THE USE OF GENERAL LINEAR MODELS FOR FAILURE DATA AND CATEGORICAL DATA

by

ROGER MARK SAUTER

B.A., Mid-America Nazarene College, 1980

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1982

Approved by:

George A. Milliken
Major Professor

ACKNOWLEDGEMENTS

I want to thank Dr. Richard Sundheim for his help with the categorical analysis, both through lectures in Categorical Data and his time spent answering my numerous questions.

I also want to thank the members of my committee, Dr. S. S. Yang, Dr. Dallas Johnson (I especially appreciated his late night help on getting a BMD program to run for an example), and Dr. George Milliken. I appreciated the fact that Dr. Milliken not only served as my major professor in helping me with this paper, but also was a very good friend during my stay here.

Most importantly I want to thank my Lord and Savior Jesus Christ, who gave me the strength to keep on going and not to get overanxious.

TABLE OF CONTENTS

Chapter 1 The Cox Regression Model

Chapter 2 The Loglinear Model

Chapter 3 The GSK Model

Chapter 4

CHAPTER 1

Cox Regression Model

1.1 Introduction

Failure time data consists of data which measures the time until a certain event occurs. For example, for a electronic component, it is the time until the component fails to function or for a person who is diagnosed to have a certain disease it is the time it takes the person to recover. There are many factors which can influence the failure time of a particular component. For the electronic component, the temperature and humidity of the environment could effect the failure time. For the person with a disease, the recovery time could depend on the age of the person, the extent the disease and the type of treatment. It is desired to construct a model to describe failure rates as a function of these independent variables. The Cox model is a model which allows the researcher to incorporate the information about the independent variables into describing the failure time. The Cox model is a failure time distribution of the form given by $h_0(t)e^{\underline{x\beta}}$, where $\underline{x}$ is a vector of explanatory variables and $\underline{\beta}$ is the vector of corresponding regression coefficients. The values of the elements of $\underline{\beta}$ are related to the effect each variables has on the failure time. Estimates of the parameters, $\underline{\beta}$, can be obtained to provide an estimate of the distribution of failure times. The following are some examples of where the Cox model can be applied.

Example 1. Lawless (1981) gave an example of survival data on 40 advanced lung cancer patients. There were a number of explanatory variables used, including the effects of two chemotherapy treatments and type and size of tumor. It is desired to know if the type of treatment has an effect on the failure time and which of the explanatory variables are making a significant contribution to the failure times. This can be done by modeling the failure times as a function of these explanatory variables using the Cox regression model.

Example 2. In a factory it is desired to know which of the variable events effect the failure time of a certain component. These explanatory variables could include years of experience of machine operator, hours of use per day, temperature and humidity. The failure times then could be modeled as a function of these explanatory variables using the Cox regression model. Estimates for $\underline{\beta}$ could be obtained, to give the researcher an idea of the effect of each variable on the failure times.

The rest of this chapter contains a literature review for the Cox model, followed by a description of the general procedure. The last section of this chaper contains a couple of examples where the Cox model is analyzed by the computer programs available in BMD and SAS.

1.2 Definitions

The Cox Regression Model

The Cox regression model was first suggested by Cox (1972) in order to analyze censored failure time data. Before looking at this model, it is necessary to understand what is meant by failure time data and censoring. The following definitions were taken from Kalbfleisch and Prentice (1980) and Lawless (1981).

Definition 1.2.1

Failure time data is data that measures the time until an event occurs.

The event does not have to be when something fails, it can also apply to other areas, for example the time until a desired response occurs such as time to recovery for individuals with a heart transplant. The principle use of this procedure does extends to deaths of individuals under study or the time to failure of components at a factory. In order to analyze this type of data, it is necessary to know something about the distribution of these failure times. This is dealt with next.

Let T be a non-negative random variable representing the failure time of an individual or a component in a population of similar individuals or components. T can be either continuous or discrete. The continuous case is considered first and starting with the definition of the probability density function of T.

Definition 1.2.2

The probability density function of T is

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t<t<t+\Delta t)}{\Delta t}$$

$$= \frac{-dF(t)}{dt}$$

Here F(t) is the cummulative distribution function of T. It gives the probability of failure before some given time t.

Definition 1.2.3

The cummulative distribution of T is

$$F(t) = Pr(T<t) = \int_0^t f(x)\, dx.$$

The survivor function is just 1-F(t) and it gives the probability of an individual not failing until after time t. It is defined as follows.

Definition 1.2.4

The survivor function for T is

$$S(t) = Pr(T>t) = \int_t^\infty f(x)\, dx.$$

One other function that is of interest for the Cox model is the hazard function. It gives the instantaneous rate of death for an individual at time t, given he is still alive at time t.

Definition 1.2.5

The hazard function for T is

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t<T<t+\Delta t)}{\Delta t}$$

$$= \frac{f(t)}{S(t)} \ .$$

This function describes the way the instantaneous probability of death for an individual changes over time. Shown in figure 1 (on the next page) are three basic hazard functions and their corresponding probability density functions. The three shown are a) monotone increasing, b) monotone decreasing and c) the U-shaped hazard functions. An example of a U-shaped hazard function is when the rate of death for an individual is observed from time of birth until old age. At birth there is a relatively high rate of death, this lowers for a period of time until about 30, then the rate of death starts to increase. The most commonly used model involves the monotone increasing function.

For discrete data the definitions are similar with identical interpretations. The definitions are summarized as follows.

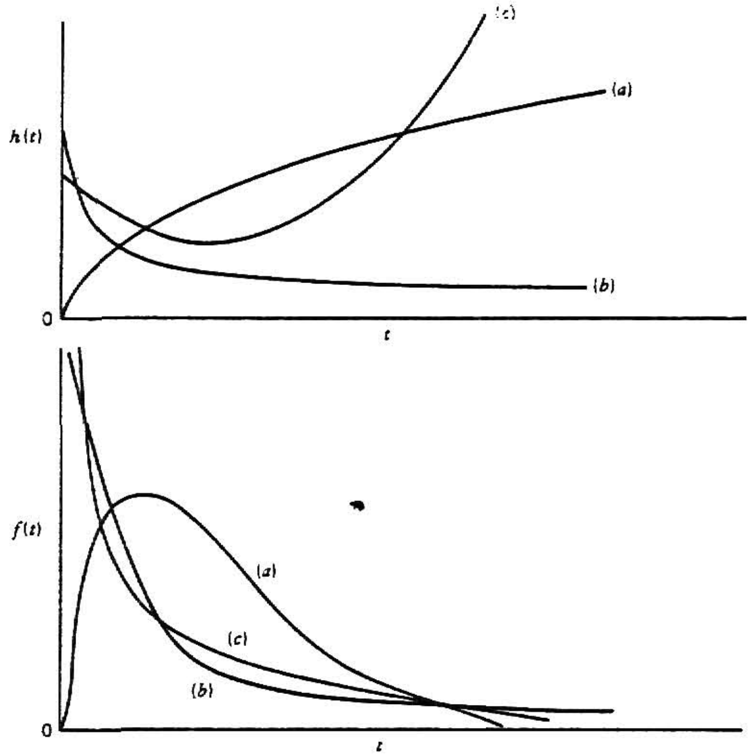

Figure 1

Definition 1.2.6

If T is a discrete random variable and it takes on the values $t_1 < t_2 \ldots < t_k$, then its probability function is

$$f(t_i) = P(T=t_i) \qquad 1 = 1,2,3,\ldots,k$$

and its survivor function is

$$S(t) = Pr(T>t) = \sum_{i \ni t_i > t}^{k} f(t_i).$$

Definition 1.2.7

If T is a discrete random variable as described above then its <u>hazard function</u> is

$$h(t_i) = Pr(T = t_i | T > t_i)$$

$$= \frac{f(t_i)}{S(t_i)} \ .$$

An example for the use of discrete models is when the failure time data is grouped or when intervals of a certain length are used (like months or years). Next to be considered is the meaning of censored data.

A major problem in the analysis of failure data is censoring. The reason the Cox model was developed was to handle censored data.

Definition 1.2.8

<u>Censoring</u> occurs when the exact times of failure are known only for part of the individuals under study, and for the rest it is only known that their failure times exceed some value or are less than some value.

It is desired to gain information from these censored failure times. There are a number of types of censoring, which need to be defined although for the analysis of the Cox model only one type of censoring is considered.

Definition 1.2.9

<u>Right censoring</u> occurs when an observation's exact failure time is not known, but that its failure time is greater than or equal to a known time L.

Definition 1.2.10

Left censoring occurs when an observation's exact failure time is not known, but that its failure time is less than or equal to a known time L.

For failure time data, right censoring is far more common than left censoring. There are two general types of right censoring which are considered next.

Definition 1.2.11

Type II censoring occurs when only the r shortest failure times are observed in a random sample of size n individuals.

Experiments of this type are often used. A total of n individuals is put on a test, when r of these individuals or components have failed the experiment is ended. It is important to select r before the start of the experiment. Another type of right censoring is type I censoring.

Definition 1.2.12

Type I censoring occurs when a predetermined time is selected and the experiment is ended at that time. Those that have not failed at this time are said to be type I censored.

For example in an experiment n individuals are put on a test, but the experiment will be ended after time L has been completed. Exact failure times are known only for those individuals with failure times less than or equal to L. A slightly more complicated type I censoring occurs when individuals enter the experiment at different times, thus when they are censored they have different censored times, denoted by L_i. Unless otherwise specified whenever censoring is mentioned it refers to right hand type I censoring. Next to be considered is the model of interest, called the proportional hazards model.

The following definitions concerning proportional hazard models were taken from Lawless (1981) and Cox (1975). The proportional hazard model is an important part of the analysis of failure data.

Definition 1.2.13

A proportional hazard family of models pertain to the class of models that different individuals have hazard functions that are proportional to each other, i.e. the ratio of $h(t|\underline{x}_1)/h(t|\underline{x}_2)$ for two individuals with regressor vectors $\underline{x}_1$ and $\underline{x}_2$, does not vary with t.

That is to say, the ratio of $h(t|\underline{x}_1)/h(t|\underline{x}_2)$ is not dependent on time t, where $\underline{x}_1$ and $\underline{x}_2$ are vectors containing the values of the corresponding regressor variables. Thus the hazard function of T, given a value for $\underline{x}$, can be written as:

$$h(t|\underline{x}) = h_0(t|\underline{x})\ g(\underline{x})$$

where h_0 and g can include unknown parameters, and h_0 is referred to as the baseline hazard function. The baseline hazard function is the hazard function for an individual with $g(\underline{x}) = 1$. For the Cox model, $g(\underline{x}) = e^{\underline{x\beta}}$, where $\underline{x} = (x_1, x_2, \ldots, x_p)$ is a vector of explanatory regression variables for an individual, which are used to give a better understanding of what is happening in the data (such as in the case of a cancer patient, $\underline{x}$ might include age, tumor size, etc.), and where $\underline{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector of regression coefficients. The vector $\underline{x}$ is then made up the independent variables that are thought to effect the length of life of an individual and the vector

$\underline{\beta}$ gives an idea of how much effect each variable has on the length of life of an individual (length of life is referred to as the response variable).

Definition 1.2.14

The Cox Model is the proportional hazard model with $g(\underline{x}) = e^{\underline{x\beta}}$, i.e. the Cox model is

$$h(t|\underline{x}) = h_0(t|\underline{x})\ e^{\underline{x\beta}} .$$

Cox (1972) introduced a new approach to the analysis of proportional hazard models. Before then it was always necessary to be able to get a good handle on the underlying distribution associated with the failure time data. If the wrong distribution was selected, that caused serious problems in the analysis. Sometimes this underlying distribution is well known, but other times it is not known. Cox suggested using an approach that is not dependent on knowing the underlying failure time distribution.

Definition 1.2.14

Distribution-free refers to the analysis of proportional hazards models without knowledge of the underlying baseline hazard function.

In order to obtain extimates for $\underline{\beta}$, using this distribution-free approach, it is necessary to incorporate the use of the partial likelihood function. The partial likelihood function is not the same as is thought of as the usual likelihood function. The *likelihood* function is the joint density of the observed values as a function of the unknown parameters. In order to define the partial likelihood function some other definitions are needed.

Definition 1.2.16

Let $\underline{y}$ be a vector of observations represented by the random variable $\underline{Y}$, which has density $f_y(\underline{y};\underline{\theta})$. Now transform $\underline{Y}$ into two random variables (V,W) in such a way that the transformation doesn't depend on $\underline{\theta}$. Then $f_v(\underline{v};\underline{\theta})$ is referred to as the marginal likelihood function based on V, and $f_{w|v}(w|v;\underline{\theta})$ is referred to as the conditional likelihood function based on W given $V = v$.

These are both special cases of the partial likelihood definition.

Definition 1.2.17

Let Y be transformed into $(X_1,S_1,X_2,S_2,\ldots,X_m,S_m)$, the full likelihood function of the sequence is

$$\prod_{j=1}^{m} f_{X_j|X^{(j-1)},S^{(j-1)}}(x_j|x^{(j-1)},s^{(j-1)};\theta) \prod_{j=1}^{m} f_{S_j|X^{(j)},S^{(j-1)}}(s_j|x^{(j)},s^{(j-1)};\theta),$$

where $x^{(j)} = (x_1,x_2,\ldots,x_j)$ and $s^{(j)} = (s_1,s_2,\ldots,s_j)$. The second product is referred to as the partial likelihood based on S in the sequence (X_j,S_j).

In general it is not easy to come up with the partial likelihood function. The purpose is to separate the parameters of interest from the nuisance parameters in the full likelihood function. In this problem it means working around the baseline hazard function.

Definition 1.2.18

For the Cox regression model, the partial likelihood function is

$$L(\underline{\beta}) = \prod_{i=1}^{k} [\exp(\underline{x}_i\underline{\beta}) / \sum_{\ell\varepsilon R_i} \exp(\underline{x}_l\underline{\beta})]$$

where k = total number of distinct time intervals,

$\underline{x}_i$ = denotes the regressor variables for the ith individual,

$\ell \varepsilon R_i$ is those individuals at risk in time interval t_i,

where $t_1 < \ldots < t_k$.

Next to be looked at is the literature review for the Cox Regression Model.

1.3 Literature Review

D. R. Cox (1972) introduced a distribution free method useful in the analysis of the proportional hazard model. The proposed approach is suited to handle censored failure times and a small number of ties. He also proposed an analysis in discrete time, designed to handle more ties. In the 1972 paper this involved the use of a conditional likelihood function to come up with estimate for the regressor coefficients.

In the discussion that followed the 1972 paper, Kalbfleisch and Prentice along with a number of other discussants raised questions about the use of the conditional likelihood function. They proposed the use of a marginal likelihood function. In 1973, they came out with an expanded version of this approach along with an alternate way of handling ties.

Finally, Cox (1975) addressed this topic once again. He introduced the partial likelihood function. As it turned out both the conditional likelihood and the marginal likelihood are special cases of the partial likelihood. The purpose of the partial likelihood is to divide the regular likelihood function so that the part that deals only with $\underline{\beta}$ can be isolated separate from the nuisance parameters.

Lawless (1981) gives a justification for the use of all three of these likelihood functions. In the following section each type of likelihood function will be looked at and some uses and examples provided.

1.4 General Procedures

The proportional hazard model proposed by Cox (1972) is

$$h\ (t;\underline{x}) = \exp\ (\underline{x}\ \underline{\beta})\ h_0(t).$$

There are several approaches as to how one can analyze the Cox regression model.

The simplest approach is to assume that $h_0(t)$ is constant, which implies an underlying failure distribution is exponential. Lawless (1981) discusses the approach in chapter 6 of his book. Another parametric approach discussed by Lawless involves the Weibull distribution. He used standard estimation methods such as maximum likelihood. The advantage of the ML approach is that both the probability density and survivor function can generally be found easily. Another possibility is to restrict $h_0(t)$ qualitatively, by assuming it to be monotone. All of these above procedures are mentioned by Cox (1972), but he also concentrates on another approach that is considered next.

The emphasis of Cox's analysis was to be able to get good estimates of the regression parameters. His approach involves leaving $h_0(t)$ arbitrary or even completely unknown. The nuisance parameters are contained in the hazard function $h_0(t)$. Cox states that it seems possible that the loss of information about $\underline{\beta}$ when $h_0(t)$ is left arbitrary is most often slight, if this is true then the procedure to be discussed here is a reasonable way of looking at $\underline{\beta}$. A major problem that would arise from leaving $h_0(t)$ arbitrary or completely unknown is there could be a major loss of efficiency in estimating $\underline{\beta}$. A couple of authors have addressed this problem and their findings are discussed later in this section.

Assume that $h_0(t)$ is arbitrary to the point where no information can be obtained about $\underline{\beta}$ in an interval in which no failures occur, as $h_0(t) = 1$. Cox uses an argument conditional on the set of t_i's. By conditioning on the risk set R_i, he defined a conditional likelihood function,

$$L_1(\underline{\beta}) = \sum_{i=1}^{k} [\exp(\underline{x}_i\underline{\beta})/ \sum_{\ell\varepsilon R_i} \exp(\underline{x}_\ell\underline{\beta})] \quad ,$$

where $\underline{x}$ is the vector of regressor variable for the ith individual.

$\underline{\beta}$ is the vector of regressor coefficients.

k the number of time intervals and R_i denotes the set of all individuals alive at time interval t_i.

Cox (1975) points out that he incorrectly referred to this as a conditional likelihood function. Thus more accurately this should be referred to as a partial likelihood function. Also Kalbfleisch and Prentice (1973) show that the above function is a marginal likelihood function of ranks under the assumptions of no censoring and that $\underline{x}$ doesn't depend on time. Cox (1975) analysis of the partial likelihood has a wider range of application than just the proportional hazards model, but that is the only case considered here.

Before going on to the actual estimation of $\underline{\beta}$ for this partial likelihood the definition for partial likelihood as given by V. T. Farewall (1979) is presented.

Efron (1977) looked at the efficiency of this partial likelihood for censored data. In section 3 of his paper, it is shown that if the class of nuisance functions $h_0(t)$ is moderately large, then inferences about $\underline{\beta}$ based on the partial likelihood are asymptotically equivalent to those based on all of the data, which solves the major problems of this procedure as were presented in the discussion that followed Cox (1972). In addition, Tsaitis (1981) proves the asymptotic consistency and normality of the maximum partial likelihood estimator. The next topic dealt with is estimation for the Cox regression model.

The unknown components for the Cox model are the regression parameters and the baseline hazard function. It is possible to think of the baseline hazard in terms of the baseline survivor function.

$$S_0(t) = \exp(-\int_0^t h_0(u)\ du) = \exp[-H_0(t)]$$

where $H_0(t)$ is the baseline cummulative hazard function. The survivor function for T, given $\underline{x}$, (Lawless (1981)) can be written as

$$S(t;\underline{x}) = \exp(-\int_0^t h(u;\underline{x})\ du) = [S_0(t)]^{\exp(\underline{x\beta})} \quad .$$

The goal then is to get an estimate for $\underline{\beta}$ when $h_0(t)$ is unknown and test equality of survivor functions in the m sample case.

The estimator of $\underline{\beta}$ is obtained by maximizing the partial likelihood equation. This function allows for Type II censoring, but adjustments are needed for ties. In order to handle more than a small number of ties, it is necessary to modify the partial likelihood function $L_1(\underline{\beta})$ as

$$L_2(\underline{\beta}) = \prod_{i=1}^{k} [\exp(\underline{S}_i\underline{\beta})/(\sum_{\ell\varepsilon R_i} \exp(\underline{x}_\ell\underline{\beta}))^{d_i}] ,$$

where d_i is the number of lifetimes equal to t_i and S_i is the sum of the regression vector $\underline{x}$ for these d_i individuals. If all d_i are equal to one then $L_2(\underline{\beta})$ reduces to $L_1(\underline{\beta})$. The log partial likelihood function is

$$\log L_2(\underline{\beta}) = \sum_{i=1}^{k} \underline{S}_i\underline{\beta} - \sum_{i=1}^{k} d_i\log(\sum_{\ell\varepsilon R_i} \exp(\underline{x}\ \underline{\beta})), \quad (3)$$

and the first derivatives of Log L_2 with respect to β_r $r = 1,\dots,p$, are

$$\frac{\partial\log L_2(\underline{\beta})}{\partial\beta_r} = \sum_{i=1}^{k} [S_{ir} - d_i \sum_{\ell\varepsilon R_i} x_{\ell r}\exp(\underline{x}_\ell\underline{\beta})/ \sum_{\ell\varepsilon R_i} \exp(\underline{x}_\ell\underline{\beta})]$$

where p = the number explanatory variables being used, and S_{ir} in the rth component in $S_i = (S_{ir}\ \dots,\ S_{ip})$. The matrix $\underline{I}$, which contains the negative of the second partial derivatives of log $L_2(\underline{\beta})$ has entries,

$$I_{rs}(\underline{\beta}) = \frac{-\partial^2\log L_2(\underline{\beta})}{\partial\beta_r\partial\beta_s} = \sum_{i=1}^{k} d_i[\sum_{\ell\varepsilon R_i} x_{\ell r}x_{\ell s}\exp(\underline{x}_\ell\underline{\beta})/ \sum_{\ell\varepsilon R_i} \exp(\underline{x}_\ell\underline{\beta})$$

$$- (\sum_{\ell} x_{\ell r}\exp(\underline{x}_\ell\underline{\beta}))(\sum_{\ell} x_{\ell s}\exp(\underline{x}_\ell\underline{\beta}))/(\sum_{\ell} \exp(\underline{x}_\ell\underline{\beta}))^2].$$

The maximum likelihood equations $\partial \log L / \partial \beta_r = 0$ $(r = 1,\ldots,p)$ can generally be solved without difficulty by using the Newton-Raphson method.

There are a number of programs available that calculate the value of $\underline{\hat{\beta}}$ for the partial likelihood function using the Newton-Raphson method. SAS has two programs that do this COXREGR and PHGLM. Both of these programs yield similar results, but PHGLM has more options, so it was used in the example that follows the next section. BMD also has a program P2L that computes the value for $\underline{\hat{\beta}}$. One other thing that is available in these two programs, is a stepwise procedure that helps to determine which explanatory variables are significant in describing the failure time of an individual. This is much the same as the stepwise procedure in ordinary regression. Because the BMD stepwise procedure is expensive to run for more than a few independent variables PHGLM was used first. Then the BMD procedure was started using the variables that were included by PHGLM. Before considering examples the procedure for comparing two or more survivor functions is discussed.

Consider the comparison of two life distributions, by testing the hypothesis

$$H_0: S_1(t) = S_2(t)$$

that the two survival distributions are the same. This can be done by assuming that all observations come from a single population and defining an indicator regressor variable x,

$$x = \begin{bmatrix} 0 & \text{if from population} & 1 \\ 1 & \text{if from population} & 2 \end{bmatrix} .$$

The resulting hazard functions are then $h_1(t) = h_0(t)$ and $h_2(t) = h_0(t)\, e^{\beta}$, Thus the two are identical if and only if $\beta = 0$. This is the same as saying

$$S_2(t) = S_1(t)^{\exp(\beta)} .$$

So by testing $\beta = 0$, it is the same as testing

$$H: S_2(t) = S_1(t) \text{ vs } H_1 = S_2(t) = S_1(t)^{\delta},\ \delta \neq 1$$

where $\delta = \exp(\beta)$.

Let K be the number of distinct observed failure times and let d_i represent the number of failures at time t_i. Also let n_{1i} and n_{2i} denote the number of individuals in the risk set R_i at time t_i, from each population and d_{1i} and d_{2i} are the number of deaths for each population at time t_i. In addition let $n_{1i} + n_{2i} = n_i$ and $d_{1i} + d_{2i} = d_i$ and $r_2 = \sum_{i=1}^{k} d_{2i}$ is the total number of deaths from population 2. Then the log likelihood of $L_2(\beta)$ becomes

$$\log L_2(\beta) = r_2 - \sum_{i=1}^{k} d_i \log(n_{1i} + n_{2i}e^{\beta}) .$$

The first derivative and negative of the second derivative of log $L_2(\beta)$ are

$$U(\beta) = \frac{\partial \log L_2(\beta)}{\partial \beta} = r_2 - \sum_{i=1}^{k} \frac{d_i n_{2i} e^{\beta}}{n_{1i} + n_{2i}e^{\beta}} ,$$

and

$$I(\beta) = \frac{-\partial^2 \log L_2(\beta)}{\partial\beta^2} = \sum_{i=1}^{k} \frac{d_i n_{1i} n_{2i} e^{\beta}}{n_{1i} + n_{2i}e^{\beta}} \quad .$$

The likelihood equation $U(\beta) = 0$ can be solved by Newton Raphson Method. Confidence intervals and test statistics can be obtained by the likelihood ratio method by treating β as approximately normal with mean β and variance $I(\beta)^{-1}$ or by treating $U(\beta)$ as normal with mean 0 and variance $I(\beta)$.

$U(\beta)$ gives a simple test for the equality of $S_1(t)$ and $S_2(t)$ without having to calculate $\hat{\beta}$. Thus for $H_0{:}\beta = 0$ the test statistic is

$$Z = \frac{U(0) - 0}{\text{standard deviation}} = \frac{U(0)}{(I(0))^{1/2}} \quad ,$$

where

$$U(0) = r_2 - \sum_{i=1}^{k} d_i \frac{n_{2i}}{n_i} \quad , \quad \text{and}$$

$$I_1(0) = \sum_{i=1}^{k} \frac{d_i n_{1i} n_{2i}}{n_i^2} \quad .$$

If there are a substantial number of ties, then $U(0)$ remains the same, but $I_1(0)$ is replaced by

$$I_2(0) = \sum_{i=1}^{k} \frac{d_i(n_i - d_i)n_{1i}n_{2i}}{n_i^2(n_i - 1)} \quad .$$

$I_2(0)$ reduces to $I_1(0)$ when all $d_i = 1$. A useful way of looking at $U(0)$ is by rewriting it as

$$U(0) = \sum_{i=1}^{k} (d_{2i} - \frac{d_i n_{2i}}{n_i}) .$$

If $\frac{d_i n_{2i}}{n_i}$ is thought of as the expected value of the number of deaths from population 2 at time t_i, then $U(0)$ is the sum of the differences between the expected and the observed values over the k distinct failure times.

To test that more than two lifetime distributions are equal, the 2 sample test can be expanded. Define a vector of $m-1$ indicator regressor variables $\underline{x} = (x_1,\ldots,x_{m-1})$ as follows, individuals in population $1,\ldots,m-1$ have vectors $\underline{x} = (1,0,\ldots,0),\ldots,(0,\ldots,0,1)$, respectively, and individuals in population m have $\underline{x} = (0,\ldots,0)$. Assume the survivor functions $S_i(t)$ are given by (3) where $i = 1,\ldots m$, so that

$$S_1(t) = S_0(t)^{\delta_1},\ldots,S_{m-1}(t) = S_0(t)^{\delta_{m-1}}$$

where $\delta_i = \exp(\beta_i)$. In testing the equality of these m distributions, it is the same as testing $\underline{\beta} = (\beta_1,\ldots,\beta_{m-1}) = \underline{0}$. Let N be the total number of individuals, and N_r is the total number in the rth population. In the total sample of N there are k distinct failure times $t_1 < \ldots < t_k$. Let n_{ri} be the number at risk at time t_i in the rth population and $\sum_i n_{ri} = n_i$ where n_i is the total number at risk at time t_i (this would be the total number alive at time t_i, except some of the observations maybe censored). Also let d_{ri} be the number of deaths in population r at time t_i and $\sum_i d_{ri} = d_r$. Thus the first derivatives and information matrix with $\underline{\beta} = \underline{0}$ is

$$U_r(\underline{0}) = \left(\frac{\partial \log L_2(\underline{\beta})}{\partial \beta_r}\right)_{\underline{\beta}=0} = \sum_{i=1}^{k} (d_{ri} - \frac{d_i n_{ri}}{n_i}) \quad r = 1,\ldots,m-1, \text{ and}$$

$$I_{rs}(\underline{0}) = \sum_{i=1}^{k} d_i \frac{n_{ri}}{n_i}(\delta_{rs} - \frac{n_{si}}{n_i}) \quad r,s = 1,\ldots,m - 1.$$

where $\delta_{rs} = 1$ if $r = s$, or $\delta_{rs} = 0$ if $r \neq s$. Under the hypothesis $\underline{\beta} = \underline{0}$, $\underline{U} = (U_1(\underline{0}),\ldots,U_{m-1}(\underline{0}))$ can be treated as approximately normal with mean $\underline{0}$ and covariance matrix $\underline{I}(\underline{0})$. The test statistic for $\underline{\beta} = \underline{0}$ is

$$\chi^2 = \underline{U}'\underline{I}(\underline{0})^{-1}\underline{U}.$$

Thus the hypothesis of equal distributions for the m populations would be rejected for values greater than $\chi^2_{\alpha}(m-1)$. As in the 2 sample case if the number of ties is substantial $\underline{I}_{rs}(\underline{0})$ needs to be adjusted by setting

$$I_{rs}(\underline{0}) = \sum_{i=1}^{k} \frac{d_i(n_i-d_i)n_{ri}}{n_i(n_i-1)}(\delta_{rs} - \frac{n_{si}}{n_i}) \quad r,s = 1,\ldots,m-1 \quad .$$

This test can be done equivalently with the SAS program SERVTEST, although this exact procedure is not followed, it does yield test statistics for equal survival functions. The section that follows gives examples of this procedure, along with an example for the procedures discussed in section 2.2 for estimation of $\underline{\beta}$.

1.5 Examples

The first example was taken from the BMD manual (1981) and originally was reported by Krall, Uthoff and Harley (1975) and is given on page 58. The data consists of the survival times of 65 multiple mayeloma patients with 16 explanatory variables. Forty-eight of the observations represent deaths and 17 individuals were censored. Initially it is desired to determine which of these explanatory variables are contributing to the estimates of the survival times. This is done by using the STEPWISE procedure available in SAS PROC PHGLM first, then using the model arrived at under this approach as the starting model for the BMD program P2L. The Key for the variables used in these programs is also given on page · 56. The PROC PHGLM has two options for model building, STEPWISE and BACKWARD. STEPWISE starts with no variables in the model and adds the most sign variables one at a time as long as $P \leq .10$ for at least one variable. If all P-values are greater than .10 it says that no variable gives significant information about the survival time for an individual. After a variable is added the procedure then goes back and checks to see if all of the variables meet the .05 level for staying in the model. In order to use PROC PHGLM, it is necessary to sort the data in descending order by the survival times. PROC SORT will sort the data in descending order simply by putting descending before the variable survival. In this example the only variable that was significant by the STEPWISE option was LOGBUN. The BACKWARD option was used next. This approach starts with all of the variables in the model and removes the least sign variables if its P-value is greater than .05. The model selected by STEPWISE in this example included only the variable LOGBUN, so the BACKWARD approach was also used. The model selected by the BACKWARD approach included the variables LOGBUN, PLATELET, INFEC, LOGWBC, PROTEIN, BJP, TSP and SGLOBIN. This program also prints out the estimates for $\underline{\beta}$, their standard errors, and their P-values for their individual significances to the model. Then this model was used as the starting model for the program BMDP2L STEPWISE procedure. This is accomplished by use of the START = statement. The STEPWISE procedure used is MPLR, which stands for the maximum partial likelihood ratio. The enter limit for this model is .10 the same as for PROC PHGLM, but the remove limit is .15. The model entered was not changed. Besides printing out estimates for $\underline{\beta}$, their standard errors and P-values as before, the procedure also prints out estimates for the survival, hazard and commulative hazard functions at $\underline{x} = \underline{\bar{x}}$. These results are on pages 66 - 67. From all of this it can be concluded that when LOGBUN, INFEC, LOGWBC, PROTEIN, BJP and TSP are increased they reduce the probability of survival for an individual. Also that as variables PLATELET and SGLOBIN are increased they add to the probability of survival. Recall that the proportional hazard model is

$$h(t|\underline{x}) = h_0(t)\, e^{\underline{x\beta}} ,$$

Thus if $e^{x_i\beta_i}$ is greater than one, it means that the ith variable increases the instantaneous rate of death at time t given that the individual is still alive at time t, i.e. $h(t|\underline{x})$ is increased. Similarly if $e^{x_i\beta_i}$ is less than one, it means that the ith variable reduces the instantaneous rate of death at time t given that the individual is still alive at time t. It is always true for this example that $e^{x_i\beta_i}$ is greater than or equal to one if β_i is greater than or equal to zero, and $e^{x_i\beta_i}$ is less than or equal to one if β_i is less than or equal to zero. Thus since the β-values for PLATELET and SGLOBIN are less than zero, these are the only values that increase the probability of survival, as the value of x_i increases for these variables.

The probability of survival at given survival times are on pages 69 - 70. Also included on these pages are some summary statistics like cumulative deaths and those that remain at risk for each time period, plus the estimates for the hazard function and cumulative hazard function evaluated at $\underline{x} = \underline{\bar{x}}$, just as the probability of survival is evaluated at $\underline{x} = \underline{\bar{x}}$. Thus when t = 6.00, the estimated probability of an individual whose regression vector is equal to $\underline{x} = \underline{\bar{x}}$, not dying until after time t is .8780. The estimated rate of instantaneous death at time t for this individual is .0561 and the estimated cumulative hazard rate is .1302. The computer programs and results for this problem are on pages 57 - 70.

The next example deals with the test for equal survivor functions. It was taken from Cox (1972). This test was done using SAS PROC SERVTEST. The procedure used is not exactly the same as the procedure described in Section 2.2, but yields similar results. The data for this problem is on the bottom of the page. The SURVTEST procedure tests for differences between for two or more survivor functions. Three tests are performed on this data:

1) Gehan-Wilcoxon test (a permutational test based on ranks)
2) logrank test (equivalent to Mantel-Haenzel)
3) likelihood ratio test (based on the exponential model).

Times of remission (weeks) of leukemia patients
(Gehan, 1965, from Freireich et al.)

Sample 0 (drug 6-MP)	6*, 6, 6, 6, 7, 9*, 10*, 10, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*
Sample 1 (control)	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

* Censored

For each test, a score sum is reported and a two-tailed test is printed on page 73, also the SAS program is printed here. As it turns out all three tests have P-values less than .01, it can be concluded that these two treatments do not have equal survivor functions. It would appear that treatment 1 does increase the survival rate over the control group. In the next chapter cross-classified categorical failure data is studied.

CHAPTER 2

The Loglinear Model

2.1 Introduction

Categorical failure data is discrete data generated by observing whether an experimental unit survived (lived) or failed (died), during the course of the study or possibly in time intervals such as 1-5 years, 6-10 years, and more than 10 years. For example, a two-way table is obtained when the researcher observes whether individuals treated by one of two treatments lived or died. The expected number of experimental units in the i,j cell, denoted by $\hat{M}_{ij}$ can be modeled or a function of the effects of each variable plus the effect of the combination of the two variables. The model then has the form $M_{ij} = e^{u + u_{1(i)} + u_{2(j)} + u_{12(ij)}}$, where u is the overall mean effect, $u_{1(i)}$ is the effect of variable 1, $u_{2(j)}$ is the effect of variable 2 and $u_{12(ij)}$ is the interaction of the effects of variable 1 and variable 2. The probability of survival is then marginal total of expected cell frequencies, M_1 , divided by the overall sample size. The following are examples of where the loglinear model can be applied.

Example 1. A researcher desires to know if vitamin C really does reduce the probability of catching a cold (a failure), this example was taken from Feinberg (1981), originally taken from Pauling (1971). The probability of catching a cold (failure) is modeled as a function of what treatment the skiers received, either placebo or vitamin C.

Example 2. It is desired to know what effect boot fittings and the blood-alcohol level of an individual have on whether a skier has an accident (failure) or no accident (survival). The rate of accidents (failures) can be modeled as a function of type of boot fittings and the blood-alcohol level of an individual using a loglinear model.

The remainder of Chapter 2 contains the literature review, general procedure and an example for the loglinear model.

2.2 Definitions

The loglinear model has as one of its uses the analysis of cross-classified categorical failure data. The following definitions, in part, come from Feinberg (1981).

Definition 2.2.1

Categorical Data is data that has been classified into distinct groups for each variable.

There are a variety of values that categorical variables can take on. For example for categorical failure data, there are two kinds that are looked at. This type of failure data is of the discrete form as opposed to the continuous type that is dealt with for the Cox regression model. Categorical variables that take on one of two values (such as survived or died) is referred to as dichotomous, one that can take on one of three or more values is referred to as polytomous.

Definition 2.2.2

Cross-classified categorical data is data that has been classified such that every level of one variable occurs with every level of each of the other variables.

Thus for the case when two variables are used, cross-classified corresponds to a usual contingency table. The cells then contain the number of individuals that fall into each particular combination of variable 1 and 2. The object of this approach is to model the expected cell frequencies using a loglinear approach. For the two-way table the test statistic for the loglinear model is asymptotically equivalent to the χ^2 test statistic for the test of independence in a two-way contingency table, where

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{observed cell frequences - expected cell frequencies})^2}{\text{expected cell frequencies}} .$$

The main difference lies in the fact that once past the two-way table to a three-way table, it is still possible to use this loglinear model test statistic to test for complete independence among all three variables, plus test for other possible types of independence between the three variables. These other types are discussed in Section 2.A.

In order to define the loglinear model for the two-way table as defined by Feinberg (1981), it is first necessary to present some notation. The expected cell frequencies under the independence assumption, denoted by $\hat{m}_{ij}$, for a two-way contingency table, where $i=1,\ldots,I$, $j=1,\ldots,J$ are $\hat{m}_{ij} = \frac{n_{i.}n_{.j}}{n_{..}}$ where $n_{i}.$, $n._{j}$, and $n..$ are shown in Table 1. By taking the logarithm of both sides of $\hat{m}_{ij} = \frac{n_{i.}n_{.j}}{n_{..}}$,

$$\log \hat{m}_{ij} = \log n_{i}. + \log n._{j} - \log n.. \qquad .$$

Definition 2.2.3

The loglinear model for categorical data from a two-way table is

$$\log m_{ij} = u + u_{i(1)} + u_{j(2)},$$

where $u = \frac{1}{IJ} \sum_i \sum_j \log m_{ij}$ is the overall mean,

$u_{1(i)} = \frac{1}{J} \sum_j \log m_{ij} - u$, corresponds to the main effect of Variable 1

$u_{2(j)} = \frac{1}{I} \sum_i \log m_{ij} - u$, corresponds to the main effect of variable 2.

Table 1

Population (factors)	Frequency Distribution of Data Category Response				
	1	2	. . .	J	Total
1	n_{11}	n_{12}		n_{iJ}	$n_{1}\cdot$
2	n_{21}	$n_{22}\cdot$	. .	n_{2J}	$n_{2}\cdot$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
3	n_{I1}	$n_{I2}\cdot$	. .	n_{IJ}	$n_{I}\cdot$
	$n\cdot_{1}$	$n\cdot_{2}\cdot$	. .	$n\cdot_{J}$	$n\cdot\cdot$

For the three way table, the expected frequencies are $\hat{m}_{ijk} = \frac{n_{i}\cdot\cdot n\cdot_{j}\cdot n\cdot\cdot_{k}}{n...}$, where $n...$ is the overall sample size and $n\cdot_{j}\cdot$ involves summing over variables 1 and 3 for each level of variable 2 and likewise for $n_{i}\cdot\cdot$ and $n\cdot\cdot_{k}$. Taking the log of both sides of the above equations for expected frequencies yields,

$$\log \hat{m}_{ijk} = \log n_{i}\cdot\cdot + \log n\cdot_{j}\cdot + \log n\cdot\cdot_{k} - \log n... \quad .$$

Definition 2.2.4

The saturated (meaning all interactions and main effects are included) loglinear model for categorical data for the three way table is

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)} \qquad (2)$$

where $i = 1,\ldots,I$, $j = 1,\ldots,J$, $k = 1,\ldots,K$, and

$u = \frac{1}{IJK} \sum\sum\sum_{ijk} \log m_{ijk}$, is the overall mean,

$u_{1(i)} = \frac{1}{JK} \sum_j \sum_k \log m_{ijk} - u$ measures the main effect of variable 1,

.
.
.

$u_{12(ij)} = \frac{1}{K} \sum_k \log m_{ijk} - (u + u_{1(i)} + u_{2(j)})$ measures the interaction between variable 1 and variable 2,

.
.
.

$$u_{123(ijk)} = \log m_{ijk} - (u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)})$$

measures the interaction between all three variables.

Restrictions are explicitly introduced into the parameter definitions as

$$\sum_{i=1}^{I} u_{1(i)} = \sum_{j=1}^{J} u_{2(j)} = \sum_{k=1}^{K} u_{3(k)} = 0$$

$$\sum_i u_{12(ij)} = \sum_j u_{12(ij)} = \sum_i u_{13(ik)} = \sum_k u_{13(ik)} = \sum_j u_{23(jk)} = 0$$

$$\sum_k u_{23(jk)} = \sum_i u_{123(ijk)} = \sum_j u_{123(ijk)} = \sum_k u_{123(ijk)} = 0.$$

Estimates can be obtained for each of the u-terms by substituting the $\hat{m}_{ijk}$'s for the m_{ijk}'s in the equations for the u-terms from (2). The above definitions can be expanded to an n-way table.

2.3 Literature Review

Loglinear Model

Fienberg (1981) deals with the analysis of cross-classified categorical data. This procedure can be used for failure time data that has been categorized, although it works for all types of cross-classified data regardless if one of the variables takes on discrete values of the time of failure of an individual or component, or not. The example that is considered in the last section of this chapter does involve a discrete failure time variable. The general procedure used is the loglinear model. Once past a three-way table to a higher order table, it becomes harder to define the types of independence present in the models. The majority of the discussion that follows concerns a three-way table.

The analysis involves the use of maximum likelihood estimates, denoted by MLE, for the expected cell frequencies, based on the selected model. The types of models that can be present in a 3-way table are

1) The model of complete independence, where there is no relationship present between any of the three variables. The model then is the same as the saturated 3-way model with $u_{12} = u_{13} = u_{23} = u_{123} = 0$.
2) The model of joint independence occurs when one variable is independ-

ent of both of the other two variables, this is when $u_{12} = u_{13} = u_{123} = 0$ for example from the saturated model and if this model fits well implies variable 1 is jointly independent of variables 2 and 3.

3) The model of conditional independence occurs when fixing one variable at any level, the other two variables are independent, one possibility of this is where $u_{12} = u_{123} = 0$, thus implying if variable 3 is fixed at any level variables 1 and 2 are independent.
4) The model of constant association occurs when each two variable interaction is uneffected by the third variable, here $u_{123} = 0$ and this implies no three-way interaction present.
5) The saturated model is the model selected if none of the above models fit the data well and implies there is three-way interaction present.

The model with the lowest number from above list that adequately fits the data is the one usually selected. In higher order tables it becomes harder to interpret the results in the above manner. Chapter 4 of Fienberg's book deals extensively with the topic of model selection, which is much the same as stepwise model selection in regression. After the model has been selected, a number of procedures for computing the MLE can be used, based on the sampling procedure utilized (this can be either Poisson, multinomial or product multinomial). All three of these procedures provide the same MLE. The method that always guarantees arriving at the MLE for the expected cell frequencies is an iterative approach. The program BMDP4F is capable of doing the iterations for computing the MLE for the expected cell frequencies and also calculating the goodness of fit test for each model of interest. Once a model is selected, then the expected cell frequencies and estimates for the u-terms can be printed out.

2.4 General Procedure

Loglinear Model

The saturated model for the three way table was given in Section 2.2. If for this model all of the two-way interactions and the three-way interaction are set equal to zero (i.e. $u_{12} = u_{13} = u_{23} = 0$), the model that is left includes only the overall mean plus the main effects for the three variables. If this model

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} \quad i = 1,\ldots,I,\ j = 1,\ldots,J \quad k = 1,\ldots,K,$$

fits the data well, then it is concluded that the three variables are completely independent. The first step in this analysis is to compute the maximum likelihood estimates (MLE) for the expected cell frequencies. This part is straight forward as it is identical to the methods used for the two-way contingency tables to obtain a test for independence. The expected cell frequencies for the two-way table are the product of the marginal probabilities times the overall sample size, or the product of the marginal frequency counts divided by the over all sample size. Then if the observed values differ too much from the expected the hypothesis of independence is rejected, using a chi-square approximation. The same is true for the test of independence for the three-way table. The initial step is to compute the MLE for the expected cell frequencies. Once again the m_{ijk}'s equal the product of the marginal probabilities and the overall sample size denoted by $n_{...}$, i.e.

$$\hat{m}_{ijk} = \frac{n_{i..}\ n_{.j.}\ n_{..k}}{n_{...}\ n_{...}\ n_{...}}\ n_{...} \qquad i = 1,\ldots,I,\ j = 1,\ldots,J\ k = 1,\ldots,K,$$

where $p_{i\cdot\cdot\cdot} = \frac{n_{i\cdot\cdot}}{n_{\cdot\cdot\cdot}}$ are the marginal probabilities for variable 1, $p_{\cdot j\cdot}$ are the marginal probabilities for variable 2 and $p_{\cdot\cdot k}$ are the marginal probabilities for variable 3. This is a rather simple case since it follows directly from the two-way contingency table for computing the expected cell frequencies. It is always true that the expected cell frequencies for a given model involve the marginals totals for the highest order u-terms present in the model, i.e. the u-terms that correspond to the highest order interactions present in the model. For this model, the marginal totals that corresponded to the main effect u-terms are used. Once the expected cell frequencies are calculated, the general procedure is the same for estimating the values of u-terms and for testing for goodness-of-fit of each model.

For another example of computing expected cell frequencies, consider the model of constant association. This is the model with $u_{123} = 0$, that tests for no three-way interaction. The model for this is

$$\log m_{ijk} = u + u_{1(i)} + u_{2(i)} + u_{3(k)} + u_{12(i)} + u_{13(ik)} + u_{23(jk)}$$

where the u-terms are as defined in Section 2.4 and $i = 1,\ldots,I$, $j = 1,\ldots,J$, $k = 1,\ldots,K$.

The following general procedure is used to get the estimated expected frequencies for all models, including this one;

1) for each variable, determine the highest order effect in the log-linear model involving that variable.
2) compute the observed marginal totals corresponding to the highest order effects in 1) - eg. , $n_{i\cdot k}$ $i = 1,\ldots,I$, $k = 1,\ldots,K$ corresponds to $u_{13(ik)}$ $i = 1,\ldots,I$, $k = 1,\ldots,K$,

3) estimate the expected values for the model using only the sets of observed marginal total from 2), or totals that can be computed from them.

In order to get the MLE for the expected cell frequencies an iterative approach is used. Using the above procedure, the $\hat{m}_{ijk}$'s are a function of $n_{ij\cdot}, n_{i\cdot k}$, and $n_{\cdot jk}$, since these are the marginal totals that correspond to the highest order effects in the model. According to Appendix II in Fienberg (1980), using the maximum likelihood method provides estimates, $\hat{m}_{ijk}$, which satisfy:

$$\begin{array}{lll} \hat{m}_{ij\cdot} = n_{ij\cdot} & \text{for all} & i,j \\ \hat{m}_{i\cdot j} = n_{i\cdot k} & \text{for all} & i,k \\ \hat{m}_{\cdot jk} = n_{\cdot jk} & \text{for all} & j,k. \end{array}$$

The following iterative procedure yields the MLE

1) Set $\hat{m}_{ijk}^{(0)} = 1$ for all i,j,k.

Then for $v = 0$ compute

2)

$$\hat{m}_{ijk}^{(3v+1)} = \frac{n_{ij\cdot}}{\hat{m}_{ij\cdot}^{(3v)}} \hat{m}_{ijk}^{(3v)}$$

the values for $\hat{m}_{ij\cdot}^{(3v)}$ are calculated by using the values of $\hat{m}_{ijk}^{(3v)}$,

3)

$$\hat{m}_{ijk}^{(3v+2)} = \frac{n_{i\cdot k}}{\hat{m}_{i\cdot k}^{(3v+1)}} \hat{m}_{ijk}^{(3v+1)}$$

the values for $\hat{m}_{\cdot jk}^{(3v+1)}$ are calculated by using the values of $\hat{m}_{ijk}^{(3v+1)}$ from 2)

4)

$$\hat{m}_{ijk}^{3(v+1)} = \frac{n_{\cdot jk}}{\hat{m}_{\cdot jk}^{(3v+2)}} \hat{m}_{ijk}^{(3v+2)}$$

the values for $\hat{m}_{\cdot jk}^{(3v+2)}$ are calculated by using the values of $\hat{m}_{ijk}^{(3v+2)}$ from 3).

The above procedure is repeated, v being increased by 1 each time through the cycle, until these three equations are satisfied to the desired decimal place. This means that $\hat{m}_{ij\cdot}^{3(v+1)}$ must be within say .1 of $n_{ij\cdot}$, $\hat{m}_{i\cdot j}^{3(v+1)}$ must be within .1 of $n_{i\cdot j}$ and $\hat{m}_{ijk}^{3(v+1)}$ must also be within .1 of $n_{\cdot jk}$ at the end of a given cycle in order for the procedure to stop.

After the $\hat{m}_{ijk}$'s for any model have been calculated, the next step is to test for goodness-of-fit for that model. The method used is the likelihood-ratio statistic, where

$$G^2 = 2 \sum_{\substack{\text{all} \\ \text{cells}}} (\text{observed}) \log(\frac{\text{observed}}{\text{expected}}) = 2 \sum_{\substack{\text{all} \\ \text{cells}}} n_{ijk} \log(\frac{n_{ijk}}{\hat{m}_{ijk}}) \ .$$

This test statistic has an approximate χ^2-distribution with degrees of freedom as follows

d.f. = (number of cells) - (number of parameters fitted).

The degrees of freedom for the model of complete independence are

d.f. = IJK - (1 + (I-1) + (J-1) + (K-1)).

The degrees of freedom for the model of constant association from above are

$$\begin{aligned} \text{d.f.} &= IJK - (1+(I-1)+(J-1)+(K-1)+(I-1)(J-1)+(I-1)(K-1)+(J-1)(K-1) \\ &= (K-1)(IJ-1). \end{aligned}$$

If the model of constant association fits the data, then u_{123} is concluded to equal zero, i.e., that there is no three-way interaction present.

Using the computer program BMD4F, it is easy to fit all possible models for the three-way table. Then select the simplest model that has an adequate fit. Once a model has been selected, the estimates for the u-terms (model parameters) can be obtained by using the $\hat{m}_{ijk}$'s from the appropriate model. These estimates can be obtained from BMDP4F.

In the case of four-way and higher order tables, it becomes costly to fit all possible models. The program BMDP4F has available a procedure called Stepwise. There are two ways to go about looking for an adequate or supposed 'best' model. First by starting with the simplest model of interest, BMDP4F stepwise will add the most significant higher order term not previously included in the model, until it comes up with a model that adequately fits the data. The other approach starts with the model of constant association and deletes the least significant u-term. The approach stops after the model no longer fits the data for some predetermined level of significance. It is then necessary to go back to the model that fits the data adequately. These approaches do not necessarily arrive at the same model. It is then up to the experimenter to choose the better model. Also by obtaining the estimates for the u-terms of the saturated model, one might be better able to get a feel for which u-terms are important. Thus by using these u-terms it can help the experimenter get a better idea of what model to choose for the start of the additive stepwise approach.

An example is given in the next section of a four-way table for the failure time of patients with breast cancer. The stepwise procedure could be used, if it was desired to come up with a 'best' model.

2.5 Examples

The first example uses lifetime data to illustrate the applications of the loglinear model as proposed by Fienberg (1980). The data was obtained from the BMDP manual (1981), (the original data comes from Morrison et. al. (1973)) and was also analyzed in Bishop et. al. (1975). The data deals with the three year survival of breast cancer patients with variables age, diagnostic center, and inflammation summarized as follows

Variable 1. Degree of inflammatory reation and appearance

1. minimal - malignant appearance
2. minimal - benign appearance
3. greater - malignant
4. greater - benign

Variable 2. Survival for three years

1. No
2. Yes

Variable 3. Age at diagnosis

1. under 50 years
2. 50 - 69 years
3. 70 or older

Variable 4. Center where patient was diagnosed

1. Tokyo
2. Boston
3. Glamorgan

It is of interest to the experimenters to know the effect of the first variable in the survival at different centers. Bishop et. al. (1975) proposed a couple of models to test this, first, model A is,

$$\log m_{ijkl} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(l)} + u_{12(ij)} + u_{23(jk)}$$

$$+ u_{14(il)} + u_{24(jl)} + u_{34(kl)} + u_{124(ikl)} \quad ,$$

where $i = 1,\dots,I$, $j = 1,\dots,J$, $k = 1,\dots,K$, $l = 1,\dots,L$.
This model has all three-way interactions, except u_{124} set equal to zero, plus $u_{1234} = 0$. If this model fits, it says that there is no three or four-way interaction present, except possibly u_{124}, the three-way interaction of interest. The next model tests for no three or four-way interaction at all.

Model B is

$$\log m_{ijkl} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(l)} + u_{12(ij)} + u_{23(jk)}$$

$$+ u_{24(jl)} + u_{34(kl)} + u_{14(il)} \quad ,$$

where $i = 1,\dots,I$, $j = 1,\dots,J$, $k = 1,\dots,K$, $l = 1,\dots,L$.
These two models differ by only the term, $u_{124(ijl)}$. Thus if both models give an adequate fit, it can be concluded that $u_{124} = 0$, and that it is not necessary for a good fit of the model. It also says that if age is fixed then the relationship between survival and inflammation doesn't differ among centers. By looking at the X^2(model B) minus X^2(model A), this difference

is a X^2 value for the significance of u_{124}, since it is the only term different in the two models. To actually test the significance of u_{124}, it is necessary to compare this X^2 value to a χ^2 percentage point with degrees of freedom equal to d.f. model B minus d.f. model A. The results of the analysis on page 77, indicates that both models provide an adequate fit. Model A has $X^2 = 36.21$ with 40 d.f. and a P-value equal to .6418 and model B has $X^2 = 41.39$ with 46 d.f. and a P-value equal to .6654, thus for both models the null hypothesis of an adequate fit can not be rejected. The X^2 value for the diffence is $41.39 - 36.21 = 5.18$ with $46 - 40 = 6$ d.f. Thus the effect of u_{124} is not significant implying $u_{124} = 0$, and it can be concluded that when age is fixed that the interaction of survival and inflammation doesn't differ among centers.

Other hypotheses of interest can be tested in a similar fashion. The X^2 values for the tests of complete independence and of constant association are also printed out. From the P-values it can be concluded that the model of complete independence does not give an adequate fit and that the model of constant association does provide and adequate fit. It is not surprising that this last model fits since it implies that there is no four-way interaction and from model B it was concluded already that no three or four-way interactions were present. The next chapter considers the GSK model.

Chapter 3

The GSK Model

3.1 Introduction

This model was proposed by Grizzle, Starmer, and Koch (1969). The t-year survival rate or probability that an individual with a disease is alive t-years from the time of diagnosis is used to evaluate the effectiveness of a given type of therapy. Clinical trials have been used to provide information for the estimation of t-year survival rates and to compare survival rates for different types of treatments. When some of the patients are not traceable for the entire t-years,due to reasons unrelated to the treatment and/or the disease, then if the lost to follow up or withdraw patients are removed from the analysis the estimates of the survival rates will be biased (Koch et. al. (1972). Also, in a clinical trial the patients can be classified according to several other variables such as sex, age, extent of disease, etc., and the survival rates may differ between classifications. The loglinear model enables the researcher to incorporate the withdrawn patients information as well as estimate the survival rates after accounting for the categorical effects of variables such as age and sex. For example, a two-way table is obtained when the researcher observes whether an individual lives, dies, or withdraws in year 1, year 2, year 3, year 4 or year 5. The expected probability of an individual being in the ith level of variable 1 and the jth level of variable 2, denoted by π_{ij}, can be modeled as a function of time (populations) by taking linear combinations of the rows and possibly looking at contrasts of these combinations. Then the model has the form $F(\underline{\pi}) = \underline{K} \log \underline{A\pi}$ where the matrix $\underline{A}$ is used to obtain linear combinations of the π_{ij}, such as to incorporate the withdraw information into the survival rates and the $\underline{K}$ matrix is used to construct desired contrasts for comparing t-year survival rates. The following are examples where the GSK model can be applied.

Example 1. Let variable 1 be whether a patient lives, dies, or withdraws from the study. Let variable 2 be the number of years of survival. Then the survival rate can be modeled as a function of time i.e. whether a person lived 1 year, 2 years up to 5 years.

Example 2. The data is the 5-year survival of women treated for breast cancer, (Koch et. al. 1972). There are three variables measured on each woman, degree of skin fixation, node status, and tumor size. The 5-year survival rate is then modeled as a function of these variables, to see if the survival rate is different for the different combinations of the variables.

3.2 Definitions

In the Grizzle, Starmer, and Koch (1969) paper, the loglinear model is used slightly different. It involves the ratio of logarithms of the expected cell probabilities. Lamm (1981) referred to this model as a logistic model since it is a comparison between the response (dependent) variable and each explanatory variable (or independent variable) separately, and no comparison is made between the independent variables. The GSK model corresponds more closely to the simple linear regression model, where the explanatory variables are assumed independent and the loglinear model corresponds more closely to an analysis of variance model, where interaction is tested for in selecting a model. For the GSK model, it is necessary to present some additional notation and give some definitions before going on to the general procedure described in Section 3.4. The frequency distribution for this categorical data is exactly the same as the frequency distribution for the loglinear model in Table 1 of section 2.1. Table 2 contains the expected cell probabilities, which are the probabilities used to define the expected cell frequencies.

Table 2

Expected Cell Probabilities

Population	Categories of Response				
(factors)	1	2	. . .	J	Total
1	π_{11}	π_{12}	. . .	π_{1J}	$\pi_{1\cdot}$
2	π_{21}	π_{22}	. . .	π_{2J}	$\pi_{2\cdot}$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
I	π_{I1}	π_{I2}	. . .	π_{I3}	$\pi_{I\cdot}$
	$\pi_{.1}$	$\pi_{.2}$	. . .	$\pi_{I.}$	$\pi_{..}$

Define

$$\underline{\pi}_i' = [\pi_{i1},\pi_{i2},\ldots,\pi_{iJ}] \quad \underline{\pi}' = [\underline{\pi}_1',\underline{\pi}_2',\ldots,\underline{\pi}_I']$$

which are the true but unknown cell probabilities,

$$p_{ij} = n_{ij}/n_{i\cdot} \; ; \; \underline{p}_i' = (p_{i1},p_{i2},\ldots,p_{iJ}); \; \underline{p}' = (\underline{p}_1',\underline{p}_2',\ldots,\underline{p}_I');$$

which are the estimates for the unknown cell probabilities with variance

$$\operatorname{var}(\underline{p}_i) = V(\underline{\pi}_i) = \frac{1}{n_{i\cdot}} \begin{bmatrix} \pi_{i1}(1-\pi_{i1}) & -\pi_{i1\;i2} & \cdots & -\pi_{i1}\pi_{iJ} \\ -\pi_{i2}\pi_{i1} & \pi_{i2}(1-\pi_{i2}) & \cdots & -\pi_{i2}\pi_{iJ} \\ \vdots & \vdots & & \vdots \\ -\pi_{iJ}\pi_{i1} & -\pi_{iJ}\pi_{i2} & \cdots & \pi_{iJ}(1-\pi_{iJ}) \end{bmatrix}$$

where the diagonal elements correspond to the variance terms for an ordinary binomial case and the off-diagonal elements correspond to the covariance between two proportions;

$\underline{V}(\underline{p}_i)$ = sample estimate of $\underline{V}(\underline{\pi}_i)$;

$\underline{V}(\underline{p})$ = block diagonal matrix having $\underline{V}(\underline{p}_i)$ on the main diagonal;

$f_m(\underline{\pi})$ = any function of the elements of $\underline{\pi}$ that have partial derivatives up to second order with respect to the π_{ij}, $m = 1,\ldots,u$, where $u < (J-1)I$;

$f_m(\underline{p}) = f_m(\underline{\pi})$ evaluated at $\underline{\pi} = \underline{p}$;

$(\underline{F}(\underline{\pi}))' = (f_1(\underline{\pi}), f_2(\underline{\pi}),\ldots,f_u(\underline{\pi}))$;

$\underline{F}' = (\underline{F}(\underline{p}))' = (f_1(\underline{p}), f_2(\underline{p}),\ldots,f_u(\underline{p}))$;

$\underline{H} = [\frac{\partial f_m(\underline{\pi})}{\partial \pi_{ij}} \Big| \pi_{ij} = p_{ij}]$; and $\underline{S} = \underline{H}\underline{V}(p)\underline{H}'$.

The matrix $\underline{S}$ is the sample estimate of the covariance matrix of $\underline{F}$.

Definition 3.2.1

The <u>logistic GSK model</u> is

$$\underset{t\times 1}{\underline{F}(\pi)} = \underset{t\times u}{K} \log \underset{u\times IJ}{\underline{A}} \underset{IJ\times 1}{\underline{\pi}}$$

where $\underline{K}$ and $\underline{A}$ are matrices of constants to be determined by the hypothesis of interest and $\underline{\pi}$ is a vector of cell probabilities. $\underline{A}$ is used to combine the desired probabilities and also to weight the probabilities differently if this is needed. The $\underline{K}$ matrix is used to make desired comparisons between the probabilities selected by $\underline{A}$ and is used to handle ordered categorical data (such as age). The function $\underline{F}(\underline{p})$ gives the estimates for the survival times for failure data in an example in Section 3.5, and since it is expected that survival rates will decrease as time goes on the $\underline{K}$ matrix is used to take this into account.

The general procedure to analyze the GSK model is discussed in Section 3.4.

3.3 Literature Review

This approach involves a logistic model to analyze categorical failure data. The GSK model was developed by Grizzle, Starmer, and Koch (1969), because of the flexibility in selecting the $\underline{K}$ and $\underline{A}$ matrices this procedure yields itself better to ordered variables than the loglinear model. In the analysis of this model, the method of least squares is used to test hypotheses of interest, instead of using the maximum likelihood methods used for the loglinear model. The program SAS FUNCAT was developed to handle this type of analysis. In the example in section 3.5, SAS FUNCAT was not used because of the singular covariance matrix that was produced by the derivative approach used by this program. Because of an idenity that can be used for the logistic GSK model, it is not necessary to use this derivative approach since SAS FUNCAT is set up for a number of linear models not just the logistic model. Thus SAS PROC MATRIX was used in the example to test for goodness-of-fit and to obtain estimates for the regression parameters. According to GSK (1969), the least squares method for estimation has been shown to be asymoticaly equivalent to the maximum likelihood methods. The GSK method does make it easier to test for marginal homogeneity (i.e. to test for differences within a variable), but harder to test for independence of variables. For the loglinear model discussed previously, this situation is reversed. Lamm (1981) analyzed a data set that had already been analyzed by Fienberg (1981), using the GSK approach and found that the resulting estimates were almost identical.

3.4 General Procedure

The model developed by Grizzle, Starmer, and Koch (1969) has many forms. The specific logistic model, as defined in Section 3.2 is just one of these linear models proposed by GSK (1969). In a paper by Kock, Johnson, and Tolley (1972), this logistic model is used to analyze failure time cross-classified categorical data. The GSK model uses a non-iterative approach to estimate parameters, via the least squares method and uses minimum chi-square statistics to test for goodness-of-fit.

The appropriate general model under this approach can be expressed as

$$\underline{F}(\underline{\pi}) = \underline{K} \log \underline{A} \ \underline{\pi}$$

where $\underline{K}$ and $\underline{A}$ are matrices of constants that are dependent on the hypothesis on interest (as described in Section 3.2) and $\underline{\pi}$ is the vector of the true, but unknown cell probabilities. For this model, GSK have shown that

$H = [\frac{\partial F}{\partial \pi} | \underline{\pi} = \underline{p}] = \underline{K} \ \underline{D}^{-1}\underline{A}$, thus $S = \underline{KD}^{-1}\underline{AV}(\underline{p})\underline{A'D}^{-1}\underline{K}'$, where $\underline{K}$ and $\underline{A}$ are as above and $\underline{D}$ is defined below;

$$\underline{D} = \text{diagonal } (\underline{A}'\underline{p}) = \begin{bmatrix} \underline{a}_1'\underline{p} & \underline{0} & \cdots & \underline{0} & \underline{0} \\ \underline{0} & \underline{a}_2'\underline{p} & \underline{0} & \cdots & \underline{0} \\ \underline{0} & \underline{0} & \cdot & & \\ \vdots & \vdots & & \ddots & \\ \underline{0} & \underline{0} & & & \underline{a}_u'\underline{p} \end{bmatrix}$$

where $\underline{a}_i'$ is the ith row of $\underline{A}'$.

Assume $\underline{F}(\underline{\beta}) = \underline{X}\ \underline{\beta}$, where $\underline{X}$ is a known design matrix and $\underline{\beta}$ is a vector of unknown parameters $v \leq u$. In the example to follow a two-way contingency table is considered, where there is one response variable and one explanatory variable. The two models have $\underline{X}_1 = (1,2,3,4)$, since the explanatory variable was ordered (discrete values for length of life of the individual under study) and $\hat{\underline{\beta}}$ is $\underline{b}$ obtained by minimizing $(\underline{F}(\underline{p}) - \underline{Xb})'\ S^{-1}(\underline{F}(\underline{p}) - \underline{Xb})$.

To test for the adequacy of the model, compare

$$SS(F(\underline{\beta}) = \underline{X\beta}) = \underline{F}'\underline{S}^{-1}\underline{F} - \underline{b}'(\underline{X}'\underline{S}^{-1}\underline{X})\underline{b}$$

with a χ^2-value with d.f. equal to d.f. = number of elements in $\underline{F}(\underline{p})$ - Rank $(\underline{X})$. Given the model is adequate from above, it is possible to test any hypothesis concerning $\underline{\beta}$, i.e. $H_0:\underline{C\beta}$, where $\underline{C}$ is a dxv matrix of arbitrary constants of full rank. The test statistic for this is

$$SS(\underline{C}\ \underline{\beta} = 0) = \underline{b}'\underline{C}'(\underline{C}(\underline{X}'\underline{S}^{-1}\underline{X})^{-1}\underline{C})^{-1}\underline{Cb}.$$

This above procedure is illustrated in the following section.

The example given is a two-way table and estimates are obtained for the marginal survival rates. In the case of a three-way table estimates can be obtained for each cell's survival rate. In essense, we are converting it to a two-way and look at each treatment combination versus the years of survival.

3.5 Example

This example was taken from a paper by Koch, Johnson and Tolley (1972). The data consists of five year survival data for 126 cancer patients. One question of interest was whether the probability of survival is characterized by an exponential curve. If this is the case then an appropriate model for the log survival rates, $\underline{F}$, is a straight line through the origin. In the problem it is also desirable to calculate the survival rate for each of the first four years, since the fourth and fifth years have the same estimate, i.e. this would cause $\underline{S}$ to be singular. The data for this example is as follows in Table 3.

Table 3

Five-year Survival Data for 126 Cases with Localized Kidney Cancer

Years after diagnosis	Survived the year	Died during the year	Withdrawn or 'lost'	Total alive at the beginning of year
0 - 1	60	47	19	126
1 - 2	38	5	17	60
2 - 3	21	2	15	38
3 - 4	10	2	9	21
4 - 5	4	0	6	10

In this problem there is an ordering present for the explanatory variable years, since it is expected that the probability of surviving five years is much less than the probability of surviving, say one year. Also the probability of surviving five years is dependent on the survival of each of the previous years. There is a special type of Type II censoring present here. It is referred to as a withdrawal or 'lost to follow-up'. This occurs when an individual does not report in for a given period and it is not known if the individual is dead or alive. All that is known is that the individual was alive at the end of the last period. The way this is handled by Koch, Johnson, and Tolley (1972), is that it is assumed for all cases that the individual that withdraws or is 'lost to follow-up' died in the middle of the corresponding time period that they didn't report in. Thus the linear model for this straight line through the origin is

$$E\{\underline{F}\} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \beta = \underline{X}\beta,$$

where the elements of $\underline{X}$ are the number of years of survival, either 1,2,3, or 4 years. $\underline{F}$ corresponds to the log survival rates. The survival rates are denoted by $\underline{G}$ where

$$G_i = \prod_{i=1}^{t} \left(\frac{p_{ij1} + .5\, p_{ij3}}{p_{ij2} + p_{ij2} + .5\, p_{ij3}}\right) \qquad i = 1,2,3,4 \quad ,$$

t then equals the number of years survived at the present time interval. G_2 corresponds to the survival rate for the first two years where $t = 2$. Then the log survival rates are

$$\underline{F} = [\log G_1, \log G_2, \log G_3, \log G_4].$$

G_i gives the survival rates based on weighting the withdrawals half as much as deaths and survivals. Taking all of the above factors into account yields the following model,

$$\underline{F}(\underline{\pi}) = \underline{K} \log \underline{A}\ \underline{\pi},$$

where $\underline{K} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & -1 & 1 & -1 & 0 & 0 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \end{bmatrix}$ and $\underline{A} = \begin{bmatrix} A^* & & \underline{0} \\ & \ddots & \\ \underline{0} & & A^* \end{bmatrix}$ $\underline{A}^* = \begin{bmatrix} 1 & 0 & .5 \\ 1 & 1 & .5 \end{bmatrix}$.

Thus the estimated survival rate for the first year, $F(\underline{p}_1) = \log\left(\frac{p_{11}+.5p_{13}}{p_{11}+p_{12}+.5p_{13}}\right)$ this is the logarithm of the probability of being alive or withdrawn over the overall probability of being alive at the start of the first year. The estimated survival rate for the second year averages over the first two years for the logarithms of the above ratio. For each year the estimated survival rate involves averaging over the previous years plus the current year, this averaging is accomplished by the use of the $\underline{K}$ matrix. The $\underline{A}$ matrix pulls out the desired ratio for each year.

Using the above model, $F(\underline{p}) = \underline{K} \log \underline{A}\ \underline{p}$ yields the following results,

$$\underline{F} = \begin{bmatrix} -.516538 \\ -1612735 \\ -.680523 \\ -.809682 \end{bmatrix}, \qquad \underline{S} = \begin{bmatrix} 56.44 & 56.44 & 56.44 & 56.44 \\ 56.44 & 74.72 & 74.72 & 74.72 \\ 56.44 & 74.72 & 97.52 & 97.52 \\ 56.44 & 74.72 & 97.52 & 178.7 \end{bmatrix} \times 10^{-4}$$

where $\underline{S} = \underline{K}\ \underline{D}^{-1}\ \underline{A}\ \underline{V}(\underline{p})\ \underline{A}'\underline{D}^{-1}\ \underline{K}'$.

If the exponential model does describe the data adequately, then the variation of the elements in $\underline{F}$ can be described by the model $E\{\underline{F}\} = \underline{X}\beta$, as shown next, where β is the constant mortality rate parameter.

$$E\ \{\underline{F}\} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \beta = \underline{X}\ \beta\ ,$$

where $\hat{\beta} = b = (\underline{X}'\underline{S}^{-1}\ \underline{X})^{-1}\ \underline{X}'\underline{S}^{-1}\ \underline{F}$. Here $b = -.147569$ and the goodness-of-fit test statistic as given in section 3.2 equals $X^2 = 27.92$ with d.f. = 3. This is significant at = .01, thus the exponential model does not adequately describe the data. Another model proposed by Koch et. al. (1972) was a suppressed exponential, meaning it included the an intercept parameter as follows

$$E\ \{\underline{F}\} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} (\underline{\beta}) = \underline{X}\ \underline{\beta}\ .$$

Now $\underline{b} = \begin{bmatrix} -.427916 \\ -.088622 \end{bmatrix}$ and the goodness-of-fit statistic is $X^2 = .47$ with d.f. = 2 which is not significant. Thus this suppressed exponential provides an adequate fit for the data. Also the value $e^{b_0} = 0.65$ is interpreted as the probability of being alive at the beginning of the experiment and $e^{b_1} = .91$ is the rate that the probability of survival decreases each additional year of risk. As stated earlier in this chapter, PROC MATRIX of SAS was used in order to carry out the above calculations. Table 4 contains the key to the variables used in this program and the program with results are printed on pages 78 - 82 . The PROC MATRIX program given with the results on the following page, deal with the suppressed exponential model. In order to use PROC MATRIX to get b and the goodness-of-fit test statistic for the first model looked at simply substitue X = 1/2/3/4; for the line that has X = 1 1/1 2/1 3/1 4;. This will give the results for

Table 4

K	=	the $\underline{K}$ matrix from the model used
A	=	the $\underline{A}$ matrix from the model used
API	=	vector of estimated cell probabilities
F	=	estimated survival rates
X	=	the design matrix from the suppressed exponential model
V	=	the matrix $\underline{V}(\underline{p})$
DA	=	the diagonal matrix $\underline{D}$ as defined earlier
DAINV	=	the inverse of $\underline{D}$
VF	=	the matrix $\underline{S}$
VFINV	=	the inverse of matrix $\underline{S}$
PART	=	$\underline{X}' \underline{S}^{-1} \underline{X}$
BETA	=	$\underline{b}$ the estimate for $\underline{\beta}$
SSFIT	=	the goodness-of-fit test statistic

the exponential model. Next to be considered is the summary, which briefly is a review of what has been covered in this paper.

Chapter 4

Summary

In this paper, the analysis of both continuous and discrete failure data was dealt with. Definitions were given first, concerning the models that were considered. This was followed by the analysis of the continuous data. The discrete data was in the form of cross-classified categorical data, and this was dealt with last.

The Cox regression model was used to analysis the continuous failure time data. The analysis involved using the partial likelihood equations, to obtain estimates for the regression parameters. The SAS program PROC PHGLM and BMDP2L were used in the model selection procedure, to determine which explanatory variables were important. BMDP2L also gave the estimates for the survival, hazard and cumulative hazard functions evaluated at $\underline{x}=\overline{\underline{x}}$, the average of the vectors of explanatory variables. SAS PROC SERVTEST was used to test for equal survivor functions for a treatment and a control in an example in section 1.5.

The loglinear model and the GSK model were used to analysis cross-classified categorical failure data. These models are found in chapters 2 and 3, respectively, with definitions, general procedures and examples.

The procedures given do not exhaust all of the ways to handle failure time data. Although they do provide a number of ways to look at this type of data. Three books that were used for the procedures considered in this paper, that would be valuable to those interested in more details and in more ways of handling failure time data are Lawless (1981), Kalbfiesch and Prentice (1980), and Fienberg (1981).

APPENDIX

Computer examples from sections 1.5, 2.5, and 3.5

KEY (for the following example) :

KASEID	Case identification
SURVIVAL	Survival time
FOLLOWUP	Censoring status (0 = censored, 1 = complete,i.e. dead)
LOGBUN	Log blood urea nitrogen
HGB	Hemoglobin
PLATELET	Platelets (0 = abnormal, 1 = normal)
INFEC	Infections (0 = none, 1 = present)
AGE	Age at diagnosis
SEX	Sex (1 = male, 2 = female)
LOGWBC	Log white blood cell count
FRAC	Fractures (0 = no, 1 = yes)
LOGPBM	Log percent plasma cells in bone marrow
PLYMPH	Percent lymphocytes in peripheral blood
PMYELOID	Percent myeloid cells in peripheral blood
PROTEIN	Proteinuria at diagnosis
BJP	Bence Jones preotein in urine
TSP	Total serum protein
SGLOBIN	Serum globin
SCALC	Serum calcium

```
1                                  S T A T I S T I C A L   A N A L Y S I S   S Y S T E M      16:07 WEDNESDAY, AUGUST 25, 1982

NOTE: THE JOB XPRS5784 HAS BEEN RUN UNDER RELEASE 79.5 OF SAS AT KANSAS STATE UNIVERSITY (C0306).

NOTE: SAS OPTIONS SPECIFIED ARE:
       SORT=4



1          OPTIONS PAGESIZE=51;
2          DATA MYELOMA;
3          TITLE MULTIPLE MYELOMA DATA;
4          INPUT KASEIC SURVIVAL FOLLOWUP LOGBUN HGB PLATELET INFEC AGE
5              SEX LOGWBC FRAC LOGPBM PLYMPH PMYELOID PROTEIN BJP TSP SGLOBIN SCALC;
6          CARDS;

NOTE: DATA SET WORK.MYELOMA HAS 65 OBSERVATIONS AND 19 VARIABLES. 122 OBS/TRK.
NOTE: THE DATA STATEMENT USED 0.66 SECONDS AND 168K.

72         PROC SORT ;
73         BY DESCENDING SURVIVAL;

NOTE: DATA SET WORK.MYELOMA HAS 65 OBSERVATIONS AND 19 VARIABLES. 122 OBS/TRK.
NOTE: THE PROCEDURE SORT USED 1.44 SECONDS AND 198K.

74         PROC PRINT;

NOTE: THE PROCEDURE PRINT USED 1.30 SECONDS AND 164K AND PRINTED PAGES 1 TO 2.

75         PROC PHGLM PRINTC OUTPUT;
76         EVENT FOLLOWUP;
77         MODEL SURVIVAL=LOGBUN HGB PLATELET INFEC AGE SEX LOGWBC FRAC LOGPBM
78                        PLYMPH PMYELOID PROTEIN BJP TSP SGLOBIN SCALC /BACKWARD;

NOTE: PHGLM IS SUPPORTED BY THE AUTHOR, NOT BY SAS.
NOTE: DATA SET WORK.DATA1 HAS 9 OBSERVATIONS AND 8 VARIABLES. 280 OBS/TRK.
AUTHOR: FRANK E. HARRELL, JR.
        DIVISION OF EPIDEMIOLOGY AND BIOSTATISTICS
        BOX 3337, DUKE UNIVERSITY MEDICAL CENTER
        DURHAM, NC 27710
NOTE: THE PROCEDURE PHGLM USED 14.41 SECONDS AND 180K AND PRINTED PAGES 3 TO 7.

NOTE: SAS USED 198K MEMORY.

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

```
                                          MULTIPLE MYELOMA DATA                  16:07 WEDNESDAY, AUGUST 25, 1982    1

OBS KASEID SURVIVAL FOLLOWUP LOGBUN HGB  PLATELET INFEC AGE SEX LOGWBC FRAC LOGPBM PLYPPH PMYELOID PROTEIN BJP TSP SGLOBIN SCALC

  1   48      92       1     1.4314 11.0     1      0   58  2  4.0755  1   1.4150  12      36        4     1   9    6     11
  2   47      89       1     1.3222 14.0     1      1   63  1  3.6532  1   1.6232   7      38        1     1   9    7      9
  3   46      88       1     1.1761 10.6     1      0   47  2  3.5563  0   1.7559  22      17       21     1   6    3      9
  4   65      77       0     1.0792 14.0     1      0   60  1  3.6812  0   0.9542   4      50        0     2   6    3     12
  5   45      67       1     1.3222 12.8     1      0   52  1  3.6453  1   1.0414  24      54        1     1  11    7     10
  6   44      66       1     1.4472  6.6     1      0   59  1  3.7853  1   1.8195   4      22        0     2   5    3      9
  7   43      58       1     1.2041 12.1     1      0   42  2  3.6990  1   1.5798   9      35       33     1   7    4     10
  8   64      57       0     1.2553 12.5     1      0   66  1  3.9685  0   1.9542   0       0        0     2   8    4     11
  9   42      54       1     1.2553  9.0     1      0   49  1  3.7243  1   1.6990   8      37        2     1   8    3     10
 10   63      53       0     1.1139 12.0     1      0   66  1  3.6128  1   2.0000   0       0        1     2   6    3     11
 11   41      52       1     1.0000 10.1     1      0   60  2  3.8573  1   1.6532   1      27        4     1  10    7     10
 12   40      51       1     1.5682  7.7     0      0   74  1  3.4150  1   1.0414  10      50        4     1  12    8     13
 13   38      41       1     1.0000 10.2     1      0   69  1  3.4771  1   1.4771   5      36        6     1  10    6     10
 14   39      41       1     1.4161  5.0     1      0   70  2  3.5185  1   1.3424  18      53        0     2   7    4      9
 15   62      41       0     1.7559 12.8     1      0   72  1  3.7243  1   1.4472   2      56        1     1   7    3      9
 16   37      37       1     1.6021 11.0     1      0   63  1  3.9542  0   1.2041  10      53        7     1   7    4      9
 17   36      35       1     1.1139  7.0     0      0   48  1  3.6532  1   1.1761  16      42        4     1  12    9     10
 18   35      32       1     1.3222 10.6     1      0   46  1  3.6990  1   1.6335   6      32        1     2  12    8      9
 19   61      28       0     1.2304  7.3     1      1   82  2  3.7482  1   1.6721   7      40        0     2   7    4      9
 20   34      26       1     1.2304 11.2     1      0   49  2  3.6201  1   2.0000   0       0       27     1   6    3     11
 21   33      25       1     1.0000 12.4     1      0   67  1  3.8195  1   1.6435   3      23        0     2   7    3     10
 22   32      24       1     1.3010 14.6     1      0   56  2  4.0959  1   0.4771  12      52        0     2   4    3      9
 23   30      19       1     1.0792 14.4     1      0   51  1  3.9191  1   2.0000   0       0        6     2  13    7     15
 24   31      19       1     1.2553  7.5     0      0   60  2  3.7924  1   1.9294   2      10        5     1   6    3      9
 25   59      19       0     1.3222 13.0     1      0   59  2  3.7709  1   2.0000   0       0        1     2   8    4     10
 26   60      19       0     1.3222 10.8     1      0   69  2  3.8903  1   1.5185   0       0        0     2  10    7     10
 27   29      18       1     1.4472  7.5     1      0   65  2  3.5682  0   0.9031  17      46        7     2   6    4      8
 28   27      17       1     1.2304 10.0     1      0   53  1  3.8808  1   1.4472   5      50        4     2  10    7      9
 29   28      17       1     1.5911 11.2     1      1   68  1  3.4314  0   1.6128   1      46        1     2   5    6     10
 30   26      16       1     1.3424  9.0     1      0   48  1  3.5345  1   2.0000   3       3        3     2   8    1     10
 31   25      16       1     1.3222  8.8     1      0   62  2  3.6990  1   0.6990   5      41       17     2   8    4     10
 32   58      16       0     1.1461 13.0     1      0   55  1  3.8573  0   0.9031  13      53        0     2   9    5      9
 33   24      15       1     1.6021 10.6     1      0   70  1  3.6902  1   1.4314  16      46        0     2   9    5     11
 34   23      14       1     1.3979 14.6     1      1   66  1  3.7243  1   1.2553   7      51        2     1   5    4     10
 35   22      13       1     0.7782  5.5     0      1   60  2  3.5798  1   1.3979  12      55        2     2  10    8     10
 36   57      13       0     1.6628  4.9     0      0   71  2  3.6435  0   1.7924   2      21        0     2  10   11      9
 37   54      12       0     1.1461 11.0     1      0   46  2  3.6435  0   1.1461   7      47        0     2   7    3      7
 38   56      12       0     1.3979  8.8     1      0   66  2  3.8388  1   1.3617   8      50        0     2   5    5      9
 39   17      11       1     1.1139 14.0     1      3   61  1  3.8808  1   1.2788   8      49        0     2   9    5     10
 40   18      11       1     1.2304 12.0     1      0   43  1  3.7709  1   1.1761   7      48        1     1  12    8      9
 41   19      11       1     1.3010 13.2     1      0   65  1  3.7993  1   1.8195   4      21        1     2   8    4     10
 42   20      11       1     1.5682  7.5     1      0   70  1  3.8865  0   1.6721   5      35        0     2   7    5     12
 43   21      11       1     1.0792  9.6     1      0   51  2  3.5051  1   1.9031   5      10        0     2   7    3      9
 44   55      11       0     1.6128 14.0     1      0   60  1  3.7324  1   1.8451   0       0        3     1   6    2      9
 45   16       9       1     1.7243  8.2     1      3   55  1  3.7993  1   1.7404   4      31        0     2   9    6     12
 46   53       9       0     1.0792  9.9     1      0   57  2  3.8325  1   1.6532   1      40        0     2   8    4      8
 47   13       7       1     1.9777  9.5     1      0   48  1  3.3617  1   1.5682   5      41        5     2   5    3     10
```

```
                                                         MULTIPLE MYELOMA DATA                       14:07 WEDNESDAY, AUGUST 25, 1982     2

OBS KASEID SURVIVAL FOLLOWUP LOGBUN  HGB PLATELET INFEC AGE SEX LOGWBC FRAC LOGPBM PLYMPH FMYELOID PROTEIN BJP TSP SGLOBIN SCALC

48    14     7.00       1    1.0414  5.1    0        0   61   2  3.7324  1   2.0000   0        0        1     2    8    4      10
49    15     7.00       1    1.1761 11.4    1        0   53   2  6.7243  1   1.5185  13       42        1     1    9    6      13
50    51     7.00       0    1.1139 12.4    1        0   48   2  3.7993  1   1.8593   1       11        0     2    6    3      10
51    52     7.00       0    1.5315 10.2    1        1   81   1  3.5911  0   1.8808   1       16        0     2   12   10      11
52     9     6.00       1    1.3617  9.0    1        1   77   1  3.5441  0   1.4624   7       51        1     2    7    5       8
53    10     6.00       1    2.1139 10.2    0        0   70   2  3.5441  1   1.3617   6       68        1     2    9    7       9
54    11     6.00       1    1.1139  9.7    1        0   60   1  3.5185  1   1.3979  11       36        0     2   11    7      10
55    12     6.00     . 1    1.4150 10.4    1        0   67   2  3.9294  1   1.6902  10       13        0     2   10    6       8
56     7     5.00       1    2.2355 10.1    1        0   50   2  4.9542  1   1.6628   8       27        4     1    6    4       9
57     8     5.00       1    1.6812  6.5    1        0   74   1  3.7324  0   1.7324   1       25        5     2   10    7       9
58    49     4.00       0    1.9542 10.2    1        0   59   1  4.0453  0   0.7782  21       51       12     1 [illegible] 3  10
59    50     4.00       0    1.9243 10.0    1        1   49   2  3.9590  0   1.6232  21       24        0     2   17   12      13
60     6     3.00       1    1.5441  6.7    1        1   46   2  4.4757  0   1.9345   3        5       12     2    7    3      10
61     3     2.00       1    1.5185  9.8    1        1   81   1  3.8751  1   2.0000   0        0        2     1   11    8      15
62     4     2.00       1    1.7482 11.3    0        0   75   1  3.8062  1   1.2553  13       52        0     2    5    6      12
63     5     2.00       1    1.3010  5.1    0        0   57   1  3.7243  1   2.0000   0        0        3     1   10    6       9
64     1     1.25       1    2.2175  9.4    1        0   67   1  3.6628  1   1.9542   0        0       12     2   11    7      10
65     2     1.25       1    1.9395 12.0    1        1   38   1  3.9868  1   1.9542   0        0       20     1    6    3      18
```

```
                                                                      MULTIPLE MYELOMA DATA                    16:07 WEDNESDAY, AUGUST 25, 1982    3

          STEPWISE PROPORTIONAL HAZARDS GENERAL LINEAR MODEL PROCEDURE

                              DEPENDENT VARIABLE: SURVIVAL

                               EVENT INDICATOR: FOLLOWUP

                      65 OBSERVATIONS
                      48 UNCENSORED OBSERVATIONS
                      52 EQUIVALENT SAMPLE SIZE WITH NO CENSORING
                       0 OBSERVATIONS DELETED DUE TO MISSING VALUES

          -2 LOG LIKELIHOOD FOR MODEL CONTAINING NO VARIABLES=  309.72

STEP  0. THE FOLLOWING VARIABLES ARE ENTERED:

LOGBUN   HGB      PLATELET INFEC    AGE      SEX      LOGWBC   FRAC
LOGPBM   PLYMPH   PMYELOID PROTEIN  BJP      TSP      SGLOBIN  SCALC

        CONVERGENCE OBTAINED IN  5 ITERATIONS.                      D=0.513.
        MAX ABSOLUTE DERIVATIVE=0.5028D-03.              -2 LOG L=  271.83.
        MODEL CHI-SQUARE=   37.88 WITH  16 D.F.                    P=0.0016.


THE FOLLOWING VARIABLES DO NOT MEET THE 0.0500
SIGNIFICANCE LEVEL FOR STAYING IN THE MODEL AND ARE REMOVED:

PLYMPH

STEP  1.

        CONVERGENCE OBTAINED IN  2 ITERATIONS.                      D=0.506.
        MAX ABSOLUTE DERIVATIVE=0.3292D-03.              -2 LOG L=  271.83.
        MODEL CHI-SQUARE=   37.88 WITH  15 D.F.                    P=0.0009.


THE FOLLOWING VARIABLES DO NOT MEET THE 0.0500
SIGNIFICANCE LEVEL FOR STAYING IN THE MODEL AND ARE REMOVED:

LOGPBM

STEP  2.

        CONVERGENCE OBTAINED IN  3 ITERATIONS.                      D=0.499.
        MAX ABSOLUTE DERIVATIVE=0.3651D-05.              -2 LOG L=  271.86.
        MODEL CHI-SQUARE=   37.85 WITH  14 D.F.                    P=0.0005.
```

```
          STEPWISE PROPORTIONAL HAZARDS GENERAL LINEAR MODEL PROCEDURE

                         DEPENDENT VARIABLE: SURVIVAL

                          EVENT INDICATOR: FOLLOWUP


THE FOLLOWING VARIABLES DO NOT MEET THE 0.0500
SIGNIFICANCE LEVEL FOR STAYING IN THE MODEL AND ARE REMOVED:

PMYELOID

STEP   3.

        CONVERGENCE OBTAINED IN   3 ITERATIONS.                      D=0.490.
        MAX ABSOLUTE DERIVATIVE=0.1487D-03.               -2 LOG L=  272.24.
        MODEL CHI-SQUARE=    37.47 WITH  13 D.F.                    P=0.0003.


THE FOLLOWING VARIABLES DO NOT MEET THE 0.0500
SIGNIFICANCE LEVEL FOR STAYING IN THE MODEL AND ARE REMOVED:

AGE

STEP   4.

        CONVERGENCE OBTAINED IN   3 ITERATIONS.                      D=0.482.
        MAX ABSOLUTE DERIVATIVE=0.1740D-02.               -2 LOG L=  272.52.
        MODEL CHI-SQUARE=    37.20 WITH  12 D.F.                    P=0.0002.


THE FOLLOWING VARIABLES DO NOT MEET THE 0.0500
SIGNIFICANCE LEVEL FOR STAYING IN THE MODEL AND ARE REMOVED:

FRAC

STEP   5.

        CONVERGENCE OBTAINED IN   3 ITERATIONS.                      D=0.469.
        MAX ABSOLUTE DERIVATIVE=0.1441D-04.               -2 LOG L=  273.47.
        MODEL CHI-SQUARE=    36.24 WITH  11 D.F.                    P=0.0002.


THE FOLLOWING VARIABLES DO NOT MEET THE 0.0500
SIGNIFICANCE LEVEL FOR STAYING IN THE MODEL AND ARE REMOVED:

SCALC
```

```
          STEPWISE PROPORTIONAL HAZARDS GENERAL LINEAR MODEL PROCEDURE

                         DEPENDENT VARIABLE: SURVIVAL

                          EVENT INDICATOR: FOLLOWUP

STEP   6.

       CONVERGENCE OBTAINED IN  4 ITERATIONS.                       D=0.455.
       MAX ABSOLUTE DERIVATIVE=0.6422D-07.                -2 LOG L=  274.66.
       MODEL CHI-SQUARE=   35.06 WITH  10 D.F.                     P=0.0001.

THE FOLLOWING VARIABLES DO NOT MEET THE 0.0500
SIGNIFICANCE LEVEL FOR STAYING IN THE MODEL AND ARE REMOVED:

SEX

STEP   7.

       CONVERGENCE OBTAINED IN  4 ITERATIONS.                       D=0.436.
       MAX ABSOLUTE DERIVATIVE=0.6907D-05.                -2 LOG L=  276.50.
       MODEL CHI-SQUARE=   33.22 WITH   9 D.F.                     P=0.0001.

THE FOLLOWING VARIABLES DO NOT MEET THE 0.0500
SIGNIFICANCE LEVEL FOR STAYING IN THE MODEL AND ARE REMOVED:

HGB

STEP   8.

       CONVERGENCE OBTAINED IN  4 ITERATIONS.                       D=0.421.
       MAX ABSOLUTE DERIVATIVE=0.6970D-07.                -2 LOG L=  277.76.
       MODEL CHI-SQUARE=   31.96 WITH   8 D.F.                     P=0.0001.

NO ADDITIONAL VARIABLES MET THE 0.1000 SIGNIFICANCE LEVEL FOR ENTRY.
```

```
                                                    MULTIPLE MYELOMA DATA                     16:07 WEDNESDAY, AUGUST 25, 1982    6
         STEPWISE PROPORTIONAL HAZARDS GENERAL LINEAR MODEL PROCEDURE

                       DEPENDENT VARIABLE: SURVIVAL

                        EVENT INDICATOR: FOLLOWUP


                        FINAL PARAMETER ESTIMATES

VARIABLE         BETA            STD. ERROR        CHI-SQUARE    P        D

LOGBUN        2.15430663        0.63519114           11.50    0.0007   0.207
PLATELET     -1.54731195        0.52066906            8.83    0.0030   0.167
INFEC         0.54740076        0.21980065            6.20    0.0128   0.124
LOGWBC        1.24581820        0.36445337           11.68    0.0006   0.210
PROTEIN       0.05776487        0.02222619            6.75    0.0094   0.133
BJP           1.27837669        0.41742550            9.38    0.0022   0.176
TSP           0.30492496        0.12244562            6.20    0.0128   0.124
SGLOBIN      -0.26552449        0.13389075            3.93    0.0474   0.082
```

```
STEPWISE PROPORTIONAL HAZARDS GENERAL LINEAR MODEL PROCEDURE

                 DEPENDENT VARIABLE: SURVIVAL

                  EVENT INDICATOR: FOLLOWUP


                COVARIANCE MATRIX OF ESTIMATES

                LOGBUN     PLATELET        INFEC         LOGWBC       PROTEIN          BJP            TSP          SGLOBIN
LOGBUN       0.4034678 0.0006768473    0.00630302    0.05760968  0.002763759     0.03924115   0.000081066    -0.0101527
PLATELET  0.0006768473    0.2710963    -0.0311579    -0.0692197   -0.0022826     -0.0698933    -0.0108679     0.0263621
INFEC       0.00630302   -0.0311579    0.04831222    0.01264317 0.0006083756 -0.000201774   0.002185843    -0.00518626
LOGWBC      0.05760968   -0.0692197    0.01264317     0.1328263  0.002375521     0.05768464   0.004610298   -0.00480838
PROTEIN    0.002763759   -0.0022826 0.0006083756   0.002375521 0.0004940037    0.004600353 0.0002034055 0.0001492917
BJP         0.03924115   -0.0698933 -0.000201774    0.05768464  0.004600353       0.174244   0.003726362 0.0004281983
TSP        0.000081066   -0.0108679  0.002185843   0.004610298 0.0002034055    0.003726362    0.01499253    -0.0131662
SGLOBIN     -0.0101527    0.0263621  -0.00518626   -0.00480838 0.0001492917 0.0004281983    -0.0131662     0.01792673
```

```
BMDP2L - REGRESSION WITH INCOMPLETE SURVIVAL DATA
BMDP STATISTICAL SOFTWARE, INC.
1964 WESTWOOD BLVD. SUITE 202
(213) 475-5700
PROGRAM REVISED APRIL     1982
MANUAL REVISED -- 1981
COPYRIGHT (C) 1982 REGENTS OF UNIVERSITY OF CALIFORNIA

TO SEE REMARKS AND A SUMMARY OF NEW FEATURES FOR
THIS PROGRAM, STATE NEWS. IN THE PRINT PARAGRAPH.

   AUGUST 23, 1982  AT 22:58:43


PROGRAM CONTROL INFORMATION

/ PROBLEM    TITLE IS 'MULTIPLE MYELOMA DATA'.
/ INPUT      VARIABLES ARE 19.    FORMAT IS FREE.
/ VARIABLE   NAMES ARE KASEID, SURVIVAL, FOLLOWUP,LOGBUN, HGB, PLATELET,
                       INFEC,AGE,SEX,LOGWBC,FRAC, LOGPBM, PLYMPH, PMYELOID,
                       PROTEIN, BJP, TSP, SGLOBIN, SCALC.
/ FORM       TIME IS SURVIVAL.  STATUS IS FOLLOWUP.
             RESPONSE IS 1.
/ REGRESSION COVARIATES ARE 4 TO 19.
             STEPWISE IS MPLR.
             START=IN,OUT,IN,IN,OUT,OUT,IN,OUT,OUT,OUT,OUT,IN,IN,IN,IN,OUT.
             MOVE=2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2.
/ PRINT      CASES = 65. SURVIVAL.
/ END

PROBLEM TITLE IS
MULTIPLE MYELOMA DATA

NUMBER OF VARIABLES TO READ IN. . . . . . . . . .       19
NUMBER OF VARIABLES ADDED BY TRANSFORMATIONS. .        0
TOTAL NUMBER OF VARIABLES . . . . . . . . . . .       19
NUMBER OF CASES TO READ IN. . . . . . . . . . . TO END
CASE LABELING VARIABLES . . . . . . . . . . . .
MISSING VALUES CHECKED BEFORE OR AFTER TRANS. . NEITHER
BLANKS ARE. . . . . . . . . . . . . . . . . . . MISSING
INPUT UNIT NUMBER . . . . . . . . . . . . . . .        5
REWIND INPUT UNIT PRIOR TO READING. . DATA. . .       NO
NUMBER OF WORDS OF DYNAMIC STORAGE. . . . . . .    29694

VARIABLES TO BE USED
     1 KASEID       2 SURVIVAL     3 FOLLOWUP     4 LOGBUN       5 HGB
     6 PLATELET     7 INFEC        8 AGE          9 SEX         10 LOGWBC
    11 FRAC        12 LOGPBM      13 PLYMPH      14 PMYELOID    15 PROTEIN
    16 BJP         17 TSP         18 SGLOBIN     19 SCALC
```

```
NUMBER OF CASES READ. . . . . . . . . . . . . .        65

                   RESPONSE CODES . . . . . . . .      1     DEAD

                   CENSORED CODES . . . . . . . .      0     CENSORED


DESCRIPTIVE STATISTICS FOR FIXED COVARIATES
-------------------------------------------


 VARIABLE                                          STANDARD
NO. N A M E      MINIMUM   MAXIMUM      MEAN   DEVIATION  SKEWNESS  KURTOSIS

  4 LOGBUN       0.7782     2.2355     1.3970     0.3111      0.82       3.25
  5 HGB          4.9000    14.6000    10.1923     2.5535     -0.28       2.38
  6 PLATELET     0.0        1.0000     0.8615     0.3481     -2.05       5.22
  7 INFEC        0.0        3.0000     0.2615     0.6194      2.89      12.12
  8 AGE         38.0000    82.0000    60.1538    10.3338      0.06       2.25
  9 SEX          1.0000     2.0000     1.4154     0.4966      0.34       1.08
 10 LOGWBC       3.3617     6.7243     3.8155     0.4368      4.79      31.12
 11 FRAC         0.0        1.0000     0.7538     0.4341     -1.15       2.32
 12 LOGPBM       0.4771     2.0000     1.5497     0.3642     -0.76       3.00
 13 PLYMPH       0.0       24.0000     6.7846     6.2612      0.89       3.07
 14 PMYELOID     0.0       68.0000    30.4000    19.9638     -0.31       1.73
 15 PROTEIN      0.0       33.0000     3.8308     6.6416      2.57       9.62
 16 BJP          1.0000     2.0000     1.6462     0.4819     -0.60       1.33
 17 TSP          4.0000    17.0000     8.6154     2.2479      0.78       4.40
 18 SGLOBIN      1.0000    12.0000     5.1692     2.2470      0.77       3.22
 19 SCALC        7.0000    18.0000    10.1231     1.8158      1.84       7.74


               STATUS CODE FREQUENCIES
                                              PERCENT
             TOTAL       DEAD       CENSORED  CENSORED
               65          48          17     0.2615
```

```
PAGE    5  BMDP2L MULTIPLE MYELOMA DATA

STEP NUMBER    0
---------------

                         LOG LIKELIHOOD =  -138.8785
                       GLOBAL CHI-SQUARE =        36.24   D.F.=    8  P-VALUE =0.0000

                                                STANDARD
    VARIABLE          COEFFICIENT                ERROR     COEFF./S.E.     EXP(COEFF.)
    --------          -----------              --------    -----------     -----------
  4 LOGBUN                 2.1543                0.6352         3.3916          8.6219
  6 PLATELET              -1.5473                0.5207        -2.9713          0.2128
  7 INFEC                  0.5474                0.2198         2.4904          1.7288
 10 LOGWBC                 1.2458                0.3645         3.4193          3.4758
 15 PROTEIN                0.0578                0.0222         2.5990          1.0595
 16 BJP                    1.2784                0.4174         3.0625          3.5908
 17 TSP                    0.3049                0.1224         2.4903          1.3565
 18 SGLOBIN               -0.2655                0.1339        -1.9831          0.7668


STATISTICS TO ENTER OR REMOVE VARIABLES
--------------------------------------
                       APPROX.           APPROX.
    VARIABLE           CHI-SQ.           CHI-SQ.                            LOG
NO. N A M E            ENTER             REMOVE          P-VALUE        LIKELIHOOD

  4 LOGBUN                                11.45           0.0007         -144.6023
  5 HGB                  1.26                             0.2622         -138.2499
  6 PLATELET                               7.86           0.0050         -142.8110
  7 INFEC                                  4.81           0.0283         -141.2848
  8 AGE                  0.07                             0.7959         -138.8451
  9 SEX                  1.12                             0.2897         -138.3181
 10 LOGWBC                                 7.49           0.0062         -142.6238
 11 FRAC                 0.94                             0.3316         -138.4072
 12 LOGPBM               0.19                             0.6620         -138.7830
 13 PLYMPH               0.09                             0.7596         -138.8317
 14 PMYELOID             0.07                             0.7909         -138.8434
 15 PROTEIN                                5.67           0.0173         -141.7114
 16 BJP                                   10.34           0.0013         -144.0483
 17 TSP                                    6.04           0.0140         -141.9000
 18 SGLOBIN                                3.87           0.0490         -140.8156
 19 SCALC                0.86                             0.3528         -138.4469


NO TERM PASSES THE REMOVE AND ENTER LIMITS (  0.1500  0.1000 ) .
```

```
PAGE    6  BMDP2L MULTIPLE MYELOMA DATA

SUMMARY OF  STEPWISE RESULTS

STEP  VARIABLE                      VARIABLE                  LOG          IMPROVEMENT          GLOBAL
 NO   ENTERED              DF       REMOVED               LIKELIHOOD  CHI-SQUARE P-VALUE  CHI-SQUARE P-VALUE
---  -------------------  ---    -------------------     ----------  ---------- -------  ---------- -------
  0                         8                              -138.875                           36.242    0.000
```

```
PAGE    7  BMDP2L MULTIPLE MYELOMA DATA

                                        TIME VARIABLE IS SURVIVAL

                                                                          KAPLAN    -----  PROPORTIONAL HAZARDS MODEL  -----
CASE     CASE                            CUM      CUM      REMAIN   MEIER     SURVIVAL    HAZARD    CUM HAZARD
LABEL    NUMBER   SURVIVAL    STATUS    DEATHS   CENSRD   AT RISK  SURVIVAL      EVALUATED FOR Z = ZBAR           RESIDUAL

             2      1.25   DEAD            1        0        64              0.9854                  0.0147     0.0631
             1      1.25   DEAD            2        0        63    0.9692    0.9854      0.0117      0.0147     0.1595
             5      2.00   DEAD            3        0        62              0.9613                  0.0395     0.0481
             4      2.00   DEAD            4        0        61              0.9613                  0.0395     0.3105
             3      2.00   DEAD            5        0        60    0.9231    0.9613      0.0330      0.0395     0.0257
             6      3.00   DEAD            6        0        59    0.9077    0.9527      0.0090      0.0485     0.9228
            50      4.00   CENSORED        6        1        58              0.9406                  0.0613     0.7308
            49      4.00   CENSORED        6        2        57              0.9406                  0.0613     0.1446
             8      5.00   DEAD            7        2        56              0.9286                  0.0741     0.1360
             7      5.00   DEAD            8        2        55    0.8753    0.9286      0.0128      0.0741     0.3542
            12      6.00   DEAD            9        2        54              0.8780                  0.1302     0.1681
            11      6.00   DEAD           10        2        53              0.8780                  0.1302     0.0548
            10      6.00   DEAD           11        2        52              0.8780                  0.1302     1.3194
             9      6.00   DEAD           12        2        51    0.8121    0.8780      0.0561      0.1302     0.0887
            15      7.00   DEAD           13        2        50              0.8343                  0.1812     0.9893
            14      7.00   DEAD           14        2        49              0.8343                  0.1812     0.3766
            13      7.00   DEAD           15        2        48    0.7644    0.8343      0.0510      0.1812     0.8462
            52      7.00   CENSORED       15        3        47              0.8343                  0.1812     0.2170
            51      7.00   CENSORED       15        4        46              0.8343                  0.1812     0.0681
            53      8.00   CENSORED       15        5        45              0.8250                  0.1924     0.0987
            16      9.00   DEAD           16        5        44    0.7474    0.8158      0.0112      0.2036     1.6585
            21     11.00   DEAD           17        5        43              0.7110                  0.3410     0.1119
            20     11.00   DEAD           18        5        42              0.7110                  0.3410     0.3035
            19     11.00   DEAD           19        5        41              0.7110                  0.3410     0.2869
            18     11.00   DEAD           20        5        40              0.7110                  0.3410     0.0776
            17     11.00   DEAD           21        5        39    0.6625    0.7110      0.0687      0.3410     1.0765
            55     11.00   CENSORED       21        6        38              0.7110                  0.3410     0.1493
            56     12.00   CENSORED       21        7        37              0.6989                  0.3583     0.3830
            54     12.00   CENSORED       21        8        36              0.6989                  0.3583     0.1613
            22     13.00   DEAD           22        8        35    0.6441    0.6869      0.0173      0.3755     0.4267
            57     13.00   CENSORED       22        9        34              0.6869                  0.3755     0.7217
            23     14.00   DEAD           23        9        33    0.6251    0.6609      0.0386      0.4141     0.2705
            24     15.00   DEAD           24        9        32    0.6067    0.6353      0.0396      0.4537     0.6258
            26     16.00   DEAD           25        9        31              0.5843                  0.5374     1.0524
            25     16.00   DEAD           26        9        30    0.5683    0.5843      0.0837      0.5374     1.4563
            58     16.00   CENSORED       26       10        29              0.5843                  0.5374     0.3416
            28     17.00   DEAD           27       10        28              0.5247                  0.6450     0.8839
            27     17.00   DEAD           28       10        27    0.5291    0.5247      0.1076      0.6450     0.5090
            29     18.00   DEAD           29       10        26    0.5095    0.4937      0.0609      0.7059     0.4689
            31     19.00   DEAD           30       10        25              0.4349                  0.8327     0.7355
            30     19.00   DEAD           31       10        24    0.4703    0.4349      0.1269      0.8327     1.3944
            60     19.00   CENSORED       31       11        23              0.4349                  0.8327     0.6356
            59     19.00   CENSORED       31       12        22              0.4349                  0.8327     0.7078
```

```
PAGE    8   BMDP2L  MULTIPLE  MYELOMA  DATA

                                                            KAPLAN   -----  PROPORTIONAL HAZARDS MODEL  -----
CASE    CASE                            CUM     CUM    REMAIN  MEIER     SURVIVAL    HAZARD    CUM HAZARD
LABEL   NUMBER   SURVIVAL     STATUS   DEATHS  INCMPL  AT RISK SURVIVAL        EVALUATED FOR Z = ZBAR        RESIDUAL

          32      24.00  DEAD            32      12      21    0.4487    0.3989    0.0173      0.9190     0.4036
          33      25.00  DEAD            33      12      20    0.4275    0.3647    0.0897      1.0086     0.4129
          34      26.00  DEAD            34      12      19    0.4062    0.3323    0.0931      1.1017     0.5663
          61      28.00  CENSORED        34      13      18              0.3207                1.1371     0.9274
          35      32.00  DEAD            35      13      17    0.3836    0.2988    0.0177      1.2079     1.0990
          36      35.00  DEAD            36      13      16    0.3610    0.2657    0.0392      1.3254     0.8677
          37      37.00  DEAD            37      13      15    0.3385    0.2339    0.0637      1.4528     0.8233
          39      41.00  DEAD            38      13      14              0.1778                1.7273     0.7132
          38      41.00  DEAD            39      13      13    0.2933    0.1778    0.0686      1.7273     0.2046
          62      41.00  CENSORED        39      14      12              0.1778                1.7273     0.9461
          40      51.00  DEAD            40      14      11    0.2689    0.1509    0.0164      1.8914     3.1938
          41      52.00  DEAD            41      14      10    0.2445    0.1202    0.2271      2.1185     0.2752
          63      53.00  CENSORED        41      15       9              0.1059                2.2449     0.7090
          42      54.00  DEAD            42      15       8    0.2173    0.0934    0.1263      2.3712     0.6336
          64      57.00  CENSORED        42      16       7              0.0714                2.6389     2.3456
          43      58.00  DEAD            43      16       6    0.1863    0.0653    0.0992      2.7281     2.1431
          44      66.00  DEAD            44      16       5    0.1552    0.0398    0.0620      3.2240     1.8009
          45      67.00  DEAD            45      16       4    0.1242    0.0200    0.6859      3.9098     0.8903
          65      77.00  CENSCRED        45      17       3              0.0120                4.4225     1.3372
          46      88.00  DEAD            46      17       2    0.0828    0.0068    0.0513      4.9863     1.4839
          47      89.00  DEAD            47      17       1    0.0414    0.0014    1.5839      6.5703     1.4202
          48      92.00  DEAD            48      17       0    0.0       0.0001    0.8028      8.9788     3.7279
```

```
1                                S T A T I S T I C A L   A N A L Y S I S   S Y S T E M      16:02 WEDNESDAY, AUGUST 25, 1982

NOTE: THE JOB XPRS5780 HAS BEEN RUN UNDER RELEASE 79.5 OF SAS AT KANSAS STATE UNIVERSITY (00306).

NOTE: SAS OPTIONS SPECIFIED ARE:
        SORT=4


1          DATA COX;
2          INPUT REMISS CENSOR TREAT;
3          CARDS;

NOTE: DATA SET WORK.COX HAS 42 OBSERVATIONS AND 3 VARIABLES. 680 OBS/TRK.
NOTE: THE DATA STATEMENT USED 0.21 SECONDS AND 160K.

46         PROC PRINT;

NOTE: THE PROCEDURE PRINT USED 0.59 SECONDS AND 162K AND PRINTED PAGE 1.

47         PROC SURVTEST;
48         CLASS TREAT;
49         VAR REMISS CENSOR;

NOTE: SURVTEST IS SUPPORTED BY THE AUTHOR, NOT BY SAS.
NOTE: THE PROCEDURE SURVTEST USED 0.52 SECONDS AND 160K AND PRINTED PAGE 2.

NOTE: SAS USED 162K MEMORY.

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

```
                 STATISTICAL ANALYSIS SYSTEM                                   1
                                                    16:02 WEDNESDAY, AUGUST 25, 1982

      OBS    REMISS    CENSOR    TREAT

        1       6         1        1
        2       6         2        1
        3       6         2        1
        4       6         2        1
        5       7         2        1
        6       9         1        1
        7      10         1        1
        8      10         2        1
        9      11         1        1
       10      13         2        1
       11      16         2        1
       12      17         1        1
       13      19         1        1
       14      20         1        1
       15      22         2        1
       16      23         2        1
       17      25         1        1
       18      32         1        1
       19      32         1        1
       20      34         1        1
       21      35         1        1
       22       1         2        2
       23       1         2        2
       24       2         2        2
       25       2         2        2
       26       3         2        2
       27       4         2        2
       28       4         2        2
       29       5         2        2
       30       5         2        2
       31       8         2        2
       32       8         2        2
       33       8         2        2
       34       8         2        2
       35      11         2        2
       36      11         2        2
       37      12         2        2
       38      12         2        2
       39      15         2        2
       40      17         2        2
       41      22         2        2
       42      23         2        2
```

```
                          S T A T I S T I C A L   A N A L Y S I S   S Y S T E M                                  2
                                                                                     16:02 WEDNESDAY, AUGUST 25, 1982

GEHAN-WILCOXON TEST FOR VARIABLE (REMISS, CENSOR)

 TREAT       N    SUM OF SCORES

     1      21         271
     2      21        -271

         CHISQ= 13.01 ON  1 DF,   PROB>CHISQ=0.0003

LOGRANK TEST FOR VARIABLE (REMISS, CENSOR)

 TREAT       N    OBSERVED     EXPECTED     (O-E)**2/E

     1      21         9         19.25          5.46
     2      21        21         10.75          9.77

         CHISQ= 15.23 ON  1 DF,   PROB>CHISQ=0.0001

LIKELIHOOD RATIO TEST FOR VARIABLE (REMISS, CENSOR)

 TREAT       N    DEATHS     LAMBDA

     1      21       9     0.025070
     2      21      21     0.115385

         CHISQ= 16.49 ON  1 DF,   PROB>CHISQ=0.0001
```

```
TWO-WAY FREQUENCY TABLES -- MEASURES OF ASSOCIATION
MULTIWAY FREQUENCY TABLES -- LOGLINEAR MODELS (INCLUDING STRUCTURAL ZEROS)

BMDP STATISTICAL SOFTWARE, INC.
1964 WESTWOOD BLVD. SUITE 202
(213) 475-5700
PROGRAM REVISED APRIL    1982
MANUAL REVISED -- 1981
COPYRIGHT (C) 1982 REGENTS OF UNIVERSITY OF CALIFORNIA

TO SEE REMARKS AND A SUMMARY OF NEW FEATURES FOR
THIS PROGRAM, STATE NEWS. IN THE PRINT PARAGRAPH.

   AUGUST 25, 1982  AT 18:35:58

PROGRAM CONTROL INFORMATION

/ PRINT       PAGE = 51.
/ PROBLEM     TITLE IS 'MORRISON BREAST CANCER DATA'.
/ INPUT       VARIABLES ARE 4.
              FORMAT IS FREE.
              TABLE IS 4,2,3,3.
/ VARIABLE    NAMES ARE 'INFL.APP', SURVIVED, AGE,CENTER.
/ CATEGORY    NAMES(1) ARE 'MIN.MAL','MIN.BEN','GRT.MAL','GRT.BEN'.
              CODES(1) ARE 1 TO 4.
              NAMES(2) ARE NO, YES.
              CODES(2) ARE 1, 2.
              NAMES(3) ARE UNDER50, '50-69', OVER69.
              CODES(3) ARE 1 TO 3.
              NAMES(4) ARE TOKYO, BOSTON, GLAMORGN.
              CODES(4) ARE 1 TO 3.
/ TABLE       INDICES ARE 'INFL.APP', SURVIVED, AGE, CENTER.
              DELTA IS 0.5.
/ FIT         MODEL IS ISC,SA,AC.
              MODEL IS SI,IC,SC,SA,AC.
              MODEL IS SI,IC,IA,SC,SA,AC.
              MODEL IS I,S,C,A.
/ END

PROBLEM TITLE IS
MORRISON BREAST CANCER DATA

NUMBER OF VARIABLES TO READ IN. . . . . . . . . .        4
NUMBER OF VARIABLES ADDED BY TRANSFORMATIONS. .        0
TOTAL NUMBER OF VARIABLES . . . . . . . . . . .        4
NUMBER OF CASES TO READ IN. . . . . . . . . . . TO END
CASE LABELING VARIABLES . . . . . . . . . . . .
MISSING VALUES CHECKED BEFORE OR AFTER TRANS. . NEITHER
BLANKS ARE. . . . . . . . . . . . . . . . . . . MISSING
INPUT UNIT NUMBER . . . . . . . . . . . . . . .        5
REWIND INPUT UNIT PRIOR TO READING. . DATA. . .       NO
NUMBER OF WORDS OF DYNAMIC STORAGE. . . . . . .    29102
```

```
*****  OBSERVED FREQUENCY TABLE  1

CENTER    AGE       SURVIVED               INFL.APP
------    -----     -----                  ------
                               MIN.MAL  MIN.BEN  GRT.MAL  GRT.BEN      TOTAL
----------------------------------------------------------------------------

TOKYO     UNDER50   NO               9        7        4        3 |      23
                    YES             26       68       25        9 |     128
                    --------------------------------------------|--------
                    TOTAL           35       75       29       12 |     151

          50-69     NO               9        9       11        2 |      31
                    YES             20       46       18        5 |      89
                    --------------------------------------------|--------
                    TOTAL           29       55       29        7 |     120

          OVER69    NO               2        3        1        0 |       6
                    YES              1        6        5        1 |      13
                    --------------------------------------------|--------
                    TOTAL            3        9        6        1 |      19

----------------------------------------------------------------------------

BOSTON    UNDER50   NO               6        7        6        0 |      19
                    YES             11       24        4        0 |      39
                    --------------------------------------------|--------
                    TOTAL           17       31       10        0 |      58

          50-69     NO               8       20        3        2 |      33
                    YES             18       58       10        3 |      89
                    --------------------------------------------|--------
                    TOTAL           26       73       13        5 |     122

          OVER69    NO               9       18        3        0 |      30
                    YES             15       26        1        1 |      43
                    --------------------------------------------|--------
                    TOTAL           24       44        4        1 |      73

----------------------------------------------------------------------------

GLAMORGN  UNDER50   NO              16        7        3        0 |      26
                    YES             16       20        8        1 |      45
                    --------------------------------------------|--------
                    TOTAL           32       27       11        1 |      71

          50-69     NO              14       12        3        0 |      29
                    YES             27       39       10        4 |      80
                    --------------------------------------------|--------
                    TOTAL           41       51       13        4 |     109

          OVER69    NO               3        7        3        0 |      13

                    YES             12       11        4        1 |      28
                    --------------------------------------------|--------
                    TOTAL           15       18        7        1 |      41

             TOTAL OF THE OBSERVED FREQUENCY TABLE IS      764
```

```
PAGE    1  BMDP4F MORRISON BREAST CANCER DATA


                                                                  INTERVAL RANGE
 VARIABLE        MINIMUM  MAXIMUM  MISSING CATEGORY CATEGORY   GREATER LESS THAN
 NO. NAME         LIMIT    LIMIT    CODE     CODE     NAME       THAN    OR = TO

   1 INFL.APP
                                           1.00000 MIN.MAL
                                           2.00000 MIN.BEN
                                           3.00000 GRT.MAL
                                           4.00000 GRT.BEN
   2 SURVIVED
                                           1.00000 NO
                                           2.00000 YES
   3 AGE
                                           1.00000 UNDER50
                                           2.00000 50-69
                                           3.00000 OVER69
   4 CENTER
                                           1.00000 TOKYO
                                           2.00000 BOSTON
                                           3.00000 GLAMORGN
```

```
*****DELTA= 0.500   IS ADDED TO EACH CELL FOR ALL ANALYSES

                                                      LIKELIHOOD-RATIO            PEARSON
    MODEL                                       D.F.  CHI-SQUARE   PROB         CHI-SQUARE    PROB        ITER.
    -----                                       ----  ----------   ----         ----------    ----        -----
    ISC,SA,AC.                                   40      36.21    0.6418           36.25    0.6399            4
    IS,IC,SC,SA,AC.                              46      41.39    0.6654           41.83    0.6674            4
    IS,IC,IA,SC,SA,AC.                           40      39.92    0.4738           40.17    0.4629            5
    I,S,C,A.                                     63     174.34    0.0000          181.39    C.0000            7

NUMBER OF INTEGER WORDS OF STORAGE USED IN PRECEDING    PROBLEM    1546
CPU TIME USED      4.430 SECONDS
```

```
1                                  S T A T I S T I C A L   A N A L Y S I S   S Y S T E M      16:02 WEDNESDAY, AUGUST 25, 1982

NOTE: THE JOB XPRS5781 HAS BEEN RUN UNDER RELEASE 79.5 OF SAS AT KANSAS STATE UNIVERSITY (00'0E).

NOTE: SAS OPTIONS SPECIFIED ARE:
        SORT=4


1            OPTIONS PAGESIZE=51;
2            PROC MATRIX;
3            K=1 -1 0 0 0 0 0 0/
4            1 -1 1 -1 0 0 0 0 /
5            1 -1 1 -1 1 -1 0 0 /
6            1 -1 1 -1 1 -1 1 -1;
7            A=
8            1 0 .5 0 0 0 0 0 0 0 0 0/
9            1 1 .5 0 0 0 0 0 0 0 0 0/
10           0 0 0 1 0 .5,0 0 0 0 0 0/
11           0 0 0 1 1 .5 0 0 0 0 0 0/
12           0 0 0 0 0 0 1 0 .5 0 0 0/
13           0 0 0 0 0 0 1 1 .5 0 0 0/
14           0 0 0 0 0 0 0 0 0 1 0 .5/
15           0 0 0 0 0 0 0 0 0 1 1 .5;
16           P=.4762 .3730 .1508 .6833 .0833 .2833 .5526 .0526 .3947 .4762 .0952 .4286;
17           API=A*P';
18           F=K*LOG(API);
19           X=1 1/1 2/1 3/1 4;
20           V=.0019796 -.0014097 -.0005699 0 0 0 0 0 0 0 0 0/
21           -.0014097 .0018561 -.0004464 0 0 0 0 0 0 0 0 0 /
22           -.0005699 -.0004464 .0010163 0 0 0 0 0 0 0 0 0/
23           0 0 0 .0038705 -.0008792 -.0029902 0 0 0 0 0 0/
24           0 0 0 -.0008792 .0012727 -.0003933 0 0 0 0 0 0/
25           0 0 0 -.0029902 -.0003933 .003384 0 0 0 0 0 0/
26           0 0 0 0 0 0 .0065061 -.0007649 -.0057398 0 0 0/
27           0 0 0 0 0 0 -.0007649 .0013114 -.0005463 0 0 0/
28           0 0 0 0 0 0 -.0057398 -.0005463 .0062872 0 0 0/
29           0 0 0 0 0 0 0 0 0 .0118778 -.0021588 -.009719/
30           0 0 0 0 0 0 0 0 0 -.0021588 .0041018 -.001943/
31           0 0 0 0 0 0 0 0 0 -.009719 -.001943 .0011662;
32           DA=DIAG(API);
33           DAINV=INV(DA);
34           VF=K*DAINV*A*V*A'*DAINV'*K';
35           VFINV=INV(VF);
36           PART=X'*VFINV*X;
37           BETA=INV(PART)*X'*VFINV*F;
38           SSFIT=F'*VFINV*F-(BETA'*(X'*VFINV*X)*BETA);
39           PRINT K A P API F;
40           PRINT X V DA VF VFINV BETA;
41           PRINT SSFIT;
42           X=1/2/3/4;
```

```
                    S T A T I S T I C A L   A N A L Y S I S   S Y S T E M                                1
                                                                          16:02 WEDNESDAY, AUGUST 25, 1982

K          COL1      COL2      COL3      COL4      COL5      COL6      COL7      COL8

ROW1          1        -1         0         0         0         0         0         0
ROW2          1        -1         1        -1         0         0         0         0
ROW3          1        -1         1        -1         1        -1         0         0
ROW4          1        -1         1        -1         1        -1         1        -1


         A             COL1      COL2      COL3      COL4      COL5      COL6
                       COL7      COL8      COL9     COL10     COL11     COL12

         ROW1             1         0       0.5         0         0         0
                          0         0         0         0         0         0

         ROW2             1         1       0.5         0         0         0
                          0         0         0         0         0         0

         ROW3             0         0         0         1         0       0.5
                          0         0         0         0         0         0

         ROW4             0         0         0         1         1       0.5
                          0         0         0         0         0         0

         ROW5             0         0         0         0         0         0
                          1         0       0.5         0         0         0

         ROW6             0         0         0         0         0         0
                          1         1       0.5         0         0         0

         ROW7             0         0         0         0         0         0
                          0         0         0         1         0       0.5

         ROW8             0         0         0         0         0         0
                          0         0         0         1         1       0.5


         P             COL1      COL2      COL3      COL4      COL5      COL6
                       COL7      COL8      COL9     COL10     COL11     COL12

         ROW1        0.4762     0.373    0.1508    0.6833    0.0833    0.2833
                     0.5526    0.0526    0.3947    0.4762    0.0952    0.4286
```

```
V                 COL1        COL2        COL3        COL4        COL5        COL6
                  COL7        COL8        COL9        COL10       COL11       COL12

ROW1         0.0019796  -0.0014097  -0.0005699           0           0           0
                     0           0           0           0           0           0

ROW2        -0.0014097   0.0018561  -0.0004464           0           0           0
                     0           0           0           0           0           0

ROW3        -0.0005699  -0.0004464   0.0010163           0           0           0
                     0           0           0           0           0           0

ROW4                 0           0           0   0.0038705  -0.0008792  -0.0029902
                     0           0           0           0           0           0

ROW5                 0           0           0  -0.0008792   0.0012727  -0.0003933
                     0           0           0           0           0           0

ROW6                 0           0           0  -0.0029902  -0.0003933    0.003384
                     0           0           0           0           0           0

ROW7                 0           0           0           0           0           0
             0.0065061  -0.0007649  -0.0057398           0           0           0

ROW8                 0           0           0           0           0           0
            -0.0007649   0.0013114  -0.0005463           0           0           0

ROW9                 0           0           0           0           0           0
            -0.0057398  -0.0005463   0.0062872           0           0           0

ROW10                0           0           0           0           0           0
                     0           0           0   0.0118778  -0.0021588   -0.005719

ROW11                0           0           0           0           0           0
                     0           0           0  -0.0021588   0.0041018   -0.001943

ROW12                0           0           0           0           0           0
                     0           0           0   -0.009719   -0.001943   0.0011662
```

```
APT             COL1

ROW1          0.5516
ROW2          0.9246
ROW3         0.32495
ROW4         0.90825
ROW5         0.74995
ROW6         0.80255
ROW7          0.6905
ROW8          0.7857


F               COL1

ROW1       -0.516538
ROW2       -0.612735
ROW3       -0.680523
ROW4       -0.809682


X               COL1        COL2

ROW1               1           1
ROW2               1           2
ROW3               1           3
ROW4               1           4
```

```
DA            COL1        COL2        COL3        COL4        COL5        COL6        COL7        COL8

ROW1        0.5516           0           0           0           0           0           0           0
ROW2             0      0.9246           0           0           0           0           0           0
ROW3             0           0     0.82495           0           0           0           0           0
ROW4             0           0           0     0.90825           0           0           0           0
ROW5             0           0           0           0     0.74995           0           0           0
ROW6             0           0           0           0           0     0.80255           0           0
ROW7             0           0           0           0           0           0      0.6905           0
ROW8             0           0           0           0           0           0           0      0.7857


                    VF                COL1          COL2          COL3          COL4

                    ROW1        0.00564435    0.00564435    0.00564435    0.00564435
                    ROW2        0.00564435    0.00747189    0.00747189    0.00747189
                    ROW3        0.00564435    0.00747189    0.00975189    0.00575189
                    ROW4        0.00564435    0.00747189    0.00975189       0.01787


                    VFINV             COL1          COL2          COL3          COL4

                    ROW1           724.352      -547.193   -1.0810E-13   -2.4587E-15
                    ROW2          -547.183       985.781      -430.597             0
                    ROW3                 0      -430.597       561.778      -123.181
                    ROW4                 0             0      -123.181       123.181


                              BETA              COL1

                              ROW1         -0.427916
                              ROW2        -0.0886223


                              SSFIT             COL1

                              ROW1          0.424197
```

BIBLIOGRAPHY

Bishop, Y. M. M., Fienberg S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Mass. The MIT Press.

Cox, D. R. (1972). Regression models and life tables (with discussion). J. R. Stat. Soc. B, Vol. 34, 187-202.

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

Dixon, W. S., editor, BMDP-81 Biomedical Computer Programs, P-Series. Los Angeles: University of Califormia Press, 1981.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. J. Am. Stat. Assoc., Vol. 72, 555-565.

Feinberg, S. E. (1981). *The Analysis of Cross-Classified Categorical Data.* Cambridge, Mass.: The MIT Press.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. (1969). Analysis of Categorical Data by Linear Models. Biometrics, Vol. 25, 489-504.

Helwig, J. T. and Council, K. A. (1979). SAS User's Guide, 1979 Edition. SAS Institute, Inc., Raleigh, North Carolina.

Helwig, J. T. and Council, K. A. (1981). SAS User's Guide, 1981 Supplemental Edition. SAS Institute, Inc., Raleigh, North Carolina.

Kalbfliesch, J. D., and Prentice, R. L., (1973). Marginal likelihood based on Cox's regression and life model. Biometrika, Vol. 60, 267-279.

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

Koch, G. G., Johnson, W. D., and Tolley, H.D., (1972). A linear model approach to the analysis of survival and extent of disease in multidimensional contingency tables. J. Am. Stat. Assoc. Vol. 67, 783-795.

Lamm, R. M., (1981). Loglinear modeling with FUNCAT. Ohio State University: Technical Report No. 230.

Lawless, J. F., (1981). *Statistical Models and Methods for Lifetime Data.* New York: Wiley.

Morrison, A.S., Black, M.M., Lowe, C.R., MacMahon, B., Yuasa, S., (1973). Some international differences in histology and survival in breast cancer. Int. J. Cancer, Vol. 11, 261-267.

Tsiatis, A. A., (1980). A large sample study of Cox's regression model. Annals of Statistics, Vol. 9, 93-108.

THE USE OF GENERAL LINEAR MODELS FOR FAILURE DATA AND CATEGORICAL DATA

by

ROGER MARK SAUTER

B.A., Mid-America Nazarene College, 1980

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1982

ABSTRACT

There are two problems regarding the analysis of failure data considered in this paper. The first problem looked at is the analysis of censored failure data. The analysis is straight forward if none of the individuals are censored. Cox (1972) suggested a distribution-free approach in the use of the proportional hazards model. Definitions are given for this procedure along with the motivation for its use and two examples. The second problem dealt with is the analysis of categorical data. The two approachs used are the loglinear model as presented by Feinberg (1981) and the GSK model given by Grizzle, Starmer, and Koch (1969). Once again definitions are given for these approaches along with the general procedure for the analysis of these models and an example of each model.