Report SAM-TR-81-20

(12) LEVEL II

AD A108599

STATISTICAL TOOLS FOR DETERMINING FITNESS TO FLY

Patrick L. Brockett, Ph.D.
Gerald A. Shea, Ph.D.
Department of Mathematics
The University of Texas at Austin
Austin, Texas 78712

DTIC
ELECTE
DEC 16 1981
S D
B

September 1981

Final Report for Period September 1978 - March 1980

Approved for public release; distribution unlimited.

Prepared for
USAF SCHOOL OF AEROSPACE MEDICINE
Aerospace Medical Division (AFSC)
Brooks Air Force Base, Texas 78235

DTIC FILE COPY

81 12 16 006

NOTICES

This final report was submitted by the Department of Mathematics, University of Texas at Austin, Austin, Texas 78712, under contract F33615-78-C-0623, job order 7755-17-24, with the USAF School of Aerospace Medicine, Aerospace Medical Division, AFSC, Brooks Air Force Base, Texas. Dr. Joel E. Michalek (USAFSAM/BRM) was the Laboratory Project Scientist-in-Charge.

When U.S. Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

Joel E. Michalek
JOEL E. MICHALEK, Ph.D.
Project Scientist

Richard A. Albanese
RICHARD A. ALBANESE, M.D.
Supervisor

Roy L. DeHart
ROY L. DEHART
Colonel, USAF, MC
Commander

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: SAM-TR-81-20

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): STATISTICAL TOOLS FOR DETERMINING FITNESS TO FLY

5. TYPE OF REPORT & PERIOD COVERED: Final Report, Sep 1978 - Mar 1980

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): Patrick L. Brockett, Ph.D.; Gerald A. Shea, Ph.D.

8. CONTRACT OR GRANT NUMBER(s): F33615-78-C-0623

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Mathematics, The University of Texas at Austin, Austin, Texas 78712

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 62202F; 7755-17-24

11. CONTROLLING OFFICE NAME AND ADDRESS: USAF School of Aerospace Medicine (BRM), Aerospace Medical Division (AFSC), Brooks Air Force Base, Texas 78235

12. REPORT DATE: September 1981

13. NUMBER OF PAGES: 43

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report): Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Proportional hazards model
Time dependent covariates
Maximum likelihood estimation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The goal of this project is to use data from regularly scheduled physical examinations to estimate the probability of an event such as a heart attack. This is done via the construction of a mathematical model. A high probability of an event as computed by the model would be evidence of a high risk case. The model developed here, termed the periodic checkup predictive model, is a survival distribution model similar to the proportional hazards model when there is little or no loss to followup and to logistic regression when the object is to predict the occurrence of an event during a fixed interval of time.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED 347833

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

TABLE OF CONTENTS

	Page
AN OVERVIEW	3
THE PERIODIC CHECKUP PREDICTIVE MODEL	5
IMPLEMENTATION OF THE MODEL	11
Use of a Subsample	12
Transformations of the Data	15
Calculation of the Initial Estimate	17
The Newton-Raphson Procedure	18
Verification by Sample Reuse	18
PROGRAM DESCRIPTION	19
Data Files	19
Library Routines	22
Sample Job Setup	23
Extensions	23
FLOW CHARTS	27
CONCLUSIONS FROM SAMPLE RUN	39
REFERENCES	42

A

STATISTICAL TOOLS FOR DETERMINING FITNESS TO FLY

AN OVERVIEW

Military personnel and other groups are routinely subject to regularly scheduled physical examinations or checkups. Beyond providing the personal benefits of continual health care, the checkups also serve to identify high risk cases. If the subject is a flyer or is responsible for dangerous equipment, a high probability of an incapacitating event may require a change of assignment to a less hazardous one. The goal of this project is to use data from regularly scheduled checkups to estimate the probability of an event such as a heart attack.

The task considered here is to construct a mathematical model to estimate the probability of an event. This model would allow the examining physician to summarize the subject's medical history in a meaningful way. A high probability of an event as computed by the model would be evidence of a high risk case. A number of models have been created for estimating the probability distribution of time to event, the so-called failure time. The most noticeable contribution in the biostatistical literature has been Cox's (3) proportional hazard model for possibly censored data using concomitant information. It has been applied to cancer data with a good deal of success, and a number of extensions of the model have appeared in the literature; cf. Breslow (2), Cox (3), Peduzzi et al. (9), Taulbee (12), Prentice and Kalbfleisch (9). However, there are qualitative differences between the populations of diagnosed cancer patients and of generally healthy military personnel. In the former, eventual failure occurs in a large proportion of the cases; in the latter, the event is relatively rare. Also, the chance of loss to followup is fairly large among the civilian population, while loss to followup is less of a problem among military personnel, especially rated career officers. Finally,

the data needed to predict survival among cancer patients is usually first collected at the time of initial diagnosis. The data to predict events among healthy subjects is limited to that routinely gathered during the periodic checkups. These differences lead us to propose a separate model which we have termed the periodic checkup predictive model. The model is a survival distribution model similar both to Cox's proportional hazard model when there is little or no loss to followup, and to logistic discrimination (Press and Wilson (10)) when the object is to predict the occurrence or nonoccurrence of an event during a fixed interval.

The chief motivation behind the model is to mimic the decision process of the examining physician at the end of a regularly scheduled checkup. The physician must decide on the strength of the current examination findings plus the subject's previous medical history whether there is sufficient risk of an event to require further tests. The horizon of the event period of primary interest is the time of the next regularly scheduled checkup. At that time, new data will be available that may change the estimate of risk.

In a similar manner, the model used here estimates the survival function of an individual from the moment of his last checkup until the time of his next exam. At the time of the next scheduled checkup a new assessment will be made. We indicate the time that the patient is at risk by τ; thus $\tau=0$ at the time of the most recent checkup and $\tau=1$ at the time of the next scheduled checkup. All covariates including past time-dependent ones can be considered to be fixed at $\tau=0$ and to remain so until $\tau=1$. Since an event must occur in the interval between two successive checkups, $0 \leq \tau_i \leq 1$ for all subjects suffering an event. The survivors, those who have never suffered an event, have $\tau_i=1$ for each time they survive over the period between checkups without an event, at which time new data becomes available and $\tau_i=0$

again. Ideally the data is collected and analyzed each period so that changes in the population are built into the model. In practice, a number of years may be clustered together so that there are a reasonable number of events. Then survivors appear repeatedly with expanded data sets at each new checkup. This leads to the problem of dependent sample members, but our research indicates that the problem may be solved by subsampling without replacement.

The chief advantage of the periodic checkup predictive model lies in the short horizon for the failure time. Loss to followups becomes less of a problem since it is necessary only to establish that the subject survived until the time of his next scheduled checkup. Reestimating the parameters each period allows changes in the population to be quickly built into the model. The model will more closely fit the observations since it need only predict over a short period using recent data. Finally the model is found to be computationally easy to implement.

THE PERIODIC CHECKUP PREDICTIVE MODEL

The model implemented here is based on the following assumptions:

(i) each individual is examined annually or, more generally, at the end of some fixed interval;

(ii) each examination consists of identical tests and readings;

(iii) records of at least two previous examinations are available at the time of the most recent examination for each individual;

(iv) the occurrence of an event (heart attack, say) before the next scheduled examination is relatively rare;

(v) once the measured covariates such as age, blood pressure, body mass, and certain fixed covariates are accounted for, all individuals are equally at risk until the next scheduled examination.

These assumptions are implemented by a proportional hazard model with constant base risk. Let t denote the age of the subject at the time of his last examination, and τ the future time where $\tau=0$ at the time of the last examination. Let $\underset{\sim}{z}(t)$ be the time-dependent covariates for the last and two previous examinations, and $\underset{\sim}{x}$ the time-independent covariates. Let $\underset{\sim}{y} \in R^m$, $m \geq 1$ denote appropriate transformations of $\underset{\sim}{z}(t)$ and $\underset{\sim}{x}$ that have been found to be informative about the chance of an event. The hazard rate $\lambda(\tau,\underset{\sim}{z}(t),\underset{\sim}{x})$ is assumed to have the form

$$\lambda(\tau,\underset{\sim}{z}(t),\underset{\sim}{x}) = \lambda_0 e^{\underset{\sim}{\beta}^T \underset{\sim}{y}} \qquad \text{for} \qquad 0 \leq \tau \leq 1 \tag{1}$$

and $P[\tau=1] = \exp\{-\lambda_0 e^{\underset{\sim}{\beta}^T \underset{\sim}{y}}\}$, where T indicates vector transpose and the parameter vector $\underset{\sim}{\beta} \in R^m$. This model is a version of one due to Taulbee (12) and as such is a generalization of Cox's model for two cancer populations' survival functions. It holds that the base population risk λ_0 is constant over the one period time interval until the next examination. At that time, the estimate for λ_0 is updated to account for any change in the population's prior risk of an event. For example, in recent years, it appears that fewer heart attacks are occurring among middle-aged men, suggesting that λ_0 should be successively lowered. The model is also a generalization of the Weibull base hazard rate model. To see this, note that the Weibull model hazard has the form

$$\lambda(t,\underset{\sim}{z}(t),\underset{\sim}{x}) = \lambda_0 \alpha (\lambda_0 t)^{\alpha-1} e^{\underset{\sim}{\gamma}_1^T \underset{\sim}{z} + \underset{\sim}{\gamma}_2^T \underset{\sim}{x}}$$

$$= (\lambda_0^{\alpha}\alpha)\exp[(\alpha-1)\log t + \underset{\sim}{\gamma}_1^T \underset{\sim}{z} + \gamma_2^T \underset{\sim}{x}] = \lambda_1 e^{\underset{\sim}{\beta}_1^T \underset{\sim}{y}} \tag{2}$$

where $\underset{\sim}{y}$ transforms age variable t to the variable $\log t$. The difference between equations 1 and 2 is that equation 1 gives the hazard in terms of future time τ, taking age at last checkup and the recorded covariates to be fixed until the next examination. On the other hand, equation 2 expresses the hazard for age t, and hence does not use age as a covariate. The advantage of equation 1 over equation 2 is that equation 1 requires fitting the data only over the interval until the next scheduled checkup while equation 2 requires a data fit essentially from birth until the age of an event or censoring. A second technical advantage is that in equation 1 we may estimate parameters by the method of maximum likelihood while it can be shown that no maximum likelihood estimates exist for the parameters of equation 2, and other less developed techniques must be utilized.

To find the maximum likelihood estimates (MLE) for $\lambda_0, \underset{\sim}{\beta}$ from equation 1, we note that

$$\lambda(\tau, \underset{\sim}{z}(t), \underset{\sim}{x}) = \lambda(\tau, \underset{\sim}{y}), \quad \text{say}$$

$$= f(\tau, \underset{\sim}{y})/(1-F(\tau, \underset{\sim}{y})).$$

This implies that the survival function is

$$S(\tau, \underset{\sim}{y}) = \exp\{-\lambda_0 e^{\underset{\sim}{\beta}^T \underset{\sim}{y}_\tau}\}.$$

For a sample of size n, suppose that r individuals, indexed by $1,\ldots,r$, have failed (suffered an event) prior to the time of their next scheduled checkup while $n-r$, indexed by $r+1,\ldots,n$, have survived through the time interval. Scale τ to equal 0 at the time of last checkup, and to 1 at the time of next scheduled checkup. Thus the failure times are $0 \leq \tau_j \leq 1$

for $j=1,\ldots,n$, with $\tau_j=1$ for $j=r+1,\ldots,n$. The likelihood function for the sample is

$$L = \lambda_0^r \exp\{\sum_{j=1}^{r} \underset{\sim}{\beta}^T \underset{\sim}{y}_j - \lambda_0 \sum_{i=1}^{n} \tau_i e^{\underset{\sim}{\beta}^T \underset{\sim}{y}_i}\}.$$

Hence the log likelihood

$$\ell = r \log \lambda_0 + \sum_{j=1}^{r} \underset{\sim}{\beta}^T \underset{\sim}{y}_j - \lambda_0 \sum_{i=1}^{n} \tau_i e^{\underset{\sim}{\beta}^T \underset{\sim}{y}_i}.$$

Taking the partial derivative of ℓ with respect to λ_0, and setting it equal to 0, we find the MLE for λ_0 to be

$$\hat{\lambda}_0 = r / \sum_{i=1}^{n} \tau_i e^{\underset{\sim}{\beta}^T \underset{\sim}{y}_i}. \tag{3}$$

Substituting $\hat{\lambda}$ into ℓ we have the log likelihood

$$\hat{\ell} = r \log r - r - r \log \sum_{i=1}^{n} \tau_i e^{\underset{\sim}{\beta}^T \underset{\sim}{y}_i} + \sum_{j=1}^{r} \underset{\sim}{\beta}^T \underset{\sim}{y}_i . \tag{4}$$

To find the MLE of $\underset{\sim}{\beta}$ from $\hat{\ell}$, we take the gradient of $\hat{\ell}$, $\partial\hat{\ell}/\partial\underset{\sim}{\beta}=\underset{\sim}{h}(\underset{\sim}{\beta})$, say. We find a convenient matrix form for $\underset{\sim}{h}(\underset{\sim}{\beta})$: let Y be the $n\times k$ matrix of sample points for k transformed covariates observed for n subjects. Let $\underset{\sim}{P}$ be the n-dimensional column vector with elements

$$p_i = \tau_i e^{\underset{\sim}{\beta}^T \underset{\sim}{y}_i} / \sum_{i=1}^{n} \tau_i e^{\underset{\sim}{\beta}^T \underset{\sim}{y}_i} \tag{5}$$

Let $\underset{\sim}{\Delta}$ be the n-dimensional column vector of failure indicators

$$\begin{cases} \delta_i = 1 & \text{if subject } i \text{ failed.} \\ \delta_i = 0 & \text{if subject } i \text{ survived.} \end{cases}$$

Then

$$\underset{\sim}{h}(\underset{\sim}{\beta}) = Y^T(\underset{\sim}{\Delta} - r\underset{\sim}{P}). \tag{6}$$

We find the zeros for this set of k simultaneous equations by the Newton-Raphson technique. Let D be the $n \times n$ diagonal matrix with diagonal elements p_i. Then the $n \times k$ matrix

$$\partial \underset{\sim}{P}/\partial \underset{\sim}{\beta} = (D - \underset{\sim}{P}\underset{\sim}{P}^T)Y.$$

Hence

$$\partial \underset{\sim}{h}(\underset{\sim}{\beta})/\partial \underset{\sim}{\beta} = -rY^T(D - \underset{\sim}{P}\underset{\sim}{P}^T)Y. \tag{7}$$

The Newton-Raphson method iteratively solves for $\hat{\underset{\sim}{\beta}}$ by updating values $\underset{\sim}{\beta}_\alpha$ to $\underset{\sim}{\beta}_{\alpha+1}$ where

$$\underset{\sim}{\beta}_{\alpha+1} = \underset{\sim}{\beta}_\alpha + [Y^T(D - \underset{\sim}{P}\underset{\sim}{P}^T)Y]^{-1}Y^T(\underset{\sim}{\Delta}/r - P). \tag{8}$$

There is a heuristic interpretation for some of these equations. In equation 3, $\hat{\lambda}_0$ is the ratio of the average number of failures, r/n, to the average value of the covariate effect $\tau \exp \underset{\sim}{\beta}^T \underset{\sim}{y}$, namely, $\sum_{i=1}^{n} \tau_i e^{\underset{\sim}{\beta}^T \underset{\sim}{y}_i}/n$, for the given sample of size n. Thus $\hat{\lambda}_0$ plays the combined role of providing an estimate for the population failure rate between examinations plus a centering estimate that leaves the $\underset{\sim}{y}_i$ invariant under uniform change of location. This role is analogous to that of the constant term in discriminant analysis which incorporates both prior probabilities and an overall mean. Setting $\underset{\sim}{h}(\underset{\sim}{\beta})$ equal to 0 in 6 and dividing by r shows that the MLE $\hat{\underset{\sim}{\beta}}_0$ occurs when

$$\overline{\underset{\sim}{y}}_r = Y^T P = E_p \underset{\sim}{y}, \quad \text{say,}$$

since each $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$. That is, the maximum likelihood occurs when the expected value for the covariates computed by using the $\underset{\sim}{\beta}$ in equation 5 equals the sample average for the failure group. If the $\underset{\sim}{y}_i$ display a general shift between the failure and the survival groups, then the $\underset{\sim}{\beta}$ will tend to reflect this shift, to give, in general, positive values for $\underset{\sim}{\beta}^T \underset{\sim}{y}_i$ in the failure group and negative values in the survival group. This shift will be magnified in cases where τ_i is small, and hence subjects who have events soon after their checkup will be weighted more heavily in determining $\underset{\sim}{\beta}$. Equation 8 will be recognized as weighted least-squared regression to compute the increment in $\underset{\sim}{\beta}$ at each iteration.

There is also a close resemblance between Cox's (3) proportional hazard model for life tables from censored data and the model introduced here. Equation 4 shows that the log likelihood of our model, up to a constant term in r, is

$$\sum_{j=1}^{r} \{\underset{\sim}{\beta}^T \underset{\sim}{y}_j - \log[\sum_{i=1}^{n} \tau_i e^{\underset{\sim}{\beta}^T \underset{\sim}{y}_i}]\}. \tag{9}$$

In our notation, the log likelihood of the Cox model (3) becomes

$$\sum_{j=1}^{r} \{\underset{\sim}{\beta}^T \underset{\sim}{y}_j - \log[\sum_{i' \varepsilon R(\tau_j)} e^{\underset{\sim}{\beta}^T \underset{\sim}{y}_{i'}}]\} \tag{10}$$

where $R(\tau_j)$ is the index set of survivors at failure time τ_j. Let us assume that $0 \leq \tau_1 \leq \tau_2 \leq \ldots \leq \tau_r \leq 1$ and recall that $\tau_i = 1$, $i = r+1, \ldots, n$. Also in the application to Air Force flyers it is reasonable to assume that no censoring occurs, since very few of the subjects, if any, will be lost to followup over the course of the period between checkups. Then $\{r+1, \ldots, n\}$ will

be contained in $R(\tau_j)$ for each j, $j=1,\ldots,r$. Let $r(\tau_j)=R(\tau_j)-\{r+1,\ldots,n\}$, the index set of failing subjects who survive past τ_j; if there are no ties among the failure times,

$$r(\tau_j) = \{j=1,\ldots,r\}.$$

The difference between expressions 9 and 10 is then

$$\sum_{j=1}^{r} \{\log[\sum_{k\epsilon r(\tau_j)} e^{\underset{\sim}{\beta}^T \underset{\sim}{y}_k} + \sum_{i=r+1}^{n} e^{\underset{\sim}{\beta}^T \underset{\sim}{y}_i}] - \log[\sum_{j=1}^{r} \tau_j e^{\underset{\sim}{\beta}^T \underset{\sim}{y}_j} + \sum_{i=r+1}^{n} e^{\underset{\sim}{\beta}^T \underset{\sim}{y}_i}]\}.$$

We see that the key distinction between the two models in the application to periodic checkups of Air Force flyers is that the Cox model uses only the information that flyer j_2 survived flyer j_1 if $\tau(j_1) < \tau(j_2)$, $1 \leq j_1 < j_2 \leq r$ while the periodic checkup model uses the actual time after checkup to weight the $e^{\underset{\sim}{\beta}^T \underset{\sim}{y}_j}$. In practice we suspect that the two models will lead to similar estimates of the $\underset{\sim}{\beta}$. Also because the number of failures per year, r, is small relative to the sample size, n, and because loss to followup is negligible, the Kaplan-Meier (6) estimates of $\lambda_0(\tau)$, the base hazard rate, will be nearly constant in our application. For these reasons, plus the mathematical tractability of the full likelihood technique, we have opted for the model proposed here over the Cox model.

IMPLEMENTATION OF THE MODEL

Five parts to the implementation of the periodic checkup model were addressed before creating the actual program: (i) the use of a subsample from the surviving population, (ii) transformations and selecting reexpressions of the data, (iii) calculation of the initial estimate, (iv) estimation of the

parameters by the Newton-Raphson procedure, and (v) verification of the procedure by reusing the subsample. These will be considered in turn.

Use of a Subsample

Because of the rarity of the event, perhaps of the order of two heart attacks per year for each 1000 at risk, it will be necessary to use a large sample size over a number of years to find a useful number of cases. The mixture of years makes the procedure less sensitive to changes in the population, such as the possible decrease in heart attacks in recent years. However, without a large program to collect data each year, it will have to be assumed that there is little or no change in the population at risk over a period sufficiently long to gather a reasonably large number of events.

Let us consider the example of the data set gathered by considering approximately 3000 flyers at risk over the period 1974-1978. To assume that there were at least two checkups previous to the checkups that began the risk period, the risk period was taken to be the two years 1976-1978. During this period, eight of the 3000 were admitted to hospitals with the primary diagnosis of acute myocardial infarction, an average of 1.33 per 1000 per year. Suppose that the 1976 and 1977 risk years are both used. The presence of a 1978 checkup merely indicates that the subject was not lost to followup during the 2-year risk period 1976-1978. Then the total number of risk cases, using both years, is roughly 6000. That is, in the direct model, 6000 cases need to be analyzed to find eight events. The proposed alternative was to sample systematically 20% of the 3000 from each of two strata based on age and to use the second-to-last checkup plus the two previous to that one, usually 1975 to 1977, only. When the small number of incomplete cases were eliminated, 553 subjects remained. However, all eight event cases would be used. To check

the precision of the estimates, we compare the ratio of the two standard deviations of the estimates from the samples. We find

$$\frac{\sigma[(1/8) + (1/561)]^{1/2}}{\sigma[(1/8) + (1/6000)]^{1/2}} = \frac{.3561}{.3538} = 1.0065.$$

We conclude that only negligible gains in precision are possible using all 6000 risk cases. To check the bias of the subsample method, we assume that each member of the stratified sample represents ten "identical clones" in the full sample. This assumption appears to be reasonable for this large subsample suggesting that a relatively small nine-member neighborhood can be constructed in the full sample for most of the subsampled cases. It is similar to the assumptions in stratified sampling. Then the maximum likelihood estimate, $\hat{\underset{\sim}{\beta}}$, for $\underset{\sim}{\beta}$ from the subsample of size n satisfies

$$r^{-1} \sum_{j=1}^{r} \underset{\sim}{y}_j = \sum_{i=1}^{n} \underset{\sim}{y}_i e^{\hat{\underset{\sim}{\beta}}^T \underset{\sim}{y}_i} / \sum_{i=1}^{n} e^{\hat{\underset{\sim}{\beta}}^T \underset{\sim}{y}_i}$$

$$= \left(\sum_{i=1}^{r} \underset{\sim}{y}_i e^{\hat{\underset{\sim}{\beta}}^T \underset{\sim}{y}_i} + \sum_{i=r+1}^{n} \underset{\sim}{y}_i e^{\hat{\underset{\sim}{\beta}}^T \underset{\sim}{y}_i} \right) / \sum_{i=1}^{n} e^{\hat{\underset{\sim}{\beta}}^T \underset{\sim}{y}_i}. \tag{11}$$

The maximum likelihood estimate from the "cloned group" satisfies

$$r^{-1} \sum_{j=1}^{r} \underset{\sim}{y}_j = \left(\sum_{i=1}^{r} \underset{\sim}{y}_i e^{\hat{\underset{\sim}{\beta}}^T \underset{\sim}{y}_i} + 10 \sum_{i=r+1}^{n} \underset{\sim}{y}_i e^{\hat{\underset{\sim}{\beta}}^T \underset{\sim}{y}_i} \right) / \left(\sum_{i=1}^{r} e^{\hat{\underset{\sim}{\beta}}^T \underset{\sim}{y}_i} + 10 \sum_{i=r+1}^{n} e^{\hat{\underset{\sim}{\beta}}^T \underset{\sim}{y}_i} \right)$$

$$= \left(10^{-1} \sum_{i=1}^{r} \underset{\sim}{y}_i e^{\hat{\underset{\sim}{\beta}}^T \underset{\sim}{y}_i} + \sum_{i=r+1}^{n} \underset{\sim}{y}_i e^{\hat{\underset{\sim}{\beta}}^T \underset{\sim}{y}_i} \right) / \left(10^{-1} \sum_{i=1}^{r} e^{\hat{\underset{\sim}{\beta}}^T \underset{\sim}{y}_i} + \sum_{i=r+1}^{n} e^{\hat{\underset{\sim}{\beta}}^T \underset{\sim}{y}_i} \right). \tag{12}$$

This suggests that if n is sufficiently large relative to r that the effect of the r first terms in the numerator and denominator of the right hand of equation 1 is small and the two maximum likelihood estimators of $\underset{\sim}{\beta}$ will be very similar. Here 8 over 553 is roughly 1 in 70, and the 8 failing cases do not display data dramatically different from the 553 control cases. We conclude that the $\underset{\sim}{\beta}$ estimates from the subsample are both reasonably accurate and precise enough to justify the use of the subsample approach.

The parameter estimate that will display bias (though not lack of precision) will be $\hat{\lambda}_0$, the MLE for the base hazard rate constant λ_0. Here

$$\hat{\lambda}_0 = r / \sum_{i=1}^{n} e^{\hat{\underset{\sim}{\beta}}^T \underset{\sim}{y}_i}$$

is composed of the average rate of events in the subsample r/n and the constant term for the covariate coefficients $e^{\hat{\underset{\sim}{\beta}}^T \underset{\sim}{y}}$. But the average rate for the population is not 8/561 but rather 8/6000. This is easily corrected by substituting .75, the average hazard rate for the subsample, for r in $\hat{\lambda}_0$. This may be verified by noting that

$$P[T>1 \mid \underset{\sim}{y}] = e^{-\lambda_0 e^{\underset{\sim}{\beta}^T \underset{\sim}{y}}}$$

is typically .995 or larger. Then we have the approximation

$$\sum_{i=1}^{n} P[T_i \leq 1 \mid \underset{\sim}{y}_i] = \sum_{i=1}^{n} [1-\exp\{-\hat{\lambda}_0 e^{\hat{\underset{\sim}{\beta}}^T \underset{\sim}{y}_i}\}]$$

$$= \sum_{i=1}^{n} \hat{\lambda}_0 e^{\hat{\underset{\sim}{\beta}} \underset{\sim}{y}_i} = \sum_{i=1}^{n} r e^{\hat{\underset{\sim}{\beta}} \underset{\sim}{y}_i} / \sum_{i=1}^{n} e^{\hat{\underset{\sim}{\beta}} \underset{\sim}{y}_i} = r.$$

Thus the expected number of failures in the subsample roughly equals the numerator of $\hat{\underset{\sim}{\lambda}}_0$; that is, r should equal .75. This correcting factor for $\hat{\lambda}_0$

is necessary when the event in question is so rare that a stratified subsampling technique is mandated, but inference of survival probabilities involves the entire risk set.

Transformations of the Data

One of the chief advantages of the periodic checkup model with its constant update of the survival probability is that the time-dependent data can be considered to be fixed at time $\tau=0$. Following Frank (5) we feel that three checkups provide sufficient information about the subject's state to limit our consideration only to those three most recent examinations. Three orthogonal time-series transformations of the data check for constant, linear, and quadratic trends in the data: (1 1 1), (-1 0 1), and (1 -2 1). We considered these as representatives of general classes of time-dependent transformations. The first class or block of constant transformations presently built into the program has the coefficient vectors: (1 1 1), (0 0 1), (0 1 0), (1 0 0), and (1 2 4). The second or linear block is composed of (0 -1 1), (-1 0 1), (-1 1 0), and (-1 -1 2). The final block has the single quadratic coefficient vector (1 -2 1). Each coefficient vector $(a_1\ a_2\ a_3)$ is used to find the corresponding trend in the data for each scalar variable. Let x_0, x_{-1}, x_{-2} represent the value of the covariate at the last and most recent checkup, at the previous one, and at the second to last examination. Then we compute the inner product $a_1x_{-2} + a_2x_{-1} + a_3x_0$ to get a transformed scalar variable y. This transformation y_i is computed for all failing subjects, $i=1,\ldots,r$, and surviving subjects, $i=r+1,\ldots,n$. The value of y as a discriminator is determined by computing a t statistic for each transformation y for the two samples of failures and survivors. The variance estimate is derived from the covariance matrix for each variable over the three

examinations from the surviving group since it is possible that the failing group may not have identically distributed cases, cf. Shea (11). Since r is small relative to n, the effect is minimal in any case. Then

$$t = (\overline{y}_r - \overline{y}_{n-r})[(1/r + 1/n-r)(a^T S a)]^{-1/2}$$

where $a^T = (a_1, a_2, a_3)$; $\overline{y}_r$ and $\overline{y}_{n-r}$ are sample averages for transformed variable y, and S is the autocovariance matrix for the variable under consideration from the surviving sample. Since the statistic has $n-r=553$ degrees of freedom for the denominator, and $r=8$, it may be safely assumed that t is close to having the standard normal distribution. Therefore any t value greater than one suggests that the chance is less than one-third that the two populations have identical mean values for this transformation.

To select from among the thirty t values computed, we imposed some prior constraints. First it was decided that age and at least one transformation of each variable would be included. This reflects the belief that age and each variable observed are recorded because experience has related each of these to the chance of an event. The transformations with highest t values are employed subject to these constraints until from six to nine such transformations have been selected. In the sample run, the maximum number of time-dependent data transformations was set to six.

The variables are modeled to be log-linear in the hazard rate and log-loglinear in the survival function. It is natural to ask if a reexpression of the data would serve to separate the two samples better. Two reexpressions were tried by taking the natural log and the inverse of each raw data value. These were motivated by the fact that most variables considered were ratios such as mm Hg/cm^2 or kg/cm^2, and also these reexpressions have traditionally been considered very fruitful in other applications; cf. Box et al. (1). No

distinctly better separations of the data were found under either reexpression except for a slight improvement from using the logarithm of the body mass index (BMI). The improvement was not deemed sufficiently large to justify the loss of flexibility in the program introduced by taking the logarithm of one variable while the others were unchanged. Therefore no reexpression of data is built into the program, but an investigator may, if he chooses, reexpress data before entering it.

The option to include new variables (e.g., smoking and family history) which are known risk factors is included in the program. These variables specified by the user together with the temporal contrasts selected by the above t-values would then be used to construct estimates of $\lambda_0, \underset{\sim}{\beta}$ and $S(t)$.

Calculation of the Initial Estimate

Formally, the iterative procedure for estimating the $\underset{\sim}{\beta}$ is similar to that for estimating logistic discriminant function coefficients; cf. Press and Wilson (10). For this reason, the linear discriminant coefficients,

$$\hat{\underset{\sim}{\beta}}_0 = S^{-1}(\overline{\underset{\sim}{X}}_r - \overline{\underset{\sim}{X}}_{n-r}),$$

where $\overline{\underset{\sim}{X}}_r$ and $\overline{\underset{\sim}{X}}_{n-r}$ are the average covariate vectors for those who failed and did not fail and S is the pooled covariance matrix of the covariate vectors, were used as the initial values for the iterative solution of the maximum likelihood equations. The sample run verified that $\hat{\underset{\sim}{\beta}}_0$ was a good initial value. Since $\hat{\underset{\sim}{\beta}}_0$ is basically independent of the sample size, this close approximation between $\hat{\underset{\sim}{\beta}}_0$ and $\hat{\underset{\sim}{\beta}}$ verified that the effect of using a subsample rather than the full sample is probably negligible.

The Newton-Raphson Procedure

The Newton-Raphson procedure for finding the MLE is based on the assumption that the zeros of the first derivative provide the global maximum of the likelihood. Since the log likelihood is differentiable everywhere on the domain of $\underset{\sim}{\beta}$, since there is a unique critical point, and since the Jacobian of the derivative, the matrix of second-order derivatives, is the negative of a positive definite matrix, the Newton-Raphson procedure indeed leads to the unique MLE. Moreover, the inverse of the matrix of second-order derivatives evaluated at $\hat{\beta}$, denoted by GINV in the final iteration of the program, is the Fisher information matrix.

Verification by Sample Reuse

The program has been written to apply the $\underset{\sim}{\beta}$ and λ_0 (corrected) MLE to the subsample cases. This allows the investigator to decide on the apparent error of a procedure that predicts an event if the probability $P[T \leq 1 | \underset{\sim}{z}(t)]$ is greater than p_0, say, and predicts no event otherwise. However, sample reuse leads to a favorable bias on the error of the procedure; cf. Lachenbruch (7). A better idea of the error can be computed from the bootstrap methods summarized by Efron (4). However, in initial trials such as this, the apparent error provides some idea of the usefulness of the procedure. The sample run here provides the following estimates of false positives (survivors who would have been predicted to fail during the interval between their last recorded checkups, usually the year 1977-1978) and false negatives (acute myocardial infarction patients who would not have been considered at risk) for various values of p.

p_0	False + (% of 553)	False - (% of 8)
.0012	39.2	0.0
.0014	33.3	12.5
.0016	25.9	25.0
.0018	21.5	37.5
.0020	17.2	50.0
.0025	10.1	62.5
.0030	6.9	87.5
.0035	4.0	100.0

This sample reuse estimate suggests that if p_0=.0016 is used, only one-fourth of the survivors and one-fourth of the event cases would be misclassified. However, even if this optimistic estimate is true, hundreds of false positives would appear among the 3000 subjects at risk each year. We conclude that the information provided by the data presently available, namely, systolic and diastolic blood pressures, age, and body mass index, is still insufficient to allow the computed probability to be anything more than a convenient summary statistic to the examining physician. As more significant variables are added to the covariate history (e.g., smoking behavior, family history of coronary heart disease, triglyceride and lipid readings, etc.) one would expect to see improved discrimination results for the procedure.

PROGRAM DESCRIPTION

Data Files

Three input data files are needed to run the program: INPUT, DATAS, and DATAF.

1. 'INPUT' file: contains some constants needed to run the program.

(a) Number of Cards in file: 13

(b) Layout of Card 1:

Field	Length	Type	Variable	
1	8	Real	EFAIL:	Average # of failures for size of control group
11	2	Integer	CYEAR:	Year of last exam
13	4	Integer	NVAR:	# of time-dependent variables to select
17	4	Integer	NAV:	# of time-independent variables to select
21	8	Real	XINC:	Increment of frequency table

(c) NVAR must be between 5 & 9, and NAV must be between 0 & 3.

Also the sum must be less than or equal to 9.

(d) Layout of Cards 2 - 11:

Field	Length	Type	Variable	
1	4	Real	A(I,1)	Weight of 1st year
5	4	Real	A(I,2)	Weight of 2nd year
9	4	Real	A(I,3)	Weight of 3rd year (year prior to last exam)

I = 1,...,10

(e) Layout of Card 12:

Field	Length	Type	Variable	
1	4	Integer	BB(1):	boundary of 1st block
5	4	Integer	BB(2):	boundary of 2nd block
9	4	Integer	BB(3):	boundary of 3rd block

(f) Layout of Card 13:

Field	Length	Type		Variable
1	4	Char.	Name(1):	name of 1st time-dependent variable
5	4	Char.	Name(2):	name of 2nd time-dependent variable
9	4	Char.	Name(3):	name of 3rd time-dependent variable
13	4	Char.	Name(4):	Always AGE

(g) A copy of sample data is included.

2. 'DATAS' file: contains records of control group.

(a) Number of records of this file does not have to be known so long as the limit is not exceeded.

(b) Number of cards per record: 2

(c) Layout of Card 1:

Field	Length	Type	Variable
1	9	Integer	ID or blank (not used in program)
10	2	Integer	Year of birth
12	2	Integer	Month of birth (not used in program)
14	2	Integer	Day of birth (not used in program)
16	2	Integer	No. of month survived after last exam.
18	2	Integer	Year of last exam prior to disease
20	4	Integer	Time-independent variable 1
24	4	Integer	Time-independent variable 2
28	4	Integer	Time-independent variable 3

(d) Fields 16-19 of Card 1 are blank in this file.

(e) There is no time-independent variables in DATAS/DATAF at present time.

(f) Sample size: 553

(g) Total number of records in DATAS & DATAF may not exceed 570; otherwise, the dimensions in the program must be increased.

(h) Layout of Card 2:

Field	Length	Type	Variable
1	8	Real	Time-dependent variable 1 of Year 1
9	8	Real	Time-dependent variable 2 of Year 1
17	8	Real	Time-dependent variable 3 of Year 1
25	8	Real	Time-dependent variable 1 of Year 2
33	8	Real	Time-dependent variable 2 of Year 2
41	8	Real	Time-dependent variable 3 of Year 2
49	8	Real	Time-dependent variable 1 of Year 3
57	8	Real	Time-dependent variable 2 of Year 3
65	8	Real	Time-dependent variable 3 of Year 3

3. 'DATAF' file: contains records of disease group.

(a) See DATAS for detail, except columns 16-19.

(b) Sample size for this data: 8

Library Routines

We use one IMSL routine LINV1F in the program, which finds the inverse of a matrix. Several Fortran functions are used. In particular, MNFLIB is loaded together with IMSLIB.

Sample Job Setup

The example runs are done on CDC Cyber 170 machine under UT2D operating system. The necessary commands are:

```
READPF, (Tape Name), MAIN, INPUT, DATAS, DATAF.
RFL,77700.
MNF,I=MAIN.
LOAD,LGO,MNFLIB,IMSLIB
```

which assumes the files are sorted at some permanent file storage. The result will be a file OUTPUT.

Extensions

1. The example run does not contain any time-independent variables. These data may be inserted in Card 1 of each record in DATAS/DATAF and set NAV (number of additional variables) accordingly.

2. The limit of NVAR + NAV $\leq$ 9 may be increased to 12 so that we can test more variables. To do so, we need to redimension the following arrays:

- Y	in	Main
- Y,XT	in	Select
- X,S,Beta,Betas,Mean,XTX,T1,G,GINV,T5	in	LSQ

Also Maxvar has to be set to 12 in the main program.

3. The limit of no more than 70 records may be increased to any reasonable number. Maxcas in main has to be set to reflect the change. The following arrays have to be redimensioned:

- X,Y,Tau,Event,ADDV	in	Main
- X,Tau,Y,Event,ADDV	in	Select
- X,P1,P,T1,Tau,XOLD,SP,Event	in	LSQ

4. The weights for variable selection may be changed by changing the corresponding data cards in INPUT file. However, the second one cannot be changed because it is used to compute the T-value of Age of last year.

5. Current block boundaries are 5,9,10,i.e.,

Block 1: 1~5

Block 2: 6~9

Block 3: 10

This may be changed by changing the data on the 12th Card.

6. Names for time-dependent variables may be changed by using a different data card at the end of INPUT file (no more than 4 characters per name).

General Structure of Algorithm

START

↓

Input Data

↓

Select Time-Dependent Variables

↓

Merge Time-Independent Variables If Any

↓

Compute Initial Estimation of Beta

↓

Find Estimation of Beta by Newton-Raphson Method

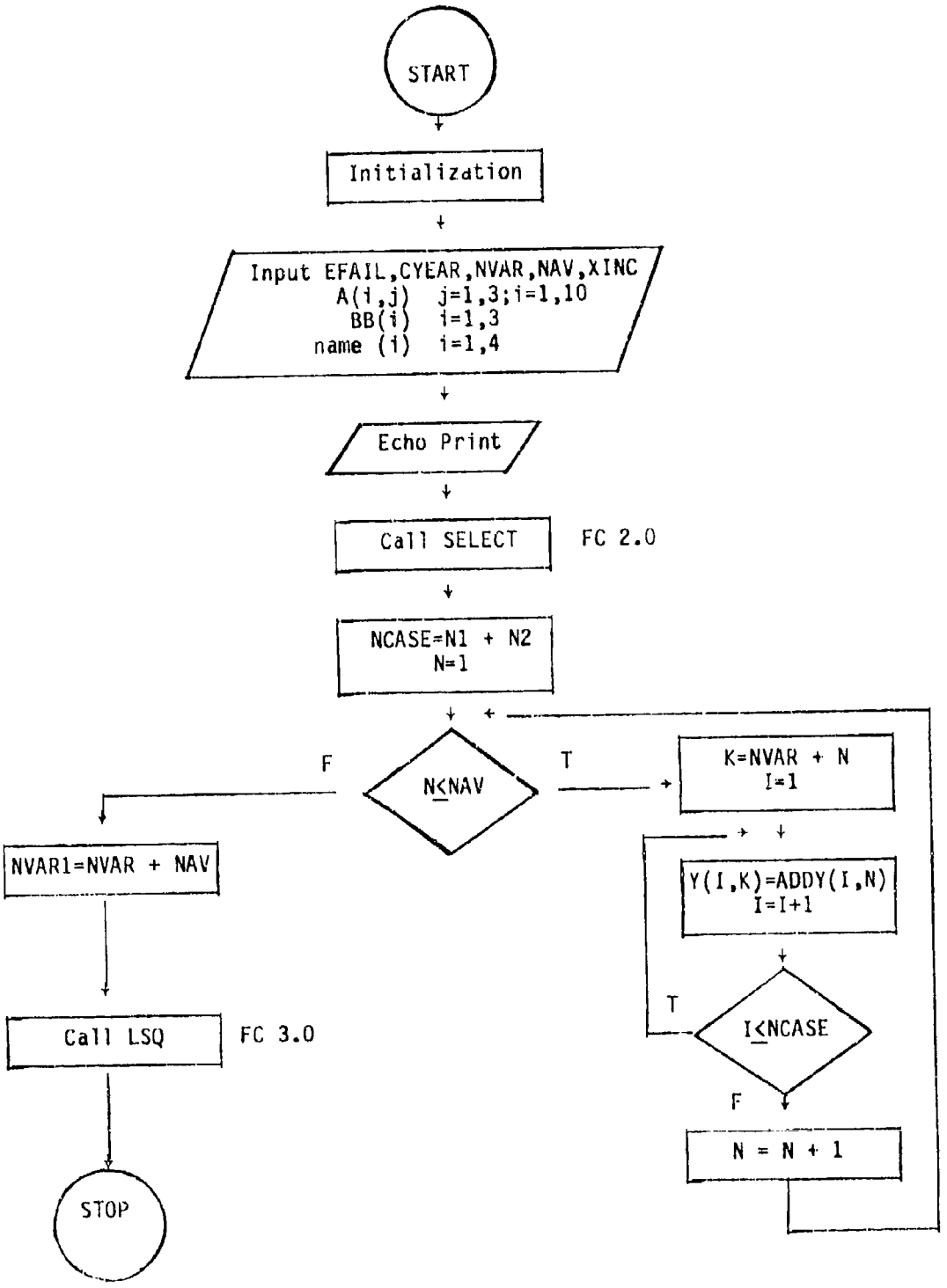
↓

Compute Survival Probability and Frequency Tables

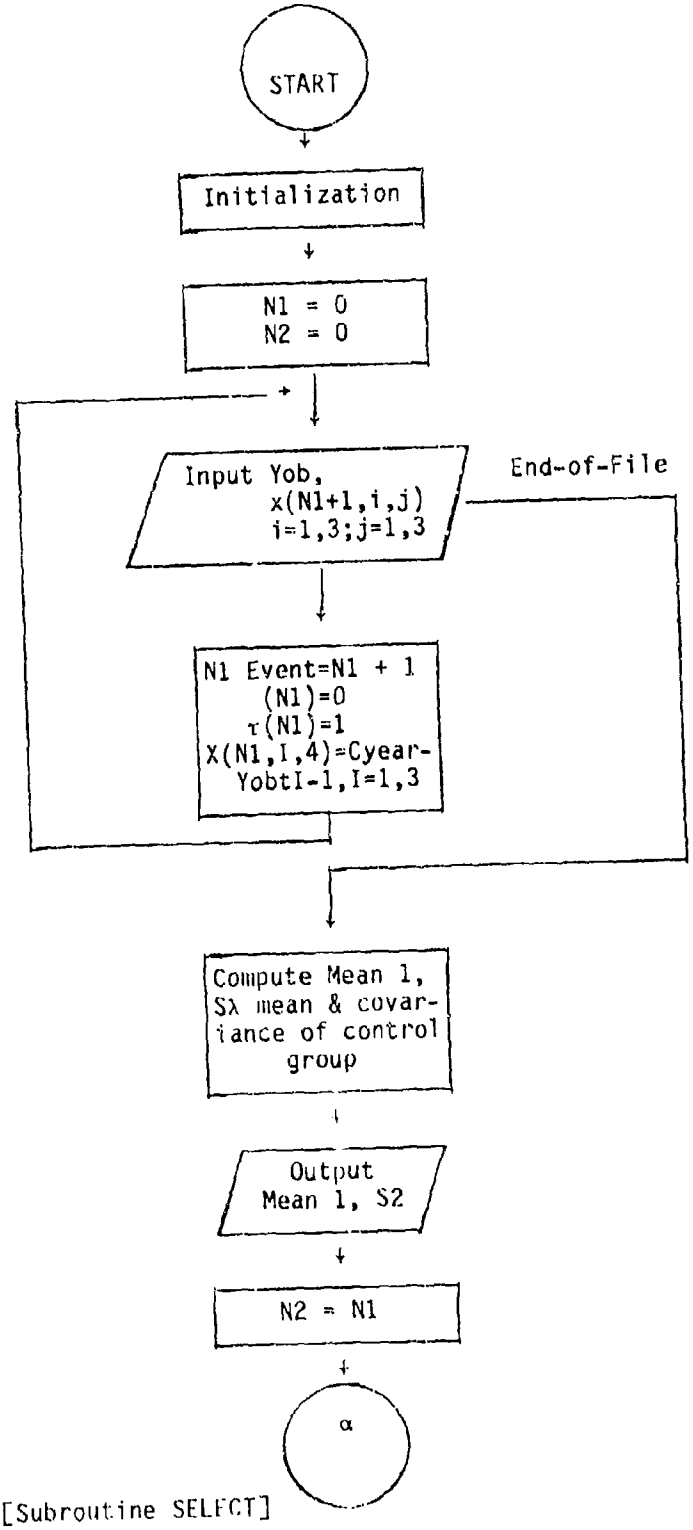
↓

END

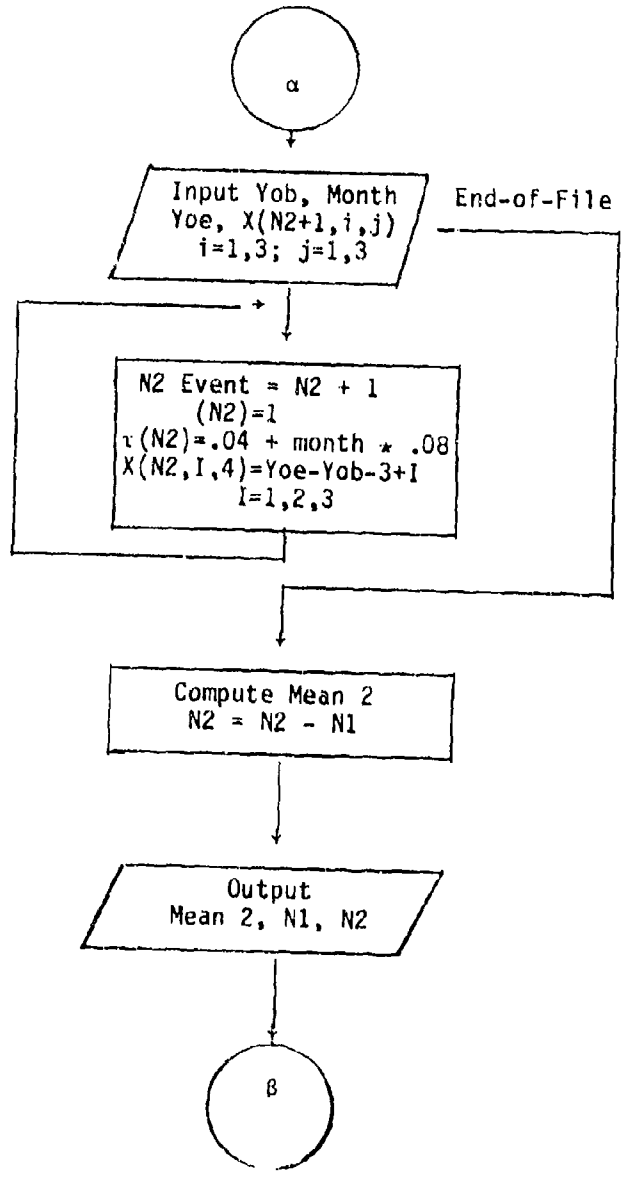
FLOW CHARTS



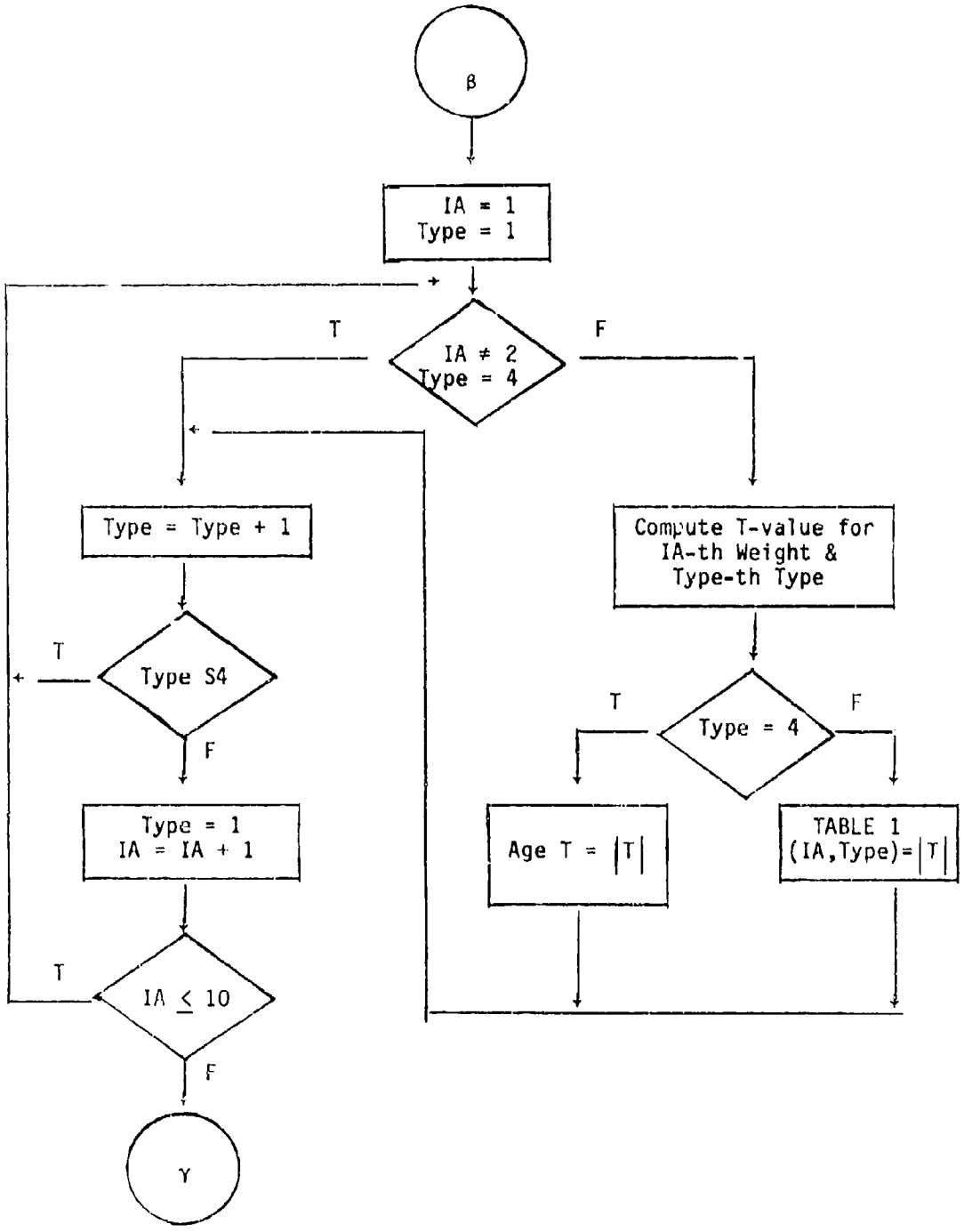
Flow Chart 1. [Main Program]



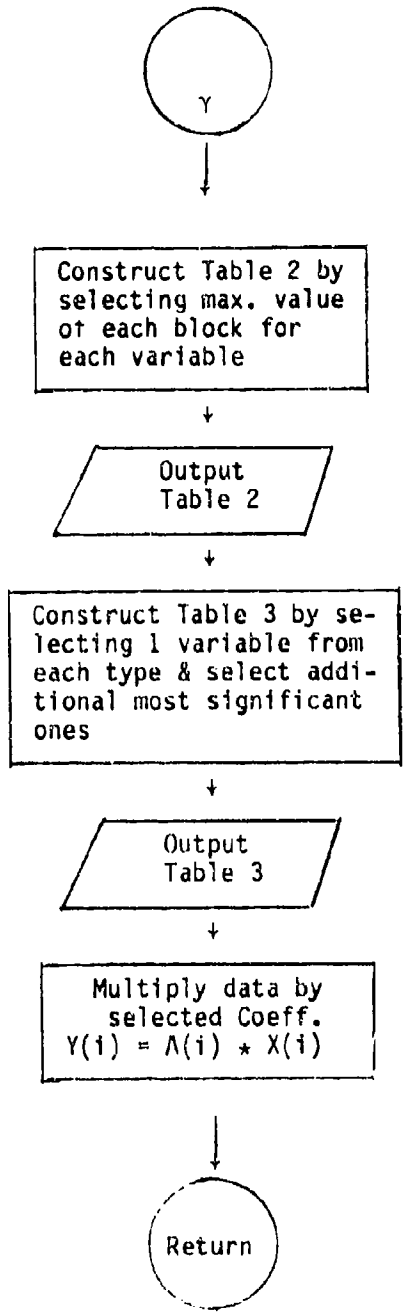
Flow Chart 2.0 [Subroutine SELECT]



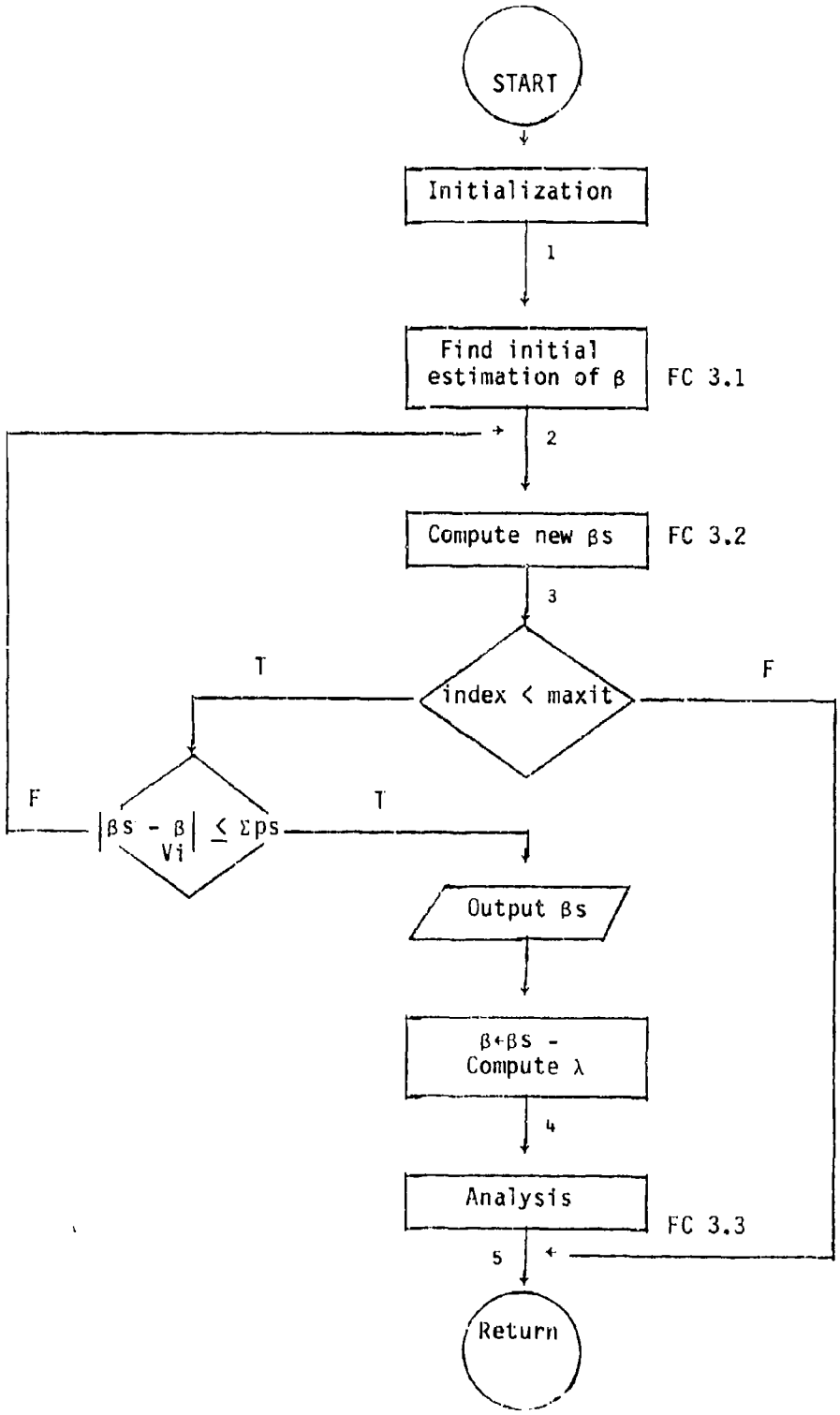
Flow Chart 2.1 [Subroutine SELECT]



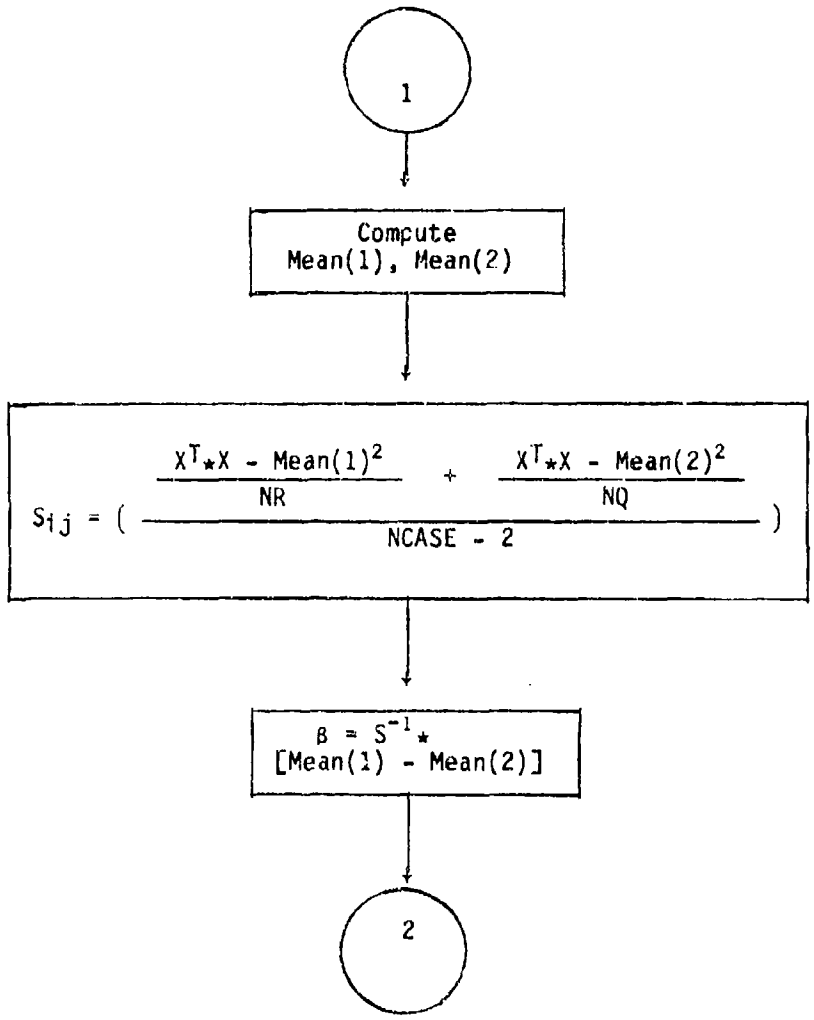
Flow Chart 2.2 [Subroutine SELECT]



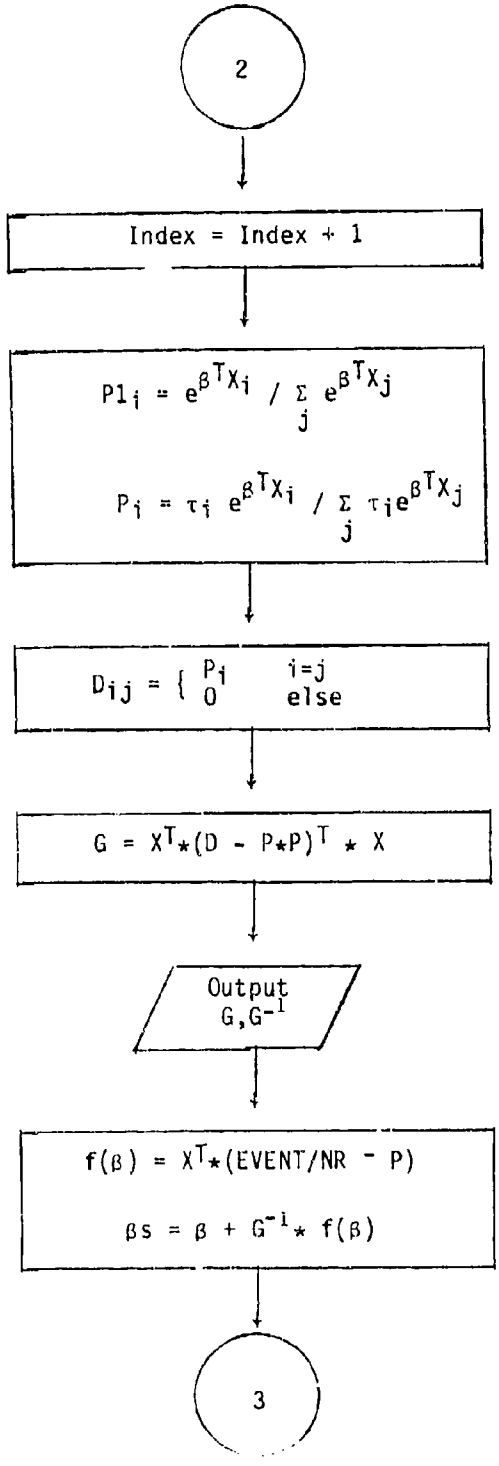
Flow Chart 2.3 [Subroutine SELECT]



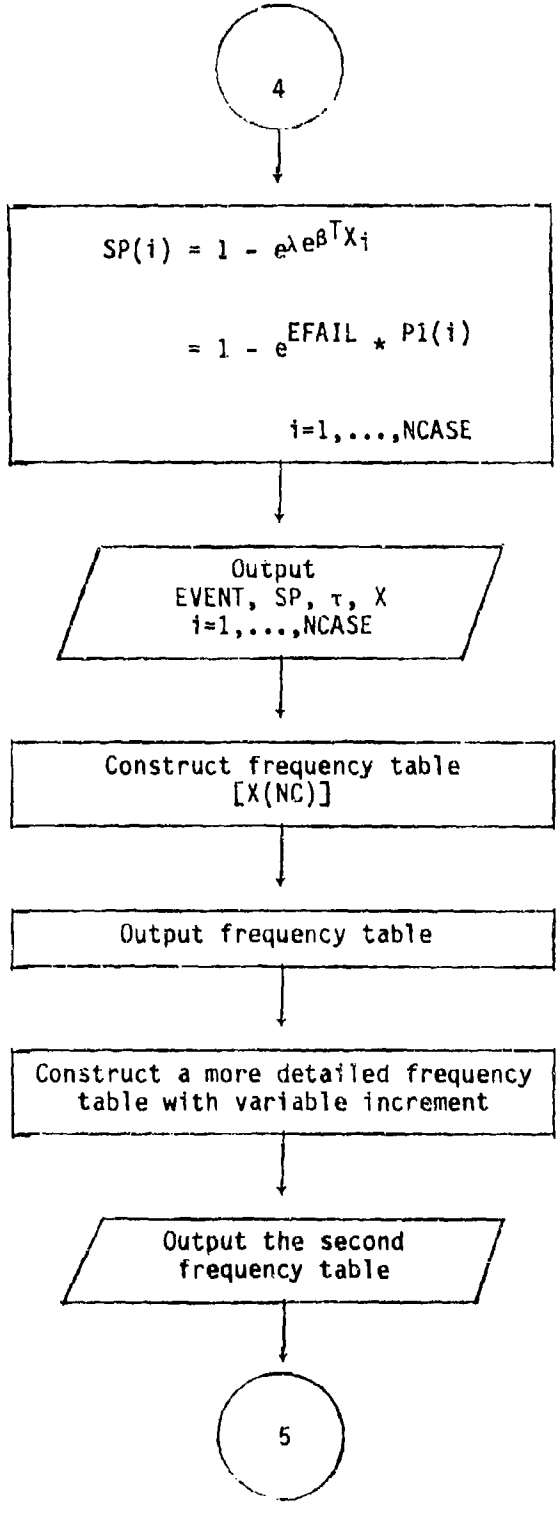
Flow Chart 3.0 [Subroutine LSQ]



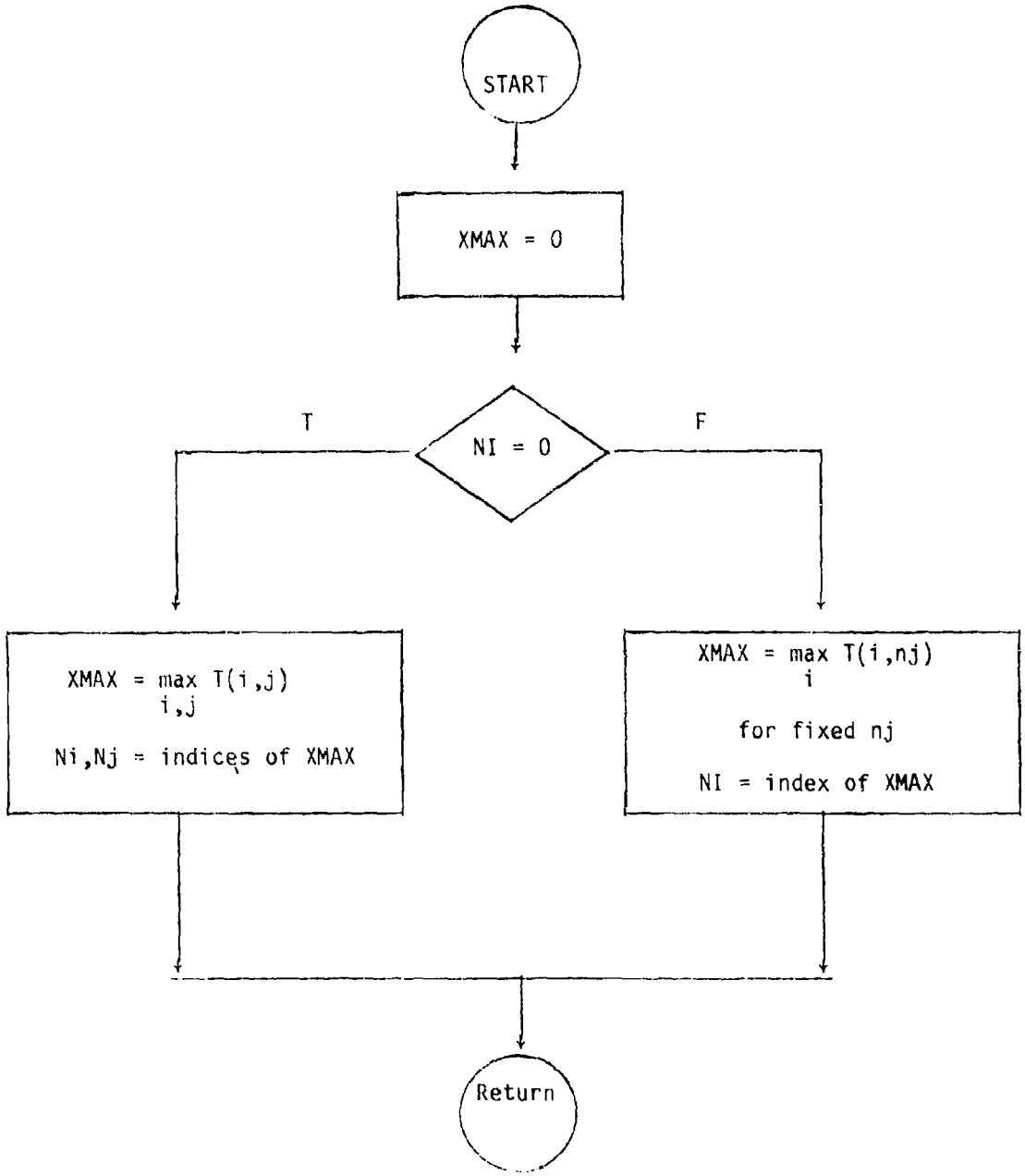
Flow Chart 3.1 [Subroutine LSQ]



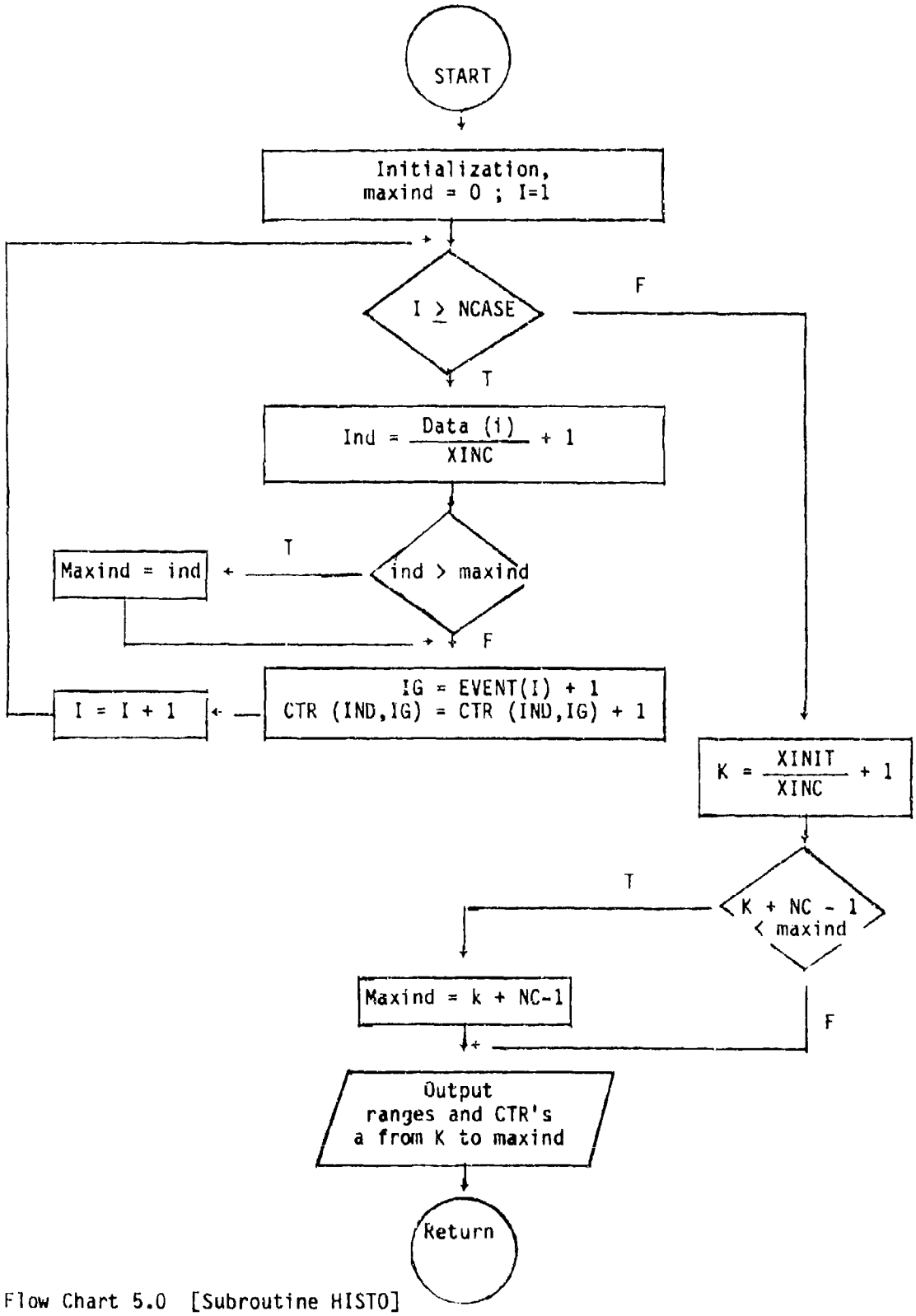
Flow Chart 3.2 [Subroutine LSQ]



Flow Chart 3.3 [Subroutine LSQ]



Flow Chart 4.0 [Subroutine MAXT]



Flow Chart 5.0 [Subroutine HISTO]

CONCLUSIONS FROM SAMPLE RUN

The sample run provided here combines eight cases of myocardial infarct and 553 systematically selected control cases from two samples stratified by age. The means of control cases for systolic and diastolic blood pressure (SBP and DBP), body mass index (BMI; weight divided by height squared), and age are presented. On the average, the disease group has higher SBP, higher DBP, lower BMI, and is younger than the control group, though none of these are significant.

The time-dependent variables SBP, DBP, and BMI were examined by means of blocks of time-series transformations. The first block, seeking constant trend, considers the 3-year average, the individual years, and an average weighting the more recent years more heavily. The second block seeks a linear trend and considers pairwise increments between the second and third years, the first and third years, and the first and second years, plus the difference between the average of the first 2 years and the last. The final transformation considers quadratic trend. Age at last checkup is automatically included, and t-tests are used to find the best separating time-series transformations of SBP, DBP, and BMI. These are most recent SBP, DBP from 2 previous years, and most recent BMI. Since the option of six time-dependent transformations was selected, the increment between the two most recent SBP and the quadratic SBP were selected by t-tests to be included in the model.

Linear discriminant analysis is used to compute the initial estimate of $\underset{\sim}{\beta}$. Three iterations of the Newton-Raphson procedure are required to find the maximum likelihood estimates of $\underset{\sim}{\beta}$ to an accuracy of 10^{-6}. Inspection of the final estimate of $\underset{\sim}{\beta}$ and the initial estimate given by discriminant analysis,

plus the rapid convergence, suggests that discriminant analysis provides a reasonable initial value for $\underset{\sim}{\beta}$.

The matrix GINV is the sample estimate for the Fisher information matrix $J(\hat{\underset{\sim}{\beta}})$. Using this result we may table the standardized z-score $(\hat{\beta}_i[(\frac{1}{n_1}+\frac{1}{n_2})J_{ii}(\hat{\underset{\sim}{\beta}})]^{-1/2})$ for each β_i component of $\underset{\sim}{\beta}$.

I	BETA (I)	VARIANCE	Z-SCORE
1	-.12875809	.1522	- .9267
2	.04446091	.0172	.9520
3	.02388326	.0238	.4347
4	-.18952202	.1931	-1.2109
5	-.01121009	.0568	- .0132
6	.01699793	.0146	.3951

We conclude that the disease group is composed of men first of lower recent BMI, then of higher SBP while being younger, then of higher DBP at the first examination with a convex quadratic trend, and inconsequentially of recent decrease in SBP, all of this relative to the control group. Therefore the data indicates that young slender men with high SBP, and to some extent with a history of high DBP and a drop, then gain, in SBP are at risk.

The summary table of the estimates for λ_0 and $\underset{\sim}{\beta}$ is based on the expected number of failures over one year for a sample of 561. Earlier we indicated that 0.75 was a reasonable estimate of expected failures for this sample; however, the various values only change the scale of the probability of an event and not the relative positions. Therefore any reasonable expected number of failures may be used. The model is checked by revising the sample to compute a frequency table for the estimates of an event during the next

year. We would expect that the estimated probabilities will be higher than average for the disease group. Each individual's probability and transformed data values are printed. The I column provides an index from 1 to 561, with the 553 control cases first, indicated by a 0 in the EVENT column, and the eight disease cases, with a 1 under EVENT, last. The probability of an event during the year between checkups is computed by

$$P[T_i \leq 1 \mid \underset{\sim}{z}_i(t_i)] = 1 - \exp\{-\hat{\lambda}_0 e^{\hat{\underset{\sim}{\beta}}^T \underset{\sim}{y}_i}\}.$$

The probability and its frequency class are found in the columns headed S and IND, respectively. The time after the most recent checkup to failure or censoring is found in the TAU column; if the subject has no event, that is, EVENT is 0, then TAU is 1.00. Otherwise TAU is .04 + (.08) (number of months between last checkup and admission to hospital with acute myocardial infarction). Finally two frequency tables are constructed, one with constant class size 0.0010 and the other with varying class size, to allow a more detailed examination of the empirical distributions for both groups.

The sample run does not allow one to conclude that the data used satisfactorily separate the control and disease samples. However, some success may be claimed if the sample reuse procedure can be believed. We normalized the rate of events at .75/561 or 1.33/1000 failures on the average over a year. Suppose we consider the cases with probability estimates greater than .0014, a convenient class boundary close to the average probability. Then 184/553 = 33.3% of the control group exceeded this critical value while 7/8 = 87.5% of the disease group exceeded it. It would be of interest to see whether or not the control subjects of highest risk have been admitted for

acute myocardial infarction in the past year. However, the individual risk of heart attack is sufficiently small that the 212 subjects of highest risk would need to be considered before the chance of an event during this past year exceeds 0.50.

Our final conclusion is that it is feasible to use a subject's medical history to estimate his probability of an event. Problems with convergence of the algorithm are overcome by establishing a good initial estimate, using an effective procedure, and limiting the data to a subsample of the large control population. An attempt to establish the full worth of the technique must await a sample with a reasonably large number of cases.

REFERENCES

1. Box, G. E. P., W. G. Hunter, and J. S. Hunter. Statistics for experimenters. New York: John Wiley and Sons, 1978.

2. Breslow, N. E. Analysis of survival data under the proportional hazards model. Int Stat Rev 43(#1):45-58 (1975).

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

4. Efron, B. Bootstrap methods. Another look at the jackknife. Ann Stat 7:1-26 (1979).

5. Frank, J. How does more information promote correct diagnosis? Presented at the joint statistical meetings of the American Statistical Association, Biometric Society, and the Institute of Mathematical Statistics, Washington, D.C., Aug. 1979.

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

7. Lachenbruch, P. A. Discriminant analysis. New York: Hafner Press, 1975.

8. Peduzzi, P. N., T. R. Holford, and R. J. Hardy. Regression methods in life table analysis with time-dependent covariates. Preprint (1978).

9. Prentice, R. L., and J. D. Kalbfleisch. Hazard Rate Models with Covariates. Biometrics 35:25-39 (1979).

10. Press, S. J., and S. Wilson. Choosing between logistic regression and discriminant analysis. J Am Stat Assoc 73:699-705 (1978).

11. Shea, G. Statistical diagnosis and tests of factor hypotheses. J Am Stat Assoc 73:346-350 (1978).

12. Taulbee, J. D. A general model for hazard rate with covariables. Preprint (1978).