This research was supported in part by an American Cancer Society Cancer Institutional Sub-Grant No. IN-15-M and by National Institutes of Health, Institute of General Medical Sciences Grants GM-70004-01, GM-0038-18, and GM-12868-08.

# SOME FURTHER REMARKS ON A LINEAR MODELS APPROACH TO THE ANALYSIS OF SURVIVAL RATES

by

Gary G. Koch, William D. Johnson, and H. Dennis Tolley

Department of Biostatistics
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina

Institute of Statistics Mimeo Series No. 794

January 1972

## ABSTRACT

Matrix algorithms are presented which generate estimates of t-year survival rates for patients with chronic disease from a categorical data approach. Weighted least squares has been applied to the resulting estimates in order to fit linear models which, together with large sample theory, provide a straightforward and unified method for testing hypotheses of interest.

# A LINEAR MODELS APPROACH TO THE ANALYSIS OF SURVIVAL AND EXTENT OF DISEASE IN MULTI-DIMENSIONAL CONTINGENCY TABLES

Gary G. Koch, William D. Johnson, and H. Dennis Tolley
University of North Carolina

## 1. INTRODUCTION

One problem of interest in the study of human subjects with chronic disease has been the statistical evaluation of the effectiveness of therapy. Historically, the t-year survival rate (or the probability that an exposed individual is alive t years subsequent to diagnosis, treatment or some other definitive zero point) has provided an objective criterion for this purpose. Clinical trials belonging to the general class of patient follow-up studies frequently have as their primary objective the estimation and comparison of survival rates for various groups of interest. Hence, individuals with certain diseases (e.g., particular types of cancer) may be observed following treatment until recurrence of the disease, death, or close of the study. Difficulties arise here since some patients are not traceable for the entire t years either because of death due to unrelated causes, lost to follow-up, or termination of the study. Moreover, if the study period is lengthy, a cohort effect may prevail. Many aspects of these and other problems in the analysis of follow-up studies are discussed in detail by Chiang [9, Chapter 12]. Other significant contributions include the studies of Greenwood [19], Frost [16], Karn [23], Boag [5], Dorn [12], Fix and Neyman [15], Berkson and Gage [3], Littell [25], Epstein and Sobel [14], Merrell and Shulman [27], Cutler and Ederer [10], Elveback [13], Armitage [2], Chiang [6,7,8] and Myers, Axtell, and Zelen [28].

In the analysis of clinical follow-up studies, individuals are frequently cross-classified according to several variables, with survival rates differing across the resulting groups. Examples of classification variables include age, social class, stage of disease at diagnosis, and treatment. Statistical techniques have been developed to adjust for the effect of such variables on overall survival rates, and thereby allow the analyst to compare survival rates for the different treatments with greater precision. One approach to this problem is formulated in terms of a distribution for survival rates (Myers, Axtell, and Zelen [28]), while another involves procedures for pairing treated patients with controls (Mantel and Haenszel [26]). However, in certain situations a simple arithmetic correlation of results is sufficient. The latter procedure is often applied for age adjustment.

This paper indicates how a general method for analyzing qualitative data can be used to estimate the survival rate for each of several groups and to test the statistical significance of the effects of the underlying variables. In particular, one can investigate whether certain factor groups (e.g., different treatments) have an overall effect on the survival rate after accounting for other categorical effects (e.g., stage of disease, age, and sex). For this purpose, primary emphasis will be given to the actuarial approach of Berkson and Gage [3]. Other authors who have used this method include Littell [25], Cutler and Ederer [10], and Elveback [13].

The effective number of individuals exposed to risk during a particular time interval (e.g., a particular year) of study poses a problem with this approach because of withdrawal and lost to follow-up patients. An individual is considered to be a withdrawal if alive until the termination date of the study which occurs, however, prior to the t-th year of follow-up. Generally, these cases are viewed as having been exposed to risk and survived all time intervals occurring prior to the termination date and half of the interval which includes the termination

date. Littell [25] has shown that this method for handling withdrawals gives a biased estimate of the t-year survival rate; and this bias is independent of sample size, increases as the t-year survival rate decreases, and decreases as the number of intervals comprising the t-year period in the analysis increases.

Patients are lost to follow-up for such reasons as emigration, change of address, enlistment in the Armed Forces, change of name by marriage, or intrinsic lack of cooperation. Since the survival status of these patients remains unknown subsequent to their last follow-up appearance, there is no single correct way of computing the effective number of individuals exposed to risk in the presence of lost to follow-up. In particular, if the researcher excludes such patients altogether from the analysis (aside from reporting their number), an underestimate of the t-year survival rate will be obtained. On the other hand, if he reasons that lost patients would surely return for treatment if necessary and counts them as survivors, an over-estimate will be produced. Thus, Cutler and Ederer [10] and Dorn [12] have implied that lost to follow-up patients can be regarded as withdrawals in the interval they were lost; i.e., they are viewed as having been exposed to risk and survived all time intervals which were completed prior to being lost and half of the interval during which lost to follow-up occurred. This approach, in some sense, provides maximum utility of the information contributed in the length of time they were studied. More recently, Chiang [8,9] suggests it is more appropriate to treat lost to follow-up as a competing risk. In this context, t-year survival rates are computed with lost to follow-up removed as a risk.

In this initial discussion of the use of general linear models to analyze categorical data with survival as a response variable, we have followed the procedure described by Cutler and Ederer [10]. Other models will be considered in future investigations.

## 2. GENERAL LINEAR MODEL APPROACH APPLIED TO SURVIVAL RATE

Grizzle, Starmer, and Koch [20] (subsequently abbreviated GSK) have described a method of using linear regression models and weighted least squares for

testing hypotheses and fitting simplified models to multi-dimensional contingency tables which arise when frequency counts are obtained for cross-classifications of qualitative variables. Their formulation offers a unified approach to the analysis of such data with the flexibility of estimating both linear and non-linear functions of the cell proportions together with a corresponding variance-covariance matrix. The resulting test statistics belong to the class of minimum modified chi-square statistics due to Neyman [29] which is equivalent to the general quadratic form criteria of Wald [30]. As a result, they have central $\chi^2$-distributions when the corresponding null hypotheses at which they are directed are true. Alternative analyses include the method based on maximum likelihood as formulated by Bishop [4] and Goodman [17,18] and the method based on minimum discrimination information as described by Ku, Varner and Kullback [24]. All three of these methods are based on BAN estimates (Neyman [29]) and as such are asympototically equivalent. We have chosen the least squares approach because of computational convenience.[1]

To illustrate the application of the GSK approach to the analysis of survival rates in the categorical data framework, we consider the hypothetical data in Table 1. These data pertain to the special case where a single decrement is involved, but where both withdrawals and lost to follow-up must be considered. If deaths unrelated to the disease of interest were present, they would be dealt with in the same manner as the withdrawal and lost to follow-up patients (although in this event, a competing risk analysis may be more appropriate). The groups in Table 1 correspond to various aggregates of individuals which have been formed on the basis of characteristics like age at diagnosis, stage of disease at diagnosis, or method of treatment. Within each particular group, an individual is identified with one or more of t

---

[1] Computer programs which permit the efficient calculation of the corresponding estimates and test-statistics are not difficult to prepare. In particular, the one used for the matrix operations in the analyses of this paper can be obtained from: Program Librarian, Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, North Carolina, 27514.

TABLE 1

HYPOTHETICAL MULTIDIMENSIONAL CONTINGENCY TABLE
WITH SURVIVAL AS A RESPONSE

| Population | | Exposure Year $j$ (1) | Survived the Year $\ell^i_{j+1}$ (2) | Died During Year $d^i_j$ (3) | Withdrawn or Lost $w^i_j$ (4) | Total Alive at Beginning of Year $\ell^i_j$ (5) |
|---|---|---|---|---|---|---|
| Group 1 | 1 | 1 | $n_{111}$ | $n_{112}$ | $n_{113}$ | $n_{11.}$ |
| | 2 | 2 | $n_{121}$ | $n_{122}$ | $n_{123}$ | $n_{12.}$ |
| | . | . | . | . | . | . |
| | . | . | . | . | . | . |
| | . | . | . | . | . | . |
| | t | t | $n_{1t1}$ | $n_{1t2}$ | $n_{1t3}$ | $n_{1t.}$ |
| Group 2 | t+1 | 1 | $n_{211}$ | $n_{212}$ | $n_{213}$ | $n_{21.}$ |
| | t+2 | 2 | $n_{221}$ | $n_{222}$ | $n_{223}$ | $n_{22.}$ |
| | . | . | . | . | . | . |
| | . | . | . | . | . | . |
| | . | . | . | . | . | . |
| | 2t | t | $n_{2t1}$ | $n_{2t2}$ | $n_{2t3}$ | $n_{2t.}$ |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| Group r | (r-1)t+1 | 1 | $n_{r11}$ | $n_{r12}$ | $n_{r13}$ | $n_{r1.}$ |
| | (r-1)t+2 | 2 | $n_{r21}$ | $n_{r22}$ | $n_{r23}$ | $n_{r2.}$ |
| | . | . | . | . | . | . |
| | . | . | . | . | . | . |
| | . | . | . | . | . | . |
| | rt | t | $n_{rt1}$ | $n_{rt2}$ | $n_{rt3}$ | $n_{rt.}$ |

populations at risk according to the number of years of exposure (i.e., either all t years or the number of years until either death, lost to follow-up or withdrawal has occurred). It follows that there are a total of rt populations that are identified by respective group x exposure year combinations.

With this general setting fixed, $n_{ijk}$ represents the number of individuals in the (ij)-th row (i.e., in the i-th group, j-th exposure year) who give the k-th response. The quantities $n_{ij1}$ in Column (2) represent the numbers known to have survived the corresponding exposure year; the quantities $n_{ij2}$ in Column (3) represent the numbers known to have died during the corresponding exposure year; and the quantities $n_{ij3}$ in Column (4) represent the numbers known to have been lost to follow-up or withdrawn during the corresponding exposure year. In this manner, an individual in the i-th group is classified into one of three cells for each year of exposure which is applicable. Finally, it can be noted that $n_{ij1}$, $n_{ij2}$ and $n_{ij3}$ correspond to the life table quantities $\ell^i_{j+1}$, $d^i_j$, and $w^i_j$ respectively.

Henceforth, we shall assume that the vector $\underset{\sim}{n}_{ij}$ where $\underset{\sim}{n}'_{ij} = (n_{ij1}, n_{ij2}, n_{ij3})$ has the multinomial distribution with parameters $n_{ij.}$ and $\underset{\sim}{\pi}_{ij}$ where $n_{ij.} = (n_{ij1}+n_{ij2}+n_{ij3})$ is the number of individuals alive at the beginning of the j-th exposure year and $\underset{\sim}{\pi}'_{ij} = (\pi_{ij1}, \pi_{ij2}, \pi_{ij3})$. With this notation, $\pi_{ij1}$ represents the probability an individual in the i-th group who is alive at the beginning of the j-th year is exposed to risk for an entire year and survives the year (so that he is alive at the beginning of the(j+1)-th year. Similarly, $\pi_{ij2}$ is the probability an individual in the i-th group who is alive at the beginning of the j-th year is exposed to risk until death at some time during that year. Finally, $\pi_{ij3}$ is the probability an individual in the i-th group who is alive at the beginning of the j-th year is exposed to risk for a period less than one year (due to termination of the study, lost to follow-up, etc.) and survives the period of exposure.[2]

---

[2] See Appendix for a more precise interpretation of $\pi_{ij3}$.

From the properties of the multinomial distribution, it follows that $p_{ijk} = (n_{ijk}/n_{ij.})$ is an unbiased estimator of $\pi_{ijk}$. Moreover, the variance-covariance matrix corresponding to the vector $\underset{\sim}{p}_{ij}$ where $\underset{\sim}{p}'_{ij} = (p_{ij1}, p_{ij2}, p_{ij3})$ is

$$\underset{\sim}{V}(\underset{\sim}{\pi}_{ij}) = \underset{\sim\sim}{\mathrm{Var}}(\underset{\sim}{p}_{ij}) = \frac{1}{n_{ij.}} \begin{bmatrix} \pi_{ij1}(1-\pi_{ij1}) & -\pi_{ij1}\pi_{ij2} & -\pi_{ij1}\pi_{ij3} \\ -\pi_{ij1}\pi_{ij2} & \pi_{ij2}(1-\pi_{ij2}) & -\pi_{ij2}\pi_{ij3} \\ -\pi_{ij1}\pi_{ij3} & -\pi_{ij2}\pi_{ij3} & \pi_{ij3}(1-\pi_{ij3}) \end{bmatrix}$$

At this point, it is appropriate to form composite vectors $\underset{\sim}{\pi}$ and $\underset{\sim}{p}$ where

$$\underset{\sim}{\pi}' = (\underset{\sim}{\pi}'_{11}, \underset{\sim}{\pi}'_{12}, \cdots, \underset{\sim}{\pi}'_{rt})$$

$$\underset{\sim}{p}' = (\underset{\sim}{p}'_{11}, \underset{\sim}{p}'_{12}, \cdots, \underset{\sim}{p}'_{rt})$$

which correspond to all group x exposure year combinations.

Accordingly, $\underset{\sim}{p}$ is an unbiased estimator of $\underset{\sim}{\pi}$. Also, since the data for the respective exposure years within each group are uncorrelated and since the experiences of each group are statistically independent, a consistent estimate for the variance-covariance matrix of $\underset{\sim}{p}$ is the block diagonal matrix $\underset{\sim}{V}$ defined in (2.1) where $\underset{\sim}{V}(\underset{\sim}{p}_{ij})$ is $\underset{\sim}{V}(\underset{\sim}{\pi}_{ij})$ with $\underset{\sim}{\pi}_{ij}$ replaced by $\underset{\sim}{p}_{ij}$.

$$\underset{\sim}{V} = \begin{bmatrix} \underset{\sim}{V}(\underset{\sim}{p}_{11}) & \underset{\sim}{0} & \cdots & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{V}(\underset{\sim}{p}_{12}) & \cdots & \underset{\sim}{0} \\ \cdots & \cdots & \cdots & \cdots \\ \underset{\sim}{0} & \underset{\sim}{0} & \cdots & \underset{\sim}{V}(\underset{\sim}{p}_{rt}) \end{bmatrix} \qquad (2.1)$$

For any set of s functions of the cell probabilities $\underset{\sim}{p}$ that can be written in the general log-linear vector form

$$\underset{\sim}{F} = \underset{\sim}{F}(\underset{\sim}{p}) = \underset{\sim}{K}[\underset{\sim\sim}{\log_e}(\underset{\sim}{A}\underset{\sim}{p})]$$

where $\underset{\sim}{A}$ is a known (u x 3rt) matrix, $\underset{\sim}{K}$ is a known s x u matrix, and $\underset{\sim\sim}{\log_e}$ is a vector operation which transforms a given vector into the corresponding vector of natural logarithms, GSK have shown that a consistent estimate of the corresponding variance-covariance matrix is

$$\underset{\sim}{V}_{\underset{\sim}{F}} = \underset{\sim}{K}\underset{\sim}{D}_{\underset{\sim}{a}}^{-1}\underset{\sim}{A}\underset{\sim}{V}\underset{\sim}{A}'\underset{\sim}{D}_{\underset{\sim}{a}}^{-1}\underset{\sim}{K}' \qquad (2.2)$$

where $\underset{\sim}{D}_{\underset{\sim}{a}}$ is a u x u diagonal matrix with the elements of the vector $\underset{\sim}{a} = \underset{\sim}{A}\underset{\sim}{p}$ on the main diagonal. By suitable choices of the $\underset{\sim}{A}$ and $\underset{\sim}{K}$ matrices associated with the functions $\underset{\sim}{F}$, the analyst has considerable flexibility in formulating estimators which are relevant to the particular structure associated with the data at which the analysis is directed.

We now construct the vectors and matrices required to produce functions of cell proportions which can be used to generate an estimate of the probability of survival for t years within each group. As in Cutler and Ederer [10], it will be assumed that all patients withdrawn or lost to follow-up in a given year were exposed to risk of dying, on the average, for half of that year. We have

$$\underset{\sim}{p}' = (p_{111}, p_{112}, p_{113}, \cdots, p_{1t1}, p_{1t2}, p_{1t3}, \cdots, p_{r11}, p_{r12}, p_{r13}, \cdots, p_{rt1}, p_{rt2}, p_{rt3})$$

and choose the (2rt x 3rt) matrix $\underset{\sim}{A}$ in (2.3) which is a block diagonal matrix with

$$\underset{\sim}{A} = \begin{bmatrix} \underset{\sim}{A}^* & \underset{\sim}{0} & \cdots & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{A}^* & \cdots & \underset{\sim}{0} \\ \cdots & \cdots & \cdots & \cdots \\ \underset{\sim}{0} & \underset{\sim}{0} & \cdots & \underset{\sim}{A}^* \end{bmatrix} \qquad (2.3)$$

blocks $\underset{\sim}{A}^*$ defined in (2.4) and the (r x 2rt) matrix $\underset{\sim}{K}$ in (2.5) which is a block

$$\underset{\sim}{A}^* = \begin{bmatrix} 1 & 0 & 0.5 \\ 1 & 1 & 0.5 \end{bmatrix} \qquad (2.4)$$

$$\underset{\sim}{K} = \begin{bmatrix} \underset{\sim}{K}^* & \underset{\sim}{0} & \cdots & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{K}^* & \cdots & \underset{\sim}{0} \\ \cdots & \cdots & \cdots & \cdots \\ \underset{\sim}{0} & \underset{\sim}{0} & \cdots & \underset{\sim}{K}^* \end{bmatrix} \qquad (2.5)$$

diagonal matrix with blocks $\underset{\sim}{K}^*$ defined by the (1 x 2t) vector in (2.6).

$$\underset{\sim}{K}^* = (\ 1 \quad -1 \quad 1 \quad -1 \quad \ldots \quad 1 \quad -1). \qquad (2.6)$$

The resulting vector of functions obtained as $\underset{\sim}{F} = \underset{\sim}{K}[\log_e(\underset{\sim}{A}\underset{\sim}{p})]$ with $\underset{\sim}{A}$ and $\underset{\sim}{K}$ being the matrices defined in (2.3) and (2.5) may be written as

$$\underset{\sim}{F}' = [F(\underset{\sim}{p})]' = [\log_e G_1, \log_e G_2, \ldots, \log_e G_r] \qquad (2.7)$$

where $G_i$ denotes the estimate of the t-year survival rate for the i-th group obtained from the expression

$$G_i = \prod_{j=1}^{t} (p_{ij1} + 0.5p_{ij3})/(p_{ij1} + p_{ij2} + 0.5p_{ij3}). \qquad (2.8)$$

The corresponding estimated covariance matrix $\underset{\sim}{V}_{\underset{\sim}{F}}$ shown in (2.9) may be obtained by the use of (2.2) where "var" means sample estimate of variance.

$$\underset{\sim}{V}_F = \begin{bmatrix} \text{var } (\log_e G_1) & 0 & \cdots & 0 \\ 0 & \text{var}(\log_e G_2) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \text{var}(\log_e G_r) \end{bmatrix} \quad (2.9)$$

As derived in this discussion, $\underset{\sim}{F}(p)$ pertains to the logarithms of the survival rates. To obtain the survival rate vector, we transform $\underset{\sim}{F}(p)$ by the exponential function, exp (log $G_i$), to get $\underset{\sim}{G}(p)$. Corresponding to this transformation, we have a diagonal covariance matrix $\underset{\sim}{W}$, for $\underset{\sim}{G}(p)$ with elements of the type

$$\text{var } G_i = G_i^2[\text{var } (\log_e G_i)]$$

These calculations can also be formulated in terms of standard vector and matrix operations and thus can be readily performed.

The variation in the elements of the resulting $\underset{\sim}{F}(p)$ (or $\underset{\sim}{G}(p)$) vector is analyzed by fitting the linear regression model $\underset{\sim\sim}{X}\beta$ where $\underset{\sim}{X}$ is a known (r x d) matrix of coefficients with rank d $\leq$ r, sometimes called the design matrix, and $\beta$ is a vector of unknown parameters. The estimate $\underset{\sim}{b}$ of $\beta$ is obtained by minimizing $(\underset{\sim}{F}(p)-\underset{\sim\sim}{X}b)'\underset{\sim}{V}_F^{-1}(\underset{\sim}{F}(p)-\underset{\sim\sim}{X}b)$. If the data are adequately described by this model, a test of the hypothesis $H_0$: $\underset{\sim\sim}{C}\beta=\underset{\sim}{0}$, where $\underset{\sim}{C}$ is a known (c x d) coefficient matrix with rank c $\leq$ d which determines a comparison of interest, is produced by the conventional methods of weighted multiple regression. Specifically, the test statistic is

$$SS(\underset{\sim\sim}{C}\beta=\underset{\sim}{0}) = \underset{\sim}{b}'\underset{\sim}{C}'[\underset{\sim}{C}(\underset{\sim}{X}'\underset{\sim}{V}_F^{-1}\underset{\sim}{X})^{-1}\underset{\sim\sim}{C}b \quad (2.10)$$

which has approximately a chi-square distribution in large samples. The degrees of freedom for this test is the rank c of the matrix $\underset{\sim}{C}$. In addition, a goodness of fit test to determine the validity of the model is

$$SS[\underset{\sim}{F}(\pi) = \underset{\sim\sim}{X}\beta] = \underset{\sim}{F}'\underset{\sim}{V}_F^{-1}\underset{\sim}{F} - \underset{\sim}{b}'(\underset{\sim}{X}'\underset{\sim}{V}_F^{-1}\underset{\sim}{X})\underset{\sim}{b}, \quad (2.11)$$

which, under the hypothesis that the model fits, has approximately a chi-square distribution in large samples with D.F. = (number of elements in $\underset{\sim}{F}(p)$) - Rank ($\underset{\sim}{X}$) =

(r-d) here. Finally, it can be noted that this type of weighted least squares analysis can be applied either to $\underset{\sim}{F}(\underset{\sim}{p})$ or $\underset{\sim}{G}(\underset{\sim}{p})$. There are advantages in an analysis using $\underset{\sim}{F}(\underset{\sim}{p})$ provided that a model on a relative (or multiplicative) scale is warranted.

The data in Table 1 have been arranged in the usual life table format. Alternatively, the same results can be obtained by viewing the data in the strict contingency table format shown in Table 2. In this event, one considers the r groups

TABLE 2

ALTERNATIVE CONTINGENCY TABLE FOR HYPOTHETICAL DATA

| | Died During Exposure Year | | | | Withdrawn or Lost During Exposure Year | | | | Alive at End of t-th Exposure Year | Total Entered |
|---|---|---|---|---|---|---|---|---|---|---|
| Group | 1 | 2 | ... | t | 1 | 2 | ... | t | | |
| 1 | $n_{112}$ | $n_{122}$ | ... | $n_{1t2}$ | $n_{113}$ | $n_{123}$ | ... | $n_{1t3}$ | $n_{1t1}$ | $n_{1..}$ |
| 2 | $n_{212}$ | $n_{222}$ | ... | $n_{2t2}$ | $n_{213}$ | $n_{223}$ | ... | $n_{2t3}$ | $n_{2t1}$ | $n_{2..}$ |
| . | . | . | ... | . | . | . | ... | . | . | . |
| . | . | . | ... | . | . | . | ... | . | . | . |
| . | . | . | ... | . | . | . | ... | . | . | . |
| r | $n_{r12}$ | $n_{r22}$ | ... | $n_{rt2}$ | $n_{r13}$ | $n_{r23}$ | ... | $n_{rt3}$ | $n_{rt1}$ | $n_{r..}$ |

as separate independent multinomial populations in which each individual is assigned to one of (2t + 1) mutually exclusive categories. The relevant model is given in (2.12)

$$\phi = \prod_{i=1}^{r} \frac{n_{i..}!}{\left[\prod_{j=1}^{t} n_{ij2}!n_{ij3}!\right] n_{it1}!} \prod_{j=1}^{t} \left[\xi_{ij2}^{n_{ij2}} \xi_{ij3}^{n_{ij3}}\right] \xi_{it1}^{n_{it1}} \qquad (2.12)$$

where $n_{i..}(=n_{i11}+n_{i12}+n_{i13})$ is the total number of individuals in the i-th group who are in the study and where $\xi_{ij2}$ is the probability that an individual in the i-th group will be exposed to risk until death which occurs at some time during the j-th year, $\xi_{ij3}$ is the probability that an individual in the i-th group will be exposed to risk until lost to follow-up or withdrawal which occurs at some time during

the j-th year[3], and $\xi_{itl}$ is the probability that an individual in the i-th group will be exposed to risk for the entire t year period and survives. The sample estimates of the respective $\xi_{ijk}$ for k = 2,3 are the corresponding quantities $y_{ijk} = (n_{ijk}/n_{i..})$ where i = 1,2,...,r and j = 1,2,...,t; while the sample estimates of the $\xi_{itl}$ are the corresponding quantities $y_{itl} = (n_{itl}/n_{i..})$ where i = 1,2,...,r. We can then form the composite [r(2t+1) x 1] vector $\underset{\sim}{y}$ where $\underset{\sim}{y}' = (\underset{\sim}{y}_1', \underset{\sim}{y}_2', \ldots, \underset{\sim}{y}_r')$ with $\underset{\sim}{y}_i' = (y_{il2}, \ldots, y_{it2}, y_{il3}, \ldots, y_{it3}, y_{itl})$. With this notation, estimates of the t year survival rates for each group may be obtained by using a [2rt x r(2t+1)] matrix $\underset{\sim}{A}$ like that in (2.3) except that now there are r blocks with $\underset{\sim}{A}^*$ as the [2t x (2t+1)] matrix defined in (2.13) instead of rt blocks with $\underset{\sim}{A}^*$ as the (2 x 3) matrix in (2.4). The $\underset{\sim}{K}$ matrix would be the same as that shown in (2.5) and (2.6).

$$\underset{\sim}{A}^* = \begin{bmatrix}
0 & 1 & 1 & \ldots & 1 & 1 & 0.5 & 1 & 1 & \ldots & 1 & 1 & 1 \\
1 & 1 & 1 & \ldots & 1 & 1 & 0.5 & 1 & 1 & \ldots & 1 & 1 & 1 \\
0 & 0 & 1 & \ldots & 1 & 1 & 0 & 0.5 & 1 & \ldots & 1 & 1 & 1 \\
0 & 1 & 1 & \ldots & 1 & 1 & 0 & 0.5 & 1 & \ldots & 1 & 1 & 1 \\
0 & 0 & 0 & \ldots & 1 & 1 & 0 & 0 & 0.5 & \ldots & 1 & 1 & 1 \\
0 & 0 & 1 & \ldots & 1 & 1 & 0 & 0 & 0.5 & \ldots & 1 & 1 & 1 \\
 & \ldots & & \ldots & & \ldots & & \ldots & & \ldots & & \ldots & \\
0 & 0 & 0 & \ldots & 0 & 1 & 0 & 0 & 0 & \ldots & 0.5 & 1 & 1 \\
0 & 0 & 0 & \ldots & 1 & 1 & 0 & 0 & 0 & \ldots & 0.5 & 1 & 1 \\
0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0.5 & 1 \\
0 & 0 & 0 & \ldots & 0 & 1 & 0 & 0 & 0 & \ldots & 0 & 0.5 & 1
\end{bmatrix} \qquad (2.13)$$

For the special case where t = 5, the appropriate $\underset{\sim}{A}^*$ matrix here is (2.13) with all of the "..." removed.

---

[3] The $\xi_{ij3}$ formulated here could be alternatively decomposed into components like those considered with respect to $\pi_{ij3}$ in the appendix.

Finally, regardless of whether the data are arranged as in Table 1 or Table 2, the complete survival profile can be estimated by using a $\underset{\sim}{K}$ matrix with $\underset{\sim}{K}^*$ as the (t x 2t) matrix defined in (2.14) instead of the (1 x 2t) matrix in (2.6).

$$\underset{\sim}{K}^* = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & -1 & 1 & -1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & -1 & 1 & -1 & \cdots & 1 & -1 \end{bmatrix} \tag{2.14}$$

## 3. AN EXAMPLE WITH ONE GROUP

In Table 3, the five year survival data are shown for an example discussed

TABLE 3

3. FIVE YEAR SURVIVAL DATA FOR 126 MALE CONNECTICUT RESIDENTS WITH LOCALIZED KIDNEY CANCER, DIAGNOSED 1946-1951 AND FOLLOWED THROUGH DECEMBER 31, 1951

| Years After Diagnosis | Survived the Year | Died During Year | Withdrawn or Lost | Total Alive at Beginning of Year |
|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) |
| 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 |

by Cutler and Ederer [10]. These data involve only one group; and as a result, we consider a single (5 x 3) contingency table of the type shown in Table 1. The corresponding estimated probability vector $\underset{\sim}{p}$ corresponds to

$$\underset{\sim}{p}' = (0.48,\ 0.37,\ 0.15,\ 0.63,\ \ldots,\ 0.60).$$

Accordingly, with $\underset{\sim}{A}$ and $\underset{\sim}{K}$ as shown in (2.3), (2.4), (2.5), and (2.6) where $t = 5$ the following results are obtained.

$$\underset{\sim}{K}[\log_e(\underset{\sim}{A}\underset{\sim}{p})] = -0.82$$

$$G = e^{-0.82} = 0.44$$

$$\text{var } G = G^2 \cdot \text{var}(\log_e G) = 0.003590$$

These results agree with those obtained by Cutler and Ederer [10] for these data. The advantage of this approach to survival rate analysis, however, is that this estimation procedure can be combined with a general method for fitting models to survival rates and testing the statistical significance of appropriate hypotheses.

For this example with a single group, one question of interest is whether the probability of survival is characterized by an exponential curve. In this event, an appropriate model for the log survival rates is a straight line through the origin. To investigate the validity of an exponential model for the data in Table 3, we use the same $\underset{\sim}{A}$ matrix based on (2.3) and (2.4) that was involved in the calculation of the 5-year rate G; but we modify the $\underset{\sim}{K}$ matrix to that shown in (3.1) which produces the estimates $F_1, F_2, F_3, F_4$ of the logs of the 1-year, 2-year,

$$\underset{\sim}{K} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 0 & 0 \end{bmatrix} \qquad (3.1)$$

3-year, 4-year survival rates. The 5-year rate has been excluded here because the existence of 0 deaths in the 5-th exposure year causes the 4-year rate and the 5-year rate to have the same estimate; in other words, inclusion of the log 5-year

rate would cause $\underset{\sim}{V}_{\underset{\sim}{F}}$ to become singular. These estimates and the corresponding estimated covariance matrix are shown in (3.2). If an exponential model is appropriate, then the variation of the elements comprising the vector $\underset{\sim}{F}$ in (3.2) can be described

$$\underset{\sim}{F} = \begin{bmatrix} -0.52 \\ -0.62 \\ -0.69 \\ -0.82 \end{bmatrix} \qquad \underset{\sim}{V}_{\underset{\sim}{F}} = \begin{bmatrix} 56.45 & 56.45 & 56.45 & 56.45 \\ & 77.14 & 77.14 & 77.14 \\ & & 99.95 & 99.95 \\ & & & 181.97 \end{bmatrix} \times 10^{-4} \quad (3$$

by the model $\underset{\sim}{E}\{\underset{\sim}{F}\} = \underset{\sim\sim}{X}\beta$ shown in (3.3) where $\beta$ is the constant mortality rate

$$\underset{\sim}{E}\{\underset{\sim}{F}\} = \underset{\sim}{E}\begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}\beta = \underset{\sim\sim}{X}\beta \qquad (3.3)$$

parameter. The goodness of fit statistic obtained by applying (2.11) to this case is $X^2 = 27.92$ with D.F. = 3. Since this result is significant at the $\alpha = 0.01$ level, an exponential model is not appropriate for these data. On the other hand, had this model been suitable the derived estimates of $b = -0.15$ and $V_b = 0.00082$ would have been of greater interest as a concise characterization of this survival data.

A model which is appropriate for these data is the truncated exponential as shown in (3.4). The goodness of fit statistic from (2.11) is $X^2 = 0.47$ with D.F. = 2

$$\underset{\sim}{E}\{\underset{\sim}{F}\} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \underset{\sim\sim}{X}\beta \qquad (3.4)$$

which is not significant. The corresponding estimates are $b_0 = -0.43$ and $b_1 = -0.09$. Finally, the quantity $e^{b_0} = 0.65$ can be interpreted as the probability of being alive subsequent to initial treatment (i.e. at time 0) and $e^{b_1} = 0.91$ can be interpreted as the rate at which the probability of survival decreases for each additional year of risk.

## REFERENCES

[1] American Joint Committee on Cancer Staging and End Results Reporting, Clinical Staging System for Cancer of the Breast, 1962.

[2] Armitage, P., "The Comparison of Survival Curves," Journal of the Royal Statistical Society, Series A, 122 (Part III 1959), 279-300.

[3] Berkson, J. and Gage, R.P., "Survival Curve for Cancer Patients Following Treatment," Journal of the American Statistical Association, 47 (September 1952), 501-15.

[4] Bishop, Y.M.M., "Full Contingency Tables, Logits, and Split Contingency Tables," Biometrics, 25 (June 1969), 383-99.

[5] Boag, J.W., "Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy," Journal of the Royal Statistical Society, Series B, 11 (No.1 1949), 15-53.

[6] Chiang, C.L., "A Stochastic Study of the Life Table and Its Applications: I. Probability Distributions of the Biometric Functions," Biometrics, 16 (December 1960), 618-35.

[7] Chiang, C.L., "A Stochastic Study of the Life Table and Its Applications: II. Sample Variance of the Observed Expectation of Life and Other Biometric Functions," Human Biology, 32 (September 1960), 221-38.

[8] Chiang, C.L., "A Stochastic Study of the Life Table and Its Applications: III. The Follow-up Study With the Consideration of Competing Risks," Biometrics, 17 (March 1961), 57-78.

[9] Chiang, C.L., Introduction to Stochastic Processes in Biostatistics, New York: John Wiley and Sons, Inc., 1968.

[10] Cutler, S.J. and Ederer, F., "Maximum Utilization of the Life Table Method in Analyzing Survival," Journal of Chronic Disease, 6(December 1958), 699-712.

[11] Cutler, S.J. and Myers, M.H., "Clinical Classification of Extent of Disease in Cancer of the Breast," Journal of the National Institute of Cancer, 39 (August 1967), 193-207.

[12] Dorn, H., "Methods of Analysis for Follow-up Studies," Human Biology, 22 (December 1950), 238-48.

[13] Elveback, L., "Estimation of Survivorship in Chronic Disease: The 'Actuarial' Method," Journal of the American Statistical Association, 53 (June 1958), 420-40.

[14] Epstein, B. and Sobel, M., "Life Testing," Journal of the American Statistical Association, 48 (September 1953), 486-502.

[15] Fix, E., and Neyman, J., "A Simple Stochastic Model of Recovery, Relapse, Death and Loss of Patients," Human Biology, 23 (September 1951), 205-41.

[16] Frost, W.H., "Risk of Persons in Familiar Contact With Pulmonary Tuberculosis," American Journal of Public Health, 23 (May 1933), 426-32.

[17] Goodman, L.A., "The Multivariate Analysis of Qualitative Data: Interaction Among Multiple Classifications," Journal of the American Statistical Association, 65 (March 1970), 226-56.

[18] Goodman, L.A., "Partitioning of Chi-Square, Analysis of Marginal Contingency Tables, and Estimation of Expected Frequencies in Multidimensional Contingency Tables," Journal of the American Statistical Association, 66 (June 1971), 339-44.

[19] Greenwood, M., "A Report on the Natural Duration of Cancer," Reports on Public Health and Medical Subjects, No. 33, 1-26. His Majesty's Stationary Office, 1926.

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

[21] International Union Against Cancer, Malignant Tumor of the Breast: Clinical Stage Classification and Presentation of Results; 1960-1964, Geneva, Switzerland, 1960.

[22] Johnson, W.D. and Koch, G.G., "A Note on the Weighted Least Squares Analysis of the Ries-Smith Contingency Table Data," Technometrics, 13 (May 1971), 438-47.

[23] Karn, M.N., "A Further Study of Methods of Constructing Life Tables When Certain Causes of Death are Elinimated," Biometrika, 25 (May 1933), 91-101.

[24] Ku, H.H., Varner, R.N., and Kullback, S., "Analysis of Multidimensional Contingency Tables," Journal of the American Statistical Association, 66 (March 1971), 55-64.

[25] Littell, A.S., "Estimation of the T-year Survival Rate From Follow-up Studies Over a Limited Period of Time," Human Biology, 24 (May 1952), 87-116.

[26] Mantel, N. and Haenszel, W., "Statistical Aspects of Analysis of Data From Retrospective Studies of Disease," Journal of the National Cancer Institute, 22 (April 1959), 719-48.

[27] Merrell, M. and Shulman, L.E., "Determination of Prognosis in Chronic Disease, Illustrated by Systematic Lupus Erythematosus," Journal of Chronic Disease, 1 (January 1955) 12-32.

[28] Myers, M., Axtell, L., and Zelen, M., "The Use of Prognostic Factors in Predicting Survival for Breast Cancer Patients," Journal of Chronic Disease, 19 (August 1966), 923-33.

[29] Neyman,J., "Contributions to the Theory of the $\chi^2$ Test," Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, Berkeley and Los Angeles: University of California Press, (239-72) 1949.

[30] Wald, A., "Test of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large," Transactions of the American Mathematical Society, 54 (November 1943), 426-82.

[31] Zippin, C., "Comparison of the International and American Systems for Staging of Breast Cancer," Journal of the National Cancer Institute, 36 (January 1966) 53-62.

# APPENDIX

The definition of $\pi_{ij3}$ used here is not entirely precise because withdrawals and lost to follow-up cases belong to entirely different categories in the life table framework. In other words, everyone in a follow-up study is eligible for being lost to the study; but only a few of the patients are eligible to being withdrawn during the j-th year due to termination of the study. Hence, on the basis of a point of view analogous to that in Chiang [9, Chapter 12], $\pi_{ij3}$ can be defined as

$$\pi_{ij3} = \lambda_{ij3}\pi_{ij31} + (1 - \lambda_{ij3})\pi_{ij32} + \lambda_{ij3}\pi_{ij33}$$

where for the population of individuals who are in the i-th group and who are alive at the beginning of the j-th year, $\lambda_{ij3}$ is the proportion who are eligible for withdrawal during that year due to termination of the study; $\pi_{ij31}$ is the conditional probability that an individual who is eligible for withdrawal survives the period from the beginning of the j-th year until the termination date of the study (which occurs prior to the beginning of the (j+1)-th year); $\pi_{ij32}$ is the conditional probability that an individual who is not eligible for withdrawal is exposed to risk for a period less than one year due to lost to follow-up and survives the period of exposure; and $\pi_{ij33}$ is the conditional probability that an individual who is eligible for withdrawal is exposed to risk (for a period less than one year) until lost to follow-up which occurs prior to the termination date of the study and survives the period of exposure.

Similarly, if $\pi_{ij11}$ represents the conditional probability that an individual who is not eligible for withdrawal is exposed to risk for an entire year and survives the period of exposure, then

$$\pi_{ij1} = (1 - \lambda_{ij3})\,\pi_{ij11} \quad ;$$

also, if $\pi_{ij21}$ represents the conditional probability that an individual who is not eligible for withdrawal is exposed to risk until death (which is a period less than one year) and if $\pi_{ij22}$ represents the conditional probability that an individual who is eligible for withdrawal is exposed to risk until death (which occurs prior to the termination date of the study) then

$$\pi_{ij2} = (1 - \lambda_{ij3})\, \pi_{ij21} + \lambda_{ij3}\pi_{ij22}$$

Other formulations of this type could be used to reflect survival until death due to certain unrelated causes.

In this paper, we have chosen to work with $\pi_{ij1}$, $\pi_{ij2}$, $\pi_{ij3}$ because it is these quantities which are relevant to the construction of the estimators of survival rates which are associated with the approach of Cutler and Ederer [10] as discussed in this paper; i.e., if $\pi_{ij11}$, $\pi_{ij12}$, $\pi_{ij22}$, $\pi_{ij31}$, $\pi_{ij32}$, $\pi_{ij33}$ had been initially defined separately, at some later stage of the analysis these would have been combined according to the linear functions $\pi_{ij1}$, $\pi_{ij2}$, $\pi_{ij3}$ with $\lambda_{ij3}$ and $(1 - \lambda_{ij3})$ being viewed as known constants.

Finally, if within this framework, it is assumed that whether an individual is a withdrawal in the j-th year is actually a random event (related to the date of diagnosis of the disease) with probability $\lambda_{ij3}$ of occurrence, then it follows that the multinomial distribution with parameters $\underset{\sim}{\pi}_{ij}$ and $n_{ij.}$ is a suitable model for the observed frequency vector $\underset{\sim}{n}_{ij.}$. Otherwise, it can be verified that the estimated log survival rate shown in (2.7) and the corresponding estimated covariance matrix shown in (2.9) are essentially unaffected by the validity of this assumption.

An alternative method, which has been described by Chiang [9, Chapter 12] for this more general situation, involves treating lost to follow-up as

a competing risk and viewing the survival of both non-withdrawal and withdrawal patients as being characterized by the same underlying exponential model (except for the period of exposure - i.e., the period of expsoure for withdrawals is approximately half the length of the period of exposure for non-withdrawals). He then uses maximum likelihood estimation. We will consider a minimum modified chi-square approach to this model in a future investigation.