parameter estimates in Table 4.3, the fitted hazard ratios for model (4.2) and for model (4.3) at two ages, 50 and 60 years, were graphed as a function of CUMEX(t). The three curves appear in Figure 4.3. Since the maximum cumulative exposure among the 29 individuals that died from lung cancer was 357.3 fiber-years/cc, CUMEX(t) ranged from 1 to 360 in this figure. Although the maximum cumulative exposure among the 768 study participants was 700 fiber-years/cc, fitting the hazard ratio beyond 360 fiber-years/cc would have been extrapolation. Numerical values for these fitted hazard ratios at selected levels of CUMEX(t) are presented in Table 4.5.

The fitted hazard ratio for the quadratic CUMEX(t) model increased to a maximum value of 4.75 at exposure level 280.2 fiber-years/cc, and then decreased with CUMEX(t) beyond 280.2. It is expected that the hazard ratio should increase with CUMEX(t) within the range established by those who died from lung cancer, i.e., for CUMEX(t) $\leq$ 357.3 fiber-years/cc. Since there was only one death from lung cancer with CUMEX(t) > 280.2 (see Table 4.2), the quadratic CUMEX(t) model's predictive ability was limited in that range.

The maximum likelihood estimate of γ_1 in the log(CUMEX(t)) model was negative; therefore, its fitted hazard ratio decreased with age for each value of CUMEX(t). Figure 4.3 displays a sharp decrease in this model's fitted hazard ratio between the ages 50 and 60. For instance, its value at CUMEX(t) = 280.2 was 11.81 at age 50 and 2.88 at

FIGURE 4.3

FITTED HAZARD RATIOS FOR THE GOMPERTZ QUADRATIC CUMEX(t) AND LOG(CUMEX(t)) MODELS

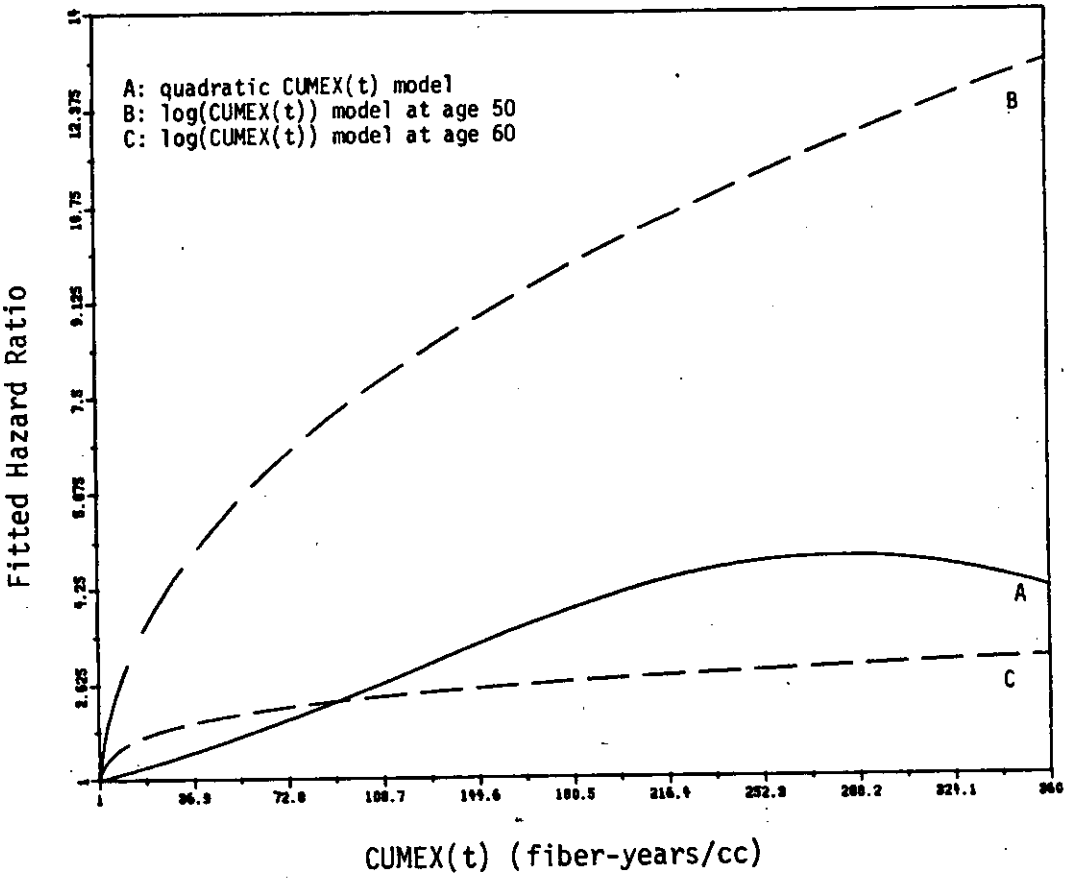


TABLE 4.5

FITTED HAZARD RATIOS FOR THE QUADRATIC CUMEX(t) AND LOG(CUMEX(t)) MODELS

CUMEX(t)	Quadratic CUMEX(t) Model	Log(CUMEX(t)) Model at Age 50 Years	Log(CUMEX(t)) Model at Age 60 Years	Quadratic CUMEX(t) Model Under Specified Conditions*
1	1.0	1.0	1.0	1.0
10	1.104	2.754	1.540	1.141
50	1.647	5.549	2.083	1.956
100	2.482	7.517	2.372	3.425
200	4.178	10.184	2.701	7.210
280.2	4.752	11.805	2.877	9.114
360	4.184	13.175	3.016	8.354

*Person-years of follow-up and deaths from lung cancer with LATENCY(t) < 15 years were excluded from the model fitting, along with the covariate log(LATENCY(t)).

age 60. For cumulative exposures greater than 100 fiber-years/cc, the log(CUMEX(t)) model's two fitted hazard ratio curves at ages 50 and 60 fell in between the quadratic CUMEX(t) model's fitted curve. For fixed CUMEX(t), the log(CUMEX(t)) model's fitted hazard ratio was maximized at age 46, the earliest lung cancer death, and it decreased to the value one by age 67.5 years. It is expected that the hazard ratio should remain greater than one within the range of ages at death from lung cancer. There were four individuals that died from lung cancer between the ages 68 and 73 (see Table 4.2), and the log(CUMEX(t)) model's fitted hazard ratio was less than one at those ages. However, with only five lung cancer deaths beyond age 65, the log(CUMEX(t)) model's predictive ability was limited in that range.

Consider once again the SMR results obtained by Dement (1980,1982) in Table 4.1. The SMR ratio for each dose category was presented, using the lowest dose group as the standard. Recall that dose-response may be estimated by the SMR ratios if the death rate for the j^{th} dose category and the k^{th} level of extraneous variables (age and calendar time) is described by a multiplicative model, where the effects of the extraneous variables in the study population are proportional to the rates in the standard population (Freeman and Holford 1980). Making this multiplicative model assumption, the SMR ratios in Table 4.1 were compared with the fitted hazard ratios obtained by the quadratic CUMEX(t) model. Since Dement's SMR analysis adjusted for LATENCY(t) by excluding all person-years of follow-up and deaths from lung cancer with LATENCY(t) < 15 years, model (4.2) was refitted under these conditions, and the covariate log(LATENCY(t)) was excluded. The maximum

likelihood estimates of β_1 and β_2 were 1.49703×10^{-2} (s.e.=0.47015×10^{-2}) and -0.25090×10^{-4} (s.e.=0.12433×10^{-4}), respectively. Their magnitudes were greater than those obtained by model (4.2) in Table 4.3, and the fitted hazard ratios obtained by these parameter estimates were also greater (see Table 4.5). In order to fit the hazard ratio to Dement's dose categories, a particular CUMEX(t) value was assigned to each stratum. The CUMEX(t) mean at the ages of death among those who died from lung cancer within a dose category was used, and the lowest dose category's value became the standard for comparison. These CUMEX(t) means and fitted hazard ratios for the four dose categories appear in Table 4.6, along with the SMR ratios obtained by Dement's analysis. Although the fitted hazard ratio was larger than the SMR ratio in each of the compared strata, their values were similar.

TABLE 4.6

COMPARISON OF DEMENT'S SMR RATIOS WITH FITTED HAZARD RATIOS OF THE QUADRATIC CUMEX(t) MODEL

Dement's Cumulative Exposure Strata	CUMEX(t)*	Observed Number of Deaths from Lung Cancer	SMR Ratio**	Fitted Hazard Ratio** for the Quadratic CUMEX(t) Model, Under Specified Conditions***
< 27.4	10.835	10	1.0	1.0
≥27.4,<109.6	79.491	6	1.60	2.39
≥109.6,<274	170.682	11	4.39	5.29
≥ 274	318.794	2	6.96	7.87

*These are the mean CUMEX(t) values at the ages of death among those who died from lung cancer (within strata).

**The lowest dose group was used as the standard for comparison.

***To conform with Dement's SMR analysis, person-years of follow-up and deaths from lung cancer with LATENCY(t) < 15 years were excluded from the model fitting, along with the covariate log (LATENCY(t)).

Plotting Goodness-of-Fit Via the Cumulative Hazard

Cumulative hazard curves were fitted by the quadratic CUMEX(t) and log(CUMEX(t)) models to the nonparametric cumulative hazard estimates in Figure 4.1. The three strata in that figure were defined only by CUMEX(t) ($\leq$ 50, > 50 but $\leq$ 125, and > 125 fiber-years/cc). Therefore,the covariates YEAR(t) and log(LATENCY(t)) were excluded from the model fitting in order to obtain comparable fitted cumulative hazard curves. The parameter estimates in Table 4.4 were used. Due to the small number of observed lung cancer deaths, joint stratification of the person-years of follow-up by CUMEX(t), YEAR(t), and LATENCY(t) was not performed. In order to fit cumulative hazard curves to the three CUMEX(t) strata, values for CUMEX(t) and log (CUMEX(t)) had to be chosen. The stratum-specific CUMEX(t) and log (CUMEX(t)) means at the ages of death among those who died from lung cancer were used. Table 4.7 presents these mean values.

The cumulative hazard goodness-of-fit plots for models (4.2) and (4.3), with the covariates YEAR(t) and log(LATENCY(t)) excluded, appear together in Figure 4.1. Notice how the quadratic CUMEX(t) model's fitted curves depict a strong CUMEX(t) effect through age 70, whereas the slopes of the log(CUMEX(t)) model's fitted curves are roughly equal by that age. The log(CUMEX(t)) model detects the catch-up in slope for the "least CUMEX(t)" stratum; however, its fitted slope for the "most CUMEX(t)" stratum is too small during the ages 53-62 years. The quadratic CUMEX(t) model's fit to the "most CUMEX(t)" stratum was very good prior to age 62 but its slope became too large beyond that age. This graphical display is consistent with

TABLE 4.7

STRATUM-SPECIFIC CUMEX(t) AND LOG(CUMEX(t)) MEANS AT THE AGES OF DEATH AMONG THOSE WHO DIED FROM LUNG CANCER

CUMEX(t) Stratum (fiber-years/cc)	Observed No. of Lung Cancer Deaths	CUMEX(t) Mean	Log(CUMEX(t)) Mean
≤ 50	10	10.83	2.102
50 - 125	9	89.83	4.460
> 125	10	216.56	5.326

the results obtained by the likelihood ratio tests. That is, given the small number of lung cancer deaths, it is unclear as to which model is more appropriate: a multiplicative model with a quadratic CUMEX(t) effect or a nonmultiplicative model using log(CUMEX(t)) in which the hazard ratio decreases with age.

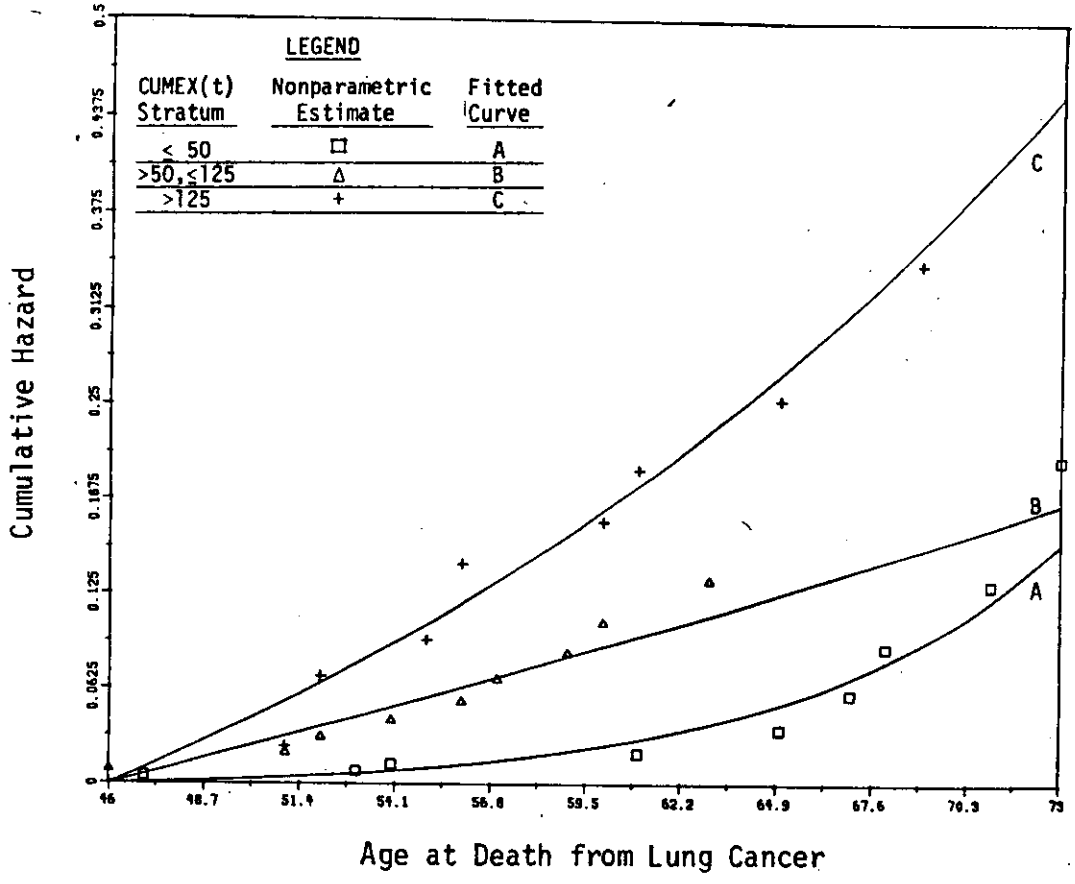
Some of the observed lack of fit for models (4.2) and (4.3) in Figure 4.1 is due to the attempted fitting of a survival model, as a function of continuous covariates, to nonparametric estimates that are functions of the same covariates, but stratified. When the covariates YEAR(t) and log(LATENCY(t)) are excluded, models (4.2) and (4.3) contain five parameters. A more general model would have allowed distinct scale and shape parameters for each CUMEX(t) stratum. Specifically,

$$\lambda(t;\text{CUMEX}(t) \in \text{stratum } i) = \beta_i \exp\{\gamma_i(t-\delta_0)\},$$

for $i=1,2,3$, and $t \geq \delta_0$. This seven parameter Gompertz model was fitted to the data, and its fitted cumulative hazard curves appear in Figure 4.4. The fits to the nonparametric estimates are much improved.

FIGURE 4.4

FITTED CUMULATIVE HAZARDS FOR A MORE GENERAL GOMPERTZ MODEL: DISTINCT SCALE AND SHAPE PARAMETERS ARE ALLOWED FOR EACH STRATUM



4.4 Comparisons with the Weibull Distribution and Cox's Model

Fitting the Weibull Distribution

The quadratic CUMEX(t) and log(CUMEX(t)) models were refitted, this time using an underlying Weibull hazard:

$$\lambda(t;\text{YEAR}(t),\ \log(\text{LATENCY}(t)),\ \text{CUMEX}(t)) =$$

$$\exp\{\beta_0+\beta_1\text{YEAR}(t)+\beta_2\log(\text{LATENCY}(t))+\beta_3(\text{CUMEX}(t))+\beta_4(\text{CUMEX}(t))^2\}\cdot(t-\delta_0)^{\alpha_0},\quad t\geq\delta_0\ , \tag{4.6}$$

and

$$\lambda(t;\text{YEAR}(t),\ \log(\text{LATENCY}(t)),\ \log(\text{CUMEX}(t))) =$$

$$\exp\{\beta_0+\beta_1\text{YEAR}(t)+\beta_2\log(\text{LATENCY}(t))+\beta_3\log(\text{CUMEX}(t))\}\cdot(t-\delta_0)^{[\alpha_0+\alpha_1\log(\text{CUMEX}(t))]},\quad t\geq\delta_0\ , \tag{4.7}$$

where $\alpha_0 > 0$ in (4.6) and $[\alpha_0+\alpha_1\log(\text{CUMEX}(t))] > 0$ for all observed values of CUMEX(t) in (4.7).

The maximized log likelihood of the quadratic CUMEX(t) model in (4.6) was only slightly less than that obtained by the Gompertz distribution in (4.2), -171.21 compared to -170.24. The maximum likelihood estimates of the parameters in model (4.6), along with their standard errors, appear in Table 4.3. Since the estimate for δ_0 was less than the first order statistic, the likelihood was a continuous function at that point, and its standard error was computed. The estimates of the covariates' effects (the parameters β_1,β_2,β_3, and β_4) and standard errors were very similar between models (4.2) and (4.6). So were the estimates of β_0 and δ_0. This consistency appears

graphically in Figure 4.5, where the quadratic CUMEX(t) model's fitted cumulative hazards were compared between the Gompertz and Weibull distributions. Again, the covariates YEAR(t) and log(LATENCY(t)) were excluded from the model fitting, and the parameter estimates in Table 4.4 were used.

The maximized log likelihood of the log(CUMEX(t)) model in (4.7) was noticeably smaller than that achieved by the Gompertz distribution in (4.3), -173.37 compared to -171.18. In fact, the likelihood ratio test of the shape parameter α_1 was not significant ($\chi_1^2 = 1.36$, $p = .24$). Nor was the likelihood ratio test of the overall log(CUMEX(t)) effect significant ($\chi_2^2 = 2.92$, $p = .23$). The maximum likelihood estimates of the parameters in model (4.7), along with their standard errors, appear in Table 4.3. Note the large standard errors for β_1 and α_1. Figure 4.6 compares the log(CUMEX(t)) model's fitted cumulative hazard curves between the Gompertz and Weibull distributions. Again, the covariates YEAR(t) and log(LATENCY(t)) were excluded from the model fitting, and the parameter estimates in Table 4.4 were used. Although the two models' fits to the nonparametric estimates were similar, the Gompertz model's fit to the "least CUMEX(t)" stratum was better than that achieved by the Weibull model. To conclude, the Weibull distribution's poorer performance in fitting the log(CUMEX(t)) model may be explained by the fact that a covariate's effect on the shape parameter is not the same for the Gompertz and Weibull distributions. The hazard ratio is a log linear function of time in the Gompertz distribution as opposed to a log linear function of log time in the Weibull distribution. The Gompertz distribution's representation of nonproportionality performed

FIGURE 4.5

FITTED CUMULATIVE HAZARDS FOR THE GOMPERTZ AND WEIBULL QUADRATIC CUMEX(t) MODELS

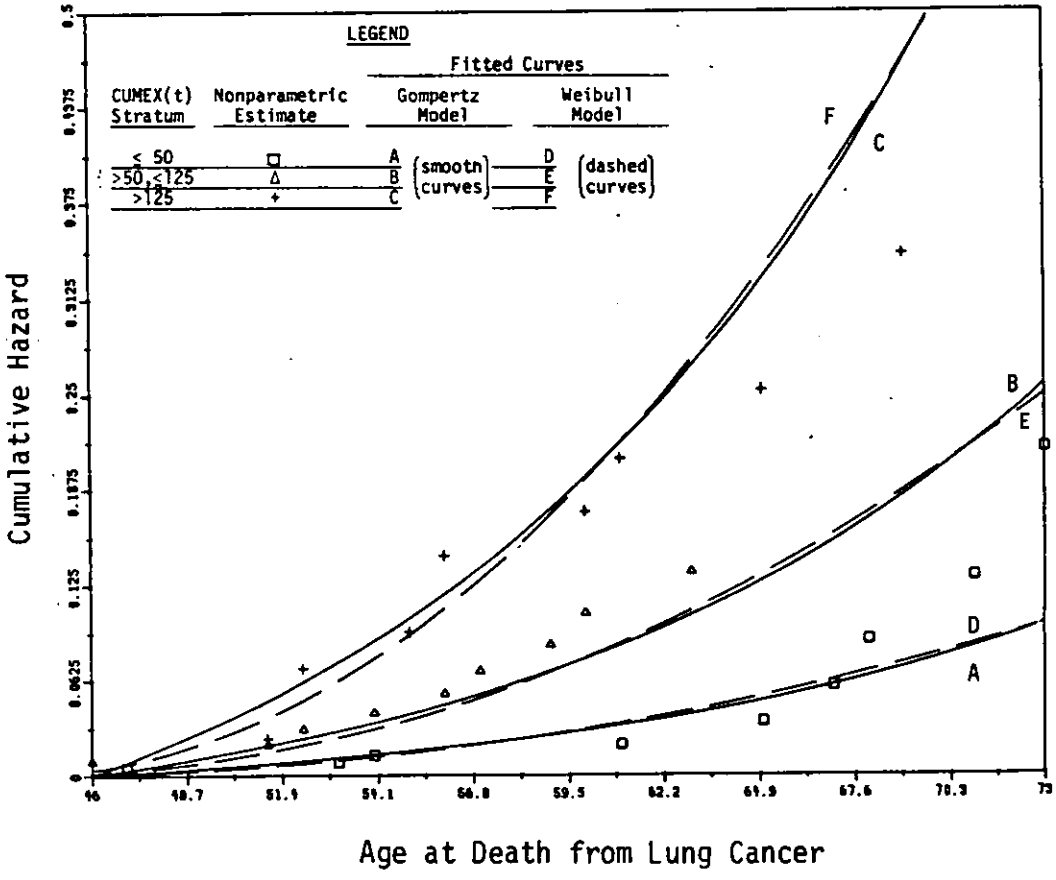
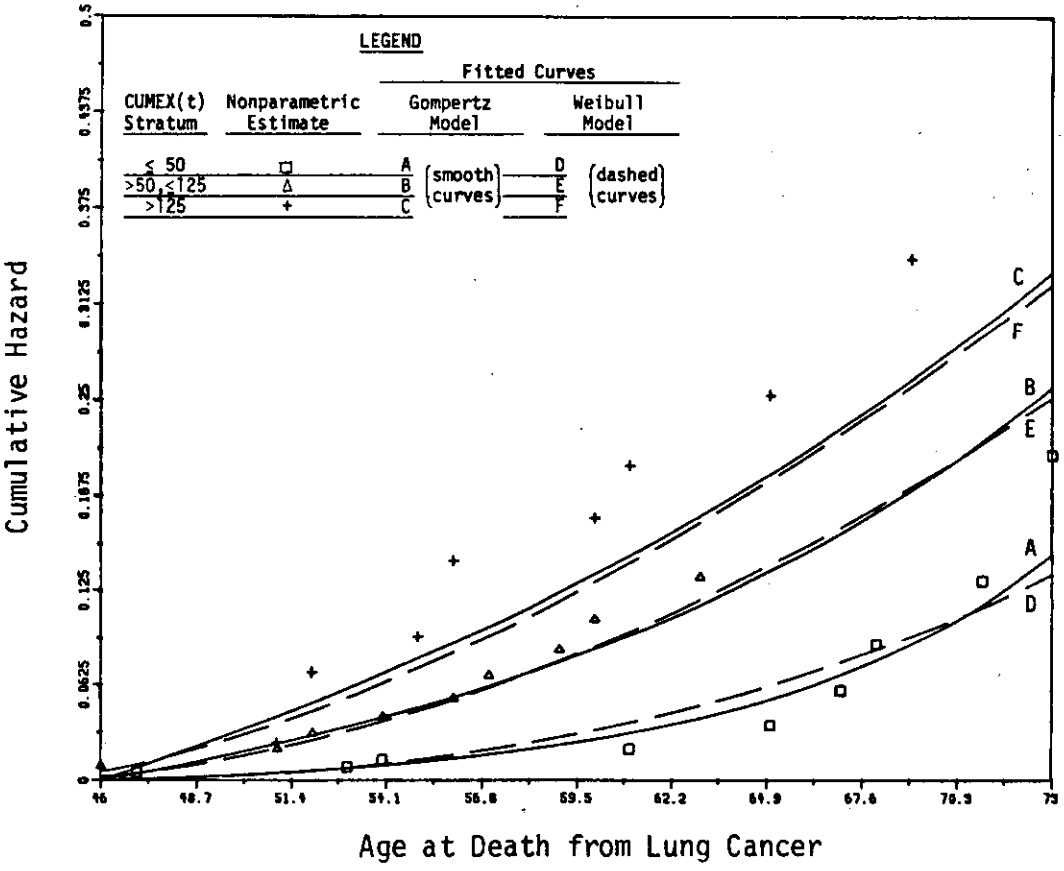


FIGURE 4.6

FITTED CUMULATIVE HAZARDS FOR THE GOMPERTZ AND WEIBULL LOG(CUMEX(t)) MODELS



better for these data.

Fitting Cox's Model

Two versions of Cox's model were fitted to the data:

$$\lambda(t;\text{YEAR}(t),\ \log(\text{LATENCY}(t)),\ \text{CUMEX}(t)) =$$

$$\lambda_0(t)\ \exp\{\beta_1\text{YEAR}(t) + \beta_2\log(\text{LATENCY}(t)) + \beta_3\text{CUMEX}(t) + \beta_4(\text{CUMEX}(t))^2\} \tag{4.8}$$

and

$$\lambda(t;\text{YEAR}(t),\ \log(\text{LATENCY}(t)),\ \log(\text{CUMEX}(t))) =$$

$$\lambda_0(t)\ \exp\{\beta_1\text{YEAR}(t) + \beta_2\log(\text{LATENCY}(t)) + \beta_3\log(\text{CUMEX}(t)) +$$

$$\gamma_1\log(\text{CUMEX}(t))(t-46)\}. \tag{4.9}$$

Models (4.8) and (4.9) correspond to the Gompertz distribution models (4.2) and (4.3), with the exceptions that $\lambda_0(t)$ is left unspecified and age t is centered by a fixed value $\hat{\delta}_0 = 46$, not the parameter δ_0. In addition, since there is no parameter γ_0 in (4.9), the parametric restriction in (4.3) that $[\gamma_0 + \gamma_1\log(\text{CUMEX}(t))] > 0$ for all observed values of CUMEX(t) is not applicable here. Using Breslow's approximation to Cox's partial likelihood, (1.35) with one cause of failure (c*=1), the maximum likelihood estimates of the parameters in models (4.8) and (4.9) were obtained by MAXLIK. These values, along with their standard errors, appear in Table 4.8. Included in this table are the parameter estimates of $\beta_1,\beta_2,\beta_3,\beta_4$, and γ_1 that were obtained by the Gompertz distribution models (4.2) and (4.3), as displayed in Table 4.3.

The quadratic CUMEX(t) model's parameter estimates and standard errors were very similar between the Cox and Gompertz approaches.

TABLE 4.8

COMPARING THE GOMPERTZ MODEL'S RESULTS
WITH THOSE OBTAINED BY COX'S MODEL

Quadratic CUMEX(t) Models

	COX'S MODEL (4.8)		GOMPERTZ MODEL (4.2)	
Parameter	M.L.E.	Standard Error	M.L.E.	Standard Error
β_1	-0.33141	0.16751	-0.32047	0.16612
β_2	1.92985	0.93319	1.80331	0.88618
β_3	1.09495×10^{-2}	0.42832×10^{-2}	1.12034×10^{-2}	0.42263×10^{-2}
β_4	-0.19297×10^{-4}	0.09375×10^{-4}	-0.19991×10^{-4}	0.09039×10^{-4}

Log CUMEX(t) Models

	COX'S MODEL (4.9)		GOMPERTZ MODEL (4.3)		MODIFIED VERION* OF MODEL (4.3)	
Parameter	M.L.E.	Standard Error	M.L.E.	Standard Error	M.L.E.	Standard Error
β_1	-0.28894	0.16986	-0.28754	0.16775	-0.28043	0.16892
β_2	1.83698	0.92208	1.73503	0.89988	1.72964	0.90487
β_3	0.63479	0.28841	0.53823	0.19830	0.60949	0.26482
γ_1	-0.03207	0.01704	-0.02505	0.00768	-0.03057	0.01439

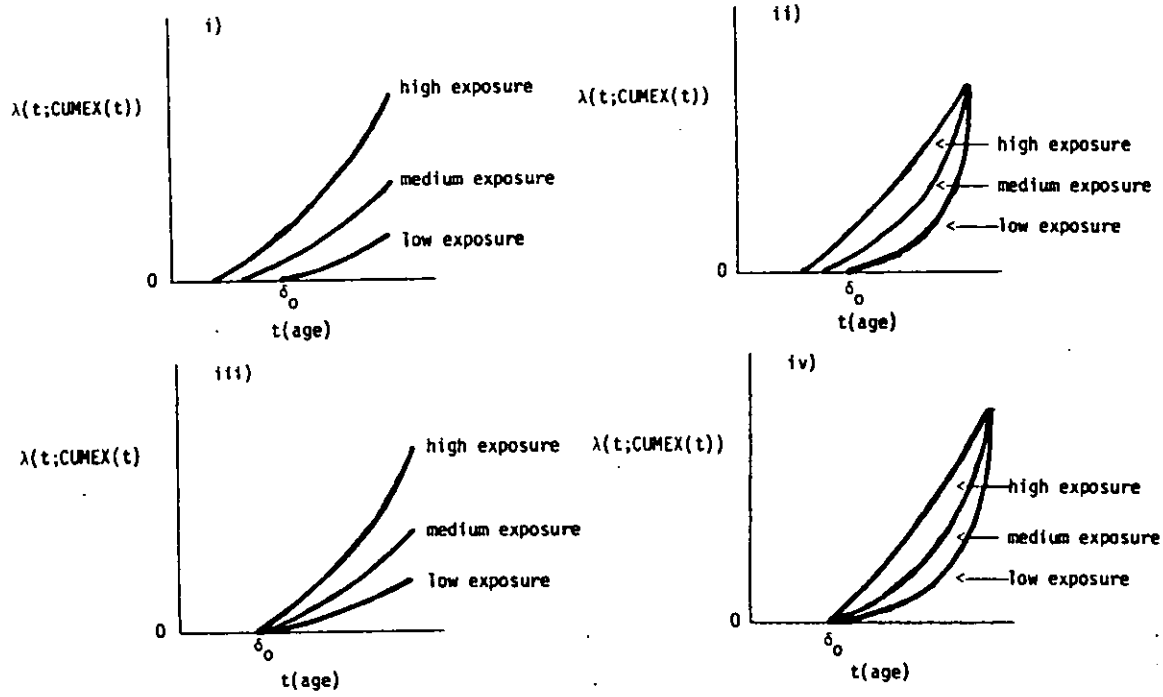
*The constraint that $[\gamma_0 + \gamma_1 \log(\text{CUMEX}(t))] > 0$ for all observed values of CUMEX(t) was released.

However, there were some discrepancies among the parameter estimates in the log(CUMEX(t)) model. The magnitudes of the parameters β_3 and γ_1, the log(CUMEX(t)) effect, were estimated to be smaller by the Gompertz model. However, when the constraint on the parameters γ_0 and γ_1 in (4.3) was released, allowing for a negative estimate of the shape parameter at very large exposures, the parameter estimates for β_3 and γ_1 became more alike those obtained by Cox's model (see Table 4.8).

To conclude, the similarity of the results obtained by the Cox and Gompertz models provided evidence in addition to the results obtained by the nonparametric cumulative hazard estimates that the choice of an underlying Gompertz distribution was reasonable for this data.

4.5 Discussion

Four scenarios of the dose-response effect of cumulative exposure to chrysotile asbestos (CUMEX(t)) on the hazard of death from lung cancer were investigated:



Scenarios i) and ii) are similar in that the individual's age at first risk of death decreases with cumulative exposure. Scenarios iii) and iv) assume that the individual's age at first risk of death is not a function of cumulative exposure, that is, individuals begin dying from lung cancer at the same age, no matter what their cumulative exposures are. Scenarios i) and iii) are similar in that the hazard curves are proportional with time beyond δ_0, that is, the dose-response effect remains constant with age beyond δ_0. Scenarios ii) and iv) are similar in that the dose-response effect for $t \geq \delta_0$ is strongest at the initial time δ_0, and then decreases with age until there is no longer any effect.

The data analysis found that the individual's age at first risk of death from lung cancer does not decrease with cumulative exposure; therefore, dose-response scenarios i) and ii) were ruled out. One might expect this result in an occupational cohort study since the textile workers received different rates of exposure at various ages throughout their follow-up. However, scenarios i) and ii) might be found in animal experiments where the exposures are administered at the same time.

The results of the quadratic CUMEX(t) and log(CUMEX(t)) Gompertz models are consistent in that a statistically significant dose-response effect of cumulative exposure to chrysotile asbestos on the hazard of death from lung cancer is found. However, it is not clear which interpretation is more appropriate. The multiplicative "quadratic CUMEX(t)" model suggests a dose-response effect that remains constant with age. This is scenario iii). The log(CUMEX(t)) model, on the other hand, suggests that the dose-response effect is strongest at the individual's

age at first risk of death from lung cancer (estimated by the Gompertz model (4.3) as the first order statistic, age 46), and then decreases with age until there is no longer an effect (estimated by the Gompertz model (4.3) to be age 67.5). This is dose-response scenario iv). A larger number of observed lung cancer deaths, especially at the older ages, may have provided the investigator with sufficient information to rule out one of these two interpretations. In addition, some other transformation of CUMEX(t) might have better explained the dose-response effect.

By defining other covariates from the workers' histories of asbestos exposure, a more detailed investigation of the dose-response effect may be performed. For instance, recall from Section 4.2 that EX_m denotes the individual's chrysotile asbestos exposure during the year following his m^{th} birthday, and $CUMEX_m = \sum_{\{i:i\leq m\}} EX_i$ denotes his cumulative exposure through age m. Two covariates of interest are then defined by

$$CUMEX_b(t) = CUMEX_{m-b} = \sum_{\{i:i\leq m-b\}} EX_i$$

and

$$EX_b(t) = \sum_{\{i:m-b<i\leq m\}} EX_i,$$

where b is an integer less than m and $t \in (m-1,m]$. The covariate $CUMEX_b(t)$ denotes the individual's cumulative exposure at b years prior to age t. The covariate $EX_b(t)$ is its complement, that being the individual's total exposure within b years of age t. Clearly, $CUMEX_b(t) + EX_b(t) = CUMEX(t)$ and $CUMEX_0(t) = CUMEX(t)$. By testing for b =1,2, etc., the significance of the $EX_b(t)$ effect given that

$CUMEX_b(t)$ is in the model, one can identify the maximum number of years b* in which the hazard of death from lung cancer at age t is not significantly influenced by an individual's asbestos exposure within b* years of t. This lag time in the dose-response effect of asbestos exposure might also be estimated by b**, that value for b in which the significance of the $CUMEX_b(t)$ effect, with $EX_b(t)$ excluded from the model, is maximized.

As another example, define

MAXEXP(t) = the individual's maximum asbestos exposure within a six-month period prior to age t,

and

AGEMAX(t) = the age at which MAXEXP(t) had occurred.

By including these covariates into a model with CUMEX(t), one may determine simultaneously the significance of two hypothesized forms of dose-response: a cumulative exposure effect versus a "one shot" exposure effect.

There were three time descriptors of an individual's death from lung cancer in this data: age at death, time since initial employment at death, and calendar time at death. Two individuals' hazards at the same age were considered to be more homogeneous than two individuals' hazards at the same time since initial employment or two individulas' hazards at the same calendar year. Based on this assumption, the individual's age at death was the most important descriptor among the three, and it was chosen as the outcome variable. Although the other two time descriptors were incorporated into the hazard as covariates, either one of them could have been used instead

as the outcome variable. In that case, the individual's age at time t would have been used as a covariate.

Finally, the parametric models that were used in this chapter assumed the hazard to be zero prior to some age δ_0, and positive beyond that age. The parameter δ_0 represents age latency of the disease, that is, the individual's age at first risk of death from lung cancer. The Gompertz model estimated δ_0 as the first order statistic, age 46 years; consequently, the hazard was estimated to be zero prior to that age. Since an individual's risk of death from lung cancer is truly positive prior to age 46, one might consider using a parametric model that allows for this biological fact. Whatever model is used, however, the main focus should be on the ascertainment of reliable dose-response estimates from the data at hand.

CHAPTER 5

SUGGESTIONS FOR FUTURE RESEARCH

In this dissertation, repeated measurements were incorporated into the hazard by defining the covariates as step functions of time, remaining constant between consecutive measurements. In the design of any follow-up study that plans to utilize such covariates, the question of how often to perform the measurements must be addressed. A balance between precision and cost will have to be found. In addition, biological implications must be considered. If one is interested in the effect of a covariate that is defined by the most recent measurement, then performing the measurements once a year will yield a different covariate than if the measurements are performed once a month. A covariate $z(t)$ that denotes the mean value within one year of time t may have a different meaning than a covariate $z(t)$ that denotes the mean value within one month of t.

The use of past history covariates - covariates defined by an individual's repeated measurements prior to his most recent value - may provide a more thorough investigation of a particular variable's effect on the hazard. For example, assume that an individual's cholesterol level is measured once every three months in a follow-up study. Then the effect of an individual's cholesterol level on his

hazard of heart disease onset may be investigated by at least two covariates: the individual's mean value within three months of time t (the most recent value) and the individual's maximum cholesterol level within five years of time t (the maximum of the most recent twenty values). The potential use for this approach appears to be vast, and it is hoped that more ground will be broken in the future.

The piecewise exponential hazard model should yield more precise estimates of the cumulative hazard than those obtained by the discrete hazard model when there is much censoring and covariate information within the intervals separating the exact failure times (see Section 1.7). Asymptotically, the two approaches may be equivalent. Formal comparisons between these two approaches should be performed for small and large samples.

Finally, computer programs that are tailored for the incorporation of repeated measurements into the hazard need to be developed.

BIBLIOGRAPHY

Bailey, R.C., Homer, L.D., and Summe, J.P. (1977). A proposal for the analysis of kidney graft survival. Transplantation 24, 309-315.

Beck, G.J. (1979). Stochastic survival models with competing risks and covariates. Biometrics 35, 427-438.

Birnbaum, Z.W. (1979). On the mathematics of competing risks. DHEW Publication No. (PHS) 79-1351. U.S. Department of Health, Education, and Welfare, pp. 1-58.

Breslow, N.E. (1972). Discussion on Professor Cox's (1972) paper. J. R. Stat. Soc. B 34, 216-217.

Breslow, N.E. (1974). Covariance analysis of censored survival data. Biometrics 30, 89-99.

Breslow, N.E. (1975). Analysis of survival data under the proportional hazards model. Int. Stat. Rev. 43, 45-48.

Burbank, F. and Fraumeni, J. (1972). U.S. cancer mortality: non-white predominance. J. Natl. Cancer Inst. 49, 649-659.

Chen, Y.Y., Hollander, M., and Langberg, N.A. (1982). Small-sample results for the Kaplan-Meier estimator. J. Am. Stat. Assoc. 77, 141-144.

Chiang, C.L. (1968). Introduction to Stochastic Processes in Biostatistics. Wiley, New York.

Cox, D.R. (1972). Regression models and life tables (with discussion). J. R. Stat. Soc. B 34, 187-220.

Cox, D.R. (1975). Partial likelihood. Biometrika 62, 269-276.

Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant data. J. Am. Stat. Assoc. 72, 27-36.

David, H.A. and Moeschberger, M. (1978). The Theory of Competing Risks. Griffin, London.

Dement, J.M. (1980). Estimation of dose and evaluation of dose-response in a retrospective cohort mortality study of chrysotile asbestos textile workers. Ph.D. dissertation in Environmental Sciences and Engineering. University of North Carolina at Chapel Hill.

Dement, J.M., Harris, R.L., Symons, M.J., and Shy, C. (1982). Estimates of dose-response for respiratory cancer among chrysotile asbestos textile workers. Ann. Occup. Hyg. 26, 869-887.

Efron, B. (1977). Efficiency of Cox's likelihood function for censored data. J. Am. Stat. Assoc. 72, 557-565.

Elandt-Johnson, R.C. (1976). Conditional failure time distributions under competing risk theory with dependent failure times and proportional hazard rates. Scand. Acta. J., 37-51.

Elandt-Johnson, R.C. (1980). Some prior and posterior distributions in survival analysis: A critical insight on relationships derived from cross-sectional data. J. R. Stat. Soc. B 42, 96-106.

Elandt-Johnson, R.C. and Johnson, N.L. (1980). Survival Models and Data Analysis. Wiley, New York.

Elandt-Johnson, R.C. (1981). On some methodological issues in the analysis of survival data from prospective-type experiments. Institute of Statistics Mimeo Series No. 1340. University of North Carolina at Chapel Hill. pp. 1-36.

Farewell, V.T. (1979). An application of Cox's proportional hazard model to multiple infection data. Applied Statistics 28, 136-143.

Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. Biometrics 21, 826-838.

Freeman, D.H. and Holford, T.R. (1980). Summary rates. Biometrics 36, 195-206.

Gail, M.H. (1975). A review and critique of some models used in competing risks analysis. Biometrics 31, 209-222.

Gail, M.H. (1981). Evaluating serial cancer marker studies in patients at risk of recurrent disease. Biometrics 37, 67-78.

Glasser, M. (1967). Exponential survival with covariance. J. Am. Stat. Assoc. 62, 561-568.

Gross, A.J. and Clark, V.A. (1975). Survival Distributions: Reliability Applications in the Biomedical Sciences. Wiley, New York.

Harris, P.J., Harrell, F.E., Lee, K.L., Behar, V.S., and Rosati, R.A. (1979). Survival in medically treated coronary artery disease. Circulation 60, 1259-1269.

Harris, P.J., Behar, V.S., Harrell, F.E., Conley, M.J., Lee, K.L., and Rosati, R.A. (1981). Multivariable prediction of outcome in coronary disease based on left ventricular ejection fraction and other important prognostic characteristics. Unpublished report by the Department of Community and Family Medicine, Duke University Medical Center.

Hazelrig, J.B., Turner, M.E., and Blackstone, E.H. (1982). Parametric survival analysis combining longitudinal and cross-sectional-censored and interval-censored data with concomitant information. Biometrics 38, 1-16.

Higgins, J.E. (1978). A model for the analysis of survival with an intervening event. Ph.D. dissertation in Biostatistics. University of North Carolina at Chapel Hill.

Hoel, D.G. (1972). A representation of mortality data by competing risks. Biometrics 28, 475-488.

Holford, T.E. (1976). Life tables with concomitant information. Biometrics 32, 587-597.

Holt, J.D. (1978). Competing risk analyses with special reference to matched pair experiments. Biometrika 65, 159-165.

Kalbfleisch, J.D. (1974). Some efficiency calculations for survival distributions. Biometrika 61, 31-38.

Kalbfleisch, J.D. and Prentice, R.L. (1973). Marginal likelihoods based on Cox's regression and life model. Biometrika 60, 267-278.

Kalbfleisch, J.D. and Prentice, R.L. (1980). The Statistical Analysis of Failure Time Data. Wiley, New York.

Kaplan, E.B. and Elston, R.C. (1972). A subroutine package for maximum likelihood estimation (MAXLIK). Institute of Statistics Mimeo Series No. 823. University of North Carolina at Chapel Hill.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. 53, 457-481.

Lagakos, S.W. (1976). A stochastic model for censored survival data in the presence of an auxiliary variable. Biometrics 32, 551-559.

Lagakos, S.W. (1979). General right censoring and its impact on the analysis of survival data. Biometrics 35, 139-156.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother. Rep. 50, 163-170.

Mantel, N. and Byar, D.P. (1974). Evaluation of response time data involving transient states: An illustration using heart transplant data. J. Am. Stat. Ass. 69, 81-86.

Nelson, W. (1972). Theory and application of hazard plotting for censored failure data. Technometrics 14, 945-966.

Prentice, R.L., Kalbfleisch, J.D., Peterson, A., Flournoy, N., Farewell, V., and Breslow, N. (1978). The analysis of failure times in the presence of competing risks. Biometrics 34, 541-554.

Prentice, R.L., Williams, B.J., and Peterson, A.V. (1981). On the regression analysis of multivariate failure time data. Biometrika 68, 373-379.

Stablein, D.M., Carter, W.H., and Wampler, G.L. (1980). Survival analysis of drug combinations using a hazards model with time dependent covariates. Biometrics 36, 537-546.

Taulbee, J.D. (1977). A general model for the hazard rate with covariables and methods for sample size determination for cohort studies. Ph.D. dissertation in Biostatistics. University of North Carolina at Chapel Hill.

Taulbee, J.D. (1979). A general model for the hazard rate with covariables. Biometrics 35, 439-450.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. Proc. Natl. Acad. Sci. 72, 20-22.

Turnbull, B.W., Brown, B.W., and Hu, M. (1974). Survivorship analysis of heart transplant data. J. Am. Stat. Assoc. 69, 169-173.

Williams, J.S. and Lagakos, S.W. (1977). Models for censored survival analysis: Constant-sum and variable-sum models. Biometrika 64, 215-224.

APPENDIX I

Maximum Likelihood Estimation of the Location Parameter in the Gompertz Distribution

Assume that n individuals' failure times, $t_1,\ldots,t_n$, follow a Gompertz (β,γ,δ) distribution, that is

$$\lambda(t) = \begin{cases} \beta e^{\gamma(t-\delta)}, & t \geq \delta,\ \beta > 0,\ \gamma > 0, \\ 0, & \text{otherwise.} \end{cases}$$

The likelihood for these uncensored observations is written as

$$L = \left\{ \prod_{i=1}^{n} \beta e^{\gamma(t_i-\delta)} e^{-\frac{\beta}{\gamma}(e^{\gamma(t_i-\delta)} - 1)} \right\} h(t_{(1)} - \delta),$$

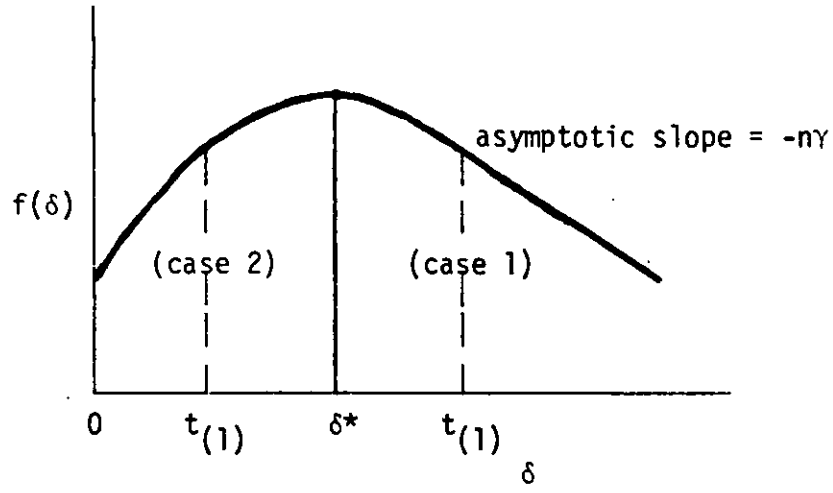
where $t_{(1)}$ is the first order statistic and

$$h(t_{(1)} - \delta) = \begin{cases} 1 & \text{if } \delta \leq t_{(1)} \\ 0 & \text{otherwise.} \end{cases}$$

If the failure times were exponentially distributed, then the likelihood would increase strictly with δ as it approaches $t_{(1)}$. Here, as δ increases towards $t_{(1)}$, the contributions to the likelihood of the individuals' survival functions increase, but their hazards decrease. The log likelihood, as a function of δ, is proportional to

$$f(\delta) = -n\gamma\delta - \left\{ \frac{\beta}{\gamma} \sum_{i=1}^{n} e^{\gamma t_i} \right\} e^{-\gamma\delta},$$

subject to the restriction that $\delta \leq t_{(1)}$. A graph of δ versus $f(\delta)$ appears below,



and $f(\delta)$ is maximized at

$$\delta^* = \frac{\ln\left\{\frac{\beta}{\gamma}\,\frac{\sum_{i=1}^{n} e^{\gamma t_i}}{n}\right\}}{\gamma}$$

If $\delta^* \le t_{(1)}$ (case 1), then it is the MLE of δ. However, if $\delta^* > t_{(1)}$ (case 2), then $f(\delta)$ is maximized at $t_{(1)}$ for $\delta \le t_{(1)}$, and $t_{(1)}$ becomes the MLE of δ. Denoting $t_{(i)}$ as the i^{th} order statistic, i=1,...,n,

$$\delta^* = \frac{\ln\left\{\frac{\beta}{\gamma n}\, e^{\gamma t_{(1)}}\left[1 + e^{\gamma(t_{(2)} - t_{(1)})} + \ldots + e^{\gamma(t_{(n)} - t_{(1)})}\right]\right\}}{\gamma}$$

$$\Longrightarrow \quad \delta^* = t_{(1)} + \frac{\ln\left\{\frac{\beta}{\gamma}\,\frac{\sum_{i=1}^{n} e^{\gamma(t_{(i)} - t_{(1)})}}{n}\right\}}{\gamma}.$$

If the argument in the log function is greater than one, then $\delta^* > t_{(1)}$. This will happen if $\beta \ge \gamma$ or if the failure times are spread out enough to compensate for $\beta < \gamma$.

APPENDIX II

Documentation of the Program that Computes the Piecewise Exponential Cumulative Hazard Estimates in Chapter 4

Using SAS (Statistical Analysis System computer package), the nonparametric cumulative hazard estimates in Figures 4.1, 4.4, 4.5, and 4.6 were obtained by the following program. Included are the actual values.

$CUMEXA_m$ denotes the total exposure within 10 years of the m^{th} birthday,

$CUMEXB_m$ denotes the total exposure between 10 and 20 years prior to the m^{th} birthday,

$CUMEXC_m$ denotes the total exposure between 20 and 30 years prior to the m^{th} birthday,

$CUMEXD_m$ denotes the cumulative exposure at 30 years prior to the m^{th} birthday,

AGE denotes the age last birthday m, where m=13,...,85,

FEMPAG denotes the age last birthday at first employment,

AGEDTH denotes the age last birthday at death from lung cancer,

GRP identifies the covariate stratum,

D denotes the observed number of deaths from lung cancer at a given age (within strata),

PRYRS denotes person-years of follow-up (person-years of follow-up prior to the first order statistic, age 46, were not used),

HAZ denotes the hazard,

LENGTH denotes the lengths of the intervals between the exact failure times, and

CUMHAZ denotes the piecewise exponential cumulative hazard estimates at the death times.

```
NOTE: THE PROCEDURE DELETE USED 0.06 SECONDS AND 60K.
1                                          PROC RELEASE DDNAME=RAT;
NOTE: 23 TRACKS ALLOCATED, 23 USED, 3 EXTENTS, BEFORE RELEASE.
NOTE: 23 TRACKS ALLOCATED, 23 USED, 3 EXTENTS, AFTER RELEASE.
NOTE: THE PROCEDURE RELEASE USED 0.08 SECONDS AND 60K.

2          DATA T; SET TAPE.CCV;
3          ARRAY LATNCY LATNCY13-LATNCY85;
4          ARRAY CUMEXA CUMEXA13-CUMEXA85;
5          ARRAY CUMEXB CUMEXB13-CUMEXB85;
6          ARRAY CUMEXC CUMEXC13-CUMEXC85;
7          ARRAY CUMEXD CUMEXD13-CUMEXD85;
8          ARRAY CUMEX CUMEX13-CUMEX85;
9          ARRAY C1X C1X13-C1X85; ARRAY C2X C2X13-C2X85; ARRAY C3X C3X13-C3X65;
10         DO OVER CUMEX;
11         CUMEX=CUMEXA+CUMEXB+CUMEXC+CUMEXD;
12         AGE= I +12;
13         IF FERTAGE<=AGE<=AGEDTH THEN DO;
14         IF CUMEX<=50 THEN C1X=1; IF 50<CUMEX<=125 THEN C2X=1; IF CUMEX>125 THEN C3X=1;
15         END;
16         END;

NOTE: DATA SET WORK.T HAS 769 OBSERVATIONS AND 663 VARIABLES. 3 OBS/TRK.
NOTE: THE DATA STATEMENT USED 2.84 SECONDS AND 148K.

17         PROC MEANS SUM; VARIABLES C1X13-C1X85 C2X13-C2X85 C3X13-C3X85;
18         OUTPUT OUT=P SUM=C1X13-C1X85 C2X13-C2X85 C3X13-C3X85;
19         %MACRO WITH
20         PY=J(ND,1,0); CUMHAZ=J(ND,1,0); HAZ=J(ND,1,0); LENGTH=J(ND,1,0);
21         DO L=1 TO ND;
22         J1=T(L,*)-T(1,*)+1; J2=T(L+1,*)-T(1,*);
23         DO M=J1 TO J2;
24         PY(L,*)=PY(L,*)+C(M,*);
25         END;
26         HAZ(L,*)=D(L,*)#/PY(L,*); LENGTH(L,*)=T(L+1,*)-T(L,*);
27         IF L=1 THEN CUMHAZ(L,*)=HAZ(L,*)#LENGTH(L,*);
28         ELSE CUMHAZ(L,*)=CUMHAZ(L-1,*)+(HAZ(L,*)#LENGTH(L,*));
29         END;
30         %

NOTE: DATA SET WORK.P HAS 1 OBSERVATIONS AND 219 VARIABLES. 10 OBS/TRK.
NOTE: THE PROCEDURE MEANS USED 2.64 SECONDS AND 166K AND PRINTED PAGES 1 TO 4.

31         PROC MATRIX ;
32         FETCH C1F DATA=P(KEEP=C1X46-C1X73);  FETCH C2F DATA=P(KEEP=C2X46-C2X63);
33         FETCH C3F DATA=P(KEEP=C3X46-C3X69);

- - - - - - - - - T O P - O F - F O R M - - - - - - - - -
2                                                S T A T I S T I C A L   A N A L Y S I S   S Y S T E M

34         C1=C1F*; C2=C2P*; C3=C3P*; FREE C1P C2P C3F;
35         T1=45/47/53/54/61/65/67/68/71/73; D1=2/1/1/1/1/1/1/1/1;
36         T2=45/46/51/52/54/56/57/59/60/63; D2=1/1/1/1/1/1/1/1/1;
37         T3=45/51/52/55/56/60/61/65/69; D3=1/2/1/2/1/1/1/1;
38         ND1=NROW(T1)-1; ND2=NROW(T2)-1; ND3=NROW(T3)-1;
39         GRP1=J(ND1,1,1); GRP2=J(ND2,1,2); GRP3=J(ND3,1,3);
40         T=T1; D=D1; C=C1; ND=ND1; WITH
41         T1=T(2:ND+1,*); D1=D; PY1=PY; HAZ1=HAZ; LENGTH1=LENGTH; CUMHAZ1=CUMHAZ;
42         Z1=GRP1||T1||D1||PY1||HAZ1||LENGTH1||CUMHAZ1;
43         FREE GRP1 T1 D1 PY1 HAZ1 LENGTH1 CUMHAZ1;
44         T=T2; D=D2; C=C2; ND=ND2; WITH
45         T2=T(2:ND+1,*); D2=D; PY2=PY; HAZ2=HAZ; LENGTH2=LENGTH; CUMHAZ2=CUMHAZ;
46         Z2=GRP2||T2||D2||PY2||HAZ2||LENGTH2||CUMHAZ2;
47         FREE GRP2 T2 D2 PY2 HAZ2 LENGTH2 CUMHAZ2;
48         T=T3; D=D3; C=C3; ND=ND3; WITH
49         T3=T(2:ND+1,*); D3=D; PY3=PY; HAZ3=HAZ; LENGTH3=LENGTH; CUMHAZ3=CUMHAZ;
50         Z3=GRP3||T3||D3||PY3||HAZ3||LENGTH3||CUMHAZ3;
51         FREE GRP3 T3 D3 PY3 HAZ3 LENGTH3 CUMHAZ3;
52         Z=Z1//Z2//Z3;
53         OUTPUT Z OUT=PP(RENAME=(COL1=GRP COL2=AGEDTH COL3=D COL4=PEYRS COL5=HAZ
54         COL6=LENGTH COL7=CUMHAZ));

NOTE: DATA SET WORK.PP HAS 26 OBSERVATIONS AND 8 VARIABLES. 280 OBS/TRK.
NOTE: THE PROCEDURE MATRIX USED 0.46 SECONDS AND 144K AND PRINTED PAGE 5.

55         DATA RAT.LCX; SET PP; IF GRP=1 THEN CUMEX='<=50      ';
56         IF GRP=2 THEN CUMEX='>50,<=125'; IF GRP=3 THEN CUMEX='>125      ';
57         KEEP GRP CUMEX AGEDTH D PEYRS HAZ LENGTH CUMHAZ;

NOTE: DATA SET RAT.LCX HAS 26 OBSERVATIONS AND 8 VARIABLES. 276 OBS/TRK.
NOTE: THE DATA STATEMENT USED 0.09 SECONDS AND 68K.

58         PROC CONTENTS DIRECTORY DATA=RAT.LCX;

NOTE: THE PROCEDURE CONTENTS USED 0.10 SECONDS AND 60K AND PRINTED PAGES 6 TO 7.

58                                                 PROC PRINT;

NOTE: THE PROCEDURE PRINT USED 0.18 SECONDS AND 124K AND PRINTED PAGE 8.

59         PROC PLOT; PLOT CUMHAZ*AGEDTH=GRP;

NOTE: THE PROCEDURE PLOT USED 0.20 SECONDS AND 118K AND PRINTED PAGE 9.

NOTE: SAS USED 154K MEMORY.

NOTE: SAS INSTITUTE INC.
      SAS CIRCLE
      BOX 8000
      CARY, N.C. 27511-8000

- - - - - - - - - T O P - O F - F O R M - - - - - - - - -
```

```
------------------------------ SOURCE STATEMENTS ------------------------------
DATA RAT.LCX; SET PP; IF GRP=1 THEN CUMEX='<=50      ';
IF GRP=2 THEN CUMEX='>50,<=125'; IF GRP=3 THEN CUMEX='>125     ';
KEEP GRP CUMEX AGEDTH D PRYRS HAZ LENGTH CUMHAZ;
```

S T A T I S T I C A L A N A L Y S I S S Y S T E M

OBS	GRP	AGEDTH	D	PRYRS	HAZ	LENGTH	CUMHAZ	CUMEX
1	1	47	2	780	0.0025641	2	0.005128	<=50
2	1	53	1	1847	0.0005414	6	0.008377	<=50
3	1	54	1	222	0.0045045	1	0.012881	<=50
4	1	61	1	998	0.0010020	7	0.019895	<=50
5	1	65	1	258	0.0038760	4	0.035399	<=50
6	1	67	1	85	0.0117647	2	0.058929	<=50
7	1	68	1	33	0.0303030	1	0.089232	<=50
8	1	71	1	72	0.0138889	3	0.130898	<=50
9	1	73	1	24	0.0416667	2	0.214232	<=50
10	2	46	1	110	0.0090909	1	0.009091	>50,<=125
11	2	51	1	474	0.0021097	5	0.019639	>50,<=125
12	2	52	1	96	0.0104167	1	0.030056	>50,<=125
13	2	54	1	179	0.0055866	2	0.041229	>50,<=125
14	2	56	1	161	0.0062112	2	0.053652	>50,<=125
15	2	57	1	68	0.0147059	1	0.068358	>50,<=125
16	2	59	1	117	0.0085470	2	0.085452	>50,<=125
17	2	60	1	48	0.0208333	1	0.106285	>50,<=125
18	2	63	1	111	0.0090090	3	0.133312	>50,<=125
19	3	51	1	252	0.0039683	6	0.023810	>125
20	3	52	2	43	0.0465116	1	0.070321	>125
21	3	55	1	124	0.0080645	3	0.094515	>125
22	3	56	2	40	0.0500000	1	0.144515	>125
23	3	60	1	141	0.0070922	4	0.172883	>125
24	3	61	1	29	0.0344828	1	0.207366	>125
25	3	65	1	88	0.0113636	4	0.252821	>125
26	3	69	1	45	0.0222222	4	0.341710	>125