This research was supported in part by an American Cancer Society Cancer Institutional Sub-Grant No. IN-15-M and a Public Health Service Research Career Development Award (No. GM 70004) from the National Institute of General Medical Sciences (G.G.K.)

# AN APPLICATION OF LINEAR MODELS TO ANALYZE CATEGORICAL DATA PERTAINING TO THE RELATIONSHIP BETWEEN SURVIVAL AND EXTENT OF DISEASE

by

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

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 770

September 1971

## ABSTRACT

The weighted least squares approach to the analysis of categorical data has been applied to the factorial arrangement of selected clinical characteristics to identify their effects on survival in women with cancer of the breast. Linear models are fitted sequentially until the statistical significance of the three factors skin fixation, node status, and tumor size and their interrelationships have been identified and essentially all of the statistically significant variation among five year survival rates has been accounted for. These results are used as the basis of a survival-oriented, statistical classification of extent of disease using an approach analogous to cluster analysis. The notion of distance together with appropriate chi-square tests is used to compare variation both within and among stages of disease.

# AN APPLICATION OF LINEAR MODELS TO ANALYZE CATEGORICAL DATA PERTAINING TO THE RELATIONSHIP BETWEEN SURVIVAL AND EXTENT OF DISEASE

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

## 1. INTRODUCTION

One area of application which has become increasingly important to statisticians and other researchers is the analysis of categorical data. Often, the principal objective in such investigations is either the testing of appropriate hypotheses or the fitting of simplified models to the multi-dimensional contingency tables which arise when frequency counts are obtained for the respective cross-classifications of specific qualitative variables. Grizzle, Starmer, and Koch [ 7 ] (subsequently abbreviated GSK) have described how linear regression models and weighted least squares can be used for this purpose. The resulting test statistics belong to the class of minimum modified chi-square due to Neyman [ 13 ] which is equivalent to the general quadratic form criteria of Wald [ 14 ]. As such, they have central $\chi^2$-distributions when the corresponding null hypotheses at which they are directed are true. Two direct competitors to this approach are that based on maximum likelihood as formulated by Bishop [ 2 ] and Goodman [5,6] and that based on minimum discrimination information as formulated by Ku and Kullback [ 11 ]. In large samples, all three of these approaches are asymptotically equivalent in the sense of being based on BAN estimates as described

by Neyman [ 13 ]. Thus, in most applications, the choice of which method to use is a matter of practical convenience.

Recently, Koch, Tolley, and Johnson [ 10 ] have indicated how the GSK approach can be used to estimate survival rates and corresponding standard errors from life tables. The resulting statistics are the same as those discussed by Cutler and Ederer [ 3 ] and hence take advantage of the available survival information for all patients including those who are lost-to-follow-up during the time period of a given clinical study. Moreover, if the various patients involved have been cross-classified according to other qualitative variables which reflect, for example, extent of disease, this same methodology can be used to examine the manner in which the variation of the survival rates depends on such factors. In particular, modified analyses of variance with adjusted tests of significance can be undertaken and used to motivate simplified models for the data. Certain powerful aspects of this type of analysis will now be illustrated in more detail.

## 2. AN EXAMPLE RELATED TO CANCER OF THE BREAST

One problem of interest in cancer research is the relationship between clinical classifications of the extent of disease and patient survival. The medical diagnosis is usually the most important determinant of treatment, and survival is necessarily the most obvious consequence of treatment. Hence, it is appropriate to identify the extent to which patients with similar characteristics and perhaps similar treatment have similar survival experience over a five or ten year period. The results of such analyses may lead to new insights in the method of treatment for certain disease classifications.

In the remainder of this paper, we shall consider data which were investigated by Cutler and Myers [ 4 ] as a basis for an empirical statistical classification of the extent of disease in cancer of the breast. In their paper, Cutler and Myers were concerned with a series of 2039 cases in women under 65 years of age. These data were originally discussed by Zippin [ 15 ] in the context of a project carried out by the End Results Group of the National Cancer Institute, in order to compare two classification schemes for breast cancer - one due to the International Union Against Cancer [ 8 ] and the other due to the American Joint Committee on Cancer Staging and End Results Reporting [ 1 ]. The basis of each of these procedures is the TNM system which involves the clinical status of the tumor, the regional lymph nodes, and the distant metastases.

From the original total of 2039 cases, Cutler and Myers [ 4 ] identified a sub-group of 1233 which could be appropriately used to explore the inter-relationship between axillary node status, tumor size, and skin fixation with respect to their effect on survival.[1] In their analysis, the cases were classified into categories corresponding to these factors in the following manner:

---

[1] Of the 806 other patients, 418 were considered as a separate group by Cutler and Myers [ 4 ] because they had characteristics which were strongly associated with high mortality. The remaining 388 were eliminated because either axillary node status and/or tumor size was unknown.

1. Degree of skin fixation (S)
   a. none ($S_0$)
   b. incomplete ($S_1$)
   c. complete ($S_2$)
2. Node status (N)
   a. clinically negative ($N_0$)
   b. palpable ($N_1$)
3. Tumor size (T)
   a. <2cm. ($T_1$)
   b. >2-4cm. ($T_2$)
   c. >4cm. ($T_3$)

Thus, cross-classification of the 1233 cases according to these three factors defines 18 groups. In Table 1, the number of cases, the five-year survival rate, and the estimated standard error of the five year survival rate for each of these 18 groups appear in Columns 1, 2, 3 respectively. These results were abstracted for us by the End Results Group, National Cancer Institute so that we could illustrate the application of the GSK methodology to survival data involving complex contingency tables. They were obtained by applying the method of Cutler and Ederer [ 3 ] to the appropriate life tables for the 18 disease classifications. However, as pointed out earlier, they could also have been obtained by appropriately re-writing the life tables as two-dimensional contingency tables and then applying the GSK approach in the manner described by Koch, Tolley, and Johnson [ 10 ].

By specifically considering five year survival rates our analysis will have a somewhat different orientation than that of Cutler and Myers [ 4 ]. Their approach was directed at estimated "average

TABLE 1

| Category | No. of Cases | 5-Yr. Surv. | Est. Std. Error | Pred. Based on $X_2$ | Pred. Based on $X_3$ | Pred. Based on $X_4$ | Pred. Based on $X_6$ | Std. Error of $X_6$ Pred. Val. | Residual of $X_6$ Pred. Val |
|---|---|---|---|---|---|---|---|---|---|
| $S_0N_0T_1$ | 195 | .88 | .024 | .89 | .90 | .92 | .89 | .020 | -.01 |
| $S_0N_0T_2$ | 226 | .77 | .028 | .77 | .76 | .75 | .76 | .018 | .01 |
| $S_0N_0T_3$ | 96 | .62 | .050 | .58 | .59 | .58 | .64 | .032 | -.02 |
| $S_0N_1T_1$ | 72 | .78 | .049 | .75 | .79 | .76 | .77 | .030 | .01 |
| $S_0N_1T_2$ | 89 | .67 | .050 | .66 | .64 | .64 | .64 | .026 | .03 |
| $S_0N_1T_3$ | 53 | .49 | .069 | .57 | .47 | .52 | .51 | .035 | -.02 |
| $S_1N_0T_1$ | 41 | .95 | .034 | .94 | .90 | .89 | .95 | .027 | .00 |
| $S_1N_0T_2$ | 114 | .74 | .042 | .74 | .76 | .72 | .72 | .017 | .02 |
| $S_1N_0T_3$ | 78 | .51 | .057 | .56 | .59 | .55 | .48 | .036 | .03 |
| $S_1N_1T_1$ | 24 | .63 | .099 | .75 | .79 | .73 | .59 | .024 | .04 |
| $S_1N_1T_2$ | 55 | .58 | .066 | .59 | .64 | .61 | .59 | .024 | -.01 |
| $S_1N_1T_3$ | 59 | .57 | .065 | .51 | .47 | .49 | .59 | .024 | -.02 |
| $S_2N_0T_1$ | 15 | .93 | .069 | .92 | .83 | .86 | .90 | .035 | .03 |
| $S_2N_0T_2$ | 30 | .67 | .086 | .65 | .68 | .69 | .67 | .031 | .00 |
| $S_2N_0T_3$ | 26 | .38 | .095 | .41 | .51 | .52 | .43 | .047 | -.05 |
| $S_2N_1T_1$ | 7 | .71 | .171 | .75 | .71 | .70 | .67 | .044 | .04 |
| $S_2N_1T_2$ | 15 | .47 | .129 | .51 | .56 | .58 | .55 | .035 | -.08 |
| $S_2N_1T_3$ | 38 | .39 | .079 | .37 | .39 | .46 | .42 | .037 | -.03 |

mortality rates" derived from an exponential model for the ten year survival experience of each of the 18 groups. In accordance with methodology due to Myers, Axtell, and Zelen [ 12 ], an unweighted analysis of variance was then applied to a logarithmic transformation of these quantities in order to ascertain the statistical significance of the factors S, N, and T. In particular, their results indicated significant ($\alpha$=.01) main effects for nodes (N) and tumor size (T) as well as a significant ($\alpha$=.05) node by tumor size (N x T) interaction. The main effect of skin fixation (S) was apparent but not significant. These results were then applied to a mathematical model which yielded adjusted estimates for the "average mortality rates", which were based on all of the data as opposed to that for a single group. These adjusted rates were then used as the basis of an alternative classification scheme for the extent of disease which was proposed by Cutler and Myers [ 4 ] and which was compared to the American and International staging rules.

## 3. AN ALTERNATIVE ANALYSIS

Since we have indicated that the five year survival rates in Table 1 could have been obtained by using a contingency table approach, it follows that their variation with respect to the factors S, N, and T can be investigated by using the GSK methodology. Thus, if we let $\underset{\sim}{p}$ denote the vector of 18 five year survival rates in Column 2 of Table 1, linear regression models of the form

$$E\{\underset{\sim}{p}\} = \underset{\sim\sim}{X}\underset{\sim}{\beta}$$

can be fitted by weighted regression where $\underset{\sim}{X}$ is a specified 18xr coefficient or design matrix of rank $r < 18$, $\underset{\sim}{\beta}$ is the corresponding vector of parameters to be estimated and weighting is with respect

to the reciprocals of the appropriate estimated variances (i.e., the squares of the quantities in column 3 of Table 1). The resulting estimate $\underset{\sim}{b}$ of $\underset{\sim}{\beta}$ is given by

$$\underset{\sim}{b} = [\underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{X}]^{-1}\underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{p}$$

where $\underset{\sim}{V}$ denotes the diagonal matrix whose elements are the respective estimated variances. If the data are adequately described by this model, a test of the hypothesis $H_0: \underset{\sim\sim}{C\beta} = 0$ where $\underset{\sim}{C}$ is a uxr coefficient matrix is produced by conventional methods of weighted regression. In particular, the test statistic is given by

$$X^2 = SS(\underset{\sim\sim}{C\beta} = \underset{\sim}{0}) = \underset{\sim}{b}'\underset{\sim}{C}'[\underset{\sim}{C}(\underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{C}']^{-1}\underset{\sim\sim}{Cb}$$

which has approximately a chi-square distribution with D.F. = u in large samples under $H_0$. Finally, a goodness of fit test to determine the validity of the model is

$$X^2 = SS(E(\underset{\sim}{p}) = \underset{\sim\sim}{X\beta}) = \underset{\sim}{p}'\underset{\sim}{V}^{-1}\underset{\sim}{p} - \underset{\sim}{b}'[\underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{X}]\underset{\sim}{b}$$

which under the hypothesis that the model fits, has approximately a chi-square distribution with D.F. = (18-r) in large samples.

The first model which was applied to $\underset{\sim}{p}$ is based on $X_1$. This is a complete

$$\underset{\sim}{X}_1 = \begin{bmatrix}
1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \\
1 & 1 & -1 & 1 & 1 & -1 & 0 & 2 & 0 & 0 & 2 & -2 & 0 & 2 & 0 & 0 & 2 & -2 \\
1 & 1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 \\
1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 \\
1 & 1 & -1 & -1 & -1 & 1 & 0 & 2 & 0 & 0 & 2 & -2 & 0 & -2 & 0 & 0 & -2 & 2 \\
1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 \\
1 & 0 & 2 & 1 & 0 & 2 & 1 & -1 & 0 & 2 & 0 & -2 & 1 & -1 & 0 & 2 & 0 & -2 \\
1 & 0 & 2 & 1 & 0 & 2 & 0 & 2 & 0 & 0 & 0 & 4 & 0 & 2 & 0 & 0 & 0 & 4 \\
1 & 0 & 2 & 1 & 0 & 2 & -1 & -1 & 0 & -2 & 0 & -2 & -1 & -1 & 0 & -2 & 0 & -2 \\
1 & 0 & 2 & -1 & 0 & -2 & 1 & -1 & 0 & 2 & 0 & -2 & -1 & 1 & 0 & -2 & 0 & 2 \\
1 & 0 & 2 & -1 & 0 & -2 & 0 & 2 & 0 & 0 & 0 & 4 & 0 & -2 & 0 & 0 & 0 & -4 \\
1 & 0 & 2 & -1 & 0 & -2 & -1 & -1 & 0 & -2 & 0 & -2 & 1 & 1 & 0 & 2 & 0 & 2 \\
1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & 1 & -1 & -1 & 0 & 2 & 0 & 0 & -2 & -2 & 0 & 2 & 0 & 0 & -2 & -2 \\
1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 \\
1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & -1 & 1 & 1 & 0 & 2 & 0 & 0 & -2 & -2 & 0 & -2 & 0 & 0 & 2 & 2 \\
1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1
\end{bmatrix}$$

model since the number of effects (i.e. columns of $\underset{\sim}{X}_1$) is equal to the dimension of $\underset{\sim}{p}$ (which is the same as the number of rows of $\underset{\sim}{X}_1$). A description of the effects corresponding to the respective columns of $\underset{\sim}{X}_1$ together with significance tests of their importance appear in the first 17 rows of Table 2 with the overall mean (Column 1 of $\underset{\sim}{X}_1$)

TABLE 2

| Source of Variation ($\underset{\sim}{X}_1$) | D.F. | $X^2$ |
|---|---|---|
| Skin Fixation (Linear):S(L) | 1 | 5.02 |
| Skin Fixation (Non-linear):S(Q) | 1 | .22 |
| Node Status: N | 1 | 12.09 |
| S(L) x N | 1 | .07 |
| S(Q) x N | 1 | .05 |
| Tumor Size (Linear):T(L) | 1 | 47.36 |
| Tumor Size (Non-linear):T(Q) | 1 | .01 |
| S(L) x T(L) | 1 | 1.72 |
| S(Q) x T(L) | 1 | 1.33 |
| S(L) x T(Q) | 1 | .34 |
| S(Q) x T(Q) | 1 | .01 |
| N x T(L) | 1 | 4.32 |
| N x T(Q) | 1 | .23 |
| S(L) x N x T(L) | 1 | 1.13 |
| S(Q) x N x T(L) | 1 | 2.36 |
| S(L) x N x T(Q) | 1 | .29 |
| S(Q) x N x T(Q) | 1 | .01 |
| Skin Fixation (Total):S | 2 | 5.48 |
| S x N | 2 | .24 |
| Tumor Size (Total):T | 2 | 48.11 |
| S x T | 4 | 2.53 |
| N x T | 2 | 4.92 |
| S x N x T | 4 | 6.37 |

being excluded. Here, it should be noted that the "linear:non-linear" partition of the main effects and interactions associated with S and T is simply a matter of convenience and that other partitions of the degrees of freedom associated with these effects can be achieved by suitable modifications of the columns of $\underset{\sim}{X}_1$. Moreover, regardless which

partition is used, the results of joint tests of significance like those in rows 18-23 of Table 2 will be the same. Thus, the reader who is uncomfortable with any effort to scale S or T can choose to look at only these tests without any loss of generality. On the other hand, since the likelihood of survival tends to decrease as either tumor size increases or degree of skin fixation becomes more complete, one can argue that S and T are at least ordinally scaled with respect to survival. In this context, one can interpret their "linear" components as being measures of the degree of monotonicity in the relationships between survival and S and T and their "non-linear" components as being measures of the variation which is not attributable to the "linear" components.

In any event, the analysis in Table 2 is analogous to that appropriate for a 3x2x3 factorial design. However, special attention must be paid to the fact that the p's are only approximately normally distributed and have different variances. Since the different variances of the p's are accounted for by the weights, the resulting test statistics have approximately chi-square distributions when the corresponding hypotheses are true provided the sample size is suitably large (i.e., there are approximately 25 observations per degree of freedom extracted from the data - see Johnson and Koch [ 9 ]; here 18 degrees of freedom are of interest and 18 x 25 = 450 < 1233) which is the case for these data.

From Table 2, it is apparent that this analysis leads to conclusions which are similar to those obtained by Cutler and Myers; i.e., highly significant ($\alpha$=.01) main effects for tumor size and nodes, significant ($\alpha$=.05) linear component of skin fixation, and significant

($\alpha$=.05) interaction between nodes and the linear component of tumor size. Since it can be noted that neither the three factor interaction SxNxT nor any of its individual components are significant ($\alpha$=.10), the results in Table 2 can be refined by fitting a new model $\underset{\sim}{X}_2$ which corresponds to the first 14 columns of $\underset{\sim}{X}_1$ (i.e., the last 4 columns of $\underset{\sim}{X}_1$ and the corresponding parameters are deleted). This is an incomplete model since $\underset{\sim}{X}_2$ has 14<18 columns. Its use must be justified by a goodness of fit test which measures the model's ability to account for the variation in the data. As indicated previously, an appropriate test statistic for this purpose is the weighted residual sum of squares. In this case, the $X^2$(D.F.=4) = 6.37 corresponds to the excluded three factor interaction and has already been found non-significant ($\alpha$=.10). The results of this analysis are given in Table 3. The conclusions here

TABLE 3

| Source of Variation ($\underset{\sim}{X}_2$) | D.F. | $X^2$ |
|---|---|---|
| Skin Fixation (Linear):S(L) | 1 | 5.04 |
| Skin Fixation (Non-linear):S(Q) | 1 | .70 |
| Node Status:N | 1 | 9.95 |
| S(L) x N | 1 | .15 |
| S(Q) x N | 1 | .10 |
| Tumor Size (Linear):T(L) | 1 | 58.36 |
| Tumor Size (Non-linear):T(Q) | 1 | .13 |
| S(L) x T(L) | 1 | 3.60 |
| S(Q) x T(L) | 1 | .18 |
| S(L) x T(Q) | 1 | .35 |
| S(Q) x T(Q) | 1 | .11 |
| N x T(L) | 1 | 3.38 |
| N x T(Q) | 1 | .45 |
| Overall Model | 13 | 164.31 |
| Residual | 4 | 6.37 |

are essentially the same as those implied by Table 2 except the node x linear component of tumor size interaction does not attain significance at the $\alpha$=.05 level. Finally, from the estimated parameters $\underset{\sim}{b}$ derived from this model, predicted (or adjusted) values can be calculated for the survival rates for each skin fixation x node status x tumor size combination by use of the formula

$$\hat{\underset{\sim}{p}} = \underset{\sim}{X}[\underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{X}]^{-1}\underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{p} .$$

Predicted values corresponding to $\underset{\sim}{X}_2$ are displayed in Column 4 of Table 1. There, it can be noted that large residuals with absolute values in excess of 0.05 occur for $S_0N_1T_3$, $S_1N_1T_1$, $S_1N_1T_3$. Thus, even though this model satisfies the criterion of a non-significant residual and involves a large number of parameters, it does not provide an entirely satisfactory explanation of the variation in the data.

Since none of the two-factor interactions in Table 3 are significant another incomplete model of interest is that corresponding to $\underset{\sim}{X}_3$ which is based only on the main effects of the three-factors.

$$\underset{\sim}{X}_3 = \begin{bmatrix} 1 & 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 & 0 & 2 \\ 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 0 & 2 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 0 & 2 & 1 & 1 & -1 \\ 1 & 0 & 2 & 1 & 0 & 2 \\ 1 & 0 & 2 & 1 & -1 & -1 \\ 1 & 0 & 2 & -1 & 1 & -1 \\ 1 & 0 & 2 & -1 & 0 & 2 \\ 1 & 0 & 2 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 0 & 2 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 & -1 \\ 1 & -1 & -1 & -1 & 0 & 2 \\ 1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix}$$

The results of this analysis are displayed in Table 4. In particular,

TABLE 4

| Source of Variation ($\underset{\sim}{X}_3$) | D.F. | $X^2$ |
|---|---|---|
| Skin Fixation:S | 2 | 3.76 |
| Node Status:N | 1 | 17.86 |
| Tumor Size:T | 2 | 97.38 |
| Overall Model | 5 | 154.38 |
| Residual | 12 | 16.30 |

the residual $X^2$(D.F. = 12) = 16.30 is non-significant ($\alpha$=.10) and thus justifies further consideration of the model. Tests on the parameters corresponding to $\underset{\sim}{X}_3$ indicate significant ($\alpha$=.01) main effects for nodes and tumor size. Finally, predicted values based on $\underset{\sim}{X}_3$ appear in column 5 of Table 1. There it can be noted that large residuals with absolute values in excess of 0.05 occur for $S_1N_0T_3$, $S_1N_1T_1$, $S_1N_1T_2$, $S_1N_1T_3$, $S_2N_0T_1$, $S_2N_0T_3$, $S_2N_1T_2$. Thus, although the goodness-of-fit test for this model is non-significant, its predictive value is not as satisfactory as one might want.

If one proceeds from $\underset{\sim}{X}_1$ in a somewhat different manner, the incomplete model corresponding to $\underset{\sim}{X}_4$ becomes of interest. The results of this

$$\underset{\sim}{X}_4 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & 0 & 0 \\ 1 & 1 & -1 & -1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & -1 & -1 \\ 1 & 0 & -1 & 1 & -1 \\ 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & -1 & -1 & 1 \\ 1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 0 & 0 \\ 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 0 & 0 \\ 1 & -1 & -1 & -1 & 1 \end{bmatrix}$$

analysis are displayed in Table 5 and indicate that the residual is

TABLE 5

| Source of Variation ($\underset{\sim}{X}_4$) | D.F. | $X^2$ |
|---|---|---|
| Skin Fixation (Linear):S(L) | 1 | 2.65 |
| Node Status:N | 1 | 16.99 |
| Tumor Size (Linear):T(L) | 1 | 66.32 |
| N x T(L) | 1 | 2.25 |
| Overall Model | 4 | 154.68 |
| Residual | 13 | 16.00 |

non-significant ($\alpha$=.10) and that the main effects of nodes and linear component of tumor size are significant ($\alpha$=.01). Predicted values based on $\underset{\sim}{X}_4$ appear in column 6 of Table 1. There it can be noted that large residuals with absolute values in excess of 0.05 occur for $S_1N_0T_1$, $S_1N_1T_1$, $S_1N_1T_3$, $S_2N_0T_1$, $S_2N_0T_3$, $S_2N_1T_2$, $S_2N_1T_3$. Hence, this model has more or less the same weakness as that specified by $\underset{\sim}{X}_3$.

At this point, it becomes worthwhile to approach the data from a somewhat different point of view. From the complete model $\underset{\sim}{X}_1$, we have already observed that the most important sources of variation are the main effects of nodes and tumor size. In addition, the main effects of skin fixation and the node x tumor size interaction should also be noted. However, given these conclusions the next objective is to find a model with as few parameters as possible, but which still accounts for most of the variation among the five year survival rates. Thus, the GSK methodology is being used in the same spirit as stepwise regression. In particular, two useful rules for this purpose are

1. The goodness-of-fit $X^2$ test for the excluded degrees of freedom should not exceed 3.84 (the $\alpha$=.05 critical value of the $\chi^2$-distribution with D.F. = 1) in order to ensure that the residual does not contain any hidden, but individually significant, component and hence that the corresponding model can be presumed to account for essentially all the variation
2. Each of the variables retained in the model should be significant at the $\alpha$=.05 level (when its effects are tested after covariance adjustment for all other retained variables).

In many situations, both of these rules can be satisfied while in others they may need to be compromised. However, any rules which apply to ordinary stepwise regression may be used in an analogous manner. Hence in this context, it is apparent that significance levels are not being used in an inferential sense as Type I error rates but in a descriptive sense as indicators of important sources of variation.

With this point of view in mind, the complete hierarchical model defined by $\underset{\sim}{X}_5$ becomes of interest. A description of the effects corresponding

$$\underset{\sim}{X}_5 = \begin{bmatrix}
1 & 1 & -1 & 1 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & -1 & 1 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & -1 & 1 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 2 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\
1 & 0 & 2 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\
1 & 0 & 2 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 \\
1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\
1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\
1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 \\
1 & -1 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\
1 & -1 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\
1 & -1 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1
\end{bmatrix}$$

to the columns of $\underset{\sim}{X}_5$ together with significance tests of their importance appear in the first 17 rows of Table 6 with the overall mean (Column 1 of $\underset{\sim}{X}_5$) being excluded. In addition, certain joint tests are given

TABLE 6

| Source of Variation ($X_5$) | D.F. | $X^2$ |
|---|---|---|
| Skin Fixation (Linear):S(L) | 1 | 5.02 |
| Skin Fixation (Non-linear):S(Q) | 1 | .22 |
| Node Status N in $S_0$ | 1 | 8.10 |
| " " N in $S_1$ | 1 | 7.19 |
| " " N in $S_2$ | 1 | 2.29 |
| Tumor Size (Linear)$^2$ in $S_0N_0$ | 1 | 22.26 |
| " " " " $S_0N_1$ | 1 | 11.77 |
| " " " " $S_1N_0$ | 1 | 43.88 |
| " " " " $S_1N_1$ | 1 | .26 |
| " " " " $S_2N_0$ | 1 | 21.87 |
| " " " " $S_2N_1$ | 1 | 2.89 |
| Tumor Size (Non-linear) in $S_0N_0$ | 1 | .26 |
| " " " " " $S_0N_1$ | 1 | .29 |
| " " " " " $S_1N_0$ | 1 | .04 |
| " " " " " $S_1N_1$ | 1 | .05 |
| " " " " " $S_2N_0$ | 1 | .02 |
| " " " " " $S_2N_1$ | 1 | .25 |
| Combined S(L) and S(Q) | 2 | 5.48 |
| Combined Node Status N | 3 | 17.58 |
| Combined Tumor Size (Linear) | 6 | 102.93 |
| Combined Tumor Size (Non-linear) | 6 | 0.90 |

in rows 18-21 of Table 6. The results of this analysis indicate that node status is very significant ($\alpha$=.01) for skin fixation categories $S_0$ and $S_1$, that the linear component of tumor size is very significant ($\alpha$=.01) for the skin fixation x node status combinations $S_0N_0$, $S_0N_1$, $S_1N_0$, and $S_2N_0$, and that the linear component of skin fixation is significant ($\alpha$=.05). The estimated effects of node status within skin fixation categories together with those for the linear and non-linear

components of tumor size within skin fixation x node status combinations are displayed in Table 7. Hence, it is apparent that there is no interaction

TABLE 7

| Additive Effects in | Tumor Size Linear | Tumor Size Non-linear | Additive Effects in | Node Status |
|---|---|---|---|---|
| $S_{\ }N_{\ }$ | .130 | .007 | $S_{\ }$ | .055 |
| $S^{o}N^{o}$ | .145 | .012 | $S^{o}$ | .070 |
| $S^{o}N^{1}$ | .220 | .003 | $S^{1}$ | .068 |
| $S^{1}N^{o}$ | .030 | -.007 | $S^{2}$ | |
| $S^{1}N^{1}$ | .275 | .005 | | |
| $S^{2}N^{o}$ | .160 | -.027 | | |
| $S^{2}N^{1}$ | | | | |

between skin fixation and node status in the sense that the node status effect is approximately the same for each of the three skin fixation categories. With respect to tumor size, all the non-linear components and the linear component within $S_1N_1$ are definitely unimportant and can be excluded from further consideration. However, the linear component does have two other facets which are definitely significant - one corresponding to $S_0N_0$, $S_0N_1$, and $S_2N_1$ and the other corresponding to $S_1N_0$ and $S_2N_0$.

A final model which reflects all of these conclusions is that specified by $\underset{\sim}{X}_6$. Since the residual sum of squares $X^2$(D.F. = 13) = 2.76 < 3.84, this model is consistent with the first of the two decision rules mentioned earlier. Moreover, the results in Table 8, indicate that the second rule is also satisfied. In particular, node status and the two facets of the linear component of tumor size are significant at the $\alpha$=.01

$$\underset{\sim}{X}_6 = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & -1 & 0 \\ 1 & 1 & -1 & 1 & 0 \\ 1 & 1 & -1 & 0 & 0 \\ 1 & 1 & -1 & -1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & -1 \\ 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 1 & -1 & 1 & 0 & 1 \\ 1 & -1 & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 & -1 \\ 1 & -1 & -1 & 1 & 0 \\ 1 & -1 & -1 & 0 & 0 \\ 1 & -1 & -1 & -1 & 0 \end{bmatrix}$$

TABLE 8

| Source of Variation ($\underset{\sim}{X}_6$) | D.F. | $X^2$ |
|---|---|---|
| Skin Fixation (Linear):S(L) | 1 | 6.37 |
| Node Status:N | 1 | 20.48 |
| Tumor Size (Linear) in $S_0$ or $S_2N_1$ | 1 | 39.31 |
| Tumor Size (Linear) in $S_1N_0$ or $S_2N_0$ | 1 | 75.25 |
| Overall Model | 4 | 167.92 |
| Residual | 13 | 2.76 |

level and the linear component of skin fixation is significant at the $\alpha$=.05 level. The estimated parameters and the corresponding estimated variance-covariance matrix pertaining to $\underset{\sim}{X}_6$ are given in Table 9.

TABLE 9

| Effect | Estimate | Variance - Covariance Matrix x $10^6$ | | | |
|---|---|---|---|---|---|
| Mean | .655 | 259.28 | -156.08 | -76.77 | -4.00 | -90.75 |
| S(L) | .047 | | 352.86 | -12.90 | -119.46 | 98.20 |
| N | .062 | | | 186.56 | -36.42 | -49.07 |
| T(L) in $S_0$ or $S_2N_1$ | .128 | | | | 417.23 | 9.60 |
| T(L) in $S_1N_0$ or $S_2N_0$ | .236 | | | | | 739.69 |

Moreover, these results can be used to derive a test to compare the two facets of linear tumor size; in particular $X^2 = (.108)^2/(.001138) = 10.25$ with D.F. = 1, thus confirming the significance ($\alpha$=.01) of the difference between these two effects.

As with the other incomplete models, predicted values derived on The basis of $\underset{\sim}{X}_6$ are given in column 7 of Table 1. In addition, the residuals for this situation appear in column 9 of Table 1 and are all less than 0.05 in absolute value except for the $S_2N_1T_2$ combination. Thus, this model provides a reasonably complete explanation of the variation of the five year survival rates as a function of the factors skin fixation, node status, and tumor size. A final by-product of this analysis is that standard errors can be estimated for the adjusted predicted values in column 7 of Table 1 by using the square roots of the diagonal elements of the matrix

$$\underset{\sim}{V}(\underset{\sim}{\hat{p}}) = \underset{\sim}{X}[\underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{X}]^{-1}\underset{\sim}{X}'$$

These quantities are exhibited in column 8 of Table 1. By comparing the values in column 8 with those in column 3, one can observe the extent to which the adjusted or predicted survival rates in column 7 represent more precise (or less subject to chance variation) estimates of the corresponding survival rates than the original estimates in column 2. The reason for this result is that the predicted values are derived from the analysis of the entire data set for 1233 cases treated as a whole. In the next section, we shall indicate how these results can be further applied to the problems involved in classifying extent of disease.

## 4. IDENTIFICATION OF STAGES

We have classified each of the 18 groups shown in Figure 1 according to stage of disease using the predicted survival rate derived from $\underset{\sim}{X}_6$ as the criterion. Each point on the graph represents the predicted five year survival rate plus or minus one standard error for a group of patients having one of 18 combinations of clinical findings as indicated below the graph. The rates appear in decreasing order to facilitate our choice of boundaries for the stages. Use of judgment in determining these boundaries is consistent with the approach of Cutler and Myers [4]. Their classification, however, is based on a graph of the "average mortality rates" determined from their analysis.

Letting stage I, $II_H$, $II_L$, $III_H$, and $III_L$ denote the most to least favorable survival rates, respectively, we selected the following definitions for stage of disease:

| | |
|---|---|
| I: | survival rate > .80 |
| $II_H$: | survival rate > .70 - .80 |
| $II_L$: | survival rate > .60 - .70 |
| $III_H$: | survival rate > .50 - .60 |
| $III_L$: | survival rate $\leq$ .50 |

This definition of stages is shown below the graph (KTJ). For purposes of comparison, the stages defined by the American Joint Committee on Cancer Staging and End Results Reporting (AJC), the International Union Against Cancer (IUAC), and Cutler and Myers (CM) are also shown.

Once stages have been determined by observing groups that, upon inspection of the graph, appear to have similar survival experiences,

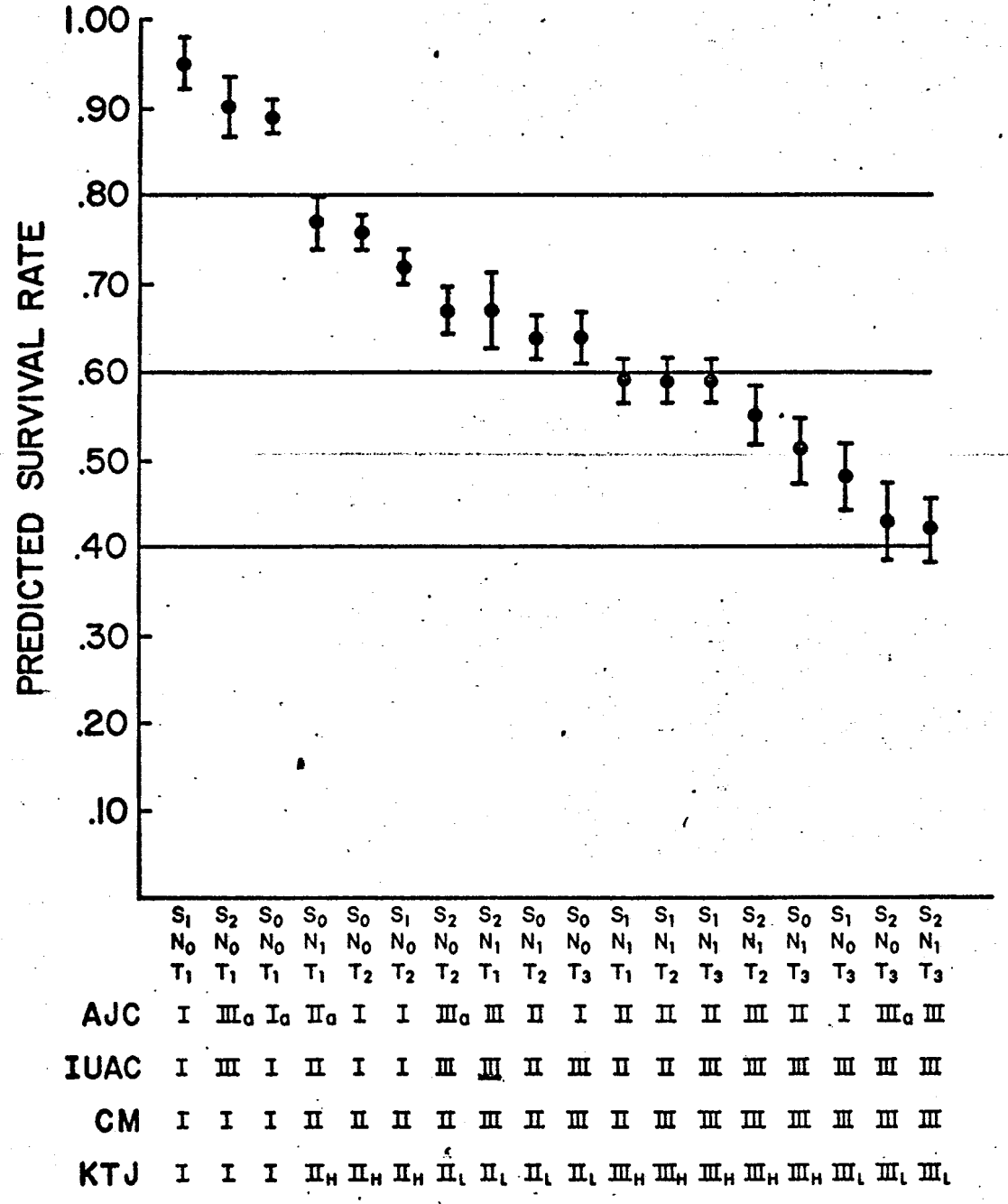
PREDICTED SURVIVAL RATE
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
S1 N0 T1
S2 N0 T1
S0 N0 T1
S0 N1 T1
S0 N0 T2
S1 N0 T2
S2 N0 T2
S2 N1 T1
S0 N1 T2
S0 N0 T3
S1 N1 T1
S1 N1 T2
S1 N1 T3
S2 N1 T2
S0 N1 T3
S1 N0 T3
S2 N0 T3
S2 N1 T3
AJC I IIIa Ia IIa I I IIIa III II I II II II III II I IIIa III
IUAC I III I II I I III III II III II II III III III III III III
CM I I I II II II II III II III II III III III III III III III
KTJ I I I IIH IIH IIH IIL IIL IIL IIL IIIH IIIH IIIH IIIH IIIH IIIL IIIL IIIL

Figure 1.

it is appealing to assess the statistical validity of such a classification scheme based on proximity. For such an assessment, it is important to note that the predicted five year survival rates based on the optimum fitting model are derived statistics and, as such, are not statistically independent. As a result, tests of significance of differences among the stages based on these values are not directly applicable. The distance between the groups as determined in relationship to the total among groups sum of squares is of some heuristic interest, however, in evaluating the judgmental staging. Moreover, once the stages have been determined, the raw five year survival rates can be analyzed in terms of chi-square tests to compare the variation both within and among the respective stages. These tests as well as the distances based on the predicted rates can be generated in a straightforward manner using the GSK approach.

The model for the five stage analysis is based on $\underset{\sim}{X}_7$.

$$\underset{\sim}{X}_7 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

The model for the three stage analysis is based on the three column matrix, $\underset{\sim}{X}_8$, which can be obtained from $\underset{\sim}{X}_7$ by adding column 3 to column 2 and column 5 to column 4. Applying these models together with the appropriate contrast matrices $\underset{\sim}{C}$, (which are constructed for comparing all groups in a pairwise manner) to the raw and five year survival rates, respectively, we obtain distances and $X^2$ tests shown in Table 10. The normalized

TABLE 10

| Between | I | IIa | IIb | II | IIIa | IIIb | III | Within |
|---|---|---|---|---|---|---|---|---|
| I | * | 0.170 | 0.308 | 0.287 | 0.663 | 0.634 | 0.861 | 0.007 |
| IIa | 24.9 | * | 0.041 | * | 0.194 | 0.279 | * | 0.006 |
| IIb | 46.7 | 8.7 | * | * | 0.031 | 0.119 | * | 0.002 |
| II | 46.0 | * | * | * | * | * | 0.262 | 0.049 |
| IIIa | 80.2 | 27.3 | 4.3 | * | * | 0.055 | * | 0.011 |
| IIIb | 98.7 | 45.0 | 14.3 | * | 3.5 | * | * | 0.003 |
| III | 146.0 | * | * | 47.0 | * | * | * | 0.069 |
| Within | 3.0 | 0.5 | 0.7 | 9.9 | 2.1 | 2.2 | 7.8 | * |
| D.F. | 2 | 2 | 3 | 6 | 4 | 2 | 7 | * |

distances between the predicted survival rates are determined as the ratio of the weighted sum of squares for the comparison to the total among groups weighted sum of squares. These appear above the main diagonal of the table; the corresponding $X^2$ values for the same contrast among the raw survival rates are given below the main diagonal. The within group distances appear in the right hand marginal and the within groups sums of squares together with their degrees of freedom are given in the lower marginal.

Focusing attention on the normalized distances, it is clear that the pairwise distances between the stages for the five stage analysis are

substantial except for the distances between stages $II_a$ and $II_b$, between $II_b$ and $III_a$, and between $III_a$ and $III_b$. The $X^2$ values based on the raw five year survival rates are consistent with these findings. Hence, it is reasonable to combine groups $II_a$ and $II_b$ and groups $III_a$ and $III_b$ for the three stage analysis. The simultaneous tests for the two analyses and the pooled distances among the groups are presented in Table 11. From this

TABLE 11

| Source of Variation | D.F. | $X^2$ | Distance |
|---|---|---|---|
| Among all five groups | 4 | 161.7 | 0.971 |
| 1. Among I, II, III | 2 | 149.4 | 0.875 |
| 2. Remainder | 2 | 12.3 | 0.096 |
| Within all five groups | 13 | 8.4 | 0.029 |
| Among Total for all 18 groups | 17 | 170.1 | 1.000 |

table it can be seen that the five group analysis accounts for 97.1% of the total distance among all 18 classifications compared to 87.5% for the three stage analysis. Moreover, the remainder $X^2$ is significant indicating that additional information can be obtained by the finer partitioning of the 18 groups of interest. The crucial issue here is to satisfy the objectives of the researcher. If a fine partitioning is desired the five stage rule could be adopted; if a rough partitioning is satisfactory the three stage rule is adequate.

In the previous section, the statistical significance of the three factors skin fixation, node status, and tumor size and their interrelationships have been identified in terms of a linear model which

explains essentially all of the statistically significant variation among the five year survival rates. These results have been now used as the basis of a survival-oriented, statistical classification of extent of disease. Alternatively, a cluster analysis approach could be applied directly to the raw five year survival rates if the primary question of interest is the determination of combinations of clinical evaluation which lead to similar survival rates and in this context have similar severity or stage of disease involvement. However, since the sample sizes and standard errors which correspond to these estimates are quite heterogeneous, we feel that use of predicted values determined from an appropriately formulated fitted model permits additional insights with respect to the boundaries for the clusters. Thus, our analysis is in the same spirit as cluster analysis but we remove as much variability from the data as is practical before determining data points in the same neighborhood.

ACKNOWLEDGMENT

The authors would like to thank the End Results Group, National Cancer Institute for providing us with the data analyzed in this paper.

## REFERENCES


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

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

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

[4.] 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 (1967), 193-207.

[5.] Goodman, L. A., "The Multivariate Analysis of Qualitative Data: Interaction Among Multiple Classifications," Journal of the American Statistical Association, 65 (March, 1970), 226-256.

[6.] "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), 339-344.

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

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

[9.] 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), 438-447.

[10.] Koch, G. G., Tolley, H. D., and Johnson, W. D., "An Application of Linear Models to Analyze Categorical Data Pertaining to the Relationship Between Survival and Extent of Disease." Submitted for publication in Journal of the American Statistical Association.

[11.] Ku, H. H. and Kullback,S., "Interaction in Multidimensional Contingency Tables: An Information Theoretic Approach," National Bureau of Standards Journal of Research, 72B (1968), 159-199.

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

[13.] 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.

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

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