A GENERAL MODEL FOR THE HAZARD RATE WITH COVARIABLES AND METHODS FOR SAMPLE SIZE DETERMINATION FOR COHORT STUDIES

by

John DeWeese Taulbee

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1154

December 1977

A GENERAL MODEL FOR THE HAZARD RATE WITH COVARIABLES AND METHODS FOR SAMPLE SIZE DETERMINATION FOR COHORT STUDIES

by

John DeWeese Taulbee

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

1977

Approved by:

Michael J. Symons
Advisor

R. C. Elandt-Johnson
Advisor

[illegible]
Reader

Kerry L. Lee
Reader

ABSTRACT

JOHN DEWEESE TAULBEE. A General Model for the Hazard Rate with Covariables and Methods for Sample Size Determination for Cohort Studies. (Under the direction of MICHAEL J. SYMONS and REGINA C. ELANDT-JOHNSON.)

This research is concerned with developing improved methods for analyzing survival data and determining appropriate sample sizes for cohort studies.

The model proposed for the hazard function incorporating covariables is a polynomial with different functions of the covariables as coefficients of the various powers of time. This model does not require the assumption that the hazards for different individuals be in constant ratio over time, and it allows for testing whether this assumption is reasonable. The model is parametric, which allows for easy specification of the survival curve and interpretation of results. At the same time, it is general enough so that the form of the hazard is not unduly restricted.

Methods for fitting the model to data, testing hypotheses about model parameters, and testing the goodness of fit of the model are presented. Use of the model and these methods is illustrated by application to three different cohorts. Results obtained using other methods are compared with those of the method proposed here for one of the cohorts.

Methods for determination of required sample sizes for cohort survival studies are developed and illustrated with examples. The

methods require that the hypothesis of interest be expressed in terms of the model parameters and that the desired test size and power required against a specific alternative be specified. A distribution of follow-up times must also be specified.

ACKNOWLEDGMENTS

It has been a special privilege during this research to work closely with my advisors, Dr. M.J. Symons and Dr. Regina C. Elandt-Johnson. Their time, encouragement, and helpful discussions are gratefully acknowledged. The other members of my advisory committee, Drs. P.K. Sen, K.L. Lee, and L.A. Ripperton have also been most helpful, and a special thanks is extended to each of them.

I especially wish to thank Anna Colosi for her skillful computer programming. All computer runs required for Chapter III, as well as other runs used to generate ideas were done by her. Gratitude is also expressed to Joyce Hill for her skillful typing of the manuscript.

The financial assistance of the Department of Biostatistics during the course of this graduate program is gratefully acknowledged.

Finally, I must give thanks to my family. My wife Anita provided unfailing encouragement, support and love. Our new daughter, Anita Ruthanne, has provided added incentive to complete this work. I am also indebted to my parents and my sister for their encouragement and support.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION AND REVIEW OF THE LITERATURE

1.1 Introduction

The aim of this research is to provide useful methods for the study of survival of cohorts, including techniques for more effective analysis and for determination of appropriate sample sizes for such studies. These methods would be applicable to morbidity and mortality studies, clinical trials, and any area where the study of time until a specific event is of interest.

Survival analysis is the area of statistical methods to which this research is directed. A review of such methods which have already been in use is presented beginning in Section 1.4. This is followed in Section 1.8 by an outline of the material presented in subsequent chapters.

A review of methods for sample size determination is included in Chapter IV.

Of special interest for application are studies of occupational cohort mortality. In view of this, we next discuss the purpose of these investigations and the type of data which is usually available. A review of methods currently used in the study of occupational mortality is then provided.

1.2 Occupational Mortality Studies: Purpose and Data

Before discussing methods for analysis of studies of occupational mortality, we will first briefly mention some of the reasons for these studies and the type of data which are generally used.

Many of the studies are undertaken with the objective of determining whether the mortality patterns for people who work in or have retired from a particular industry or company are different from those of another group or population of individuals. If differences are noted, another investigation may be undertaken to determine if particular jobs, job types or work areas within the industry or company are associated with these differences. Most often, the latter is conducted and analyzed in a manner similar to the former. That is, groups of workers are formed on the basis of characteristics of their jobs and the mortality of these groups is compared with that of all groups considered together or with some other population.

The data used for analysis most often consists of information about each member of a cohort of workers. The cohort usually consists of a group of workers with some characteristics in common. For example, one or more of the criteria in the following list are most often used to identify a cohort:

1. all persons working in a particular plant or industry as of a given date or period of time;
2. all retirees of a plant living as of a given date or time period; or
3. all those entering a workforce in a certain time period.

Other criteria may be used to form a cohort, but the three listed above, in some combination, appear to be the most frequently used. Indeed, virtually all of the studies discussed in the next section used some combination of the above criteria for identification of the cohort studied, the major exception being those using proportional mortality analysis. For this type of study, only those workers and/or retirees of a plant or industry who have died in a specified time period are considered.

The data available on each member of the study population usually include at least time of observation, (i.e., length of time that the individual was in the study), age, sex, and for deaths, time of death. Some or all of the following information may also be included: race, date of hire, and, where appropriate date and type of retirement, cause of death and some information about type of job(s) held within the plant or industry.

The studies discussed in the next section are too numerous for a description of the criteria used to identify the study group and the data available to be given for each one, but the above description provides an accurate picture of how the study populations are formed and what data are used.

1.3 Currently Used Methods of Analysis of Occupational Cohort Mortality Data

Methods currently used for analysis of occupational mortality data might be called standard epidemiological methods for analysis of mortality data. These are discussed next.

1.3.1 Proportional Mortality Analyses

Proportional mortality analysis is employed when it is difficult or undesirable to obtain information concerning mortality and life status (living or dead) for the population under study. Only deaths from the population of interest are identified. Generally, for the specific cause of interest, the proportion of all deaths due to that cause in the standard population for a specific age, sex, race group is multiplied by the total number of deaths in the corresponding age, sex, race group in the population of interest, resulting in an *expected* number of deaths. Of course, grouping can be done on other variables as well as age, sex, and race. The standard population is usually the population of the nation or some local governmental unit. Observed and expected numbers of deaths are summed over the groups, and the ratio of the sums (observed/expected) is called the standardized proportional mortality ratio (SPMR).

Moss, Scott, and Atherley [1972] studied mortality from lung cancer and bronchitis among newspaper workers by this method. Other studies using this method include a study of oral and pharyngeal cancers in textile workers by Moss and Lee [1974] and a study of causes of death among construction machinery operators by Decoufle, Lloyd, and Salvin [1977].

The main justification for using this method of analysis is that if the population of interest and the standard population have similar overall mortality, the SPMR generally relates the same information regarding a specific cause of death as the standardized mortality ratio, which will be discussed later.

1.3.2 Computation of Death Rates

Some investigators have observed the mortality experience of an occupational cohort, computed *age - specific death rates* for the cause(s) of interest, and compared these age specific rates with those from another industry or a national or local population. Enterline [1974] used this method to compare cancer mortality among chromate workers with that among oil refinery workers as part of a broader study of respiratory cancer among chromate workers. Mancuso, Ciocco, and El-Attar [1968] also studied rubber worker mortality in this way. The method appears useful only for making general comparisons and tests of hypothesis are not usually applied.

The age specific death rate is the number of deaths occuring in one year among those in a particular (usually five-year) age range divided by the average population in that age range for a one-year period. (The rate is usually expressed as some number per 1000 or per 100,000.)

The practice of comparing the direct age - adjusted death rates (adr) of populations seems to have fallen into disuse, at least insofar as occupational mortality studies are concerned. Suppose we have n age groups indexed by $i = 1, 2, \ldots, n$ (or, more generally age-sex-race groups and so on). A standard mid-year population is chosen with K_i individuals in age group i. Let m_{ji} be the age specific death rate for age group i for study population j ($j = 1, 2$, for illustration). Then the age adjusted death rate for population j is

$$adr_j = \left[\sum_{i=1}^{n} (m_{ji} K_i)\right] \left[\sum_{i=1}^{n} K_i\right]^{-1} .$$

Instead of comparing separately the n pairs (m_{1i}, m_{2i}), $i = 1, 2, \ldots, n$, we would compare adr_1 with adr_2. Equality of adr_1 with adr_2 implies

$$\sum_{i=1}^{n} m_{1i}K_i = \sum_{i=1}^{n} m_{2i}K_i,$$

or

$$\left[\sum_{i=1}^{n} (m_{1i} - m_{2i})K_i\right] = 0.$$

Of course, this last equality can hold even if $m_{1i} \gg m_{2i}$ for some i, and $m_{1i} \ll m_{2i}$ for other i. In the absence of further assumptions, comparison of the adr_j is not very informative about the mortality patterns of the two groups. However, if we can assume that the age specific rates are in the same ratio for all age groups, that is, $m_{2i} = cm_{1i}$ for all i then $adr_1 = adr_2$ implies

$$\sum_{i=1}^{n} [m_{1i} - cm_{1i}]K_i = 0 = (1 - c)\left(\sum_{i=1}^{n} m_{1i}K_i\right),$$

whence $c = 1$ and $m_{2i} = m_{1i}$ for all i. If we can assume a constant difference between m_{1i} and m_{2i}, then equality of the adjusted rates will also imply equality of the age specific rates for all age groups. Some such assumptions must be made, then, for useful comparison of adjusted death rates. Often these assumptions may not hold, and this may account for the disuse of the method.

The assumptions made are somewhat analogous to assuming a constant ratio of hazards over time in survival analysis, which will be discussed later.

1.3.3 Analysis Using the Standardized Mortality Ratio (SMR)

Computation of SMRs facilitates comparison of death rates in much the same way that SPMR computation facilitates comparison of the proportional mortality of two groups.

Generally, a cohort of workers is identified as of a point in time and the mortality experience of the cohort is observed. The number of person years accumulated in each of several age-sex-race groups by the cohort members is multiplied by the death rate for that age-sex-race group in the standard population, resulting in an expected number of deaths. Observed and expected numbers of deaths are summed over the groups, and the ratio of the sums (observed/expected) is called the standardized mortality ratio, or SMR.

Investigations using this method of analysis seem to be the most frequently done in studying occupational mortality. Many such studies use the national population of a comparable time period as the standard population, including the following: Fox, Lindars, and Owen's [1974] study of cancer in the rubber and cablemaking industry, studies of the mortality of asbestos workers by Enterline [1965] and Enterline, Decoufle, and Henderson [1972], and other investigations by Enterline and Henderson [1975], Tabershaw and Gaffey [1974], Ingalls and Risquez-Iribarren [1961], Case and Hosker [1954], Doll [1952], and Ortmeyer, *et.al*. [1974].

Other studies producing SMRs have used state and local populations as the standard in the hope of gaining a more accurate picture of the change in mortality associated with work in an industry. Among these are the following: Dorn and Baum's [1955] study of mortality in cigarette factory workers, Lloyd and Ciocco's [1969] study of mortality among steelworkers, as well as studies by Henderson and Enterline [1973], Wagoner, et.al. [1964], Mancuso and Coulter [1959] and [1963], and Monson and Nakano [1976a] and [1976b].

Because of problems in interpretation of SMRs computed using national, state, and local populations as the standard arising from the healthy worker effect, some authors have used the industrial population of interest as the standard in investigating the mortality of subset(s) of that population. This was done in a study by Andjelkovich, Taulbee, Symons, and Williams [1977] wherein the mortality of 28 mutually exclusive subgroups, based on within plant job experience, was examined using the whole plant population as the standard. Other studies using industrial populations as a standard include those by Lloyd, Lundin, Redmond, and Geiser [1970], Reid and Buck [1956], Doll, et.al. [1965], Pasternack and Ehrlich [1972] and Doll, et.al [1972].

In studying mortality experience of uranium workers, Scott, Bahler, DeLa Garza, and Lincoln [1972] employed yet a different form of standard population and a different method for obtaining the expected number of deaths. Instead of age-specific mortality rates, information on the Bureau of Vital Statistics standard population was available in life-table form. Age at first observation, sex, race, and length of observation for each individual in the study population was used to obtain the probability of death in the observation period for a person of the same age, race, and sex from the life-table. The sum of these probabilities was considered to be the expected number of deaths for the population of interest, and the ratio of observed to expected deaths was given. Such a method would be useful when the mortality experience of the desired standard population is embodied in a life table rather than in the form of age-specific death rates.

Presentation of the data in terms of survival probabilities was uncommon in these studies. Two studies presented empirical survival curves in addition to SMR analyses. Elmes and Simpson [1971] presented a comparison of empirical survival curves (percent surviving versus age), for different ages of entry into industry and Taylor [1966] presented empirical survival curves for various levels of cumulative experience of chromate workers. In both cases, the information presented in the survival curves was discussed only briefly, and the SMR analysis was the main thrust of the investigation.

1.3.4 Shortcomings of the SMR

Gaffey [1975] observed that when the results of a study of occupational mortality are analyzed, the choice of a standard population against which to compare the results and the choice of a methodology for carrying out the comparison are the two elements involved.

As can be seen from earlier discussion, several alternatives are available for the standard population. However, as Gaffey further notes and as we have observed, "when the appropriate data are available, the Standardized Mortality Ratio (SMR) is almost universally employed" for analysis of the data.

Gaffey goes on to illustrate three characteristics of the SMR which may make it undesirable. First, "there is essentially no relationship between the value of the SMR and life expectancy". Two populations with the same number of person years in each group and the same number of deaths will have the same SMR regardless of whether the deaths are distributed differently in the two populations. Gaffey concludes that "within rather wide limits, populations with different life expectancies may have the same SMR, and populations with the same life expectancies may have different SMRs". Gaffey's statements in regard to life expectancies could be made about survival distributions, also. Two other characteristics of the SMR, discussed a.d illustrated by Gaffey, which may make it undersirable are 1) that the SMR and relative risk are not, in general, equal, with the size of the discrepancy depending on the age of the study population and 2) that at older ages the SMR is subject to limitations on its possible values, more or less independently of any hazard to which the study population may be exposed.

As a result of the first characteristic mentioned in the preceding paragraph, we note that it is not useful to compare SMR's of two different study populations even if the same standard population has been used.

Gaffey concludes that the SMR "has the advantage of being explainable, at least at a superficial level, to persons who are outside the field, and the negative advantage that there are at present no other generally accepted techniques".

From the preceding discussion, the need to develop better methods of analysis for study of occupational mortality seems clear.

1.4 Survival Analysis Methods

Methods for survival analysis have long been available. The methods discussed in this section consider survival time only. Other information about study subjects is not utilized. These methods are discussed and evaluated with consideration also given to their usefulness for the problem of analysis of occupational cohort mortality data.

1.4.1 Actuarial Method

The actuarial method has been used to study the survival of groups of individuals for a long time (see, e.g., Berkson and Gage [1950]). The method generally entails division of the follow-up period into intervals with the important data being number alive at the start of each interval, number dying during the interval, and number withdrawn alive during the interval. Estimates for conditional probabilities of dying within an interval or surviving an interval, as well as the unconditional probability of surviving to the start of an interval are often calculated. Estimates of the hazard may also be calculated for each interval.

1.4.2 Product-Limit Method

Kaplan and Meier [1958] describe a method, called the product limit method, which produces a non-parametric maximum likelihood estimate of the survival function. The survival function for a specified time t is the product over all failure times less than or equal to t of the number alive just after a particular failure time divided by the number alive just before. This method calls for individual observation with each response time recorded, rather than the grouped data used with the actuarial method.

1.4.3 Parametric Methods

There are many parametric forms proposed as useful in describing the survivorship of various populations. These can be characterized by a hazard function, $\lambda(t)$, which uniquely specifies both the survival curve, $S(t)$, and the density $f(t)$ through the following relationships:

$$S(t) = \exp\left[-\int_0^t \lambda(u)\,du\right], \quad \lambda(u) \geq 0, \quad u \geq 0$$

$$f(t) = -\frac{dS(t)}{dt} .$$

For the exponential distribution, $\lambda(t)$ is a constant. According to Clark [1975] this distribution may be appropriate for human survival if *the observation time is short* and if the individuals are middle-aged, when the conditional probability of death from all causes is relatively constant. Other possibilities are the linear hazard or Rayleigh distribution (see Kodlin [1967]), which is actually a special case of the generalized Rayleigh distribution employed by Krane [1963], the Gompertz distribution discussed by Garg, et.al. [1970], and the log-normal, Weibull, and gamma distributions (see, e.g., Gross and Clark [1975]).

Considerable effort has been devoted to finding procedures for distinguishing which of the parametric models best fit the data. Comparison of the exponential and log-normal distributions is found in an article by Irwin [1942]. Antle and Klimko [1975] present a test for distinguishing the normal from the log-normal distribution. Techniques for discriminating among various groups of survival distributions are also found in articles by Gehan and Siddiqui [1973], Parr and Webster [1965], Dumonceaux and Antle [1973], and Farewell and Prentice [1977].

A considerable amount of work then, has been done in the area of describing survival distributions depending on time only. Use of these techniques for heterogeneous groups, such as an occupational cohort, is not really satisfactory. Such groups vary widely with respect to distribution of such variables as age, race, and sex, in addition to measures of work exposure that are of interest as controllable determinants of survival. One approach might be to form small, homogeneous groups from such a cohort, and analyze them separately. However, analysis may not be reliable for the small sized groups and interpretation of results would then become difficult.

Models which allow for forming homogeneous groups based on values of covariates and then comparing survival across these groups have been proposed, and are discussed in the following section.

1.5 Categorical Approaches

One approach to analyzing the effect of covariables on survival and comparing the survival of different groups is to form categories of individuals based on the values of the covariables of interest.

Mantel [1966] extends the earlier work of Mantel [1963] and Mantel and Haenszel [1959] to compare survival of two groups of individuals and indicates that extension to comparison of more than two groups is possible. The procedure, when covariates are involved, is to set up, for each of several groups formed by categorization of covariate values and simultaneously each of several study time intervals, a contingency table of the form

	Died	Survived
Group I	r_{1i}	$N_{1i} - r_{1i}$
Group II	r_{2i}	$N_{2i} - r_{2i}$

where i indicates the covariate category-time interval of the table. Mantel [1966] suggests a statistic for testing the null hypothesis that there is no difference in survival between the two groups I and II. Mantel and Haenzel [1959] indicate that the test is useful against the alternative where there is *consistent association in the same direction* over the various covariate category-time interval subclassifications between the disease and the study factors. Under the null hypothesis, the test statistic

$$(|\sum_i r_{1i} - \sum_i E(r_{1i})| - \tfrac{1}{2})^2 / \sum_i \mathrm{Var}(r_{1i})$$

is approximately distributed as chi-square with one degree of freedom. Formulas for $E(r_{1i})$ and $\mathrm{Var}(r_{1i})$ are given by Mantel [1966].

Koch, Johnson, and Tolley [1972] use the Grizzle-Starmer-Koch (GSK) method to allow for comparison of survival over t years of r groups which may be formed on the basis of categorizing covariate values. They visualize the following contingency table:

Group	Died during year 1	2	...	t	Withdrew during year 1	2	...	t	Alive at end t - year	Total in group
1	m_{112}	m_{122}	$\cdots$	m_{1t2}	m_{113}	m_{123}	$\cdots$	m_{1t3}	m_{1t1}	$m_{1\cdot\cdot}$
2	m_{212}	m_{222}	$\cdots$	m_{2t2}	m_{213}	m_{223}	$\cdots$	m_{2t3}	m_{2t1}	$m_{2\cdot\cdot}$
$\vdots$	$\vdots$				$\vdots$				$\vdots$	$\vdots$
r	m_{r12}	m_{r22}	$\cdots$	m_{rt2}	m_{r13}	m_{r23}	$\cdots$	m_{rt3}	m_{rt1}	$m_{r\cdot\cdot}$

The GSK method may then be used to test hypotheses concerning differences in survival of the r groups.

The form of the survival distribution is not given much attention by these methods. This is not a problem if differences in survival are the only point of interest, but in some cases the form of the survival distribution is of considerable interest. Truett, Cornfield and Kannel [1967] and Glasser [1967] both point out the fact that simultaneous categorization of a large number of covariables leads to such a large number of groups that analysis would be difficult if not impossible with sample sizes most often met in practice. This points to the need for analytical techniques to handle covariates, especially continuous ones, in a more efficient manner than categorization. Regression approaches to handling covariates would seem to be promising, and these are discussed in the next sections. The first to be discussed is the logistic model, which is apparently one of the first methods to incorporate covariates with a regression approach.

1.6 The Logistic Model

The logistic model has been used to determine the effects of covariables on the probability of "response", where response might be development of disease, death, or some other event of interest.

With covariable vector $\underline{z}_i = (z_{1i}, z_{2i}, \ldots, z_{ki})'$ for individual i the model has been expressed as

$$\log\left\{\frac{\Pr(\text{response} \mid z_{1i}, \ldots, z_{ki})}{\Pr(\text{no response} \mid z_{1i}, \ldots, z_{ki})}\right\} = g(\underline{z}_i, \underline{B})$$

where $\underline{B}$ is a vector of unknown parameters, $\underline{B} = (B_1, B_2, \ldots, B_k)$.

Most often,

$$g(\underline{z}_i, \underline{B}) = \alpha + \sum_{j=1}^{k} B_j z_{ji} = \alpha + \underline{B}\,\underline{z}_i .$$

Truett, Cornfield, and Kannel [1967] used this model in an analysis of the risk of coronary heart disease in Framingham. Myers, Hankey, and Mantel [1973], Hankey and Mantel [1974], and Brown [1975] have done some work in extending the model to allow for assessment of response probability changes over intervals of time. Prentice [1976] extends and applies the model to a case - control study. Although the work extending the model to handle time interval data improves prospects for its use in analysis of survival data when the form of the survival function over time is of interest, the model seems best suited for situations where response probability is the main point of interest and the time over which observations are made is chosen by the investigator to suit the intent of his study. For example, a five-year period might be chosen to study the effect of a cancer therapy, with the response being remission, death, or some other event of interest.

Methods which simultaneously incorporate time and covariates into the survival function would seem to be more appropriate for analysis of cohort mortality data than use of the logistic model, since changes in survival over time are of interest. Such methods are discussed in the following section.

1.7 Regression Approaches to Incorporation of Concomitant Information (covariables) in Survival Models

1.7.1 Usual Form for the Hazard Function

Suppose we have information on the values of s covariables for each individual in the population of interest.

Let $\underline{z}_i = (z_{i1}, z_{i2}, \ldots, z_{is})'$ be the column vector of covariates for individual i. Then a resonable approach might be to assume that the hazard at time t for individual i is some function of t *and* $\underline{z}_i$, $\lambda(t, \underline{z}_i)$, say. There are, of course, a large number of possibilities for the form of $\lambda(t, \underline{z}_i)$, the major restriction being that $\lambda(t, \underline{z}_i) \geq 0$ at all points $(t, \underline{z}_i)$.

Let $\underline{B} = (B_1, B_2, \ldots, B_s)$ be a row vector of unknown parameters and let $\lambda_0(t)$ be a function of time not depending on $\underline{z}_i$ or $\underline{B}$ such that $\lambda_0(t) \geq 0$ for $0 \leq t < \infty$. Also, let $h(\underline{z}_i, \underline{B})$ be a function of $\underline{z}_i$ and $\underline{B}$ in which $h(\cdot, \cdot)$, $\underline{z}_i$ and $\underline{B}$ may, but usually do not depend upon time. With most regression-type models for incorporation of covariables in the survival function, the hazard function for individual i can be expressed in the following form:

$$\lambda(t, \underline{z}_i) = \lambda_0(t) h(\underline{z}_i, \underline{B}) \tag{1.7.1}$$

where $\lambda(t, \underline{z}_i) \geq 0$ for $0 \leq t < \infty$. The survival curve for individual i is then given by

$$S(t, \underline{z}_i) = \exp\left[-\int_0^t \lambda(u, \underline{z}_i)\,du\right], \tag{1.7.2}$$

with density function

$$f(t, \underline{z}_i) = -\frac{dS(t, \underline{z}_i)}{dt}$$

in the continuous case.

In the models to be discussed in the remainder of Section 1.7, $\lambda_0(t)$ is referred to as the "underlying hazard", and for all these models, $h(\underline{z}_i, \underline{B}) = 1$ when $\underline{z}_i = \underline{0}$ so that $\lambda(t, \underline{z}_i)$ becomes $\lambda_0(t)$, the underlying hazard, whenever all covariates are zero.

1.7.2 Models Assuming a Parametric Form for $\lambda_0(t)$

The use of a survival model with a regression-type approach toward inclusion of covariates specified by a hazard of the form of equation (1.7.1) was apparently first suggested in 1965, when Feigl and Zelen [1965] proposed that the hazard for individual i with one covariate x_i could be given by $(a+bx_i)^{-1}$. This corresponds to equation (1.7.1) with $\lambda_0(t) = 1/a$ and $h(\underline{z}_i, \underline{B}) = (1+\underline{B}\,\underline{z}_i)^{-1}$ with $\underline{B} = b/a$ and $\underline{z}_i = x_i$. They suggest the model for the case when no censoring is observed, i.e., all survival times are known, and demonstrate the maximum likelihood method for estimation of the parameters and the variance - covariance matrix. An example is given concerning survival from acute myelogenous leukemia with the covariate being the log of the white blood count at time of diagnosis.

Axtell, et.al. [1972] extended this model to include three covariates in studying survival times of patients with Hodgkins' disease, in which the bx_i of Feigl and Zelen's hazard is replaced by $B_1z_{1i} + B_2z_{2i} + B_3z_{3i}$.

Zippin and Armitage [1966] extend the method by description of a method for maximum likelihood estimation of parameters in the case of censoring, illustrated with the same data used by Feigl and Zelen [1965] with some survival times randomly selected to be censored. Campbell, Reade, and Radden [1974] use the method of Zippin and Armitage to compare survival of mice subjected to several treatment methods for transplanted malignant thymoma.

Feigl and Zelen suggest two possible alternatives for the hazard: $(a + bx_i)$ and $(ae^{bx_i})^{-1}$. The first corresponds to equation

(1.7.1) with $\lambda_0(t) = a$ and $h(\underline{z}_i, \underline{B}) = (1 + \underline{B}\,\underline{z}_i)$ where $\underline{B} = b/a$ and $\underline{z}_i = x_i$. The second corresponds to equation (1.7.1) with $\lambda_0(t) = 1/a$ and $h(\underline{z}_i, \underline{B}) = e^{\underline{B}\,\underline{z}_i}$ where $\underline{B} = -b$ and $\underline{z}_i = x_i$. Byar et.al. [1974] use an extended version of the former with bx_i replaced by $\sum_{j=1}^{k} B_j x_{ji}$ in analyzing survival from prostatic cancer. Glasser [1967] uses the latter to suggest a way to study differences in survival among groups, and suggests the extension to more than one covariable wherein $\underline{B}$ and $\underline{z}_i$ become vectors, not just scalars.

Whether $h(\underline{z}_i, \underline{B})$ is taken to be $e^{\underline{B}\,\underline{z}_i}$, $(1 + \underline{B}\,\underline{z}_i)$, or $(1 + \underline{B}\,\underline{z}_i)^{-1}$, $\lambda_0(t)$ is taken to be a constant in this approach. This means that survival time is assumed exponentially distributed for all individuals, although the hazard can be different for different individuals. Feigl and Zelen [1965] give the following justification for use of the exponential distribution in studying survival from chronic diseases such as leukemia: often the diagnosis of such diseases is accidental, for example, an individual involved in an auto accident or being examined for a routine surgical procedure may be discovered to have the disease, and survival time is measured from date of diagnosis until date of death. Of course, such a situation is often not the case so some method of allowing for the hazard to change with time should be found. One way would be to allow $\lambda_0(t)$ to take on the parametric form of the hazard for some distribution other than the exponential. Prentice [1973] demonstrated this possibility using a Weibull hazard for $\lambda_0(t)$. Another way would be to allow $\lambda_0(t)$ to be more general in a non-parametric approach, and this will be discussed in the next section.

Before investigating other techniques, however, it is worthwhile to list two important advantages of these techniques, noted by Feigl and Zelen [1965]: the techniques allow for comparison of survival curves for different groups after adjustment for concomitant information thought to be important, and instead of obtaining a mean or median survival time as many earlier techniques have provided for, this method allows examination and assessment of the entire survival distribution for each individual. This is much more informative in that the life chances of an individual can be specified for any subsequent point in time.

1.7.3 Models Assuming a Non-Parametric Form for $\lambda_0(t)$

Cox [1972] suggests this form for the hazard function:

$$\lambda(t, \underline{z}_i) = \lambda_0(t)\exp(\underline{B}\,\underline{z}_i) \quad ,$$

where $\lambda_0(t)$ is arbitrary. He then states that no information about $\underline{B}$ can be obtained in time intervals during which no failures occur, because $\lambda_0(t)$ might be zero in such intervals. So, Cox constructs a likelihood conditional on the times at which failures occur, a factor at each time of death being the ratio of the hazard for the individual observed to have died divided by the sum of the hazards of those alive just before the time of death being considered (if all individuals have distinct times of death). Hence, $\lambda_0(t)$ cancels out of each factor, and is not involved in the *conditional likelihood*. If there is multiplicity at a time of death, the factor contributed to the likelihood at that time is the product of hazards for the m individuals observed to have died divided by the sum of the products of hazards for each possible combination of m individuals alive just before the time of death being considered. Again, $\lambda_0(t)$ does not appear in the "conditional" likelihood. The likelihood is thus based on the probability of observing what was actually observed at each failure time "conditional" on a risk set for that failure time.

There has been discussion of exactly what Cox's "conditional" likelihood is by Kalbfleisch and Prentice [1972] and [1973]. In the 1972 discussion, criticism is given that Cox's likelihood forms a reference set attaching positive probability to events in which the same individual fails several times. In the 1973 paper, they demonstrate that if no ties exist in times of death *and* if only the order of deaths, not the times of death are considered, *and* if no covariates are time dependent, then the marginal likelihood for $\underline{B}$ has the same form as Cox's likelihood. They note, however that the assumptions imply that lengths of time between successive failures are irrelevant to inference on $\underline{B}$, and that apparently, the likelihood given by Cox cannot be derived from a conditional argument. The marginal likelihood for the case of tied death times derived by Kalbfleisch and Prentice [1973] is somewhat different from that given by Cox for data with ties. Again, it is emphasized that their results can be obtained only if there are no time dependent covariables. Cox [1975] gives some justification for his likelihood by showing that it can be considered a partial likelihood containing virtually all information in the sample relevant to the estimation of $\underline{B}$.

At any rate, Cox's method is relatively easy to apply if there are no ties in the data, and he gives the necessary equations for maximum likelihood estimation of $\underline{B}$ by iterative techniques and also a test of the hypothesis $\underline{B} = \underline{0}$.

Kalbfleisch [1974a] studied the relative efficiency for estimation of B a scalar (i.e., only one covariable is used) of Cox's method against the cases where $\lambda_0(t)$ is a constant and a Weibull hazard assuming no censoring. At $B = 0$, the relative efficiency in both comparisons was good, being asymptotically one. The relative efficiency was found to decrease as B moved away from zero.

Efron [1977] states that Cox's method has full asymptotic efficiency in many realistic situations. He also states that if there is no censoring, and if there is only one covariable so that B is a scalar, then the Fisher information for B based on Cox's likelihood has, asymptotically, the ratio of unity with that based on all the data. However, he warns that this conclusion may be seriously distorted if a large proportion of items have the same fixed censoring time. Several authors have used Cox's method in comparison of survival data for different groups. Cangir, George, and Sullivan [1975] employed the method as a way of determining which variables were most important in survival duration. They then used a logistic regression model to order pretreatment characteristics according to their predictive value for response. Rubens, Knight, and Hayward [1975], Berry and Wagner [1976], and Enker and Jacobitz [1976] all used Cox's method to compare survival of treatment groups, with the only covariates used being those identifying which group a subject was in. Ramirez, et.al. [1975] compared two treatments for breast cancer in a similar way, but also included other covariables besides those indicating which treatment group the patient was in.

For estimation of the survival function, Cox provides for a step-function estimate by making the rather strong assumption that

$\lambda_0(t)$ is identically zero except at points where failures have occurred. As Kalbfleisch and Prentice [1973] point out, a continuous estimate of the survival function might be preferable, "especially for suggesting a parametric form for $\lambda_0(t)$ or for communicating information to non-statisticians".

1.7.4 Modifications of Cox's Method

Breslow [1972] modified Cox's model by allowing the hazard to remain constant between failure times. That is, if $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$ are the ordered failure times and $t_{(0)} = 0$, he would specify $\lambda_0(t)$, left arbitrary by Cox, as

$$\lambda_0(t) = \lambda_i \quad \text{for} \quad t_{(i-1)} < t \le t_{(i)} \ ,$$

$i = 1, 2, \ldots, n$. This provides for estimation of a continuous survival distribution.

Holford [1976] proposed a hazard of the same form,

$$\lambda(t, \underline{z}_i) = \lambda_0(t) \exp(\underline{B}\,\underline{z}_i) \ ,$$

with the period of observation broken-up into intervals $(t_j, t_{j+1}]$ and $\lambda_0(t)$ given by

$$\lambda_0(t) = \lambda_j \quad \text{for} \quad t_j < t \le t_{j+1} \ ,$$

which implies that survival times within the interval are exponentially distributed. He points out that if the intervals are fixed in advance, this method is essentially an actuarial-type analysis with covariates.

Holford also notes that the method is equivalent to Glasser's [1967] method if there is only one interval, to Breslow's [1972] method if times of death form the end points of the intervals, and that if the intervals become arbitrarily short, the likelihood function for estimation of $\underline{B}$ is similar to that of Cox [1972].

The methods of Holford [1976] and Breslow [1972] avoid problems of tied observations by making mild assumptions about $\lambda_0(t)$.

1.7.5 Advantages and Disadvantages of Current Regression-Type Methods

Regression-type methods for survival analysis can be advantageous, recalling the benefits noted by Feigl and Zelen [1965] and cited at the end of Section 1.7.2 regarding improvement of comparability of groups and the ability for characterization of the survival curve of an individual.

Further, Joysey, et.al. [1973] note that use of a regression covariate to identify which group a subject is in assesses differences in group survival over the whole time period, while actuarial methods and logistic methods usually only examine specific points in time for differences. The specific case they examined was difference in survival between recipients of kidneys from: 1) blood group 0 donors, and 2) other matched blood groups. Actuarial techniques did not indicate a significant difference in survival at any point, but use of Cox's model indicated a difference in overall survival.

There are, however, some questionable aspects of the methods as they are now developed. The models, as they now stand with $\underline{z}_i$ not time dependent, require that the hazards for individuals having different covariate values be in constant ratio over time regardless of the form of the underlying hazard $\lambda_0(t)$.

To illustrate, let us consider the case where survival of two groups is of interest, with only the covariable $z_{i1} = 0$ for those in group I and $z_{i1} = 1$ for all in group II. Assume the model

$\lambda(t, \underline{z}_i) = \lambda_0(t)e^{\underline{B}\,\underline{z}_i}$. Then regardless of $\lambda_0(t)$, the ratio of the group II hazard to the group I hazard is a constant e^B. To account for possibly differing ratio over time, Cox [1972] proposes inclusion of another covariate $z_{i2} = z_{i1}t$, where t is time. Then with $\underline{B} = (B_1, B_2)$, the ratio of the group II hazard to the group I hazard becomes $e^{B_1 + B_2 t}$, which does depend on time. This may not, however be a desirable approach since, as we have seen, the good relative efficiency of Cox's model for $\underline{B}$ estimation and perhaps even some of the justification for his conditional or partial likelihood may be predicated on none of the covariates depending on time, which, in this case, z_{i2} does.

Kalbfleisch [1974b] takes the approach of considering only covariates for which the constant hazard ratio holds as elements of $\underline{z}_i$. Other variables are called factors and divided into s strata with the hazard for the jth stratum given by

$$\lambda_j(t, \underline{z}_i) = \lambda_j(t)e^{\underline{B}\,\underline{z}_i}, \quad j = 1, 2, \ldots, s.$$

This is a return to categorization discussed in Section 1.5 and does not take full advantage of the features of a regression approach. Neither of these approaches proposed for allowing non-constancy of the ratio of hazards for subjects with different covariate values seems very satisfactory.

Another question is what form is best for incorporation of covariates. Thus far $h(\underline{z}_i, \underline{B})$ has been given as $e^{\underline{B}\,\underline{z}}$, $(1 + \underline{B}\,\underline{z})$, or $(1 + \underline{B}\,\underline{z})^{-1}$. Kalbfleisch [1974b] recommends $e^{\underline{B}\,\underline{z}}$ since $\underline{B}$ can take on any values with that function, whereas with the other two functions $\underline{B}$ must be restricted so that $\lambda(t, \underline{z}_i) \geq 0$. However, when Breslow [1974] analyzed leukemia survival data using both $e^{\underline{B}\,\underline{z}}$ and $(1 + \underline{B}\,\underline{z})^{-1}$ with $\lambda_0(t) = \lambda$, a constant, the agreement in terms of test statistics between the two models was good. This question has not received much attention.

In regard to choosing between non-parametric and parametric forms for $\lambda_0(t)$, both methods have advantages. Allowing $\lambda_0(t)$ to be quite general, as in the non-parametric approach, has the advantage of providing reasonably good relative efficiency for the estimation of $\underline{B}$ without having to make assumptions about $\lambda_0(t)$. However, the good relative efficiency of these methods is apparently dependent upon certain assumptions. Efron [1977] states that if a large proportion of items have the same fixed censoring time, the relative efficiency of Cox's method may not be so good. Unfortunately, just such a situation is very often the case in occupational cohort mortality studies. A large group of individuals is typically identified as of a point in time, followed over some period observing deaths, and at the end of the period a large portion of the cohort still alive is censored by closing the study. It is frequently impractical to follow the cohort for a period long enough to avoid this circumstance. Also, in applying analysis of survivorship to occupational cohorts, it is virtually imperative that the method used be well suited to *providing estimates of the survival function*, which aids in interpreting the results of the study. It is felt that methods not meeting this criterion will not be acceptable to non-statisticians interested in such studies. Non-parametric methods are not nearly as well suited for this as are parametric methods.

The use of a parametric model for $\lambda_0(t)$, such as the exponential, Weibull, gamma, lognormal, or any other distribution would provide for ready estimation of the survival function. There exist methods, discussed earlier in Section 1.4.3, for choosing among these models when no covariates are involved. However, with the incorporation of covariates, there is little known about choosing among the models.

Neither the non-parametric nor the parametric models for incorporation of covariates into a method for survival analysis seems especially well suited for analysis of cohort mortality data. An approach incorporating the better features of both techniques would be desirable.

1.8 Outline of Subsequent Chapters

In the following chapter, a general model for the hazard incorporating covariables is proposed. The model is parametric, allowing for easy specification of the survival curve and interpretation of results. However, the model does not unduly restrict the form of the hazard. Further, it does not require the assumption that hazards for different individuals be in constant ratio over time. Fitting the model to data, testing goodness of fit, and testing hypotheses about model parameters are also discussed. Three different cohorts are analyzed in Chapter III to provide examples of the methods proposed in Chapter II.

After a review of the appropriate literature in Chapter IV, the model is used in constructing methods for determination of appropriate sample sizes for survival studies. Examples of sample size determination are also provided.

Finally, in Chapter V, suggestions for further research on related topics are made.

CHAPTER II

A MODEL FOR THE SURVIVAL DISTRIBUTION WITH DISCUSSION OF HYPOTHESIS TESTING AND FITTING TO THE DATA

2.1 Introduction

From previous discussion of survival models with covariates, it is apparent that both parametric and non-parametric formulations for the underlying hazard have advantages. In searching for a new model, an attempt should be made to incorporate the advantages of both formulations into one insofar as it is possible to do so. A parametric model is desirable because of the ease with which the survival curve can then be specified, but in order to avoid unwarranted assumptions, it should be general enough to be considered "almost non-parametric".

In consideration of survival analysis without covariables, the generalized Rayleigh distribution might be thought of as being in this category. The hazard for this model can be expressed as

$$\lambda(t) = \lambda_0 + \lambda_1 t + \cdots + \lambda_m t^m = \sum_{k=0}^{m} \lambda_k t^k$$

where m is some integer greater than or equal to zero, and λ_0 is a constant, not to be confused with $\lambda_0(t)$, the underlying hazard discussed earlier. As Krane [1963] states, the use of this distribution is in some sense a non-parametric technique. He admits that 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. Two hazard functions which are special cases of this one are the constant hazard leading to the exponential distribution and the linear hazard discussed by Kodlin [1967], leading to the Rayleigh distribution. These two distributions are quite often reasonable to use: the exponential, for example, is suggested by Feigl and Zelen [1965] for studies of survival after diagnosis of cancer and the Rayleigh is suggested by Gross and Clark [1975] for studying the survival of middle aged groups. Further, as Clark [1975] states, the Weibull hazard was proposed because the hazard can be expressed as a polynomial in time under fairly general conditions. In other cases, as Krane [1963] indicates, many parametric forms of the hazard may be expanded in a converging power series. For example, consider the Gompertz distribution with hazard $\lambda(t) = Re^{at}$. Let $\lambda^{(n)}(0)$ be the n-th derivative of $\lambda(t)$ evaluated at $t = 0$. Here $\lambda^{(n)}(0) = Ra^n$, so a Taylor series expansion of $\lambda(t)$ is given by

$$\lambda(t) = \sum_{n=0}^{\infty} \frac{Ra^n t^n}{n!} .$$

If a is sufficiently small and t not too large, eventually higher order terms will contribute virtually nothing to the hazard, and it can be expressed as a polynomial. In fitting a polynomial hazard to data, this suggests that if higher order terms make significant additions to the hazard, a Gompertz distribution may fit the data.

In general then, several advantages of the generalized Rayleigh distribution in our situation are as follows: 1) two commonly used distributions that apply in certain circumstances are special cases; 2) other distributions can be approximated quite well, especially if the time interval is reasonably short; 3) the use of this distribution can be regarded as a non-parametric technique; and 4) at the same time, it provides the advantage of a parametric model that allows easy specification of the survival curve.

If the approaches discussed previously for incorporation of covariates were applied to this model, we would have

$$\lambda(t, \underline{z}_i) = \left[\sum_{k=0}^{m} \lambda_k t^k\right] h(\underline{z}_i, \underline{B}) \tag{2.1.1}$$

as the hazard for an individual with covariate vector $\underline{z}_i$. However, this assumes a constant ratio of hazards for individuals with different covariate values. To avoid this assumption, one might use Cox's [1972] suggestion for inclusion of a time-dependent covariate or Kalbfleisch's [1974b] suggestion of partitioning the data, but as discussed in Chapter I, these methods have their drawbacks.

To assume a constant ratio of hazards over time for two individuals with different characteristics (i.e., different covariate values) regardless of the form of the hazard function has major consequences. The implication can be understood by considering the effect of this assumption on the survival function. Let $S_i(t)$ be the survival function for two individuals, $i = 1$ and 2. Assuming that the hazard functions for the two individuals are in constant ratio over time implies that for *all* values of t such that $0 < t < \infty$, *one* of the following three relationships must hold:

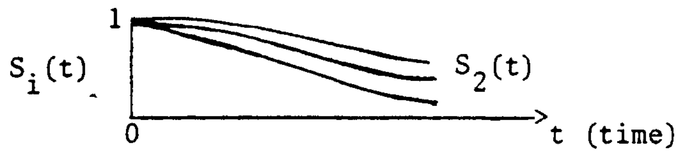
1) $S_1(t) > S_2(t)$,

2) $S_1(t) = S_2(t)$,

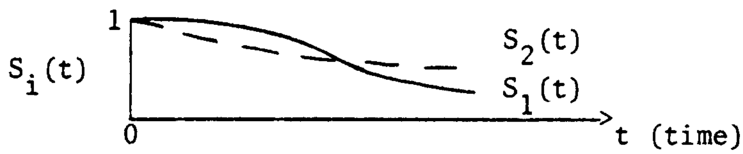
or 3) $S_1(t) < S_2(t)$.

Of course, $S_i(0) = 1$ and $S_i(\infty) = 0$ for $i = 1, 2$. That is, by assuming hazards in constant ratio, we imply that the survival curves can *never cross*. A similar result would apply if a constant difference between hazards, rather than a constant ratio, were presumed.

To illustrate, the first figure below shows the possibilities for the relationship where a constant ratio (or difference) of two hazards is assumed. If $S_2(t)$ is the middle



curve, the three curves shown are essentially the possibilities for $S_1(t)$. In particular, a relationship like the one in the next figure below is not possible. Considering that



this relationship may well occur in practice, a formulation which allows for modelling it, and which will at the same time provide for testing whether or not the assumptions discussed here are valid seems highly desirable.

2.2 Consideration of a General Model for the Hazard Function Incorporating Covariables

Suppose that instead of multiplying the underlying hazard by a constant (in time) function of the covariables, we were to multiply it by a function of the covariables *and* time wherein the covariables are still not time dependent. Then, instead of $\lambda_0(t)c(\underline{z}_i)$, we would have $\lambda_0(t)c(t, \underline{z}_i)$. Again, reasoning that functions of time can often be expressed as a polynomial, suppose

$$c(t, \underline{z}_i) = c_0(\underline{z}_i) + c_1(\underline{z}_i)t + c_2(\underline{z}_i)t^2 + \cdots$$

where each $c_j(\underline{z}_i)$, $j = 0, 1, \ldots$ is a possibly different function of $\underline{z}_i$. Now, if the underlying hazard is also a polynomial, the resultant hazard,

$$\lambda(t, \underline{z}_i) = \lambda_0(t)c(t, \underline{z}_i)$$

would also be a polynomial in time, where the coefficient of each power of t would be the product of constants from $\lambda_0(t)$ and various constant (in time) functions of $\underline{z}_i$, the $c_j(\underline{z}_i)$, $j = 0, 1, \ldots$. We can now express the hazard for an individual with covariate vector $\underline{z}_i$ as

$$\lambda(t, \underline{z}_i) = g_0(\underline{z}_i) + g_1(\underline{z}_i)t + \cdots + g_m(\underline{z}_i)t^m ,$$

where m is some non-negative integer and the $g_k(\underline{z}_i)$, $k = 0, 1, \ldots, m$ are possibly different functions of $\underline{z}_i$. However, we choose to let

$$g_k(\underline{z}_i) = \lambda_k h(\underline{z}_i, \underline{B}_k), \quad k = 0, 1, \ldots, m ,$$

in order to facilitate testing the hypothesis that the hazards for individuals with different covariate vectors are in constant ratio over time. Also, we choose to let $h(\underline{z}_i, \underline{B}_k)$ assume one of the three forms discussed in Chapter I: $e^{\underline{B}_k\underline{z}_i}$, $(1 + \underline{B}_k\underline{z}_i)$, or $(1 + \underline{B}_k\underline{z}_i)^{-1}$. Note that for $k > 0$, $e^{\underline{B}_k\underline{z}_i}$ may not be a satisfactory choice for $h(\underline{z}_i, \underline{B}_k)$ in certain cases. As an example, suppose that the hazard for an individual with covariate vector $\underline{z}_1$ is linearly increasing with time, while the hazard for another individual with covariate vector $\underline{z}_2$, $\underline{z}_1 \neq \underline{z}_2$ is linearly decreasing with time. In such a case, we would require that if $\lambda_1 > 0$, then $h(\underline{z}_1, \underline{B}_1) > 0$ and $h(\underline{z}_2, \underline{B}_1) < 0$. Clearly, in such a case, $h(\underline{z}_i, \underline{B}_1)$ could not be given by $e^{\underline{B}_1\underline{z}_i}$, since this quantity can never be negative. With this in mind, choosing one of the suggested forms for $h(\underline{z}_i, \underline{B}_k)$ is left to the investigator.

Thus, the new approach proposed here is to consider the hazard function for individuals with covariate vector $\underline{z}_i$ to be given by

$$\lambda_m(t, \underline{z}_i) = \lambda_0 h(\underline{z}_i, \underline{B}_0) + \lambda_1 h(\underline{z}_i, \underline{B}_1)t + \cdots + \lambda_m h(\underline{z}_i, \underline{B}_m)t^m$$

$$= \sum_{k=0}^{m} \lambda_k h(\underline{z}_i, \underline{B}_k)t^k , \qquad (2.2.1)$$

for m some integer greater than or equal to zero. Let $\underline{B}_k = (B_{k1}, B_{k2}, \ldots, B_{ks})$ for $k = 0, 1, \ldots, m$ be a row vector of unknown parameters, $\underline{z}_i = (z_{i1}, z_{i2}, \ldots, z_{is})'$ be the column vector of covariates for individual i, λ_k be an unknown scalar parameter, $k = 0, 1, \ldots, m$, and $h(\underline{z}_i, \underline{B}_k)$ be a function of $\underline{z}_i$ and $\underline{B}_k$ with none of these dependent on time. Then $\lambda_m(t, \underline{z}_i)dt$ may be interpreted as the probability that an individual with covariate vector $\underline{z}_i$ expires in the small time interval $(t, t + dt)$ given that he has survived up to time t.

The survival function is then given by

$$S_m(t, \underline{z}_i) = \exp\left\{-\left[\sum_{k=0}^{m} \frac{\lambda_k}{k+1} h(\underline{z}_i, \underline{B}_k)t^{k+1}\right]\right\} , \qquad (2.2.2)$$

where $S_m(t, \underline{z}_i)$ is the probability that an individual with covariate vector $\underline{z}_i$ will survive at least until time t.

The density function can be expressed as

$$f_m(t, \underline{z}_i) = \lambda_m(t, \underline{z}_i)S_m(t, \underline{z}_i) \qquad (2.2.3)$$

and $f_m(t, \underline{z}_i)dt$ can be interpreted as the probability that a person with covariate vector $\underline{z}_i$ dies in the short interval $(t, t + dt)$. Actually, it can be observed that we have a family of distributions, with the particular distribution specified by the values of m, where

m may be any non-negative integer, and of the $\underline{B}_k$, $k = 0, 1, \ldots, m$.

This model has the capability of allowing non-constancy of hazard ratio over time between individuals with different covariate vectors. This is accomplished by allowing a different parameter vector in the function $h(\cdot,\cdot)$ for different powers of t. Essentially, subject to whatever limitations the form of $h(\cdot,\cdot)$ imposes, this means that each coefficient of the powers of t in the model can be different for each set of covariate values in vector $\underline{z}_i$. That is, if we consider that $\lambda_k h(\underline{z}_i, \underline{B}_k)$, $k = 0, 1, \ldots, m$ could be regarded as the parameters of the hazard function of equation (2.2.1), then each parameter can depend in a different way upon $\underline{z}_i$. If and only if $\underline{B}_0 = \underline{B}_1 = \cdots = \underline{B}_m = \underline{B}$ will the hazards for individuals be in constant ratio over time. We note that if, indeed, $\underline{B}_0 = \underline{B}_1 = \cdots = \underline{B}_m = \underline{B}$, the hazard of equation (2.2.1) reduces to exactly that of equation (2.1.1), the formulation of earlier methods for accounting for covariables.

The approach proposed here has several advantages over previously suggested methods:

1) It allows considerable flexibility in the form of the hazard function and at the same time retains the parametric advantage of allowing easy specification of the survival curve.

2) It allows for the hazard functions of individuals with different covariate values to be in differing ratio over time, and specifically it allows for the possibility that their survival curves cross each other. This is accomplished without stratifying the population as Kalbfleisch [1974b] suggested or constructing an artificial time

dependent covariate, as Cox [1972] suggested.

3) Fitting an appropriate member of this family of distributions to the data and testing a variety of hypotheses about it is reasonably straightforward, as we shall see in the next sections.

2.3 Testing Hypotheses about the Model

In the manner of most other approaches, estimation of parameters will be done by maximum likelihood estimation. The likelihood for a cohort is given by

$$L = \prod_{i \,\varepsilon\, \text{deaths}} \lambda_m(t_i, \underline{z}_i) \prod_{\text{all } j} S_m(t_j, \underline{z}_j). \qquad (2.3.1)$$

where t_i, t_j are observed times of death for deaths and censoring time for survivors. Only if $m = 0$ and $\underline{B}_0 = \underline{0}$ can explicit maximum likelihood estimates be obtained. Otherwise, iterative methods must be used. A description of an estimation procedure is found in section 2.5.

Hypothesis testing can be done by taking into account two considerations:

1) If Ω is a parameter space and ω is a subspace of Ω to which p constraints have been applied, and if $\hat{\underline{\theta}}_\Omega$ and $\hat{\underline{\theta}}_\omega$ are maximum likelihood estimators of the parameter vector $\underline{\theta}$ in Ω and ω, respectively, then

$$-2\ell n[L(\hat{\underline{\theta}}_\omega)/L(\hat{\underline{\theta}}_\Omega)]$$

is asymptotically distributed as chi-square with p degrees of freedom, where $L(\cdot)$ is the likelihood and $\ell n(\cdot)$ denotes the natural logarithm.

2) The maximum likelihood estimate is asymptotically normally distributed with variance given by the appropriate element of the inverse of the Fisher information matrix.

We note that the two results stated above hold, for the case where there is no censoring, under certain regularity conditions (see Kendall and Stuart [1973], sections 18.26 and 24.7). It seems that these results have been assumed by several authors to hold, even if some observations are censored. Cox [1972] states that "... to be rigorous extension of the usual conditions for maximum - likelihood formulae and theory would be involved to cover censoring, but there is little doubt some such justification could be given." An indication that these two results may indeed hold true under certain regularity conditions even when censoring occurs is given in a paper by Sen [1976]. Also of concern is the fact that although the random variables t_i have distributions of the same form, the distributions are not identical because of differing values of the covariate vector, $\underline{z}_i$.

The question of whether the two above results will hold in the presence of censoring and failure of the random variables to be identically distributed, even though their distributions are of the same form, is not straightforwardly resolved. It is assumed here that the results will hold under these circumstances.

Verification that the necessary regularity conditions are met assuming no censoring and identical distribution are given in Appendix 1 for the case $m = 0$ with one covariate.

In choosing a member of the family of distributions proposed to fit the data, a step-up or step-down procedure somewhat analogous to those used by Krall, Uthoff, and Harley [1975] or Greenberg, Bayard, and Byar [1974] for selection of covariates can be used. Let $L(m=n)$ be the maximized likelihood, L, of equation (2.3.1) for $m=n$. Using a step-up procedure, find $L(m=0)$ and $L(m=1)$ and compute $-2\ln[L(m=0)/L(m=1)]$, this quantity being asymptotically χ_1^2 if the likelihoods are not significantly different. If they are significantly different, compute $L(m=2)$ and $-2\ln[L(m=1)/L(m=2)]$ and ascertain the significance. This procedure would be continued until no significant difference is found between $L(m=n)$ and $L(m=n+1)$ for $n=0, 1, \ldots$ and the model chosen would be the member of the family with $m=n$. Alternatively, or in addition, we could get maximum likelihood estimates for parameters $\lambda_0, \lambda_1, \ldots, \lambda_n$ for $m=n$, $n=0, 1, 2, \ldots$, and at each step test H_0: $\lambda_n=0$ using asymptotic normality of the estimates. When for a particular n, we cannot reject the null hypothesis, choose $m=n-1$ as the best fit. Using both methods would give added security that the best fit has been found. To provide further assurance that the best fit has been found, it would be wise to require two consecutive non-significant results by either of the above procedures, rather than just one, before deciding on an appropriate value for m. Note that the first procedure could also be used to test whether fitting r extra parameters is better. We would evaluate $-2\ln[L(m=n)/L(m=n+r)]$, asymptotically distributed χ_r^2, to determine if the added parameters improved the likelihood.

A possible drawback of the method suggested above for selection of the appropriate value of m is that the number of significance tests to be performed is not known in advance. Therefore, the overall level of significance for all tests together cannot be specified in advance. To avoid this problem, it is suggested that some positive integer, m_0 say, be selected as the highest value of m to be tested. Then, to assure that the overall significance level will be no greater than α, each hypothesis in the stepwise procedure could be tested with level of significance α/m_0. We are then left with the problem of choosing an appropriate value for m_0. In general, three will be a sufficiently large value for m_0.

After a particular appropriate value for m has been chosen, i.e., a member of the family of distributions has been selected, hypotheses of interest about the model can be tested.

To test whether there is a time - constant ratio between hazards of individuals with different covariate vectors,

$$-2\ell n\left[\frac{L(\text{with } \underline{B}_0 = \underline{B}_1 = \cdots = \underline{B}_m = \underline{B})}{L(\text{with } \underline{B}_k\text{'s not restricted to be equal})}\right]$$

could be computed and compared against a chi-square distribution with $m \times s$ degrees of freedom, where s is the number of covariates. To test other hypotheses about the parameters, it is suggested that either a likelihood ratio test or a different test assuming asymptotic normality of the maximum likelihood estimates be used.

2.4 A Test of Goodness of Fit

Once a model has been chosen, the question of its goodness of fit to the data arises. A test for goodness of fit at a particular point in time can be devised. Let $\delta_i(t)$ be a random variable such that

$$\delta_i(t) = \begin{cases} 0 & \text{if individual } i \text{ has died by time } t \\ 1 & \text{if individual } i \text{ is alive at time } t. \end{cases}$$

Let $S_m(t, \underline{z}_i)$ be the value of the survival function chosen at time t. If the fit is good, we should have $\Pr\{\delta_i(t) = 1\} = S_m(t, \underline{z}_i)$ and this can be regarded as the null hypothesis, under which $E(\delta_i(t)) = S_m(t, \underline{z}_i)$ and $\mathrm{Var}(\delta_i(t)) = S_m(t, \underline{z}_i)\ [1 - S_m(t, \underline{z}_i)]$. If the number of individuals is large enough, the quantity

$$\frac{\sum_{\text{all } i} \delta_i(t) - \sum_{\text{all } i} S_m(t, \underline{z}_i)}{\sqrt{\sum_{\text{all } i} \left\{S_m(t, \underline{z}_i)\ [1 - S_m(t, \underline{z}_i)]\right\}}} = \chi \tag{2.4.1}$$

is distributed approximately normally with mean zero and variance one under the null hypothesis. Also, then, the square of χ in equation (2.4.1) is distributed as χ^2 with one degree of freedom. However, the sum of squared χ of equation (2.4.1) for times $t_1, t_2, \ldots, t_n$ is not necessarily χ^2 with n degrees of freedom, since $\delta_i(t_j)$ is not independent of $\delta_i(t_{j'})$. Nonetheless, if separate evaluation of equation (2.4.1) at each of several values of t, say $t_1, t_2, \ldots, t_n$, does not produce any significant differences, this could be taken to imply that the model fits well. Appropriate values of the t_j, $j = 1, 2, \ldots, n$ should be well spaced throughout the period of observation. Beyond that, values of t_j to be used must be chosen by the investigator to suit the particular study at hand.

2.5 Fitting the Model to Data

As previously mentioned, unless $m = 0$ and $\underline{B}_0 = \underline{0}$, some iterative technique must be used to obtain maximum likelihood estimates. The subroutine MAXLIK by Kaplan and Elston [1972] has been used here. A requirement of this program and most other iterative procedures is that initial estimates of the parameters be supplied. For small data sets, the accuracy of these initial estimates is not of great concern. However, with large data sets, each iteration can be quite expensive. So, for these data sets, having initial estimates which are reasonably close to the maximum likelihood estimates can save considerable amounts of computer time by requiring fewer iterations of whatever routine is used. A method of obtaining good initial estimates is presented next.

We first categorize the covariate vector into n categories, letting $\underline{z}_i$ be a representative value for the vector in category i for $i = 1, 2, \ldots, n$. Then choose p values of t so that the t_j, $j = 1, 2, \ldots, p$, are somewhat evenly distributed over the time of observation for the cohort. We emphasize that here, n is the number of *categories* of the covariate vector and p is the number of specific values of t chosen.

A life table is then formed for those in the cohort with covariates in the i-th category, from which values of $\tilde{S}_L(t_j, \underline{z}_i)$ for $j = 1, 2, \ldots, p$ may be obtained. We let $\tilde{S}_L(t_j, \underline{z}_i)$ denote the empirical survival probability at time t_j for the i-th covariate category. This is done for each of the n categories of the covariate vector, so that we have values of $\tilde{S}_L(t_j, \underline{z}_i)$ for $j = 1, 2, \ldots, p$ and $i = 1, 2, \ldots, n$.

For most of our work, the model with $h(\underline{z}_i, \underline{B}_k)$ chosen to be $e^{\underline{B}_k \underline{z}_i}$ will be used, at least in the early stages of investigation. So, the survival function is given by

$$S_m(t_j, \underline{z}_i) = \exp\left[-\sum_{k=0}^{m} \frac{\lambda_k}{k+1} e^{\underline{B}_k \underline{z}_i} t_j^{k+1}\right].$$

Assuming $\underline{B}_k = \underline{B}_0$ for $k = 0, 1, \ldots, m$ and noting that

$\underline{B}_0\underline{z}_i = \sum_{\ell=1}^{s} B_{0\ell} z_{i\ell}$, the previous expression for $S_m(t_j, \underline{z}_i)$ can be shown to imply the following:

$$\ell n\left[-\ell n S_m(t_j, \underline{z}_i)\right] = \ell n\left[\sum_{k=0}^{m} \frac{\lambda_k}{k+1} t_j^{k+1}\right] + \sum_{\ell=1}^{s} B_{0\ell} z_{i\ell} .$$

This requires that $S_m(t_j, \underline{z}_i)$ can be neither zero nor one, $j = 1, 2, \ldots, p$, $i = 1, 2, \ldots, n$.

Now, taking $\tilde{S}_L(t_j, \underline{z}_i)$ to be the life table values mentioned earlier, if $\tilde{S}_L(t_j, \underline{z}_i) \neq 0$ and $\tilde{S}_L(t_j, \underline{z}_i) \neq 1$, we let

$$y_{ji} = \ell n[-\ell n \tilde{S}_L(t_j, \underline{z}_i)] . \tag{2.5.1}$$

Also, let

$$a_j = \ell n\left[\sum_{k=0}^{m} \frac{\lambda_k}{k+1} t_j^{k+1}\right] . \tag{2.5.2}$$

If our hypothesized model is reasonable, we should have

$$y_{ji} = a_j + \sum_{\ell=1}^{s} B_{0\ell} z_{i\ell} + \varepsilon_{ji} , \tag{2.5.3}$$

$$j, i \ni \tilde{S}_L(t_j, \underline{z}_i) \neq 0 \text{ or } 1$$

$$j = 1, 2, \ldots, p, \quad i = 1, 2, \ldots, n,$$

where ε_{ji} is an error term. Equation (2.5.3) can be regarded as a sort of multiple regression equation. Although the y_{ji} are not mutually independent and their distribution is not known, the usual closed-form least squares solutions provide appropriate initial estimates for the $B_{0\ell}$, $\ell = 1, 2, \ldots, s$ and estimates for the a_j, which can be used to obtain initial estimates for the λ_k, $k = 0, 1, \ldots, m$ in the manner described below.

If we let $\hat{a}_j$ be the estimate obtained as described for a_j, and let

$$Y_j = e^{\hat{a}_j}/t_j$$

and (2.5.4)

$$X_{kj} = t_j^k/(k+1) ,$$

we find that a reasonable multiple regression-type equation for obtaining initial estimates for the λ_k is given by

$$Y_j = \sum_{k=0}^{m} \lambda_k X_{kj} + \varepsilon_j , \tag{2.5.5}$$

where ε_j is an error term. Since $X_{0j} = 1$ for all j, equation (2.5.5) can be written as

$$Y_j = \lambda_0 + \lambda_1 X_{1j} + \cdots + \lambda_m X_{mj} + \varepsilon_j ,$$

$j = 1, 2, \ldots, p$ such that

$\tilde{S}_L(t_j, \underline{z}_i) \neq 0$ or 1 .

The usual closed-form least squares solution for equation (2.5.5) provides reasonable initial estimates for the λ_k, $k = 0, 1, \ldots, m$.

We note that because of the number of parameters being estimated, the following should be true: the number of time points (p) chosen should exceed the degree of the polynomial fit (m) by at least two; and the number of covariate categories (n) chosen should exceed the number of covariates (s). The time point zero should not be chosen, and values of (j, i) for which $\tilde{S}_L(t_j, \underline{z}_i)$ is zero or one cannot be used, since $\ell n(0) = \ell n[-\ell n(1)]$ is not bounded. It is felt that more refined methods of obtaining solutions to the multiple regression-type equations (2.5.3) and (2.5.5), such as weighted least squares, are not necessary since our aim is to obtain reasonable initial estimates. The distributions or other properties of these initial estimates are not of importance here. However, this method does provide reasonable initial estimates for λ_k, $k = 0, 1, \ldots, m$ and $\underline{B}_0 = (B_{01}, B_{02}, \ldots, B_{0s})$. The initial estimates for the $\underline{B}_k$, $k = 1, 2, \ldots, m$ are the same as those for $\underline{B}_0$.

Using initial estimates obtained as suggested above or by other methods in an iterative subroutine, such as MAXLIK by Kaplan and Elston [1972], will furnish maximum likelihood estimates of the model parameters, an estimated variance-covariance matrix, and the logarithm of the likelihood. These results then provide the information necessary for the hypothesis tests described in section 2.3.

Use of the method described here for obtaining initial estimates is justified because it can provide a substantial saving of computer time and money in the analysis of large cohorts.

CHAPTER III

ANALYSIS OF COHORT SURVIVAL DATA USING THE GENERAL MODEL

3.1 Introduction

In this chapter, three cohorts are studied using the methods described in the previous chapter. The first to be discussed is a clinical trial involving leukemia patients. This particular trial has been analyzed using several different methods. Results obtained by the other methods are presented for comparison with results using our method. The second cohort to be discussed involves a group of heart patients undergoing two different treatments. It is discussed to illustrate a case where the proportional hazard assumption does not hold. Finally, a cohort of industrial workers is analyzed to illustrate the use of our model in a case where the covariate is a continuous variable, rather than a dichotomous one identifying treatment groups only.

Analysis of each of the three cohorts is intended only to illustrate use of the model proposed. In addition, comparison of results with those obtained by other methods is presented for the leukemia clinical trial.

Unless otherwise specified, all tests of hypothesis are performed using the statistic

$$-2\ell n[L(\hat{\underline{\theta}}_\omega)/L(\hat{\underline{\theta}}_\Omega)] \quad ,$$

which, by our assumptions and description in Section 2.3, is approximately distributed as chi-square with the appropriate degrees of freedom under the null hypothesis.

3.2 Analysis of a Clinical Trial Involving Leukemia Patients

Clinical trial data reported by Freireich, et.al. [1963] and used by Gehan [1965] and Cox [1972] are analyzed here. The data are remission times, in weeks, of leukemia patients under two different treatments. The remission times for each of the two treatment groups are shown in Table 3.1. There are a total of 42 patients in the trial, 21 in each of the two groups.

Using the methods of Chapter II, we first choose an appropriate value for m_0, the largest value of m to be considered, where m is the degree of the polynomial chosen for the hazard of equation (2.2.1).

TABLE 3.1

Times of Remission in Weeks of Leukemia Patients
(Gehan, 1965, from Freireich, et.al., 1963)

Treatment Group	Remission Times
Drug 6-MP	6*, 6, 6, 6, 7, 9*, 10*, 10, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*
Control	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

*censored

With $h(\underline{z}_i, \underline{B}_k)$ taken to be $e^{\underline{B}_k \underline{z}_i}$ and one covariate, this hazard function is

$$\lambda_m(t, z_i) = \sum_{k=0}^{m} \lambda_k e^{B_k z_i} t^k . \qquad (3.2.1)$$

We take m_0 to be two. In the analysis, we use a single covariate with those receiving the drug 6-MP having $z_i = 0$, while those in the control group have $z_i = 1$. Fitting the various members of our family of distributions to the data, we obtain the log-likelihoods shown in Table 3.2.

Using the information in Table 3.2, we can test several hypotheses. Specifically, we wish to choose the appropriate member of the family of distributions for the data and to examine whether the hazard rates for the two treatment groups are significantly different.

TABLE 3.2

Log Likelihoods for Selected Members of the Family of Distributions for the 42 Patient Clinical Trial

Value of m	Restrictions on the B_k's	Log Likelihood (ℓnL)
2	none	$\ell nL = -106.2566$
1	none	$\ell nL = -107.1443$
0	none	$\ell nL = -108.5240$
0	$B_0 = 0$	$\ell nL = -116.7666$

The hypotheses of interest, along with appropriate test statistics and p-values are given in Table 3.3. In choosing the appropriate member of the family of distributions we choose the overall level of significance, α, to be 0.05. Thus, each of the two tests involved for $m_0 = 2$ would be carried out with level of significance $\alpha/2 = 0.025$. As discussed in Chapter II, this assures that the overall level of significance for the m_0 tests to be performed is no greater than α regardless of whether the tests are independent.

TABLE 3.3

Test Statistics and p-values for Hypotheses of Interest for the 42 Patient Clinical Trial

Null Hypothesis	Alternative Hypothesis	Chi-Square Value	Degrees of Freedom	p-value
$m = 0$	$m = 1$	2.7594	1	$p = 0.09668$
$m = 1$	$m = 2$	1.7754	1	$p = 0.18272$
for $m = 0$ $B_0 = 0$	for $m = 0$ $B_0 \neq 0$	16.4852	1	$p = 0.00005$

Examination of the first two rows of Table 3.3 reveals that we would choose $m = 0$. Hence, the hazard rate for individual i is given by

$$\lambda(t, z_i) = \lambda_0 e^{B_0 z_i} . \qquad (3.2.2)$$

Maximum likelihood estimates for the parameters of equation (3.2.2) along with estimated standard errors are given in Table 3.4.

TABLE 3.4

Estimates and Estimated Standard Errors for the Parameters of Equation (3.2.2), the Hazard Function Chosen for the 42 Patient Clinical Trial

Parameter	Maximum Likelihood Estimate of Parameter	Estimated Standard Error of Estimator
λ_0	0.02507	0.00836
B_0	1.52664	0.39841

To test whether the hazard rates are different for the two groups, we would perform the test shown in the last row of Table 3.3. We conclude that B_0 is significantly different from zero, that is, there is a significant difference between the hazard rates of the two groups.

Several other methods have been used by other investigators to test whether the hazard rates or survival curves of these two groups are significantly different. Gross and Clark [1975] (see pages 241-249) present the Cox F-test and Gehan's modified Wilcoxon test for analysis of these data. The p-values for testing differences in survival between the two groups for these two tests are, respectively, 0.00002 and 0.00031. In analyzing the same data, Cox [1972] provides two tests for the null hypothesis of equal hazard rates for the two groups against the general alternative, with resulting p-values 0.00004 and 0.00012. Our p-value, 0.00005, for the hypothesis of equal hazard rates is similar to these other values.

Since we have concluded that the hazard rates for the two treatment groups are constant in time, we also conclude that the hazard rates for the two groups are in constant ratio over time. Similarly, Cox [1972] states that "... any marked departure from the proportional hazard model is not likely to be a smooth monotonic change with t.", where t is survival time.

Using the results of our model fitting and hypothesis testing to interpret the data, we conclude that both groups have a constant hazard rate, with the hazard rate being significantly lower for the group treated with the drug 6-MP than for the control group.

It is felt that the proposed model has provided for straightforward analysis and interpretation of these data. Results were also shown to be similar to those of other methods.

3.3 Analysis of Observational Data for Heart Patients

We now analyze survival data for individuals treated with one of two treatments for coronary artery disease. Assignment of patients to a treatment was not randomized, so that factors other than treatment could be responsible for any differences observed. However, for purposes of illustration, it will be assumed that the patients were randomly assigned to treatment from a fairly homogeneous group. These data were obtained from the Coronary Artery Disease Data Bank at Duke University Medical Center through the courtesy of Dr. Kerry Lee, with whose permission they are used here.

A total of 2187 patients were catheterized and found to have at least one coronary artery which was at least seventy percent occluded. The time of catheterization was taken to be the origin for survival time in this study. Of the 2187 patients, 929 were treated surgically and 1258 were treated with medication only. Here, we compare survival over the first six months after catheterization for the surgically treated group with that for the group treated with medication only. Of the 1258 patients treated with medication only, 69 were censored in the first month due to not having yet reached the six-month follow-up anniversary, and 100 died during the six-month observation period. Of the 929 surgically treated patients, 61 were censored in the first month and 99 died during the six-month period.

We use one covariate, z_i, for each individual with $z_i = 0$ if individual i was treated with medication only and $z_i = 1$ if individual i was treated surgically. We choose m_0 to be three. Fitting the various members of our family of distributions to the data, we obtain the log likelihoods shown in Table 3.5.

TABLE 3.5

Log Likelihoods for Selected Members of the Family of Distributions for the Heart Patient Data

Value of m	Restrictions on the B_k's	Log Likelihood (ℓnL)
3	none	$\ell nL = -406.9672$
2	none	$\ell nL = -407.5407$
1	none	$\ell nL = -454.3360$
0	none	$\ell nL = -508.0759$
2	$B_0 = B_1 = B_2$	$\ell nL = -420.2572$
2	$B_0 = B_1 = B_2 = 0$	$\ell nL = -423.5308$

Using the information in Table 3.5 we can do the following: choose the appropriate member of the family of distributions for the data; test whether the proportional hazard model is suitable; and test whether the hazard rates for the two treatment groups are the same. The hypotheses of interest, along with appropriate test statistics and p-values are found in Table 3.6. In choosing the appropriate member of the family of distributions we use overall level of significance $\alpha = 0.05$. Each of the three tests involved for $m_0 = 3$ is carried out at level of significance $\alpha/3$, which is approximately 0.017. Examination of the first three rows of Table 3.6 reveals that the proper model is that with $m = 2$. From the fourth row of Table 3.6, it can be seen that the null hypothesis $B_0 = B_1 = B_2$ is rejected. Equivalently, we can say that the hazard rates for the two treatment groups are *not* in constant ratio over time, that is, the proportional hazard model is *not* appropriate. The last row of Table 3.6 reveals that the hazard rates for the two groups are significantly different.

TABLE 3.6

Test Statistics and p-values for Hypotheses of Interest for the Heart Patient Data

Null Hypothesis	Alternative Hypothesis	Chi-Square Value	Degrees of Freedom	p-value
$m = 0$	$m = 1$	107.4798	1	$p < 0.000005$
$m = 1$	$m = 2$	93.5906	1	$p < 0.000005$
$m = 2$	$m = 3$	1.1470	1	$p = 0.28418$
for $m = 2$ $B_0 = B_1 = B_2$	for $m = 2$, B_k's not all equal, $k = 0,1,2$	25.4330	2	$p < 0.000005$
for $m = 2$ $B_0 = B_1 = B_2 = 0$	for $m = 2$, B_k's not all zero, $k = 0,1,2$	31.9802	3	$p < 0.000005$

The hazard rate for individual i at time t is given by

$$\lambda(t, z_i) = \lambda_0 e^{B_0 z_i} + \lambda_1 e^{B_1 z_i} t + \lambda_2 e^{B_2 z_i} t^2 , \qquad (3.3.1)$$

where the t are in units of years. Maximum likelihood estimates for the parameters of equation (3.3.1) along with estimated standard errors are given in Table 3.7.

TABLE 3.7

Estimates and Estimated Standard Errors for the Parameters of Equation (3.3.1), the Hazard Function Chosen for the Heart Patient Data

Parameter	Maximum Likelihood Estimate of Parameter	Estimated Standard Error of Estimator
λ_0	0.42337	0.08258
λ_1	-2.10076	0.64921
λ_2	3.29571	1.14954
B_0	1.00080	0.18671
B_1	1.22914	0.27319
B_2	1.22820	0.30641

We conclude that the hazard rate functions for the two groups are significantly different. Further, the proportional hazard model does not apply. Although the hazard rate for surgical patients is initially higher, probably because of the initial surgical mortality, it decreases more rapidly with time than does the hazard rate for patients treated with medication only. By the third month, the hazard rate for surgical patients is lower than that for patients treated with medication only, by examination of the hazard function using the maximum likelihood estimates.

This analysis is intended only to illustrate the use of the general model. Differences between the two treatment groups with respect to risk factors not included in the analysis could be responsible for the differences observed in the hazard rates.

3.4 Analysis of an Occupational Cohort Study

We now examine a retrospective cohort study of industrial workers. The data analyzed here were obtained from the Occupational Health Studies Group at the School of Public Health, University of North Carolina at Chapel Hill through the courtesy of the Group's Director, Dr. Robert L. Harris, Jr., with whose permission they are used. The cohort consists of 7942 white males actively employed in an industrial plant as of January 1, 1964. This cohort was observed through December 31, 1971. During the eight years of observation, 313 individuals died. For this analysis, only active work experience was observed. That is, an individual was considered to be censored at the time of retirement or leaving the work force for other reasons. We note that the cohort observed can therefore be expected to experience considerably lower mortality than a general adult population.

Time is measured in years beginning at the time of first observation, January 1, 1964. The covariable for each individual i, z_i, is age in years as of January 1, 1964. For this study we choose m_0 to be one, because the starting time of the study is completely arbitrary and the hazard rate should be fairly stable in this healthy population. This situation is quite different from that of the heart patient study, where the cohort members were in poorer health and events strongly influencing survival, such as catheterization, new medical treatment, or surgery occurred at the beginning of the trial.

For this large cohort, we use the regression-type method described in Section 2.5 of Chapter II to obtain *initial* estimates of the model parameters. To obtain initial estimates, eight time points were chosen $(p = 8)$ as follows: $t_j = j = 1, 2, \ldots, 8$. Nine categories of the age at first observation (the covariate) were chosen $(n = 9)$ as follows: less than 25, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, and 60-64. The representative value of the covariate was chosen to be 22.5 for the first category, and the midpoint of the interval for all other categories. The initial estimates obtained are as follows: for $m = 0$, $\tilde{\lambda}_0 = 0.00021$; for $m = 1$, $\tilde{\lambda}_0 = 0.00016$, $\tilde{\lambda}_1 = 0.00002$; and, in each case, $\tilde{B}_0 = 0.06871$, with the ~ indicating that the value is an initial estimate. Comparison of the initial estimates for the chosen model with $m = 1$ with the maximum likelihood estimates given in Table 3.10 shows that, for these data, the initial estimates are quite good. In fact, comparison of the log likelihood of the initial estimates, -1815.8742, with that of the maximum likelihood estimates, -1815.7560, reveals that the initial estimates are *not* significantly different from the maximum likelihood estimates.

The log likelihoods obtained by fitting the various members of our family of distributions to the data are found in Table 3.8. Using the information in Table 3.8 we can choose the appropriate member of the family of distributions for the data.

TABLE 3.8

Log Likelihoods for Selected Members of the Family of Distributions for the Occupational Cohort

Value of m	Restrictions on the B_k's	Log Likelihood (ℓnL)
1	none	ℓnL = -1815.3719
0	none	ℓnL = -1824.2188
1	$B_0 = B_1$	ℓnL = -1815.7560

By examining the maximum likelihood estimates of the parameters for the chosen model and their estimated standard errors, we can test whether the covariate, age at first observation, has a significant effect on the hazard rate. The hypotheses of interest, along with appropriate test statistics and p-values, are found in Table 3.9. Examining the p-values in Table 3.9, we would clearly choose the model with $m = 1$ and $B_0 = B_1$. Also, we conclude that the covariate, age at first observation, has a definite effect on the hazard rate.

TABLE 3.9

Test Statistics and p-values for
Hypotheses of Interest for the Occupational Cohort

Null Hypothesis	Alternative Hypothesis	Chi-Square Value	Degrees of Freedom	p-value
$m = 0$	$m = 1$	17.6938	1	$p = 0.00003$
for $m = 1$ $B_0 = B_1$	for $m = 1$ $B_0 \neq B_1$	0.7682	1	$p = 0.38077$
for $m = 1$ $B_0 = B_1 = 0$	for $m = 1$ $B_0 = B_1 \neq 0$	104.227*	1	$p < 0.000005$

*Test Statistic is $[\hat{B}/(\hat{\mathrm{Var}}\ \hat{B})^{1/2}]^2$

The hazard rate for an individual with covariate z_i at time t is given by

$$\lambda(t, z_i) = (\lambda_0 + \lambda_1 t)e^{Bz_i}, \tag{3.4.1}$$

since we have chosen $m = 1$ and $B_0 = B_1 = B$ as the particular version of equation (3.2.1) for these data. Maximum likelihood estimates for the parameters of equation (3.4.1) along with estimated standard errors are given in Table 3.10.

TABLE 3.10

Estimates and Estimated Standard Errors
for the Parameters of Equation (3.4.1),
the Hazard Function Chosen for the Occupational Cohort

Parameter	Maximum Likelihood Estimate of Parameter	Estimated Standard Error of Estimator
λ_0	0.00016	0.00006
λ_1	0.00003	0.00001
B	0.06687	0.00655

We conclude that age at first observation has a significant effect on the hazard rate, as would be expected. Further, the hazard rate increases approximately linearly with time and the proportional hazard model applies. This is what would be expected for a reasonably healthy middle-aged population observed for a fairly short time period according to Gross and Clark [1975] (see page 152).

We now examine the question of goodness of fit using the method of Section 2.4 of Chapter II. Goodness of fit is examined for one, two, three, ..., and eight years of follow-up. Test statistics and p-values for each follow-up time $t_j = j = 1, 2, \ldots, 8$ are given in Table 3.11, using the maximum likelihood estimates of the parameters of the chosen model with $m = 1$ and $B_0 = B_1 = B$ (see Table 3.10). Also given are the observed and expected numbers of survivors for each time of follow-up. Note that $\sum_i \delta_i(t_j)$ is the observed number of survivors at time t_j and $\sum_i S_m(t_j, z_i)$ is the expected number of survivors at time t_j, $j = 1, 2, \ldots, 8$, while χ^2 is the square of χ, these quantities being as in equation (2.4.1).

TABLE 3.11

Observed and Expected Numbers of Survivors, Test Statistics and p-values for Testing Goodness of Fit for the Occupational Cohort of the Chosen Model with $m = 1$ and $B_0 = B_1$

$j = t_j$	$\sum_i \delta_i(t_j)$	$\sum_i S_m(t_j, z_i)$	χ^2	p-value
1	7902	7894.11	1.5854	0.20798
2	7856	7853.54	0.0764	0.78224
3	7820	7811.77	0.5669	0.45149
4	7775	7769.01	0.2245	0.63563
5	7726	7728.95	0.0439	0.83404
6	7692	7690.46	0.0102	0.91955
7	7660	7653.46	0.1605	0.68870
8	7629	7618.70	0.3565	0.55046

If we desire an overall level of significance no greater than α, we could evaluate the test at each time point with level of significance $\alpha/8$. Suppose we choose $\alpha = 0.05$. Then $\alpha/8 = 0.00625$. Clearly, the fit of the model chosen is adequate. For comparison, let us examine the goodness of fit of the model with $m = 0$. The maximum likelihood estimates of the parameters are $\hat{\lambda}_0 = 0.00030$ and $\hat{B}_0 = 0.06430$. Using the same procedure as before, the results are given in Table 3.12. The p-values for $t_j = 1$, 2, and 3 are all lower than 0.00625, so we would conclude that the model with $m = 0$ does *not* fit well.

Comparisons of occupational cohort survival data such as those used here are often made with an external population, such as the United States general population. Here, we examine such a comparison by obtaining estimates for the model parameters from the 1969-1971

U.S. life table for white males. These estimates are obtained by the regression-type method of Section 2.5 of Chapter II.

TABLE 3.12

Observed and Expected Numbers of Survivors, Test Statistics and p-values for Testing Goodness of Fit for the Occupational Cohort of the Model with m = 0

$j = t_j$	$\sum_i \delta_i(t_j)$	$\sum_i S_m(t_j, z_i)$	χ^2	p-value
1	7902	7875.55	12.1089	0.00050
2	7856	7824.46	9.2623	0.00234
3	7820	7779.19	11.0603	0.00088
4	7775	7738.18	7.1736	0.00740
5	7726	7703.66	2.2574	0.13298
6	7692	7673.47	1.3793	0.24022
7	7660	7646.76	0.6420	0.42299
8	7629	7623.68	0.0965	0.75607

Values of $\tilde{S}_L(t_j, z_i)$ are the life table probabilities of surviving to at least age $t_j + z_i$ given survival to age z_i. This is done for $t_j = j = 1, 2, \ldots, 8$ for $z_i = 20, 21, \ldots, 64$. Values of t_j and z_i for which $t_j + z_i$ is greater than 65 are *not* used since in the occupational cohort mandatory retirement occurs at age 65 whereupon the individual is censored in our study of mortality among active workers. Estimates obtained in this manner for the U.S. white male population, 1969-1971, for λ_0, λ_1 and $B_0 = B_1 = B$ are, respectively, 0.00010, 0.00002, and 0.09286. The log likelihood for these estimates in the occupational cohort is -1942.9013. Since the maximized log likelihood for the cohort is -1815.7560, we compute the test statistic as $-2[-1942.9013 + 1815.7560]$, or 254.2906. Under the null hypothesis that the parameter values for the occupational cohort are not significantly different from those obtained for the life table population, this test statistic is asymptotically distributed as chi-square with three degrees of freedom. We thus obtain a p-value of less than 0.000005 for this null hypothesis. Therefore, we conclude that the hazard rate in the occupational cohort is significantly different from that in the 1969-1971 U.S. white male life table population for the comparable age range. Further examination of the hazard rates reveals that the hazard rate estimated for the occupational cohort is lower than that estimated for the life table population for all points in time and for all values of the covariate, age at first observation, which are of interest ($20 \le z_i \le 64$).

One final comment is in order about the method of obtaining estimates of the model parameters for a comparison population which has its survival information in life table form. The methods of Section 2.5, Chapter II were suitable in the present situation since the proportional hazard assumption seems valid. However, this may not always be true. *If this assumption is not considered valid,* we suggest the method described below for obtaining parameter estimates from life table data.

After obtaining the values of $\tilde{S}_L(t_j, \underline{z}_i)$ as described in Section 2.5, Chapter II, choose as parameter estimates those values which minimize the following quantity:

$$\sum_i \sum_j [\tilde{S}_L(t_j, \underline{z}_i) - S_m(t_j, \underline{z}_i)]^2 ,$$

where $S_m(t_j, \underline{z}_i)$ is a function of the parameters given by equation

(2.2.2). Estimates so obtained could be considered, in some sense, least squares estimates of the model parameters. While use of this method does not require that the proportional hazard assumption hold, it does, in general, require iterative procedures for obtaining the estimates. Such procedures are not required with the other technique. If it is not known whether the proportional hazard model holds, perhaps some goodness of fit criterion could be used to determine which method produces better estimates.

A point of interest with regard to the length of follow-up for studies such as this one is the question of length of time required for changes in the hazard rate to be detectable. To examine this question, the models with $m = 0$ and $m = 1$ were fit to the data with observation times 1, 2, ..., and 8 years. Testing at the 0.05 level of significance in each case, we find that for follow-up times of three years or less, the model with $m = 0$ is chosen. For follow-up times of four or more years, the model with $m = 1$ is preferred. For data such as those used here then, it appears that a follow-up period of at least four years is required if one desires to detect changes in the hazard with time.

CHAPTER IV

DETERMINATION OF APPROPRIATE SAMPLE SIZE FOR COHORT STUDIES

4.1 Introduction

Most of the work in determining length of follow-up and sample size for cohort studies has been done in regard to controlled clinical trials. In fact, Feinstein [1975] claims the "favorite indoor sport" of statistical theoreticians is contriving new ways to gauge sample size for controlled clinical trials.

Virtually all this work has assumed a control and experimental group of *equal size* to be compared. Schlesselman [1974a] gives the well known formula for sample size of each of two groups

$$N = \frac{1}{(p_1 - p_2)^2} \left\{ z_\alpha \sqrt{2\bar{p}\,\bar{q}} + z_\beta \sqrt{p_1(1-p_1) + p_2(1-p_2)} \right\}^2 ,$$

where p_1 is probability of event given no exposure, p_2 is probability of event given exposure, $\bar{p}$ is the average of p_1 and p_2, and $\bar{q} = 1 - \bar{p}$, and α and β are probabilities of type I and II errors, respectively. In another paper, Schlesselman [1974b] gives tables of required sample size for various values of p_1, p_2/p_1, and type I and II errors.

Pasternack [1972] assumes a constant hazard for both case and control groups, and an independent constant hazard for loss to follow-up and produces tables which yield sample size for each group given five-year survival rate for control, hypothesized increase in survival rate for cases, and type I and II errors of interest.

In regard to length of follow-up, Pasternack and Gilbert [1971] make similar assumptions to those of Pasternack [1972] above, and give tables of required duration of study based on anticipated median survival time for cases and controls, total number of subjects to be entered each year, and type I and II errors desired.

Schork and Remington [1967] allow for the possibility that individuals may move back and forth between treatment and control groups during the study period. For a study L time units long, they let p_1 denote the probability of event in a single time unit given no exposure (control); $p_2 = kp_1$ denote the probability of event in a single time unit given exposure (treatment); and R_t and R_t' denote the proportions shifting from the control and treatment groups to the other group, respectively, in time unit t, $t = 1, 2, \ldots, L$. They then provide a formula for determining sample size for the desired values of L, p_1, p_2, the R_t and R_t', $t = 1, 2, \ldots, L$ and type I and II errors.

George and Desu [1974] assume a constant hazard λ_c for the control group, a possibly different hazard λ_E for the experimental (treatment) group, and a fixed duration for the trial. They then provide a formula for the approximate sample size required, given values of trial duration, λ_c, λ_E, and type I and II errors.

Halperin, Rogot, Gurian, and Ederer [1968] assume a constant hazard rate and no drop-out for the control group. For the treatment group, they assume a hazard which declines linearly over f years to a constant rate and also a constant drop-out rate. They then produce tables of required sample size for specified values of trial duration, f, type I and II errors, and the three rates mentioned above.

None of these methods is appropriate for the study of occupational cohort mortality. Often, one group is first identified and studied, only later to be compared with another.

Sample size and duration of these frequently retrospective studies have often been chosen to be the maximum number studied for the longest time that seems acceptably convenient to the investigator based on length of time needed to gather and prepare the data for analysis. Consideration is frequently given to the size and length of other studies previously published. In none of the studies of occupational cohorts reviewed in Chapter I was there any evidence that sample size and length of follow-up were chosen using criteria more objective than those just mentioned.

The method for sample size determination proposed in the following sections is suitable for occupational cohort mortality studies. It should also be useful for other cohort studies of time until a specified event, including clinical trials.

4.2 Asymptotic Distribution of the Estimators of the Model Parameters

The approach here is to use the model and the asymptotic distribution of the maximum likelihood estimators of model parameters discussed in Chapter II to obtain sample size requirements for testing specific hypotheses about the model parameters, both individually and jointly. This approach requires the following assumptions: 1) the maximum likelihood estimators of the model parameters are asymptotically unbiased and normally distributed, with the variance-covariance matrix given by the inverse of the Fisher information matrix, as discussed in Section 2.3 of Chapter II; 2) in the cohort to be studied, the covariate vector has a known distribution, with $\underline{z}_i$ denoting the covariate vector for individual i; 3) the censoring times have a known distribution, possibly conditional on the $\underline{z}_i$, with τ_i denoting the censoring time for individual i; 4) an indicator variable denoted by I_i, with $I_i = 1$ if individual i failed while under observation and $I_i = 0$ otherwise, has a known distribution conditional on the $\underline{z}_i$ and τ_i; and 5) the observation times have a known distribution conditional on the $\underline{z}_i$, τ_i, and I_i, with t_i denoting the observation time for individual i. To reflect these assumptions, we express the conditional cumulative distribution function (c.d.f.) of t_i as $F_0(t_i|\underline{z}_i, \tau_i, I_i)$; the conditional c.d.f. of I_i as $F_1(I_i|\underline{z}_i, \tau_i)$; the conditional c.d.f. of τ_i as $F_2(\tau_i|\underline{z}_i)$; and, the c.d.f. of $\underline{z}_i$ as $F_3(\underline{z}_i)$.

We will use the model of Chapter II with m chosen to be zero and one covariate, so that the hazard for an individual with covariate z_i is given by $\lambda_0 e^{B_0 z_i}$, where λ_0 and B_0 are scalar parameters. Extension of the method to the model with more than two parameters is straightforward.

The logarithm of the likelihood for the model, denoted by ℓnL, is then

$$\ell nL = \sum_{i=1}^{N} [I_i(\ell n\lambda_0 + B_0 z_i) - \lambda_0 e^{B_0 z_i} t_i] , \qquad (4.2.1)$$

where for individual i, $i = 1, 2, \ldots, N$, z_i is the covariate value, t_i is the censoring time for survivors, for whom $I_i = 0$, and t_i is the time of death for failures, for whom $I_i = 1$.

We find that

$$\frac{\partial \ell nL}{\partial \lambda_0} = \sum_{i=1}^{N} [I_i \lambda_0^{-1} - e^{B_0 z_i} t_i] ,$$

$$\frac{\partial \ell nL}{\partial B_0} = \sum_{i=1}^{N} [I_i z_i - \lambda_0 e^{B_0 z_i} t_i z_i] ,$$

$$-\frac{\partial^2 \ell nL}{\partial \lambda_0^2} = \sum_{i=1}^{N} I_i \lambda_0^{-2} ,$$

$$-\frac{\partial^2 \ell nL}{\partial \lambda_0 \partial B_0} = \sum_{i=1}^{N} e^{B_0 z_i} t_i z_i \quad , \text{ and}$$

$$-\frac{\partial^2 \ell nL}{\partial B_0^2} = \sum_{i=1}^{N} \lambda_0 e^{B_0 z_i} t_i z_i^2 .$$

So, the Fisher information matrix is given by

$$V^{-1} = E \begin{bmatrix} \sum_{i=1}^{N} I_i \lambda_0^{-2} & \sum_{i=1}^{N} e^{B_0 z_i} t_i z_i \\ \sum_{i=1}^{N} e^{B_0 z_i} t_i z_i & \sum_{i=1}^{N} \lambda_0 e^{B_0 z_i} t_i z_i^2 \end{bmatrix} , \qquad (4.2.2)$$

where the expectation is with respect to the random variables t_i, $i = 1, 2, \ldots, N$, and N is the number of individuals in the cohort, or, for our purposes, the sample size.

Now, let

$$A = \iiint I_i dF_1(I_i | z_i, \tau_i) dF_2(\tau_i | z_i) dF_3(z_i), \qquad (4.2.3)$$

$$C = \iiiint e^{B_0 z_i} t_i z_i \, dF_0(t_i|z_i,\tau_i,I_i) dF_1(I_i|z_i,\tau_i) dF_2(\tau_i|z_i) dF_3(z_i), \quad (4.2.4)$$

and

$$D = \iiiint e^{B_0 z_i} t_i z_i^2 \, dF_0(t_i|z_i,\tau_i,I_i) dF_1(I_i|z_i,\tau_i) dF_2(\tau_i|z_i) dF_3(z_i), \quad (4.2.5)$$

where the integrals are Riemann-Stieltjes integrals over the appropriate region. We may now express the Fisher information matrix as

$$V^{-1} = N \begin{bmatrix} \lambda_0^{-2} A & C \\ C & \lambda_0 D \end{bmatrix} . \quad (4.2.6)$$

Now, if we wish to test a hypothesis about λ_0 assuming B_0 fixed, the Fisher information is just $N\lambda_0^{-2}A$. Hence, if $\hat{\lambda}_0$ is the maximum likelihood estimator of λ_0, then $\hat{\lambda}_0$ is asymptotically normally distributed with mean λ_0 and variance $(N\lambda_0^{-2}A)^{-1}$, that is

$$(N\lambda_0^{-2}A)^{1/2}(\hat{\lambda}_0 - \lambda_0) \sim N(0, 1) . \quad (4.2.7)$$

Similarly, if we wish to test a hypothesis about B_0 assuming λ_0 fixed, the Fisher information is $N\lambda_0 D$. So, if $\hat{B}_0$ is the maximum likelihood estimator of B_0, we have

$$(N\lambda_0 D)^{1/2} (\hat{B}_0 - B_0) \sim N(0, 1) . \quad (4.2.8)$$

These results can be used to determine approximately the required sample size for a test of size α $(0 < \alpha < 1)$ requiring power of at least $1 - \beta$ $(0 < \beta < 1)$ against a specific alternative.

In order to show the result necessary to obtain sample sizes required for joint hypotheses about λ_0 and B_0, we require additional notation. Let $\underline{\theta} = (\lambda_0, B_0)'$, $\hat{\underline{\theta}} = (\hat{\lambda}_0, \hat{B}_0)'$, and let $\underline{\theta}_0$ and $\underline{\theta}_1$ be the null hypothesis and alternate hypothesis values of the parameter vector, respectively. Then, under our assumptions, $(\underline{\hat{\theta}} - \underline{\theta}_0)'V^{-1}(\underline{\hat{\theta}} - \underline{\theta}_0)$ is asymptotically distributed as chi-square with two degrees of freedom, with non-centrality parameter zero if $\underline{\theta} = \underline{\theta}_0$, and with non-centrality parameter $\eta = (\underline{\theta}_1 - \underline{\theta}_0)'V^{-1}(\underline{\theta}_1 - \underline{\theta}_0)$ if $\underline{\theta} = \underline{\theta}_1$ (see, e.g., Section 24.7 of Kendall and Stuart [19,3]). This information will permit us to obtain approximate sample sizes necessary for testing joint hypotheses about λ_0 and B_0, as we shall see in the next section. We note that the approximation is not good unless $(\underline{\theta}_1 - \underline{\theta}_0)$ is small. That is, the parameter values under the null and alternative hypotheses should be close to each other in order for the approximation to be good. Choice of values of $\underline{\theta}_0$ and $\underline{\theta}_1$ which are sufficiently close to each other can be made by the investigator considering the meaning of the parameters in the model under consideration.

4.3 Methods for Determining Sample Size

In this section, explicit formulas for required sample sizes are given where possible. However, for some hypotheses, it is possible only to produce a formula which can be used iteratively, or by trial and error with different values of N, to determine appropriate sample size. First, however, we must state our assumptions about the distributions of I_i, t_i, τ_i, and z_i, and use this information to evaluate A, C, and D of equations (4.2.3), (4.2.4), and (4.2.5), respectively.

4.3.1 Distributional Assumptions for z_i, τ_i, I_i, and t_i and Evaluation of the Fisher Information Matrix

If the covariables z_i and censoring times τ_i have known continuous distributions, then these can be used in equations (4.2.3), (4.2.4), and (4.2.5) for evaluation of A, C, and D.

Suppose, however, that z_i and τ_i are discrete random variables. Let z_i assume a finite number of values denoted by Z_j, $j = 1, 2, \ldots, J$. We take $\Pr(z_i = Z_j) = p_j$, $j = 1, 2, \ldots, J$ and note that the condition $\sum_{j=1}^{J} p_j = 1$ must hold. Also, let τ_i assume a finite number of values denoted by T_k, $k = 1, 2, \ldots, K$. Take $\Pr(\tau_i = T_k | z_i = Z_j) = p_{jk}$, $j = 1, 2, \ldots, J$ and $k = 1, 2, \ldots, K$, and note that we must have $\sum_{k=1}^{K} p_{jk} = 1$ for $j = 1, 2, \ldots, J$. Of course, if the z_i or τ_i are discrete but can attain a countably infinite number of values, the above would be true with J or K infinite. Also, if z_i or τ_i is a continuous variable but the true distribution is not known, the variable could be categorized into intervals (J intervals for z_i, K intervals for τ_i) and a representative value of z_i in the j-th interval taken to be Z_j, $j = 1, 2, \ldots, J$; and a representative value of τ_i in the k-th interval taken to be T_k, $k = 1, 2, \ldots, K$. Values for p_j and p_{jk} for $j = 1, 2, \ldots, J$, $k = 1, 2, \ldots, K$ would have to be chosen based on whatever information is available in this situation.

We choose to regard z_i and τ_i as discrete variables, each with a finite number of possible values for purposes of this discussion. Since the hazard is then $\lambda_0 e^{B_0 Z_j}$ for an individual with covariate Z_j, we find that

$$p_{jk1} = \Pr(I_i = 1 | Z_j, T_k) = \int_0^{T_k} \lambda_0 e^{B_0 Z_j} \exp[-\lambda_0 e^{B_0 Z_j} t] dt$$

or

$$p_{jk1} = 1 - \exp[-\lambda_0 e^{B_0 Z_j} T_k]$$

is the probability that an individual with covariate Z_j and censoring time T_k is observed to fail, that is, has $I_i = 1$. Therefore the probability that an individual with covariate Z_j and censoring time T_k is *not* observed to fail, that is, has $I_i = 0$ is given by

$$p_{jk0} = \exp[-\lambda_0 e^{B_0 Z_j} T_k] \ .$$

Now, let us examine the observation time t_i for individual i. For an individual with covariate Z_j and censoring time T_k, if $I_i = 0$, that is, the individual has not failed, we find

$$\Pr(t_i = t | Z_j, T_k, I_i = 0) = \begin{cases} 1 & \text{for} \quad t = T_k \\ 0 & \quad\quad\ \ t \neq T_k \end{cases} \tag{4.3.1}$$

If an individual with covariate Z_j and censoring time T_k has $I_i = 1$, that is, the individual has failed, then the probability density function of t_i is given by

$$f(t_i | Z_j, T_k, I_i = 1) = \begin{cases} \dfrac{\lambda_0 e^{B_0 Z_j} \exp[-\lambda_0 e^{B_0 Z_j} t_i]}{1 - \exp[-\lambda_0 e^{B_0 Z_j} T_k]} & 0 \le t_i \le T_k \\ 0 \quad \text{elsewhere} \ . \end{cases} \tag{4.3.2}$$

In the evaluation of C and D of equations (4.2.4) and (4.2.5), it is necessary to evaluate the following quantity:

$$\iint t_i dF_0(t_i|z_i,\ \tau_i,\ I_i)dF_1(I_i|z_i,\ \tau_i)\ .$$

This quantity is given by the following expression:

$$p_{jk0}T_k + p_{jk1}\int_0^{T_k} t\, f(t|Z_j, T_k, I_i = 1)dt\ .$$

Evaluating the integral in the above expression by parts, we find that

$$\iint t dF_0(t|Z_j,\ T_k,\ I_i)dF_1(I_i|Z_j,\ T_k) =$$

$$p_{jk0}T_k + p_{jk1}\left\{\frac{\lambda_0^{-1}e^{-B_0Z_j}[1 - (1+\lambda_0 e^{B_0Z_j}T_k)\exp(-\lambda_0 e^{B_0Z_j}T_k)]}{1 - \exp(-\lambda_0 e^{B_0Z_j}T_k)}\right\}.$$

Upon substitution of the values for p_{jk0} and p_{jk1}, the above expression becomes

$$\iint t dF_0(t|Z_j,\ T_k,\ I_i)dF_1(I_i|Z_j,\ T_k) =$$

$$T_k\exp(-\lambda_0 e^{B_0Z_j}T_k) + \lambda_0^{-1}e^{-B_0Z_j}[1 - (1+\lambda_0 e^{B_0Z_j}T_k)\exp(-\lambda_0 e^{B_0Z_j}T_k)]. \quad (4.3.3)$$

We can now evaluate A, C, and D of equations (4.2.3), (4.2.4), and (4.2.5), respectively. We find that

$$A = \sum_{j=1}^{J} p_j \sum_{k=1}^{K} p_{jk}[1 - \exp(-\lambda_0 e^{B_0Z_j}T_k)]\ , \quad (4.3.4)$$

$$C = \sum_{j=1}^{J} p_j Z_j e^{B_0Z_j} \sum_{k=1}^{K} p_{jk}T_k\exp(-\lambda_0 e^{B_0Z_j}T_k)$$

$$+ \lambda_0^{-1} \sum_{j=1}^{J} p_j Z_j \sum_{k=1}^{K} p_{jk}[1 - (1+\lambda_0 e^{B_0Z_j}T_k)\exp(-\lambda_0 e^{B_0Z_j}T_k)]\ , \quad (4.3.5)$$

and

$$D = \sum_{j=1}^{J} p_j Z_j^2 e^{B_0 Z_j} \sum_{k=1}^{K} p_{jk} T_k \exp(-\lambda_0 e^{B_0 Z_j} T_k)$$

$$+ \lambda_0^{-1} \sum_{j=1}^{J} p_j Z_j^2 \sum_{k=1}^{K} p_{jk} [1 - (1+\lambda_0 e^{B_0 Z_j} T_k) \exp(-\lambda_0 e^{B_0 Z_j} T_k)] . \qquad (4.3.6)$$

Using the above information, we can now produce formulas for determination of sample size. In all of the determinations, λ_0, B_0, A, C, D, V, and V^{-1} will be subscripted with a zero (0) or one (1) to denote that the value is that under the null or specific alternative hypothesis of interest, respectively.

4.3.2 Formulas for Determination of Sample Size for Tests of Hypothesis about One Parameter

Before specifically examining hypotheses about λ_0 or B_0, we will give the sample size requirement for a somewhat more general case. Suppose we have a random variable x which is distributed normally with mean μ and variance $[Nf(\mu)]^{-1}$, where $f(\cdot)$ is some unspecified non-negative valued function. Further, suppose that we wish to test the hypothesis

$$H_0: \ \mu = \mu_0 \qquad \text{vs.} \qquad H_1: \ \mu > \mu_0$$

and that we require a test of size α with power of at least $1 - \beta$ against the specific alternative $\mu = \mu_1$, $\mu_1 > \mu_0$. Now under H_0, the random variable

$$x_0 = [Nf(\mu_0)]^{1/2} (x - \mu_0) \qquad (4.3.7)$$

is normally distributed with mean zero, variance one. We now define z_γ to be such that for $0 < \gamma < 1$,

$$\Pr(z \le z_\gamma) = \gamma$$

where z is a standard normal variate. Hence, we reject H_0 if

$$x_0 > z_{1-\alpha}$$

and the power requirement is given as

$$\Pr[x_0 > z_{1-\alpha} | \mu = \mu_1] \geq 1 - \beta .$$

If we let

$$x_1 = [Nf(\mu_1)]^{1/2} (x - \mu_1) \tag{4.3.8}$$

then it can be shown that

$$x_0 > z_{1-\alpha}$$

implies that

$$x_1 > z_{1-\alpha}[f(\mu_0)]^{-1/2}[f(\mu_1)]^{1/2} + [Nf(\mu_1)]^{1/2} (\mu_0 - \mu_1) . \tag{4.3.9}$$

Since for $\mu = \mu_1$, x_1 has the standard normal distribution, the power requirement is given by requiring the right-hand side of inequality (4.3.9) to be less than z_β. It can then be shown that the power requirement is satisfied if

$$N > (\mu_1 - \mu_0)^{-2}\left\{z_{1-\alpha}[f(\mu_0)]^{-1/2} + z_{1-\beta}[f(\mu_1)]^{-1/2}\right\}^2 . \tag{4.3.10}$$

If we wish to test a hypothesis similar to the one above except that we have H_1: $\mu < \mu_0$ with specific alternative $\mu = \mu_1$, $\mu_1 < \mu_0$, the sample size requirement is still given by inequality (4.3.10).

Now, suppose that we wish to test the hypothesis

$$H_0: \quad \mu = \mu_0 \qquad \text{vs.} \qquad H_1: \quad \mu \neq \mu_0$$

and that we require a test of size α with power of at least $1 - \beta$ against the specific alternative $\mu = \mu_1$. With x_0 as given in equation (4.3.7), we reject H_0 if

$$x_0 < z_{\alpha/2} \qquad \text{or} \qquad x_0 > z_{1-\alpha/2} .$$

The power requirement is then

$$\Pr[x_0 < z_{\alpha/2} | \mu = \mu_1] + \Pr[x_0 > z_{1-\alpha/2} | \mu = \mu_1] \geq 1 - \beta \ .$$

If we let

$$x_L = z_{\alpha/2}[f(\mu_1)]^{1/2}\,[f(\mu_0)]^{-1/2} + [Nf(\mu_1)]^{1/2}\,(\mu_0 - \mu_1) \ , \tag{4.3.11}$$

and

$$x_u = z_{1-\alpha/2}[f(\mu_1)]^{1/2}\,[f(\mu_0)]^{-1/2} + [Nf(\mu_1)]^{1/2}\,(\mu_0 - \mu_1) \ , \tag{4.3.12}$$

then the power requirement can be shown to be that N must be of sufficient size so that

$$\Pr[z < x_L] + \Pr[z > x_u] \geq 1 - \beta \tag{4.3.13}$$

where x_L is given by equation (4.3.11), x_u by (4.3.12), and z is a standard normal variate.

Note that an explicit solution of expression (4.3.13) for N is not given. The proper value of N is the smallest positive integer resulting in values of x_L and x_u which satisfy (4.3.13), and this value must be found by trial and error in the general case. However, we will frequently find in practice that either $\Pr[z < x_L]$ or $\Pr[z > x_u]$ will be virtually zero, as we shall see in Section 4.4, in which case an explicit solution for N can be found.

Using the above results, we can now easily state sample size requirements for hypotheses about λ_0 assuming B_0 fixed and about B_0 assuming λ_0 fixed. All hypothesis tests discussed will be presumed to be of size α requiring power of at least $1 - \beta$ against the specific alternative of interest.

Note that for B_0 fixed, under our assumptions $\hat{\lambda}_0$, the maximum likelihood estimator of λ_0, is asymptotically normally distributed with mean λ_0 and variance $[N\lambda_0^{-2}A]^{-1}$. Hence, in the more general development of sample size, λ_0 assumes the role of μ and $\lambda_0^{-2}A$ assumes the role of $f(\mu)$. So, for a test of the hypothesis

$$H_0: \lambda_0 = \lambda_{00} \quad \text{vs.} \quad H_1: \lambda_0 > \lambda_{00}$$

with specific alternative of interest $\lambda_0 = \lambda_{01}$, $\lambda_{01} > \lambda_{00}$, the sample size requirement is given by

$$N > (\lambda_{01} - \lambda_{00})^{-2}[z_{1-\alpha}(\lambda_{00}^{-2}A_0)^{-1/2} + z_{1-\beta}(\lambda_{01}^{-2}A_1)^{-1/2}]^2 . \tag{4.3.14}$$

If we desire to have H_1: $\lambda_0 < \lambda_{00}$ with specific alternative of interest $\lambda_0 = \lambda_{01}$, $\lambda_{01} < \lambda_{00}$, the sample size requirement is still given by (4.3.14).

Suppose now that we want to test the hypothesis

$$H_0: \lambda_0 = \lambda_{00} \quad \text{vs.} \quad H_1: \lambda_0 \neq \lambda_{00}$$

with specific alternative of interest $\lambda_0 = \lambda_{01}$. We let

$$x_L = z_{\alpha/2}(\lambda_{00}^{-2}A_0)^{-1/2}(\lambda_{01}^{-2}A_1)^{1/2} + [N\lambda_{01}^{-2}A_1]^{1/2}(\lambda_{00} - \lambda_{01}), \tag{4.3.15}$$

and

$$x_u = z_{1-\alpha/2}(\lambda_{00}^{-2}A_0)^{-1/2}(\lambda_{01}^{-2}A_1)^{1/2} + [N\lambda_{01}^{-2}A_1]^{1/2}(\lambda_{00} - \lambda_{01}). \tag{4.3.16}$$

The sample size requirement is that N be the smallest positive integer such that

$$\Pr[z < x_L] + \Pr[z > x_u] \geq 1 - \beta . \tag{4.3.17}$$

Now, for λ_0 fixed, under our assumptions $\hat{B}_0$, the maximum likelihood estimator of B_0, is asymptotically normally distributed with mean B_0 and variance $[N\lambda_0 D]^{-1}$. Hence, in the more general development of sample size, B_0 assumes the role of μ, and $\lambda_0 D$ assumes the role of $f(\mu)$. Note that D can be regarded as a function of B_0, and, hence, so can $\lambda_0 D$. So, for a test of the hypothesis

$$H_0\colon\ B_0 = B_{00} \quad \text{vs.} \quad H_1\colon\ B_0 > B_{00}$$

with specific alternative of interest $B_0 = B_{01}$, $B_{01} > B_{00}$, the sample size requirement is given by

$$N > (B_{01} - B_{00})^{-2}[z_{1-\alpha}(\lambda_0 D_0)^{-1/2} + z_{1-\beta}(\lambda_0 D_1)^{-1/2}]^2 \ . \tag{4.3.18}$$

If we desire to have $H_1\colon\ B_0 < B_{00}$ with specific alternative of interest $B_0 = B_{01}$, $B_{01} < B_{00}$, the sample size requirement is still given by (4.3.18).

Suppose now that we want to test the hypothesis

$$H_0\colon\ B_0 = B_{00} \quad \text{vs.} \quad H_1\colon\ B_0 \neq B_{00}$$

with specific alternative of interest $B_0 = B_{01}$. We let

$$x_L = z_{\alpha/2}(\lambda_0 D_0)^{-1/2}(\lambda_0 D_1)^{1/2} + [N\lambda_0 D_1]^{1/2}(B_{00} - B_{01}) \ , \tag{4.3.19}$$

and

$$x_u = z_{1-\alpha/2}(\lambda_0 D_0)^{-1/2}(\lambda_0 D_1)^{1/2} + [N\lambda_0 D_1]^{1/2}(B_{00} - B_{01}) . \tag{4.3.20}$$

The sample size requirement is that N be the smallest positive integer such that

$$\Pr[z < x_L] + \Pr[z > x_u] \geq 1 - \beta \ . \tag{4.3.21}$$

4.3.3 Formulas for Determination of Sample Size for Tests of Hypothesis about Both Parameters

We can use the information about the joint asymptotic distribution of the maximum likelihood estimators of λ_0 and B_0 mentioned at the end of Section 4.2 to obtain a useful procedure for finding required sample size for a test of size α of the hypothesis:

$$H_0: \ (\lambda_0, B_0) = (\lambda_{00}, B_{00}) \quad \text{vs.} \quad H_1: \ (\lambda_0, B_0) \neq (\lambda_{00}, B_{00}),$$

requiring power of at least $1-\beta$ against the specific alternative hypothesis $(\lambda_0, B_0) = (\lambda_{01}, B_{01})$.

Under the null hypothesis the quantity

$$x_0 = N(\hat{\lambda}_0 - \lambda_{00}, \ \hat{B}_0 - B_{00}) \begin{bmatrix} \lambda_{00}^{-2}A_0 & C_0 \\ C_0 & \lambda_{00}D_0 \end{bmatrix} \begin{pmatrix} \hat{\lambda}_0 - \lambda_{00} \\ \hat{B}_0 - B_{00} \end{pmatrix} \tag{4.3.22}$$

is asymptotically distributed as chi-square with two degrees of freedom. If a random variable x is distributed as chi-square with ν degrees of freedom and non-centrality parameter η, let $\chi^2_{\nu,\gamma,\eta}$ be such that

$$\Pr[x \le \chi^2_{\nu,\gamma,\eta}] = \gamma \ .$$

If $\eta = 0$, it is dropped as a subscript. We reject H_0 if $x_0 > \chi^2_{2,\ 1-\alpha}$. So, the power requirement can be stated as

$$\Pr[x_0 > \chi^2_{2,\ 1-\alpha} \mid (\lambda_0, B_0) = (\lambda_{01}, B_{01})] \ge 1 - \beta \ . \tag{4.3.23}$$

Simplifying the expression (4.3.22) for x_0, we find

$$x_0 = N[(\hat{\lambda}_0 - \lambda_{00})^2\lambda_{00}^{-2}A_0 + 2(\hat{\lambda}_0 - \lambda_{00})(\hat{B}_0 - B_{00})C_0 + (\hat{B}_0 - B_{00})^2\lambda_{00}D_0] \ . \tag{4.3.24}$$

The distribution of x_0 with $(\lambda_0, B_0) = (\lambda_{01}, B_{01})$ is not known. So, in order to find required sample size, we must be able to express the probability statement (4.3.23) as an equivalent probability statement about a random variable with known distribution.

For $(\lambda_0, B_0) = (\lambda_{01}, B_{01})$, the random variable

$$x_1 = N[(\hat{\lambda}_0 - \lambda_{00})^2 \lambda_{01}^{-2} A_1 + 2(\hat{\lambda}_0 - \lambda_{00})(\hat{B}_0 - B_{00}) C_1 + (\hat{B}_0 - B_{00})^2 \lambda_{01} D_1] , \tag{4.3.25}$$

is asymptotically distributed chi-square with two degrees of freedom and non-centrality parameter

$$\eta = N[(\lambda_{01} - \lambda_{00})^2 \lambda_{01}^{-2} A_1 + 2(\lambda_{01} - \lambda_{00})(B_{01} - B_{00}) C_1 + (B_{01} - B_{00})^2 \lambda_{01} D_1] . \tag{4.3.26}$$

So, if we can obtain a probability statement about x_1 which is equivalent to expression (4.3.23), we can obtain the sample size requirement.

We now attempt to find a relationship between x_0 and x_1 of equations (4.3.24) and (4.3.25), respectively. We let

$$r_1 = \frac{\lambda_{01}^{-2} A_1}{\lambda_{00}^{-2} A_0} , \quad r_2 = \frac{C_1}{C_0} , \quad \text{and} \quad r_3 = \frac{\lambda_{01} D_1}{\lambda_{00} D_0} . \tag{4.3.27}$$

If $r_1 = r_2 = r_3 = r$, then we have $x_1 = r x_0$ and the power requirement of expression (4.3.23) is equivalent to the following:

$$\Pr[x_1 > r\chi^2_{2,\ 1-\alpha}] \ge 1 - \beta, \quad \text{or}$$

$$\Pr[x_1 \le r\chi^2_{2,\ 1-\alpha}] \le \beta. \tag{4.3.28}$$

Since, under the specific alternative of interest, x_1 is asymptotically distributed chi-square with two degrees of freedom and non-centrality parameter η, we must have

$$r\chi^2_{2,\ 1-\alpha} < \chi^2_{2,\beta,\eta} , \tag{4.3.29}$$

where η is given by expression (4.3.26). So, the sample size required is the smallest positive integer N which results in the non-centrality parameter, η, being such that inequality (4.3.29) is satisfied.

However, the assumption that r_1, r_2, and r_3 of equations (4.3.27) are equal is not, in general, true. If they are not equal, a result known about quadratic forms (see, e.g., Rao [1965], pages 37-38) can be used to determine an adequate sample size requirement. With x_0 and x_1 as given by equations (4.3.24) and (4.3.25), respectively, we have that stochastically

$$\rho_2 x_1 \le x_0 \le \rho_1 x_1 \quad , \tag{4.3.30}$$

where ρ_1 and ρ_2 are, respectively, the largest and smallest roots of the matrix $V_0^{-1}V_1$, where

$$V_0^{-1}V_1 = \begin{bmatrix} \lambda_{00}^{-2}A_0 & C_0 \\ C_0 & \lambda_{00}D_0 \end{bmatrix} \begin{bmatrix} \lambda_{01}^{-2}A_1 & C_1 \\ C_1 & \lambda_{01}D_1 \end{bmatrix}^{-1} . \tag{4.3.31}$$

The roots of $V_0^{-1}V_1$ of equation (4.3.31) are

$$\begin{aligned} \rho = {} & [2(\lambda_{01}^{-1}A_1D_1 - C_1^2)]^{-1}\Big\{(\lambda_{00}^{-2}A_0\lambda_{01}D_1 + \lambda_{01}^{-2}A_1\lambda_{00}D_0 - 2C_0C_1) \\ & \pm [(\lambda_{00}^{-2}A_0\lambda_{01}D_1 - \lambda_{01}^{-2}A_1\lambda_{00}D_0)^2 + 4(\lambda_{01}^{-2}A_1\lambda_{01}D_1C_0^2 \\ & + \lambda_{00}^{-2}A_0\lambda_{00}D_0C_1^2 - \lambda_{00}^{-2}A_0\lambda_{01}D_1C_0C_1 - \lambda_{01}^{-2}A_1\lambda_{00}D_0C_0C_1)]^{1/2}\Big\} . \end{aligned} \tag{4.3.32}$$

Note that ρ_1, the largest root, is found by using the plus (+) sign from the plus or minus ($\pm$) sign in equation (4.3.32), while ρ_2, the smallest root, is found by using the minus (-) sign.

Now, using expression (4.3.30), we find

$$\Pr[x_0 > \chi^2_{2,\ 1-\alpha}] \ge \Pr[\rho_2 x_1 > \chi^2_{2,\ 1-\alpha}] . \tag{4.3.33}$$

So, the power requirement is satisfied if for $\rho_2 > 0$,

$$\Pr[x_1 > \rho_2^{-1}\chi^2_{2,\ 1-\alpha}] \geq 1 - \beta \ . \tag{4.3.34}$$

Inequality (4.3.34) implies that for $\rho_2 > 0$,

$$\rho_2^{-1}\chi^2_{2,\ 1-\alpha} < \chi^2_{2,\beta,\eta} \ , \tag{4.3.35}$$

where η is given by equation (4.3.26). So, the sample size required is the smallest positive integer N which results in the non-centrality parameter, η, being such that inequality (4.3.35) is satisfied. It can be seen from examination of inequalities (4.3.33) and (4.3.34) that this procedure is conservative. That is, the sample size obtained may result in a power greater than $1 - \beta$. However, if r_1, r_2, and r_3 of equations (4.3.27) are all equal to r, then equation (4.3.32) has only one solution, $\rho_1 = \rho_2 = r^{-1}$. In this case, inequality (4.3.35) is equivalent to (4.3.29), the result already derived for the special case $r_1 = r_2 = r_3 = r$.

4.4 Examples of Sample Size Determination

In this section we will use the results of the previous section to determine sample size requirements for hypotheses which would be of interest in studying a cohort. The parameter values and distribution of a single covariable will be chosen to be similar to those which would arise from an occupational cohort like the one studied in Chapter III.

In all tests, α will be taken to be 0.05 and β will be 0.10. The notation used will be that of Section 4.3. The covariable is age, and, although continuous, it will be grouped into intervals and considered discrete. The distribution of the covariable is given in Table 4.1. We will have a censoring time of one year for all individuals, so that if $T_1 = 1$, we have $p_{j1} = 1$ for $j = 1, 2, 3, 4, 5$ and $p_{jk} = 0$ for $k \neq 1$, $j = 1, 2, 3, 4, 5$. Values of λ_0 and B_0 of interest as null or alternative hypotheses will be $\lambda_0 = 0.0002$ or $\lambda_0 = 0.0003$ and $B_0 = 0.07$ or $B_0 = 0.09$.

TABLE 4.1

Distribution of the Covariable
Age, Grouped with J = 5

Range of z_i	j	Value of Z_j	p_j
15 - 24	1	20	0.1852
25 - 34	2	30	0.1759
35 - 44	3	40	0.2222
45 - 54	4	50	0.2315
55 - 64	5	60	0.1852

In order that the differences in survival probabilities arising from changes in λ_0 and B_0 can be seen, one year survival probabilities for each of the five values of Z_j in Table 4.1 are given in Table 4.2 for each of the four possible combinations of λ_0 and B_0 to be considered.

We now have all the information necessary to calculate A, C, and D of equations (4.3.4), (4.3.5), and (4.3.6), respectively, for various combinations of values of λ_0 and B_0 by substitution of the quantities given above into the formulas. These calculations have been done and the results are found in Table 4.3. Using this information, we will now give sample size requirements for tests of onesided and two-sided alternatives for λ_0 with B_0 fixed and for B_0 with λ_0 fixed.

TABLE 4.2

One Year Survival Probabilities for Five Values of Z_j and Four Values of the Parameter Vector (λ_0, B_0)

Value of Z_j	Value of (λ_0, B_0)			
	(0.0002, 0.07)	(0.0002, 0.09)	(0.0003, 0.07)	(0.0003, 0.09)
20	0.9992	0.9988	0.9988	0.9982
30	0.9984	0.9970	0.9976	0.9955
40	0.9967	0.9927	0.9951	0.9891
50	0.9934	0.9822	0.9901	0.9734
60	0.9868	0.9567	0.9802	0.9357

Also, the required sample size for a test of a joint hypothesis will be given.

TABLE 4.3

Values of A, C, and D for Two Values Each of λ_0 and B_0 with Z_j Distributed as in Table 4.1 and Censoring Time One Year for All Individuals

λ_0	B_0	A	C	D
0.0002	0.07	0.00514863	1322.1360	70697.005
0.0002	0.09	0.01451959	3874.1116	211790.62
0.0003	0.07	0.00770583	1318.9203	70513.429
0.0003	0.09	0.02161528	3832.5837	209958.99

Suppose that the hypothesis of interest is

$$H_0\colon\ \lambda_0 = 0.0002 \quad \text{vs} \quad H_1\colon\ \lambda_0 > 0.0002$$

with $\lambda_{01} = 0.0003$ the specific alternative and B_0 fixed at 0.07. The first and third rows of Table 4.3 provide the necessary information under the null and alternative hypotheses, respectively. Using expression (4.3.14), we find

$$N > (0.0001)^{-2}[1.645(0.00278730) + 1.282(0.00341752)]^2$$

or

$$N > 8039.57\ .$$

So the sample size required would be

$$N \geq 8040\ .$$

Now, suppose that the hypothesis we wish to test is identical to the one above, except that we change H_1 to be

$$H_1\colon\ \lambda_0 \neq \lambda_{00}\ .$$

We use expression (4.3.17) with x_L and x_u as given in (4.3.15) and (4.3.16), respectively. We find

$$x_L = -1.96(0.00278730)(0.00341752)^{-1} + N^{1/2}(0.00341752)^{-1}(-0.0001)$$

or

$$x_L = -1.599 - N^{1/2}(0.02926)$$

and, so

$$x_u = 1.599 - N^{1/2}(0.02926).$$

Notice that for large N, $\Pr[z < x_L]$ will be very small. So let the first value of N we examine be such that

$$\Pr[z > x_u] \geq 1 - \beta\ ,$$

which implies that

$$x_u < z_\beta \ ,$$

or

$$N > [1.599 + 1.282]^2 (0.02926)^{-2}$$

which becomes

$$N > 9694.77$$

or

$$N \geq 9695 .$$

To assure that this N is not larger than necessary, we examine the value of expression (4.3.17) for $N = 9695$. For this value of N, x_L is -4.4800 and x_u is -1.2820, so

$$0 < \Pr[z < -4.4800] < 0.001$$

and

$$\Pr[z > -1.2820] = 0.90 \ ,$$

so that the power for $N = 9695$ is just slightly over the required 0.90, and $N \geq 9695$ is an adequate sample size requirement.

To ascertain sample size requirement for a test of

$$H_0: \ B_0 = 0.07 \quad \text{vs.} \quad H_1: \ B_0 > 0.07$$

with B_{01} the specific alternative of interest being 0.09 and λ_0 fixed at 0.0002, we would use inequality (4.3.18). The first and second rows of Table 4.3 provide the necessary information and we find

$$N > (0.02)^{-2} [1.645(0.26594051) + 1.282(0.15364967)]^2$$

or

$$N > 1006.32,$$

so that the sample size is required to be

$$N \geq 1007 .$$

Now, suppose the hypothesis we wish to test is identical to the one above, except that we change H_1 to be H_1: $B_0 \neq B_{00}$. We use expression (4.3.21) with x_L and x_u as given in equations (4.3.19) and (4.3.20), respectively. We find

$$x_L = -1.96(0.26594051)(0.15364967)^{-1} + N^{1/2}(0.15364967)^{-1}(-0.02)$$

or

$$x_L = -3.3924 - 0.13017N^{1/2}$$

and

$$x_u = 3.3924 - 0.13017N^{1/2} \quad .$$

Temporarily assume $\Pr[z < x_L] < 0.001$. Then we must have $x_u < z_\beta$ or

$$N > (3.3924 + 1.282)^2(0.13017)^{-2},$$

$$N > 1289.53.$$

Let us now evaluate expression (4.3.21) with $N = 1290$. We find that $x_L = -8.068$ and $x_u = -1.283$, and so

$$0 < \Pr[z < -8.068] < 0.001$$

and

$$\Pr[z > -1.283] = 0.90,$$

so that for $N = 1290$, the power is just slightly over the required 0.90 and $N \geq 1290$ is an adequate sample size requirement.

Finally, we use the method of Section 4.3.3 to determine required sample size for the joint hypothesis

$$H_0: \ (\lambda_0, B_0) = (0.0002, 0.07) \quad \text{vs.} \quad H_1: \ (\lambda_0, B_0) \neq (0.0002, 0.07),$$

with the specific alternative of interest being $(\lambda_0, B_0) = (0.0003, 0.09)$.

The first and fourth rows of Table 4.3 provide the necessary information under the null and alternative hypotheses, respectively. First, let us evaluate r_1, r_2, and r_3 of equations (4.3.27),

$$r_1 = \frac{\lambda_{01}^{-2} A_1}{\lambda_{00}^{-2} A_0} = \frac{(0.0003)^{-2}(0.02161528)}{(0.0002)^{-2}(0.00514863)} = \frac{240169.78}{128715.75} = 1.8658927 ,$$

$$r_2 = \frac{C_1}{C_0} = \frac{3832.5837}{1322.1360} = 2.8987817 , \qquad \text{and}$$

$$r_3 = \frac{\lambda_{01} D_1}{\lambda_{00} D_0} = \frac{62.987697}{14.139401} = 4.4547642 .$$

Clearly, the three values are not equal, although they are of the same order of magnitude. We first evaluate ρ_2, the lesser value of equation (4.3.32), and find $\rho_2 = 0.053452$. We find the proper value of η given by equation (4.3.26) by iteration using inequality (4.3.35). We note that if x is distributed as chi-square with two degrees of freedom and non-centrality parameter η, then (4.3.35) is equivalent to requiring

$$\Pr[x < \rho_2^{-1} \chi^2_{2,\ 1-\alpha}] < \beta . \tag{4.4.1}$$

A computerized version of the chi-square distribution may be used iteratively to find the smallest η satisfying expression (4.4.1). Since $\alpha = 0.05$, $\chi^2_{2,\ 1-\alpha} = 5.991$ and $\rho_2^{-1} \chi^2_{2,\ 1-\alpha} = 112.08$. So, we evaluate $\Pr[x < 112.08]$ for various values of η. We find that for $\eta = 140$, $\Pr[x < 112.08] = 0.0985$ and for $\eta = 139$, $\Pr[x < 112.08] = 0.1061$, so we know $139 < \eta < 140$. Further iteration shows that for $\eta = 139.798$, $\Pr[x < 112.08] = 0.1000$. We may now evaluate equation (4.3.26) with $\eta = 139.798$ and solve for N. We find $\eta = N(0.04292711)$, so that $N = 3256.64$. So, we would choose $N = 3257$ as an adequate sample size.

4.5 Discussion of the Method for Determination of Sample Size

The key assumption underlying this method of sample size determination is that the maximum likelihood estimators of the parameters of the survival time distribution are asymptotically unbiased and normally distributed with the variance-covariance matrix given by the inverse of the Fisher information matrix. Using four distribution functions and expressing the Fisher information matrix as a multiple of cohort size, N, the required sample size for a given test size and power requirement can be determined for the types of hypotheses discussed here. The four distribution functions, which can be chosen by the investigator to reflect properties of the cohort being studied, are as follows: 1) the distribution of observation times, conditional on whether the observation was an event of interest or censored, censoring time and the covariate(s); 2) the distribution of the variable indicating whether the observation was an event or censored, conditional on censoring time and the covariate(s); 3) the distribution of censoring time conditional on the covariate(s); and 4) the distribution of the covariate(s).

Provision for the distribution of censoring times to be conditional on the covariates is made because censoring time may be quite different for different covariate values in some cases. For instance, suppose that the covariate indicates whether the subject in a clinical trial is in the experimental or control group. If the experimental treatment is much more rigorous than the control treatment, one might reasonably expect considerably earlier censoring times in the experimental group. However, it is not required that censoring time depend on covariate values, as was demonstrated in the example in Section 4.4.

The method does not depend on choosing the particular form used here for the four distributions; it can be used with whatever forms of the distributions the investigator feels are appropriate. We emphasize that the number of parameters to be estimated for the survival time distribution is not restricted. Two parameters were used in this case, but the method can be extended to any number. However, it must be emphasized that the distributions chosen should reflect the conditions in the cohort to be studied insofar as possible. If the actual distributions in the cohort differ greatly from those assumed, then it is possible that the sample size chosen will not produce the desired power for the hypothesis test of interest.

We note that the method developed here can be used to select required follow-up time for a given sample size. This can be done by fixing N and allowing the distribution of censoring times to vary upward until the power requirement is satisfied. Of course, if N is too small, then no censoring time, even infinity, will be adequate.

Finally, since our results depend on the distribution of maximum likelihood estimates being known asymptotically, the question of when N is big enough so that asymptotic results can be appropriately used must be considered. If N is less than 30 or 40, then there are serious doubts about the appropriateness of the method. Since censoring may occur before all survival times are observed, the expected number of uncensored observations for a given choice of N and the four distributions required should be considered. Depending on the distributions chosen, even a very large N may be expected to produce only a very few uncensored observations. This is especially true if the hazard is small and the censoring times are short. To understand the importance of the expected number of uncensored observations, note that for B_0 fixed, $\mathrm{Var}(\hat{\lambda}_0) = [N\lambda_0^{-2}A]^{-1}$, which is just the inverse of λ_0^{-2} times the expected number of uncensored observations. So, even with large N, $\mathrm{Var}(\hat{\lambda}_0)$ may be quite large if A is small. For instance, if $\lambda_0 = 0.0002$ as in the example for sample size determination, and we have B_0 fixed at 0.07, the expected number of uncensored observations in one year with a sample size of 9695 is $9695(0.00514863) = 49.92$, and the standard deviation of $\hat{\lambda}_0$ is 0.00003. Based on experience and intuition, it is felt that the number of uncensored observations expected should be at least 10 to 20.

CHAPTER V

SUGGESTIONS FOR FUTURE RESEARCH

5.1 Forms for the Hazard Function

The approach in this dissertation has been to model the hazard function for an individual with covariate vector $\underline{z}_i$ as one of a family of functions given by equation (2.2.1), repeated below:

$$\lambda_m(t, \underline{z}_i) = \sum_{k=0}^{m} \lambda_k h(\underline{z}_i, \underline{B}_k) t^k . \tag{5.1.1}$$

Three forms for $h(\underline{z}, \underline{B})$ have been given here, as follows: $e^{\underline{B}\,\underline{z}}$, $(1 + \underline{B}\,\underline{z})$, and $(1 + \underline{B}\,\underline{z})^{-1}$. Each of these three forms, with one covariable, is presented by Feigl and Zelen [1965], using somewhat different notation (see Section 1.7.2, Chapter I).

Breslow [1974] used both $e^{\underline{B}\,\underline{z}}$ and $(1 + \underline{B}\,\underline{z})^{-1}$ in analyzing the same data. He found good agreement of test statistics from the two models. Apparently little other work has been done in choosing a form for $h(\underline{z}, \underline{B})$. So, it is felt that methods for choosing among these and possibly other forms for $h(\underline{z}, \underline{B})$ should be developed. A possible criterion for selection is comparison of goodness of fit.

Development of other models for the hazard function incorporating covariables would be useful. For example, the Burr distribution with cumulative distribution function

$$F(t) = 1 - (1 + t^c)^{-k}, \quad t \geq 0; \quad c, k > 0 \tag{5.1.2}$$

might be adapted for use with covariates. This function is quite general in that different values of c and k result in a wide range of distributions.

Further comparisons of results obtained by using the hazard function (5.1.1) with those obtained by other methods are of interest.

5.2 Incorporation of Events Important to Survival

A study of industrial workers by Andjelkovic, Taulbee, and Symons [1976] indicates that mortality patterns are changed by retirement, in a different way for each of three types of retirement: early, disability, and normal. Methods for allowing a change in the hazard function at the time of non-fatal events important to survival are needed.

Tallis, Sarfaty, and Leppard suppose that the hazard is changed at the time of development of breast cancer by addition of a time dependent function to the existing hazard. Lagakos [1976] assumes that an event called "progression" alters the hazard for death from lung cancer and proposes a stochastic approach to analysis. These methods or others might be used with the model proposed here to account for changes in the hazard related to non-fatal events.

5.3 Distributional Assumptions

In the present paper, the maximum likelihood estimates of the parameters of the proposed model are assumed to be asymptotically unbiased and normally distributed, with variance-covariance matrix given by the inverse of the Fisher information matrix. It is felt that the discussion and results in Appendix 1 provide justification for this assumption. However, some further theoretical justification would be of interest.

The suggestions regarding number of subjects and failures required for good approximation using the asymptotic theory are based on intuition and experience with applications to data. Simulation techniques might be used to provide further information on this point.

APPENDIX 1

REGULARITY CONDITIONS

We want to be able to use the two statements listed below in hypothesis testing:

1) If Ω is a parameter space and ω is a subspace of Ω to which p constraints have been applied, and if $\hat{\underline{\theta}}_\Omega$ and $\hat{\underline{\theta}}_\omega$ are maximum likelihood estimators of the parameter vector $\underline{\theta}$ in Ω and ω, respectively, then, $-2\ell n[L(\hat{\underline{\theta}}_\omega)/L(\hat{\underline{\theta}}_\Omega)]$ is asymptotically distributed as chi-square with p degrees of freedom, where $L(\cdot)$ is the likelihood and $\ell n(\cdot)$ denotes the natural logarithm.

2) The maximum likelihood estimate is asymptotically normally distributed with variance given by the appropriate element of the inverse of the Fisher information matrix.

The above statements are true for independent random variables x_i identically distributed with probability density function (p.d.f.) $f(x; \theta_1, \ldots, \theta_k)$ is, according to Gross and Clark [1975] (see page 229), the following four conditions hold:

i) It is possible to interchange the operations of integration with respect to x and differentiation with respect to θ_j, $j = 1, 2, \ldots, k$.

ii) The term $\partial^2 f/\partial\theta_i\partial\theta_j$ has finite variance, where $f \equiv f(x;\ \theta_1, \ldots, \theta_k)$.

iii) The likelihood function $L(\Omega)$ has bounded derivatives of the third order with respect to $\theta_1, \ldots, \theta_k$.

iv) The derivative of $L(\Omega)$ with respect to $\theta_1, \ldots, \theta_k$ vanishes at its maximum.

It should be pointed out that the general conditions required for our results 1) and 2) are less restrictive than the four mentioned above. However, since the density function is bounded, it is sufficient to verify these conditions. Of course, in our case, the random variables are not identically distributed because the distribution depends upon one or more covariables and we are allowing for censoring. Inagaki [1973] has produced results dealing with the non-identically distributed case, and Sen [1976] has examined the case where censoring occurs. Since the derivatives of the likelihood are bounded in our case, it is felt that verification of the four conditions listed will, in light of the results of Sen [1976] and Inagaki [1973], be sufficient to indicate that the desired results hold.

So, we now verify that the four conditions stated previously hold for the model with hazard $\lambda_0 e^{B_0 z_i}$, where λ_0 and B_0 are parameters and z_i is the covariate for individual i. We have

$$f \equiv f(x;\ \lambda_0,\ B_0) = \lambda_0 e^{B_0 z_i} \exp[-\lambda_0 e^{B_0 z_i} x] \,, \qquad x \geq 0, \quad \lambda_0 > 0, \quad -\infty < B_0 < \infty \,. \tag{1}$$

Letting $\theta_1 = \lambda_0$, $\theta_2 = B_0$, condition i) above is that

$$\frac{\partial}{\partial\theta_j}\left\{\int_a^b f\,dx\right\} = \int_a^b \left\{\frac{\partial f}{\partial\theta_j}\right\} dx\ , \quad j = 1,2 \quad . \tag{2}$$

We first find the partial derivatives of f with respect to λ_0 and B_0:

$$\frac{\partial f}{\partial\lambda_0} = e^{B_0 z_i}\exp[-\lambda_0 e^{B_0 z_i}x](1 - \lambda_0 e^{B_0 z_i}x)$$

$$\frac{\partial f}{\partial B_0} = \lambda_0 z_i e^{B_0 z_i}\exp[-\lambda_0 e^{B_0 z_i}x](1 - \lambda_0 e^{B_0 z_i}x)\ .$$

Integrating by parts, we find that

$$\int_a^b \left\{\frac{\partial f}{\partial\lambda_0}\right\} dx = e^{B_0 z_i}x\,\exp[-\lambda_0 e^{B_0 z_i}x]\Big|_a^b\ , \quad \text{or}$$

$$\int_a^b \left\{\frac{\partial f}{\partial\lambda_0}\right\} dx = be^{B_0 z_i}\exp[-\lambda_0 e^{B_0 z_i}b]$$
$$-ae^{B_0 z_i}\exp[-\lambda_0 e^{B_0 z_i}a]\ , \tag{3}$$

and

$$\int_a^b \left\{\frac{\partial f}{\partial B_0}\right\} dx = \lambda_0 z_i e^{B_0 z_i}x\,\exp[-\lambda_0 e^{B_0 z_i}x]\Big|_a^b\ , \quad \text{or}$$

$$\int_a^b \left\{\frac{\partial f}{\partial B_0}\right\} dx = b\lambda_0 z_i e^{B_0 z_i}\exp[-\lambda_0 e^{B_0 z_i}b]$$
$$-a\lambda_0 z_i e^{B_0 z_i}\exp[-\lambda_0 e^{B_0 z_i}a] \quad . \tag{4}$$

Now we evaluate the left hand side of equation (2), and find

$$\int_a^b f dx = -\exp[-\lambda_0 e^{B_0 z_i} b] + \exp[-\lambda_0 e^{B_0 z_i} a] \ .$$

So,

$$\frac{\partial}{\partial \lambda_0} \int_a^b f dx = b e^{B_0 z_i} \exp[-\lambda_0 e^{B_0 z_i} b] - a e^{B_0 z_i} \exp[-\lambda_0 e^{B_0 z_i} a], \tag{5}$$

and

$$\frac{\partial}{\partial B_0} \int_a^b f dx = b \lambda_0 z_i e^{B_0 z_i} \exp[-\lambda_0 e^{B_0 z_i} b] - a \lambda_0 z_i e^{B_0 z_i} \exp[-\lambda_0 e^{B_0 z_i} a] \ . \tag{6}$$

We see that equations (3) and (5) are equivalent as are equations (4) and (6). Hence, condition i) is satisfied.

To verify condition ii), we evaluate the second partials and the squares of the second partials. We find

$$\frac{\partial^2 f}{\partial \lambda_0^2} = e^{2B_0 z_i} \exp[-\lambda_0 e^{B_0 z_i} x](-2x + \lambda_0 x^2 e^{B_0 z_i}) \ ,$$

$$\frac{\partial^2 f}{\partial \lambda_0 \partial B_0} = z_i e^{B_0 z_i} \exp[-\lambda_0 e^{B_0 z_i} x](1 - 3\lambda_0 x e^{B_0 z_i} + \lambda_0^2 x^2 e^{2B_0 z_i})$$

and

$$\frac{\partial^2 f}{\partial B_0^2} = \lambda_0 z_i^2 e^{B_0 z_i} \exp[-\lambda_0 e^{B_0 z_i} x](1 - 3\lambda_0 x e^{B_0 z_i} + \lambda_0^2 x^2 e^{2B_0 z_i}). \tag{7}$$

After finding the expected values of the squares of these partial derivatives and the squares of their expected values, we find the following values for their variances:

$$\mathrm{Var}(\partial^2 f/\partial\lambda_0^2) = (47e^{2B_0 z_i})/(1296\lambda_0^2),$$

$$\mathrm{Var}(\partial^2 f/\partial\lambda_0 \partial B_0) = 11z_i^2 e^{2B_0 z_i}/81 ,$$

and
$$\mathrm{Var}(\partial^2 f/\partial B_0^2) = 11\lambda_0^2 z_i^4 e^{2B_0 z_i}/81 . \tag{8}$$

For finite values of the covariable z_i, the variances of equation (8) are non-negative, finite-valued quantities since we have $\lambda_0 > 0$ and $-\infty < B_0 < \infty$. Hence, condition ii) is satisfied.

In verifying conditions iii) and iv) we will use t_i as the random variable instead of x since we will be working with the likelihood function rather than the p.d.f. The likelihood function is

$$L(\Omega) = \prod_{i=1}^{N} \left\{ \lambda_0 e^{B_0 z_i} \exp[-\lambda_0 e^{B_0 z_i} t_i] \right\} , \tag{9}$$

where N is the number of individuals, t_i and z_i are, respectively, the failure time and covariate value for individual i, $i = 1, 2, \ldots, N$. Denoting $L(\Omega)$ simply by L, the likelihood function of equation (9) can be expressed as

$$L = \lambda_0^N e^{B_0 \Sigma z_i} \exp[-\Sigma\lambda_0 e^{B_0 z_i} t_i] , \tag{10}$$

where Σ will always be the sum over index i, $i = 1, 2, \ldots, N$.

The first, second, and third order partial derivatives of L with respect to λ_0 and B_0 are given below:

$$\frac{\partial L}{\partial\lambda_0} = \lambda_0^{N-1} e^{B_0 \Sigma z_i} \exp[-\Sigma\lambda_0 e^{B_0 z_i} t_i] \left\{ N - \Sigma\lambda_0 e^{B_0 z_i} t_i \right\} , \tag{11}$$

$$\frac{\partial L}{\partial B_0} = \lambda_0^N e^{B_0 \Sigma z_i} \exp[-\Sigma\lambda_0 e^{B_0 z_i} t_i] \left\{ \Sigma z_i - \Sigma\lambda_0 e^{B_0 z_i} t_i z_i \right\} , \tag{12}$$

$$\frac{\partial^2 L}{\partial \lambda_0^2} = \lambda_0^{N-2}\ e^{B_0 \Sigma z_i} \exp[-\Sigma\lambda_0 e^{B_0 z_i} t_i] \Big\{ N(N-1) - 2N\Sigma\lambda_0 e^{B_0 z_i} t_i + (\Sigma\lambda_0 e^{B_0 z_i} t_i)^2 \Big\} , \tag{13}$$

$$\frac{\partial^2 L}{\partial \lambda_0 \partial B_0} = \lambda_0^{N-1}\ e^{B_0 \Sigma z_i} \exp[-\Sigma\lambda_0 e^{B_0 z_i} t_i] \Big\{ N\Sigma z_i - (\Sigma z_i)(\Sigma\lambda_0 e^{B_0 z_i} t_i) - (N+1)(\Sigma\lambda_0 e^{B_0 z_i} t_i z_i) + (\Sigma\lambda_0 e^{B_0 z_i} t_i)(\Sigma\lambda_0 e^{B_0 z_i} t_i z_i) \Big\} , \tag{14}$$

$$\frac{\partial^2 L}{\partial B_0^2} = \lambda_0^{N}\ e^{B_0 \Sigma z_i} \exp[-\Sigma\lambda_0 e^{B_0 z_i} t_i] \Big\{ (\Sigma z_i)^2 - 2(\Sigma z_i)(\Sigma\lambda_0 e^{B_0 z_i} t_i z_i) - \Sigma\lambda_0 e^{B_0 z_i} t_i z_i^2 + (\Sigma\lambda_0 e^{B_0 z_i} t_i z_i)^2 \Big\} , \tag{15}$$

$$\frac{\partial^3 L}{\partial \lambda_0^3} = \lambda_0^{N-3}\ e^{B_0 \Sigma z_i} \exp[-\Sigma\lambda_0 e^{B_0 z_i} t_i] \Big\{ N(N-1)(N-2) - 3N(N-1)\Sigma\lambda_0 e^{B_0 z_i} t_i + 3N(\Sigma\lambda_0 e^{B_0 z_i} t_i)^2 - (\Sigma\lambda_0 e^{B_0 z_i} t_i)^3 \Big\} , \tag{16}$$

$$\frac{\partial^3 L}{\partial \lambda_0^2 \partial B_0} = \lambda_0^{N-2}\, e^{B_0 \Sigma z_i} \exp[-\Sigma\lambda_0 e^{B_0 z_i} t_i] \Big\{ N(N-1)\Sigma z_i - N(N+1)\Sigma\lambda_0 e^{B_0 z_i} t_i z_i$$

$$- 2N(\Sigma z_i)(\Sigma\lambda_0 e^{B_0 z_i} t_i) + 2(N+1)(\Sigma\lambda_0 e^{B_0 z_i} t_i)(\Sigma\lambda_0 e^{B_0 z_i} t_i z_i)$$

$$+ (\Sigma z_i)(\Sigma\lambda_0 e^{B_0 z_i} t_i)^2 - (\Sigma\lambda_0 e^{B_0 z_i} t_i)^2 (\Sigma\lambda_0 e^{B_0 z_i} t_i z_i) \Big\}, \tag{17}$$

$$\frac{\partial^3 L}{\partial \lambda_0 \partial B_0^2} = \lambda_0^{N-1}\, e^{B_0 \Sigma z_i} \exp[-\Sigma\lambda_0 e^{B_0 z_i} t_i] \Big\{ N(\Sigma z_i)^2 - 2(N+1)(\Sigma z_i)(\Sigma\lambda_0 e^{B_0 z_i} t_i z_i)$$

$$- (\Sigma z_i)^2 (\Sigma\lambda_0 e^{B_0 z_i} t_i) - (N+1)\Sigma\lambda_0 e^{B_0 z_i} t_i z_i^2 + (N+2)(\Sigma\lambda_0 e^{B_0 z_i} t_i z_i)^2$$

$$+ (\Sigma\lambda_0 e^{B_0 z_i} t_i)(\Sigma\lambda_0 e^{B_0 z_i} t_i z_i^2) + 2(\Sigma z_i)(\Sigma\lambda_0 e^{B_0 z_i} t_i)(\Sigma\lambda_0 e^{B_0 z_i} t_i z_i)$$

$$- (\Sigma\lambda_0 e^{B_0 z_i} t_i)(\Sigma\lambda_0 e^{B_0 z_i} t_i z_i)^2 \Big\}, \tag{18}$$

and

$$\frac{\partial^3 L}{\partial B_0^3} = \lambda_0^{N}\, e^{B_0 \Sigma z_i} \exp[-\Sigma\lambda_0 e^{B_0 z_i} t_i] \Big\{ (\Sigma z_i)^3 - 3(\Sigma z_i)^2 (\Sigma\lambda_0 e^{B_0 z_i} t_i z_i)$$

$$- 3(\Sigma z_i)(\Sigma\lambda_0 e^{B_0 z_i} t_i z_i^2) - \Sigma\lambda_0 e^{B_0 z_i} t_i z_i^3 + 3(\Sigma z_i)(\Sigma\lambda_0 e^{B_0 z_i} t_i z_i)^2$$

$$+ 3(\Sigma\lambda_0 e^{B_0 z_i} t_i z_i)(\Sigma\lambda_0 e^{B_0 z_i} t_i z_i^2) - (\Sigma\lambda_0 e^{B_0 z_i} t_i z_i)^3 . \tag{19}$$

Inspection of equations (16) through (19) reveals that for $\lambda_0 > 0$, $-\infty < B_0 < \infty$, $N < \infty$, and for $0 \le t_i < \infty$ and $-\infty < z_i < \infty$, $i = 1, 2, \ldots, N$, the right hand side of each equation is bounded, so that condition iii) is satisfied.

If we can show that setting $\partial L/\partial\lambda_0 = \partial L/\partial B_0 = 0$ results in solutions $\hat{\lambda}_0$ and $\hat{B}_0$ which maximize L, then condition iv) is satisfied. Let $\hat{\lambda}_0$ and $\hat{B}_0$ be solutions to the equations $\partial L/\partial\lambda_0 = 0$, $\partial L/\partial B_0 = 0$. Also, let A, B, and C, respectively, be $\partial^2 L/\partial\lambda_0^2$, $\partial^2 L/\partial\lambda_0\partial B_0$, and $\partial^2 L/\partial B_0^2$ evaluated at the point $(\hat{\lambda}_0, \hat{B}_0)$. Then, if $AC - B^2 > 0$, the point $(\hat{\lambda}_0, \hat{B}_0)$ results in L being maximized, so that condition iv) is satisfied.

Setting $\partial L/\partial\lambda_0$ equal to zero implies that

$$N = \sum_{i=1}^{N} \hat{\lambda}_0 e^{\hat{B}_0 z_i} t_i \ , \tag{20}$$

and setting $\partial L/\partial B_0 = 0$ implies that

$$\sum_{i=1}^{N} z_i = \sum_{i=1}^{N} \hat{\lambda}_0 e^{\hat{B}_0 z_i} t_i z_i \ . \tag{21}$$

Under these conditions,

$$A = \hat{\lambda}_0^{N-2} e^{\hat{B}_0 \Sigma z_i} \exp[-N]\{-N\} \ , \tag{22}$$

$$B = \hat{\lambda}_0^{N-1} e^{\hat{B}_0 \Sigma z_i} \exp[-N]\left\{-\Sigma\hat{\lambda}_0 e^{\hat{B}_0 z_i} t_i z_i\right\} \ , \tag{23}$$

and

$$C = \hat{\lambda}_0^{N} e^{\hat{B}_0 \Sigma z_i} \exp[-N]\left\{-\Sigma\hat{\lambda}_0 e^{\hat{B}_0 z_i} t_i z_i^2\right\} \ . \tag{24}$$

So, we have

$$AC - B^2 = \hat{\lambda}_0^{2N-2} e^{2\hat{B}_0 \Sigma z_i} \exp[-2N] \left\{ N\Sigma\hat{\lambda}_0 e^{\hat{B}_0 z_i} t_i z_i^2 - (\Sigma\hat{\lambda}_0 e^{\hat{B}_0 z_i} t_i z_i)^2 \right\} . \quad (25)$$

Now, $AC - B^2 > 0$ implies that $N\Sigma\hat{\lambda}_0 e^{\hat{B}_0 z_i} t_i z_i^2 > (\Sigma\hat{\lambda}_0 e^{\hat{B}_0 z_i} t_i z_i)^2$.

Using equation (20), we express the inequality as

$$\left(\sum_{i=1}^{N} \hat{\lambda}_0 e^{\hat{B}_0 z_i} t_i \right) \left(\sum_{j=1}^{N} \hat{\lambda}_0 e^{\hat{B}_0 z_j} t_j z_j^2 \right) > \left(\sum_{i=1}^{N} \hat{\lambda}_0 e^{\hat{B}_0 z_i} t_i z_i \right) \left(\sum_{j=1}^{N} \hat{\lambda}_0 e^{\hat{B}_0 z_j} t_j z_j \right) ,$$

or

$$\sum_{i=1}^{N} \sum_{j=1}^{N} e^{\hat{B}_0 (z_i + z_j)} t_i t_j z_j^2 > \sum_{i=1}^{N} \sum_{j=1}^{N} e^{\hat{B}_0 (z_i + z_j)} t_i t_j z_i z_j ,$$

or

$$\sum_{i=1}^{N} \sum_{j=1}^{N} e^{\hat{B}_0 (z_i + z_j)} t_i t_j (z_j^2 - z_i z_j) > 0 . \quad (26)$$

Now let $w_{ij} = e^{\hat{B}_0 (z_i + z_j)} t_i t_j$. So, we must show that

$$S = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (z_j^2 - z_i z_j) > 0 .$$

Note that $w_{ij} = w_{ji}$ for all i, j so that

$$S = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (z_i^2 - z_i z_j) .$$

Then we have

$$2S = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (z_i^2 - 2 z_i z_j + z_j^2)$$

or

$$2S = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(z_i - z_j)^2 .$$

Therefore, *unless* $t_i = 0$ or $z_i = z$ for *all* $i = 1, 2, \ldots, N$, where z is some constant, we have that $S > 0$ and the desired result obtains, so condition iv) is satisfied.

Of course, $t_i = 0$ for all i implies that all subjects die at the exact time of first observation, so we would have to conclude that the hazard approaches infinity at $t_i = 0$, and this case is of no importance.

If $z_i = z$ for all i, the covariate does not vary at all, so there is no sense in attempting to estimate B_0. If $z_i = z$ for all i, setting equations (11) and (12) equal to zero results in *one* solution,

$$\hat{\lambda}_0 e^{\hat{B}_0 z} = N/(\sum_{i=1}^{N} t_i) ,$$

which is just the maximum likelihood estimate of the hazard under the assumptions of a constant hazard and no covariates, as would be expected.

We have demonstrated that the conditions required for two useful properties of maximum likelihood estimators to hold are met for a model with hazard $\lambda_0 e^{B_0 z_i}$, assuming no censoring and identical distributions of the random variables t_i. The *form* of the distributions is the same. However, differing covariate values result in differing distributions for the t_i. Nonetheless, it is felt that the results obtained here considered together with those of Sen [1976] and Inagaki [1973] indicate that the desired properties of the maximum likekihood estimators hold true for the cases of interest in this paper.

BIBLIOGRAPHY

Andjelkovic, D., Taulbee, J., and Symons, M. (1976). Mortality Experience of a Cohort of Rubber Workers, 1964-1973, *Journal of Occupational Medicine 18*, 387-94.

Andjelkovich, D., Taulbee, J., Symons, M., and Williams, T. (1977). Mortality of Rubber Workers with Reference to Work Experience, *Journal of Occupational Medicine 19*, 397-405.

Antle, C.E., and Klimko, L.A. (1975). Choice of Model for Reliability Studies and Related Topics III, *Report Number ARL 75-0133*, Aerospace Research Laboratories, Wright-Patterson Air Force Base, Ohio.

Axtell, L.M., Myers, M.H., Thomas, L.H., Berard, C.W., Kagan, A.R., and Newell, G.R. (1972). Prognostic Indicators in Hodgkins Disease, *Cancer 29*, 1481-88.

Berkson, J., and Gage, R.P. (1950). Calculation of Survival Rates for Cancer, *Proceedings of the Staff Meetings of the Mayo Clinic 25*, 270-86.

Berry, G., and Wagner, J.C. (1976). Effect of Age at Inoculation of Asbestos on Occurrence of Mesotheliomas in Rats, *International Journal of Cancer 17*, 477-83.

Breslow, N. (1972). Comment on D.R. Cox (1972) paper, *Journal of the Royal Statistical Society Series B 34*, 216-17.

Breslow, N. (1974). Covariance Analysis of Censored Survival Data, *Biometrics 30*, 89-99.

Brown, C.C. (1975). On the Use of Indicator Variables for Studying the Time-Dependence of Parameters in a Response-Time Model, *Biometrics 31*, 863-72.

Byar, D.P., Huse, R., Bailar, III, J.C., and the Veterans Administration Cooperative Urological Research Group (1974). An Exponential Model Relating Censored Survival Data and Concomitant Information for Prostatic Cancer Patients, *Journal of the National Cancer Institute 52*, 321-26.

Campbell, N.R., Reade, P.C., and Radden, B.G. (1974). Effect of Cysteine on Survival of Mice with Transplanted Malignant Thymoma, *Nature 251*, 158-59.

Cangir, A., George, S., and Sullivan, M. (1975). Unfavorable Prognosis of Acute Leukemia in Infancy, *Cancer 36*, 1973-78.

Case, R.A.M., and Hosker, M.E. (1954). Tumour of the Urinary Bladder as an Occupational Disease in the Rubber Industry in England and Wales, *British Journal of Preventive and Social Medicine 8*, 39-50.

Clark, V.A. (1975). Survival Distribution, *Annual Review of Biophysics and Bioengineering 4*, 431-48.

Cox, D.R. (1972). Regression Models and Life-Tables, *Journal of the Royal Statistical Society Series B 34*, 187-202.

Cox, D.R. (1975). Partial Likelihood, *Biometrika 62*, 269-76.

Decoufle, P., Lloyd, J.W., and Salvin, L.G. (1977). Causes of Death among Construction Machinery Operators, *Journal of Occupational Medicine 19*, 123-28.

Doll, R. (1952). The Causes of Death among Gas-Workers with Special Reference to Cancer of the Lung, *British Journal of Industrial Medicine 9*, 180-85.

Doll, R., Fisher, R.E.W., Gammon, E.J., Gunn, W., Hughes, G.O., Tyrer, F.H., and Wilson, W. (1965). Mortality of Gas Workers with Special Reference to Cancers of the Lung and Bladder, Chronic Bronchitis, and Pneumoconiosis, *British Journal of Industrial Medicine 22*, 1-12.

Doll, R., Vessey, M.P., Beasley, R.W.R., Buckley, A.R., Fear, E.C., Fisher, R.E.W., Gammon, E.J., Gunn, W., Hughes, G.O., Lee, K., and Norman-Smith, B. (1972). Mortality of Gasworkers - Final Report of a Prospective Study, *British Journal of Industrial Medicine 29*, 394-406.

Dorn, H.F., and Baum, W.S. (1955). Mortality among Workers in Cigarette Factories, *Industrial Medicine and Surgery 24*, 239-41.

Dumonceaux, R., and Antle, C.E. (1973). Discrimination Between the Log-Normal and the Weibull Distributions, *Technometrics 15*, 923-26.

Efron, B. (1977). The Efficiency of Cox's Likelihood Function for Censored Data, *Journal of the American Statistical Association 72*, 557-65.

Elmes, P.C., and Simpson, M.J.C. (1971). Insulation Workers in Belfast 3. Mortality 1940-1966, *British Journal of Industrial Medicine 28*, 226-36.

Enker, W.E., and Jacobitz, J.L. (1976). The Definitive Value of Active Specific Immunotherapy for Experimental Carcinoma of the Colon, *Surgery 80*, 164-70.

Enterline, P.E. (1965). Mortality among Asbestos Products Workers in the United States, *Annals of the New York Academy of Sciences 132*, 156-65.

Enterline, P.E. (1974). Respiratory Cancer among Chromate Workers, *Journal of Occupational Medicine 16*, 523-26.

Enterline, P.E., Decoufle, P., and Henderson, V. (1972). Mortality in Relation to Occupational Exposure in the Asbestos Industry, *Journal of Occupational Medicine 14*, 897-903.

Enterline, P.E., and Henderson, V. (1975). The Health of Retired Fibrous Glass Workers, *Archives of Environmental Health 30*, 113-16.

Farewell, V.T., and Prentice, R.L. (1977). A Study of Distributional Shape in Life Testing, *Technometrics 19*, 69-75.

Feigl, P., and Zelen, M. (1965). Estimation of Exponential Survival Probabilities with Concomitant Information, *Biometrics 21*, 826-38.

Feinstein, A.R. (1975). Clinical Biostatistics XXXIV. The Other Side of Statistical Significance: Alpha, Beta, Delta and the Calculation of Sample Size, *Clinical Pharamacology and Theraputics 18*, 491-505.

Fox, A.J., Lindars, D.C., and Owen, R. (1974). A Survey of Occupational Cancer in the Rubber and Cablemaking Industries: Results of Five-Year Analysis, 1967-71, *British Journal of Industrial Medicine 31*, 140-51.

Freireich, E.J., Gehan, E., Frei, E., III, Schroeder, L.R., Wolman, I.J., Anbari, R., Burgert, E.O., Mills, S.D., Pinkel, D., Selawry, O.S., Moon, J.H., Gendel, B.R., Spurr, C.L., Storrs, R., Haurani, F., Hoogstraten, B., and Lee, S. (1963). The Effect of 6-Mercaptopurine on the Duration of Steroid-Induced Remissions in Acute Leukemia: A Model for Evaluation of Other Potentially Useful Therapy, *Blood 21*, 699-716.

Gaffey, W.R. (1975). A Critique of the Standardized Mortality Ratio, Paper Presented at the Annual Meeting of the American Occupational Medicine Association. San Francisco, April 17, 1975.

Garg, M.L., Rao, B.R., and Redmond, C.K. (1970). Maximum-likelihood Estimation of the Parameters of the Gompertz Survival Function, *Applied Statistics 19*, 152-59.

Gehan, E.A. (1965). A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-censored Samples, *Biometrika 52*, 203-23.

Gehan, E.A., and Siddiqui, M.M. (1973). Simple Regression Methods for Survival Time Studies, *Journal of the American Statistical Association 68*, 848-56.

George, S.L., and Desu, M.M. (1974). Planning the Size and Duration of a Clinical Trial Studying the Time to Some Critical Event, *Journal of Chronic Diseases 27*, 15-24.

Glasser, M. (1967). Exponential Survival with Covariance, *Journal of the American Statistical Association 62*, 561-68.

Greenberg, R.A., Bayard, S., and Byar, D. (1974). Selecting Concomitant Variables Using a Likelihood Ratio Step-Down Procedure and a Method of Testing Goodness of Fit in an Exponential Survival Model, *Biometrics 30*, 601-8.

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

Halperin, M., Rogot, E., Gurian, J., and Ederer, F. (1968). Sample Sizes for Medical Trials with Special Reference to Long-Term Therapy, *Journal of Chronic Diseases 21*, 13-24.

Hankey, B.F., and Mantel, N. (1974). A Logistic Regression Analysis of Response-Time Data where the Hazard Function is Time Dependent, (submitted for publication).

Henderson, V., and Enterline, P.E. (1973). An Unusual Mortality Experience in Cotton Textile Workers, *Journal of Occupational Medicine 15*, 717-19.

Holford, T.R. (1976). Life Tables with Concomitant Information, *Biometrics 32*, 587-97.

Inagaki, N. (1973). Asymptotic Relations Between the Likelihood Estimating Function and the Maximum Likelihood Estimator, *Annals of the Institute of Statistical Mathematics 25*, 1-26.

Ingalls, T.H., and Risquez-Iribarren, R. (1961). Periodic Search for Cancer in the Carbon Black Industry, *Archives of Environmental Health 2*, 429-33.

Irwin, J.O. (1942). The Distribution of the Logarithm of Survival Times when the True Law is Exponential, *Journal of Hygiene 42*, 328-33.

Joysey, V.C., Roger, J.H., Evans, D.B., and Herbertson, B.M. (1973). Kidney Graft Survival and Matching for HL-A and ABO Antigens, *Nature 246*, 163-65.

Kalbfleisch, J.D. (1974a). Some Efficiency Calculations for Survival Distributions, *Biometrika 61*, 31-38.

Kalbfleisch, J.D. (1974b). Some Extensions and Applications of Cox's Regression and Life Model, paper presented at the Tallahassee, Florida Meetings of the Biometric Society, Institute for Mathematical Statistics, and American Statistical Association, March 20-22, 1974.

Kalbfleisch, J., and Prentice, R.L. (1972). Comment on D.R. Cox (1972) paper, *Journal of the Royal Statistical Society Series B 34*, 215-16.

Kalbfleisch, J.D., and Prentice, R.L. (1973). Marginal Likelihoods Based on Cox's Regression and Life Model, *Biometrika 60*, 267-78.

Kaplan, E.B., and Elston, R.C. (1972). A Subroutine Package for Maximum Likelihood Estimation (MAXLIK), *Institute of Statistics Mimeo Series No. 823*, Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, N.C.

Kaplan, E.L., and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations, *Journal of the American Statistical Association 53*, 457-81.

Kendall, M.G., and Stuart, A. (1973). *The Advanced Theory of Statistics*, Volume 2, Third Edition, Hafner Publishing Company, New York.

Koch, G.G., Johnson, W.D., and Tolley, H.D. (1972). A Linear Models Approach to the Analysis of Survival Data and Extent of Disease in Multidimensional Contingency Tables, *Journal of the American Statistical Association 67*, 783-96.

Kodlin, D. (1967). A New Response Time Distribution, *Biometrics 23*, 227-39.

Krall, J.M., Uthoff, V.A., and Harley, J.B. (1975). A Step-up Procedure for Selecting Variables Associated with Survival, *Biometrics 31*, 49-57.

Krane, S.A. (1963). Analysis of Survival Data by Regression Techniques, *Technometrics 5*, 161-74.

Lagakos, S. (1976). A Stochastic Model for Censored-Survival Data in the Presence of an Auxiliary Variable, *Biometrics 32*, 551-59.

Lloyd, J.W., and Ciocco, A. (1969). Long Term Study of Steelworkers I. Methodology, *Journal of Occupational Medicine 11*, 299-310.

Lloyd, J.W., Lundin, Jr., F.E., Redmond, C.K., and Geiser, P.B. (1970). Long-Term Mortality Study of Steelworkers IV. Mortality by Work Area, *Journal of Occupational Medicine 12*, 151-57.

Mancuso, T.F., Ciocco, A., and El-Attar, A.A. (1968). An Epidemiological Approach to the Rubber Industry: a Study Based on Departmental Experience, *Journal of Occupational Medicine 10*, 213-32.

Mancuso, T.F., and Coulter, E.J. (1959). Methods of Studying the Relation of Employment and Long-Term Illness - Cohort Analysis, *American Journal of Public Health 49*, 1525-36.

Mancuso, T.F., and Coulter, E.J. (1963). Methodology in Industrial Studies, *Archives of Environmental Health 6*, 210-26.

Mantel, N. (1963). Chi-Square Tests with One Degree of Freedom; Extensions of the Mantel - Haenszel Procedure, *Journal of the American Statistical Association 58*, 690-700.

Mantel, N. (1966). Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration, *Cancer Chemotherapy Reports 50*, 163-70.

Mantel, N., and Haenszel, W. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease, *Journal Of the National Cancer Institute 22*, 719-48.

Monson, R.R., and Nakano, K.K. (1976a). Mortality among Rubber Workers I. White Male Union Employees in Akron, Ohio, *American Journal of Epidemiology 103*, 284-96.

Monson, R.R., and Nakano, K.K. (1976b). Mortality among Rubber Workers II. Other Employees, *American Journal of Epidemiology 103*, 297-303.

Moss, E., and Lee, W.R. (1974). Occurrence of Oral and Pharyngeal Cancers in Textile Workers, *British Journal of Industrial Medicine 31*, 224-32.

Moss, E., Scott, T.S., and Atherley, G.R.C. (1972). Mortality of Newspaper Workers from Lung Cancer and Bronchitis, *British Journal of Industrial Medicine 29*, 1-14.

Myers, M.H., Hankey, B.F., and Mantel, N. (1973). A Logistic-Exponential Model for Use with Response-Time Data Involving Regressor Variables, *Biometrics 29*, 257-69.

Ortmeyer, C.E., Costello, J., Morgan, W.K.C., Swecker, S., and Peterson, M. (1974). The Mortality of Appalachian Coal Miners, 1963-71, *Archives of Environmental Health 29*, 67-72.

Parr, V.B., and Webster, J.T. (1965). A Method for Discriminating Between Failure Density Functions Used in Reliability Predictions, *Technometrics 7,* 1-10.

Pasternack, B.S. (1972). Sample Sizes for Clinical Trials Designed for Patient Accrual by Cohorts, *Journal of Chronic Diseases 25,* 673-81.

Pasternack, B.S., and Ehrlich, L. (1972). Occupational Exposure to an Oil Mist Atmosphere. A 12-Year Mortality Study, *Archives of Environmental Health 25,* 286-94.

Pasternack, B.S., and Gilbert, H.S. (1971). Planning the Duration of Long-Term Survival Time Studies Designed for Accrual by Cohorts, *Journal of Chronic Diseases 24,* 681-700.

Prentice, R.L. (1973). Exponential Survivals with Censoring and Explanatory Variables, *Biometrika 60,* 279-88.

Prentice, R.L. (1976). Use of the Logistic Model in Retrospective Studies, *Biometrics 32,* 599-606.

Ramirez, G., Klotz, J., Strawitz, J.G., Wilson, W.L., Cornell, G.N., Madden, R.E., Minton, J.P., and the Central Oncology Group, (1975). Combination Chemotherapy in Breast Cancer: A Randomized Study of 4 Versus 5 Drugs, *Oncology 32,* 101-108.

Rao, C.R. (1965). *Linear Statistical Inference and Its Applications,* John Wiley and Sons, New York.

Reid, D.D., and Buck, C. (1956). Cancer in Coking Plant Workers, *British Journal of Industrial Medicine 13,* 265-69.

Rubens, R.D., Knight, R.K., and Hayward, J.L., (1975). Chemotherapy of Advanced Breast Cancer: A Controlled Randomized Trial of Cyclophosphamide Versus a Four-Drug Combination, *British Journal of Cancer 32,* 730-36.

Schlesselman, J.J. (1974a). Sample Size Requirements in Cohort and Case-Control Studies of Disease, *American Journal of Epidemiology 99,* 381-84.

Schlesselman, J.J. (1974b). Tables of the Sample Size Requirement for Cohort and Case-Control Studies of Disease, (Unpublished paper).

Schork, M.A., and Remington, R.D. (1967). The Determination of Sample Size in Treatment-Control Comparisons for Chronic Disease Studies in which Drop-out or Non-Adherence is a Problem, *Journal of Chronic Diseases 20,* 233-39.

Scott, L.M., Bahler, K.W., DeLaGarza, A., and Lincoln, T.A. (1972). Mortality of Uranium and Non-Uranium Workers, *Health Physics 23*, 555-57.

Sen, P.K. (1976). Weak Convergence of Progressively Censored Likelihood Ratio Statistics and Its Role in Asymptotic Theory of Life Testing, *The Annals of Statistics 4*, 1247-57.

Tabershaw, I.R., and Gaffey, W.R. (1974). Mortality Study of Workers in the Manufacture of Vinyl Chloride and its Polymers, *Journal of Occupational Medicine 16*, 509-18.

Tallis, G.M., Sarfaty, G., and Leppard, P. The Use of a Probability Model for the Construction of Age Specific Life Tables for Women with Breast Cancer (Unpublished paper).

Taylor, F.H. (1966). The Relationship of Mortality and Duration of Employment as Reflected by a Cohort of Chromate Workers, *American Journal of Public Health 56*, 218-29.

Truett, J., Cornfield, J., and Kannel, W. (1967). A Multivariate Analysis of the Risk of Coronary Heart Disease in Framingham, *Journal of Chronic Diseases 20*, 511-24.

Wagoner, J.K., Archer, V.E., Carroll, B.E., Holaday, D.A., and Lawrence, P.A. (1964). Cancer Mortality Patterns among U.S. Uranium Miners and Millers, 1950 through 1962, *Journal of the National Cancer Institute 32*, 787-801.

Zippin, C., and Armitage, P. (1966). Use of Concomitant Variables and Incomplete Survival Information in the Estimation of an Exponential Survival Parameter, *Biometrics 22*, 665-72.