ELSEVIER SCIENTIFIC PUBLISHERS IRELAND

Computer Methods and Programs in Biomedicine 40 (1993) 279–293

computer methods and programs in biomedicine

MULCOX2: a general computer program for the Cox regression analysis of multivariate failure time data

D.Y. Lin

Department of Biostatistics, SC-32, University of Washington, Seattle, Washington 98195, USA

(Received 8 April 1993; accepted 1 June 1993)

Abstract

Multivariate failure time data is commonly encountered in biomedicine, because each study subject may experience multiple events or because there exists clustering of subjects such that failure times within the same cluster are correlated. MULCOX2 implements a general statistical methodology for analyzing such data. This approach formulates the marginal distributions of multivariate failure times by Cox proportional hazards models without specifying the nature of dependence among related failure times. The baseline hazard functions for the marginal models may be identical or different. A variety of statistical inference can be made regarding the effects of (possibly time-dependent) covariates on the failure rates. Although designed primarily for the marginal approach, MULCOX2 is general enough to implement several alternative methods. The program runs on any computer with a FORTRAN compiler. The running time is minimal. Two illustrative examples are provided.

Key words: Censoring; Clustering; Correlated survival times; FORTRAN; Matched pair; Multiple endpoints; Proportional hazards; Recurrence; Survival analysis

1. Introduction

In 1990 I published in this journal a computer program called MULCOX [3], which implements the methods of Wei, Lin and Weissfeld [6] for the Cox regression analysis of covariate effects on failure rates when each study subject may experience multiple events or failures. Since that publication I have received a large number of requests for the program. The great demand for such a program is apparently due to the common occurrence of multiple-events data in biomedical investigations. Examples of such multivariate failure time data include times to development of physical symptoms or diseases in several body systems and the time sequence of tumor recurrences, infection episodes or asthmatic attacks in individual patients.

There are several limitations with the MULCOX program. First of all, it does not accommodate time-dependent covariates. Allowing covariates to vary over time not only enables one to study time-varying risk factors but also provides a simple way of adjusting for non-proportional hazards. Time-dependent covariates are particularly useful in the

0169-2607/93/$06.00
SSDI 0169-2607(93)01395-I

setting of multivariate failure times, because one is often interested in assessing, say, how the prior history of infections influences the risk for subsequent infections. The main 'complaint' from my correspondents on the MULCOX program has been its lack of the time-dependent covariate capability. A second serious limitation of MULCOX is that it does not deal with another kind of multivariate failure time data, namely the clustered data. Clustered failure time data arises when there is natural or artificial clustering of subjects that induces dependence among failure times of the same cluster. Examples of such data also abound, including times to blindness in the left and right eyes, times to tumor appearance in a litter-matched carcinogenicity experiment and ages at onset of a genetic disease among family members.

To provide a more versatile tool for analyzing multivariate failure time data, I have developed MULCOX2, which is a much-improved revision of MULCOX. This new program allows arbitrary patterns of time-dependent covariates. It handles not only the multiple events data but also clustered failure-time data. Furthermore, a subject is allowed to be at risk for failure in arbitrary time intervals, which is useful for dealing with non-standard situations like delayed study entries as well as for implementing certain methods on recurrence data.

The procedures implemented by MULCOX2 were described at great length in a companion statistical paper [4]. The main methodology is the marginal hazard approach due to Wei, Lin and Weissfeld [6] and Lee, Wei and Amato [2], which formulates the marginal distributions of multivariate failure times with proportional hazards models while leaving the nature of dependence for related failure times completely unspecified. The familiar partial-likelihood methods were modified to properly account for the dependence. MULCOX2 can also be used to implement several alternative approaches, including those of Andersen and Gill [1] and Prentice, William and Peterson [5] for analyzing recurrent events.

In the next section I will present the statistical methods and computational details for MULCOX2. The computer program itself will be described in Section 3. Two biomedical examples will be provided in Section 4 to illustrate the main features of the program.

2. Methods

Suppose that there are n units and that each unit can potentially experience K types of failure. The unit corresponds to the subject in the case of multiple-events data and to the cluster for clustered data. There is generally a clear distinction of different failure types in the former case, whereas the ordering is often arbitrary in the latter. If there are unequal numbers of members among the clusters, we let K be the size of the largest cluster.

Let X_{ik} be the observation time for the ith unit with respect to the kth type of failure, and let Δ_{ik} indicate, by the values 1 vs. 0, whether X_{ik} is a failure or censored observation. Also, let $Z_{ik}(t) = \{Z_{1ik}(t),..., Z_{pik}(t)\}'$ denote a $p \times 1$ vector of possibly time-dependent covariates for the ith unit with respect to the kth type of failure. If X_{ik} or Z_{ik} is missing, we set $X_{ik} = 0$ and $\Delta_{ik} = 0$. Naturally such cases will have no contributions to the statistics to be calculated.

We formulate the marginal distribution for each type of failure with a proportional hazards model. Depending on whether the baseline hazard functions are identical or are different among the K types of failures, the hazard function for the ith unit with respect to the kth type of failure takes the form

$$\lambda_k(t; Z_{ik}) = \lambda_0(t) e^{\beta' Z_{ik}(t)} \tag{2.1}$$

or

$$\lambda_k(t; Z_{ik}) = \lambda_{0k}(t) e^{\beta' Z_{ik}(t)} \tag{2.2}$$

where $\lambda_0(t)$ and $\lambda_{0k}(t)$ $(k = 1,...,K)$ are unspecified baseline hazard functions, and $\beta = (\beta_1..., \beta_p)'$ is a $p \times 1$ vector of unknown regression parameters. In the case of multiple-events data it is typically necessary to allow $\lambda_{0k}(t)$ $(k = 1,...,K)$ to be different, whereas for clustered data it is generally

sufficient to assume a common baseline hazard function. In both models 2.1 and 2.2 we let β be the same among the marginal models. This structure can always be achieved by introducing appropriate type-specific covariates.

Let $Y_{ik}(t)$ indicate, by the values 1 vs. 0, whether or not the ith unit is at risk (i.e. under observation) for the kth type of failure at time t. In most applications $Y_{ik}(t)$ takes the value 1 in the time interval $(0, X_{ik}]$ and takes the value 0 after X_{ki}. It is convenient to introduce the notation

$$S_k^{(0)}(\beta,t) = \sum_{i=1}^{n} Y_{ik}(t)e^{\beta' Z_{ik}(t)},$$

$$\bar{S}^{(0)}(\beta,t) = \sum_{k=1}^{K} S_k^{(0)}(\beta,t),$$

$$S_k^{(1)}(\beta,t) = \sum_{i=1}^{n} Y_{ik}(t)e^{\beta' Z_{ik}(t)} Z_{ik}(t),$$

$$\bar{S}^{(1)}(\beta,t) = \sum_{k=1}^{K} S_k^{(1)}(\beta,t),$$

$$S_k^{(2)}(\beta,t) = \sum_{i=1}^{n} Y_{ik}(t)e^{\beta' Z_{ik}(t)} Z_{ik}(t)Z_{ik}(t)',$$

$$\bar{S}^{(2)}(\beta,t) = \sum_{k=1}^{K} S_k^{(2)}(\beta,t).$$

The estimator $\hat{\beta}$ is obtained by solving $\{U(\beta) = 0\}$, where

$$U(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K} \Delta_{ik}\left\{Z_{ik}(X_{ik}) - \frac{\bar{S}^{(1)}(\beta,X_{ik})}{\bar{S}^{(0)}(\beta,X_{ik})}\right\}$$

under model (2.1) and

$$U(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K} \Delta_{ik}\left\{Z_{ik}(X_{ik}) - \frac{S_k^{(1)}(\beta,X_{ik})}{S_k^{(0)}(\beta,X_{ik})}\right\}$$

under model (2.2). The derivative matrix of $U(\beta)$ with respect to β is

$$A(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K} \Delta_{ik}\left\{\frac{\bar{S}^{(2)}(\beta,X_{ik})}{\bar{S}^{(0)}(\beta,X_{ik})} - \frac{\bar{S}^{(1)}(\beta,X_{ik})\bar{S}^{(1)}(\beta,X_{ik})'}{\bar{S}^{(0)}(\beta,X_{ik})^2}\right\}$$

under model (2.1) and

$$A(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K} \Delta_{ik}\left\{\frac{S_k^{(2)}(\beta,X_{ik})}{S_k^{(0)}(\beta,X_{ik})} - \frac{S_k^{(1)}(\beta,X_{ik})S_k^{(1)}(\beta,X_{ik})'}{S_k^{(0)}(\beta,X_{ik})^2}\right\}$$

under model (2.2). Because $A(\beta)$ is positive definite, the Newton-Raphson algorithm is used to obtain the unique root to $\{U(\beta) = 0\}$.

The statistic $U(\beta)$ is approximately p-dimensional normal with mean 0 and with covariance matrix $B(\hat{\beta}) = \Sigma_{i=1}^{n}\ \Sigma_{k=1}^{K}\ \Sigma_{l=1}^{K}\ W_{ik}(\hat{\beta})\ W_{il}(\hat{\beta})'$, where

$$W_{ik}(\beta) = \Delta_{ik}\left\{Z_{ik}(X_{ik}) - \frac{\bar{S}^{(1)}(\beta,X_{ik})}{\bar{S}^{(0)}(\beta,X_{ik})}\right\} - \sum_{j=1}^{n}\sum_{l=1}^{K} \frac{\Delta_{jl}Y_{ik}(X_{jl})e^{\beta' Z_{ik}(X_{jl})}}{\bar{S}^{(0)}(\beta,X_{jl})}\left\{Z_{ik}(X_{jl}) - \frac{\bar{S}^{(1)}(\beta,X_{jl})}{\bar{S}^{(0)}(\beta,X_{jl})}\right\}$$

and

$$W_{ik}(\beta) = \Delta_{ik}\left\{Z_{ik}(X_{ik}) - \frac{S_k^{(1)}(\beta,X_{ik})}{S_k^{(0)}(\beta,X_{ik})}\right\}$$

$$-\sum_{j=1}^{n}\frac{\Delta_{jk}\,Y_{ik}(X_{jk})\,e^{\beta' Z_{ik}(X_{jk})}}{S_k^{(0)}(\beta,X_{jk})}$$

$$\left\{Z_{ik}(X_{jk})-\frac{S_k^{(1)}(\beta,X_{jk})}{S_k^{(0)}(\beta,X_{jk})}\right\}$$

under models (2.1) and (2.2), respectively. Furthermore, the estimator $\hat{\beta}$ is approximately p-dimensional normal with mean β and with covariance matrix $D(\hat{\beta}) = A^{-1}(\hat{\beta})B(\hat{\beta})A^{-1}(\hat{\beta})$. We refer to $A^{-1}(\hat{\beta})$ and $D(\hat{\beta})$ as, respectively, the naive and robust variance-covariance estimators for $\hat{\beta}$ and call $U'(0)A^{-1}(0)U(0)$ and $U'(0)B^{-1}(0)U(0)$ the naive and robust log rank statistics, respectively. The naive statistics are invalid unless failure times within the same unit are independent.

Table 1
The input of control parameters

Parameter	Type
Problem title	character
File name of data input	character
File name of program output	character
n	integer
K	integer
M	integer
p	integer
Name of the 1st covariate	character
⋮	⋮
Name of the pth covariate	character
Format of data input	character
Indicator for model (2.1) vs. model (2.2)	integer
Number of multivariate hypotheses	integer
Row dimension of L for the first hypothesis	integer
⋮	⋮
Row dimension of L for the last hypothesis	integer
Matrix L for the first multivariate hypothesis	real
⋮	⋮
Matrix L for the last multivariate hypothesis	real

The robust log-rank statistic $U'(0)B^{-1}(0)U(0)$ can be used to test the global null hypothesis H_0: $\beta = 0$. Inference about individual covariate effects is drawn by referring the standardized parameter estimates to the standard normal distribution. The multivariate general linear hypothesis involving several components of β is expressed as H_0: $L\beta = 0$, where L is a $r \times p$ matrix of constants. The robust Wald statistic for testing H_0 is $(L\hat{\beta})'\{LD(\hat{\beta})L'\}^{-1}(L\hat{\beta})$, which has an approximate χ^2distribution with r degrees of freedom.

```
 1   1   0  47   0   0   1
 1   2   0  42   0   0   0
 1   3   0  21   0   0   1
 1   4   0  20   0   0   1
 1   5   0  19   0   0   1
 2   1   0  29   1   0   1
 2   2   0   6   0   0   1
 2   3   0  47   1   0   0
 2   4   0  33   0   0   0
 3   1   0  63   0   0   1
 3   2   0  57   0   0   0
 3   3   0  52   0   0   1
 3   4   0  48   0   0   1
 3   5   0  47   0   0   1
 3   6   0  45   0   0   1
 3   7   0  30   1   0   1
 3   8   0  53   0   0   0
 3   9   0  50   0   0   0
 3  10   0  40   0   0   0
 3  11   0  37   0   0   0
                .
                .
                .
91   1   0  55   0   1   1
91   2   0  47   1   1   0
91   3   0  28   0   1   1
91   4   0  27   0   1   0
92   1   0  56   0   0   1
92   2   0  55   0   0   0
92   3   0  28   0   0   1
92   4   0  26   0   0   1
92   5   0  22   0   0   1
92   6   0  20   0   0   1
92   7   0  29   0   0   0
92   8   0  24   0   0   0
93   1   0  69   0   0   1
93   2   0  18   0   0   1
93   3   0  16   0   0   1
93   4   0  15   0   0   0
93   5   0  52   1   0   0
```

Fig. 1. Data input for the schizophrenia study: *spn. dat.*

As mentioned earlier, if the ith unit has incomplete information on the kth type of failure, we set $X_{ik} = \Delta_{ik} = 0$ (and set Z_{ik} to some arbitrary values) in our formulas. In the data file, however, it is not necessary to create such dummy records for the ith unit if this unit has complete information on the first n_i ($\leq K$) types of failure and contains missing values on the last $(K - n_i)$ types. In that scenario one may simply exclude the last $(K - n_i)$ failure types from the data file. Thus when the ordering of failure types is arbitrary, as is generally true for clustered data, it is only necessary to input the complete cases.

To allow arbitrary patterns of time-dependent covariates and at-risk indicators, we represent all the information for the ith unit with respect to the kth type of failure, i.e.

$$\{Y_{ik}(t),\ X_{ik},\ \Delta_{ik},\ Z_{1ik}(t), \ldots, Z_{pik}(t)\} \quad (t > 0),$$

by a series of records in the data file:

$$(s_{ik,1},\ t_{ik,1},\ \Delta_{ik,1},\ Z_{1ik,1}, \ldots, Z_{pik,1}), \ldots,$$

$$(s_{ik,n_{ik}},\ t_{ik,n_{ik}},\ \Delta_{ik,n_{ik}},\ Z_{1ik,n_{ik}}, \ldots, Z_{pik,n_{ik}}).$$

For $l = 1, \ldots, n_{ik}$, the ith unit is at risk for the kth type of failure in the (left-opened and right-closed) time interval $(s_{ik,l}, t_{ik,l}]$, and $(Z_{1ik,l},\ldots,Z_{pik,l})$ is the covariate vector over that interval. We define $\Delta_{ik,n_{ik}} = \Delta_{ik}$ and set $\Delta_{ik,l}$ to any arbitrary values, say 0, for $l = 1,\ldots, n_{ik} - 1$ (if $n_{ik} > 1$). The intervals $(s_{ik,l}, t_{ik,l}]$ ($l = 1,\ldots,n_{ik}$) are constructed by first identifying all the time periods during which $Y_{ik}(t) = 1$ and then, if necessary, subdividing them into finer intervals within each of which the covariate vector $Z_{ik}(t)$ is constant. If $Z_{ik}(t)$ is continuous in t (for computational purposes) it suffices to regard $Z_{ik}(t)$ as being constant between

```
talon% MULCOX2
  PLEASE ENTER THE TITLE OF THIS RUN
The Schizophrenia Study
  ENTER THE NAME OF DATA FILE
spn.dat
  ENTER THE NAME OF OUTPUT FILE
spn.out
  ENTER THE NUMBER OF UNITS
93
  ENTER THE MAXIMUM NUMBER OF FAILURE TYPES
12
  ENTER THE MAXIMUM NUMBER OF TIME INTERVALS
1
  ENTER THE NUMBER OF COVARIATES IN THE MODEL
2
  ENTER THE NAME OF COVARIATE  1
Age
  ENTER THE NAME OF COVARIATE  2
Gender
  ENTER THE FORMAT OF DATA FILE
free
  ENTER THE INDICATOR (1 VS. 2) FOR MODEL1 VS. MODEL2
1
  ENTER THE NUMBER OF MULTIVARIATE HYPOTHESES
0
  PLEASE WAIT !
talon%
```

Fig. 2. Control parameters input for the schizophrenia study.

two adjacent distinct ordered failure time points, the covariate vector over the interval being $Z_{ik}(t)$ evaluated at the second failure time point. In most applications, covariates are fixed and subjects are at risk from $t = 0$ until they fail or are censored. Then only a single record $(0, X_{ik}, \Delta_{ik}, Z_{1ik}, \ldots, Z_{pik})$ is needed for the ith unit with respect to the kth failure type. Because of possibly multiple records, it is necessary to indicate the unit and failure type for each record in the data file.

```
* * * * * * * * * * * * * * * * * * * * * * * * * * * * *
*                                                         *
*            Cox Regression Analysis of                   *
*                                                         *
*          Multivariate Failure Time Data                 *
*                                                         *
* * * * * * * * * * * * * * * * * * * * * * * * * * * * *


PROBLEM TITLE IS: The Schizophrenia Study
DATA FILE IS: spn.dat
OUTPUT FILE IS: spn.out


ASSUMING COMMON BASELINE HAZARD FUNCTION


Number of Units =   93
Number of Failure Types =   12

Numbers of Uncensored Failure Times

    Failure type:    1   2   3   4   5   6   7   8   9  10  11  12
    Number:          3   8   4   2   7   2   3   1   0   1   0   0


Naive  Log Rank Stat =       10.55303  P-value =   0.00511
Robust Log Rank Stat =        7.57447  P-value =   0.02266

PARAMETER          ESTIMATE       NAIVE S.E.     EST/N.S.E.     ROBUST S.E.    EST/R.S.E.
---------          --------       ----------     ----------     -----------    ----------
Age                  -0.23832        0.48863       -0.48773        0.51722       -0.46077
Gender               -1.24401        0.41106       -3.02633        0.40832       -3.04661

NAIVE COVARIANCE MATRIX FOR BETA
--------------------------------

  0.239E+00  0.432E-02
  0.432E-02  0.169E+00

ROBUST COVARIANCE MATRIX FOR BETA
---------------------------------

  0.268E+00 -0.861E-02
 -0.861E-02  0.167E+00
```

Fig. 3. Output for the schizophrenia study: *spn.out*.

3. Computer program

3.1. General description

The MULCOX2 computer program was written in standard FORTRAN-77 with double arithmetic precision. The source program consists of 914 lines of code and requires 27 kb of disk storage. No external subroutines or functions are used. The program can run on any computer with a FORTRAN compiler.

The amount of CPU time used by MULCOX2 depends on the computer installation and the size of data. In general, the time consumption is minimal on a mainframe, even for large data-sets.

The program allows arbitrary values of n, p and K. The number of time intervals for each combination of i and k, i.e. n_{ik}, is also arbitrary. Let M denote the maximum of the n_{ik}. The matrices of data and computational results are stored in a single one-dimensional array, A. (This array is partitioned according to the values of n, p, K and M.) The dimension of A may be modified by the user if necessary.

3.2. Input

The program MULCOX2 requires two separate groups of input: the data input and the control parameters input. The records in the data file provide the values for the following variables in the order shown:

$$(ID1_i,\ ID2_k,\ s_{ik,l},\ t_{ik,l},\ \Delta_{ik,l},\ Z_{1ik,l},\ \ldots,\ Z_{pik,l})$$

$(i = 1, \ldots, n;\ k = 1, \ldots, n_i;\ l = 1, \ldots, n_{ik})$, where ID1 and ID2 identify the unit and failure type. (Note that we allow the possibility of n_i being smaller than K to accommodate the situation where the last $(K - n_i)$ failure types are missing on the ith unit.) It is required that the records for the same unit be placed consecutively in a non-decreasing order of $ID2_k$.

The control parameters are described in Table 1. For simplicity, these parameters are read from the keyboard on execution of MULCOX2. The source codes can be easily modified to input the control parameters from a batch file.

When specifying the format of the data file, the

```
174054  1   293  0   1
174077  1   255  0   0
174109  1   213  0   0
174111  1   203  0   0
204001  1   219  1   1
204001  2   373  1   1
204001  3   414  0   1
204002  1     8  1   0
204002  2    26  1   0
204002  3   152  1   0
        .
        .
        .
336025  1   146  1   0
336025  2   316  0   0
336026  1   316  0   0
336027  1   315  0   1
```

Fig. 4. Original data file for the CGD study.

```
174054  1   0  293  0   1 0 0
174054  2   0  293  0   0 1 0
174054  3   0  293  0   0 0 1
174077  1   0  255  0   0 0 0
174077  2   0  255  0   0 0 0
174077  3   0  255  0   0 0 0
174109  1   0  213  0   0 0 0
174109  2   0  213  0   0 0 0
174109  3   0  213  0   0 0 0
174111  1   0  203  0   0 0 0
174111  2   0  203  0   0 0 0
174111  3   0  203  0   0 0 0
204001  1   0  219  1   1 0 0
204001  2   0  373  1   0 1 0
204001  3   0  414  0   0 0 1
204002  1   0    8  1   0 0 0
204002  2   0   26  1   0 0 0
204002  3   0  152  1   0 0 0
            .
            .
            .
336025  1   0  146  1   0 0 0
336025  2   0  316  0   0 0 0
336025  3   0  316  0   0 0 0
336026  1   0  316  0   0 0 0
336026  2   0  316  0   0 0 0
336026  3   0  316  0   0 0 0
336027  1   0  315  0   1 0 0
336027  2   0  315  0   0 1 0
336027  3   0  315  0   0 0 1
```

Fig. 5. Data input for the marginal analysis of the CGD study: *cgd.d0*.

user may enter *FREE* or *free* if the required data items are separated by spaces or commas; otherwise a FORTRAN format expression such as (I5, I3, 2F5.1, F2.0, 5X, 3F6.2,4X, F3.0) is required. Note that the first two variables are integer-valued and the remaining ones are real.

3.3. Output

Computational results are written to the output file specified by the user. The output contains the estimators and test statistics described in Section 2. The results are self-explanatory.

4. Applications

In this section we illustrate the use of MULCOX2 with data taken from two biomedical studies. The studies and analyses were described in detail by Lin [4]. Here we show how the results reported there were obtained from MULCOX2.

4.1. The Schizophrenia Study

In a genetic epidemiological study of schizophrenia conducted by Ann E. Pulver of Johns Hopkins University, 487 first-degree relatives of 93 female schizophrenic probands were enrolled. The number of the relatives in a family ranges from one to twelve. Thus $n = 93$ and $K = 12$. The main question is whether the risk of effective illness of the relatives is associated with the age at onset of schizophrenia of probands. The gender of the relative was expected to be predictive. We consider model (2.1) with $Z_{ik} = (Z_{1ik}, Z_{2ik})'$ $(i = 1, \ldots, 93;\ k = 1, \ldots, 12)$, where

```
talon% MULCOX2
  PLEASE ENTER THE TITLE OF THIS RUN
CGD Study, Marginal Approach for Separate Treatment Effects
  ENTER THE NAME OF DATA FILE
cgd.d0
  ENTER THE NAME OF OUTPUT FILE
cgd.o0
  ENTER THE NUMBER OF UNITS
128
  ENTER THE MAXIMUM NUMBER OF FAILURE TYPES
3
  ENTER THE MAXIMUM NUMBER OF TIME INTERVALS
1
  ENTER THE NUMBER OF COVARIATES IN THE MODEL
3
  ENTER THE NAME OF COVARIATE  1
trt_1
  ENTER THE NAME OF COVARIATE  2
trt_2
  ENTER THE NAME OF COVARIATE  3
trt_3
  ENTER THE FORMAT OF DATA FILE
free
  ENTER THE INDICATOR (1 VS. 2) FOR MODEL1 VS. MODEL2
2
  ENTER THE NUMBER OF MULTIVARIATE HYPOTHESES
1
  ENTER THE DIMENSION OF MULTIVARIATE HYPOTHESIS  1
2
  ENTER THE CONTRAST MATRIX FOR MULTIVARIATE HYPOTHESIS  1
1 -1  0
1  0 -1
  PLEASE WAIT !
talon%
```

Fig. 6. Control parameters input for the marginal analysis of the CGD study.

$$Z_{1ik} = \begin{cases} 1 \text{ if the age at onset of the } i\text{th proband} \le 16, \\ 0 \text{ otherwise} \end{cases} \qquad Z_{2ik} = \begin{cases} 1 \text{ if the } k\text{th relative of the } i\text{th proband is male,} \\ 0 \text{ if the } k\text{th relative of the } i\text{th proband is female.} \end{cases}$$

and

```
* * * * * * * * * * * * * * * * * * * * * * * * * * * * *
*                                                       *
*            Cox Regression Analysis of                 *
*                                                       *
*          Multivariate Failure Time Data               *
*                                                       *
* * * * * * * * * * * * * * * * * * * * * * * * * * * * *

PROBLEM TITLE IS: CGD Study, Marginal Approach for Separate Treatment Effects
DATA FILE IS: cgd.d0
OUTPUT FILE IS: cgd.o0

ASSUMING NON-COMMON BASELINE HAZARD FUNCTIONS

Number of Units =  128
Number of Failure Types =    3

Numbers of Uncensored Failure Times

    Failure type:    1   2   3
    Number:         44  17   8

Naive  Log Rank Stat =       22.48218  P-value =   0.00005
Robust Log Rank Stat =       12.27974  P-value =   0.00648

PARAMETER          ESTIMATE       NAIVE S.E.     EST/N.S.E.     ROBUST S.E.    EST/R.S.E.
---------          --------       ---------      ----------     -----------    ----------
trt_1                -1.09398        0.33479       -3.26768        0.33506       -3.26500
trt_2                -1.23078        0.55237       -2.22819        0.53814       -2.28710
trt_3                -2.06287        1.06937       -1.92905        1.01943       -2.02355

NAIVE COVARIANCE MATRIX FOR BETA
--------------------------------

  0.112E+00  0.000E+00  0.000E+00
  0.000E+00  0.305E+00  0.000E+00
  0.000E+00  0.000E+00  0.114E+01

ROBUST COVARIANCE MATRIX FOR BETA
---------------------------------

  0.112E+00  0.958E-01  0.977E-01
  0.958E-01  0.290E+00  0.288E+00
  0.977E-01  0.288E+00  0.104E+01


MULTIVARIATE HYPOTHESIS 1
-------------------------

The contrast matrix is as follows:

   1.  -1.   0.
   1.   0.  -1.

Wald statistic =   1.01424
Degrees of freedom =  2
P-value =   0.60223
```

Fig. 7. Output for the marginal analysis of the CGD study: *cgd.o0*.

```
174054  1    0  293  0    1
174077  1    0  255  0    0
174109  1    0  213  0    0
174111  1    0  203  0    0
204001  1    0  219  1    1
204001  2  219  373  1    1
204001  3  373  414  0    1
204002  1    0    8  1    0
204002  2    8   26  1    0
204002  3   26  152  1    0
          .
            .
              .
336025  1    0  146  1    0
336025  2  146  316  0    0
336026  1    0  316  0    0
336027  1    0  315  0    1
```

Fig. 8. Data input for the Andersen-Gill Markov model and Prentice et al. total time model on the CGD study: *cgd. dl.*

The data file is shown in Fig. 1. (The middle portion of the file was suppressed.) There is (only) one record for each of the 487 relatives, which consists of family identification number, (arbitrary) identification number for the relative, beginning of follow-up (0), observation time for the relative, failure indicator for the relative (1 for illness, 0 for censored), indicator for the age at onset of the proband and indicator for the gender of the relative. The input of the control parameters and the output from the computer run are displayed in Fig. 2 and Fig. 3, respectively.

4.2. The CGD Study

A placebo controlled trial was carried out to study the ability of gamma interferon to reduce the rate of infections in patients with chronic granulomatous disease (CGD). By the end of the trial, 30 of 65 placebo patients and 14 of 63 patients on gamma interferon had experienced one or more infections. The data from this study is presented in Fig. 4. The five columns correspond to patient's identification number, infection number, observation time, failure indicator (1 for infection, 0 for censored) and treatment indicator (1 for gamma interferon, 0 for placebo). There is one record for each infection number; the last record for each patient corresponds to the final blinded study visit by the patient.

We will confine our attention to the first three infections only. Thus we have $n = 128$ and $K = 3$.

```
talon% MULCOX2
  PLEASE ENTER THE TITLE OF THIS RUN
CGD Study, AG Model with Markov Assumption
  ENTER THE NAME OF DATA FILE
cgd.dl
  ENTER THE NAME OF OUTPUT FILE
cgd.ol
  ENTER THE NUMBER OF UNITS
128
  ENTER THE MAXIMUM NUMBER OF FAILURE TYPES
3
  ENTER THE MAXIMUM NUMBER OF TIME INTERVALS
1
  ENTER THE NUMBER OF COVARIATES IN THE MODEL
1
  ENTER THE NAME OF COVARIATE  1
Treatment
  ENTER THE FORMAT OF DATA FILE
free
  ENTER THE INDICATOR (1 VS. 2) FOR MODEL1 VS. MODEL2
1
  ENTER THE NUMBER OF MULTIVARIATE HYPOTHESES
0
  PLEASE WAIT !
talon%
```

Fig. 9. Control parameters input for the Andersen-Gill Markov model on the CGD study.

For the marginal approach described in Section 2, a patient is considered to be at risk for each of the three infections on entering the study. If a particular infection time is censored, then all the subsequent infection times on the same patient are treated as being censored (not missing!) at the same time. Thus we should have one record on each of the three infections for every patient. The left end-points of all the at-risk intervals are 0.

To estimate separate treatment effects for the three types of infection we fit model (2.2) with $Z_{i1} = (R_i, 0, 0)'$, $Z_{i2} = (0, R_i, 0)'$ and $Z_{i3} = (0, 0, R_i)'$ ($i = 1,...,128$), where $R_i = 1$ if the ith patient was on gamma interferon and $R_i = 0$ otherwise. Fig. 5 shows the data file constructed from the original one (see Fig. 4) for this analysis. Fig. 6 displays the input of the control parameters. Note that the multivariate hypothesis being tested is

```
* * * * * * * * * * * * * * * * * * * * * * * * * * * * *
*                                                         *
*              Cox Regression Analysis of                 *
*                                                         *
*           Multivariate Failure Time Data                *
*                                                         *
* * * * * * * * * * * * * * * * * * * * * * * * * * * * *


PROBLEM TITLE IS: CGD Study, AG Model with Markov Assumption
DATA FILE IS: cgd.d1
OUTPUT FILE IS: cgd.o1


ASSUMING COMMON BASELINE HAZARD FUNCTION


Number of Units =  128
Number of Failure Types =    3

Numbers of Uncensored Failure Times

    Failure type:    1   2   3
    Number:         44  17   8


Naive  Log Rank Stat =       15.89108  P-value =   0.00007
Robust Log Rank Stat =       10.69262  P-value =   0.00108


PARAMETER          ESTIMATE       NAIVE S.E.     EST/N.S.E.     ROBUST S.E.    EST/R.S.E.
---------          --------       ----------     ----------     -----------    ----------
Treatment            -1.02015        0.26679       -3.82378        0.30665      -3.32674


NAIVE COVARIANCE MATRIX FOR BETA
--------------------------------

  0.712E-01


ROBUST COVARIANCE MATRIX FOR BETA
---------------------------------

  0.940E-01
```

Fig. 10. Output for the Andersen-Gill Markov model on the CGD study: *cgd.o1*.

```
174054  1    0  293  0    1   0
174077  1    0  255  0    0   0
174109  1    0  213  0    0   0
174111  1    0  203  0    0   0
204001  1    0  219  1    1   0
204001  2  219  279  0    1   1
204001  2  279  373  1    1   0
204001  3  373  414  0    1   1
204002  1    0    8  1    0   0
204002  2    8   26  1    0   1
204002  3   26   86  0    0   1
204002  3   86  152  1    0   0
                .
                .
                .
336025  1    0  146  1    0   0
336025  2  146  206  0    0   1
336025  2  206  316  0    0   0
336026  1    0  316  0    0   0
336027  1    0  315  0    1   0
```

Fig. 11. Data input for the Andersen-Gill semi-Markov model on the CGD study: *cgd.d3*.

whether or not the treatment effects are the same on the three infections. The output from this run is given in Fig. 7.

To estimate an overall treatment effect we fit model (2.2) with $Z_{ik} = R_i$ ($i = 1,..., 128$; $k = 1, 2, 3$). The data file and the input of control parameters for this analysis are the same as those shown in Figs. 5 and 6 except that there is now only one single covariate instead of three type-specific covariates. The estimate for the common regression parameter turns out to be -1.215 with robust standard error estimate of 0.353.

As mentioned earlier, MULCOX2 can also be used to implement the methods of Andersen and Gill [1] and Prentice et al. [5]. Under the Andersen-Gill model the risk of a recurrent event for a subject satisfies the usual proportional hazards model, and it is unaffected by any earlier events that occurred to the same subject unless terms that capture such dependence are included explicitly in the model as covariates. Figs. 8–10 display the input and output for fitting the Andersen-Gill model with the treatment indicator

```
talon% MULCOX2
  PLEASE ENTER THE TITLE OF THIS RUN
CGD Study, AG Model with Semi-Markov Assumption
  ENTER THE NAME OF DATA FILE
cgd.d3
  ENTER THE NAME OF OUTPUT FILE
cgd.o3
  ENTER THE NUMBER OF UNITS
128
  ENTER THE MAXIMUM NUMBER OF FAILURE TYPES
3
  ENTER THE MAXIMUM NUMBER OF TIME INTERVALS
2
  ENTER THE NUMBER OF COVARIATES IN THE MODEL
2
  ENTER THE NAME OF COVARIATE  1
Treatment
  ENTER THE NAME OF COVARIATE  2
Inf_hist.
  ENTER THE FORMAT OF DATA FILE
free
  ENTER THE INDICATOR (1 VS. 2) FOR MODEL1 VS. MODEL2
1
  ENTER THE NUMBER OF MULTIVARIATE HYPOTHESES
0
  PLEASE WAIT !
talon%
```

Fig. 12. Control parameters input for the Andersen-Gill semi-Markov model on the CGD study.

```
* * * * * * * * * * * * * * * * * * * * * * * * * * * * *
*                                                       *
*            Cox Regression Analysis of                 *
*                                                       *
*          Multivariate Failure Time Data               *
*                                                       *
* * * * * * * * * * * * * * * * * * * * * * * * * * * * *
```

```
PROBLEM TITLE IS: CGD Study, AG Model with Semi-Markov Assumption
DATA FILE IS: cgd.d3
OUTPUT FILE IS: cgd.o3

ASSUMING COMMON BASELINE HAZARD FUNCTION

Number of Units =  128
Number of Failure Types =    3

Numbers of Uncensored Failure Times

    Failure type:    1   2   3
    Number:         44  17   8

Naive  Log Rank Stat =       22.77996  P-value =   0.00001
Robust Log Rank Stat =       11.33547  P-value =   0.00346
```

PARAMETER	ESTIMATE	NAIVE S.E.	EST/N.S.E.	ROBUST S.E.	EST/R.S.E.
Treatment	-0.94318	0.26932	-3.50209	0.29225	-3.22728
Inf_hist.	0.76374	0.32845	2.32528	0.27552	2.77196

```
NAIVE COVARIANCE MATRIX FOR BETA
--------------------------------

  0.725E-01  0.136E-01
  0.136E-01  0.108E+00


ROBUST COVARIANCE MATRIX FOR BETA
---------------------------------

  0.854E-01  0.631E-02
  0.631E-02  0.759E-01
```

Fig. 13. Output for the Andersen-Gill semi-Markov model on the CGD study: *cgd.o3*.

```
* * * * * * * * * * * * * * * * * * * * * * * * * * * * *
*                                                       *
*            Cox Regression Analysis of                 *
*                                                       *
*          Multivariate Failure Time Data               *
*                                                       *
* * * * * * * * * * * * * * * * * * * * * * * * * * * * *


PROBLEM TITLE IS: CGD Study, PWP Total Time Model with Common Treatment Effect
DATA FILE IS: cgd.d1
OUTPUT FILE IS: cgd.o4


ASSUMING NON-COMMON BASELINE HAZARD FUNCTIONS


Number of Units =  128
Number of Failure Types =    3

Numbers of Uncensored Failure Times

    Failure type:    1   2   3
    Number:         44  17   8


Naive  Log Rank Stat =        9.91524  P-value =   0.00164
Robust Log Rank Stat =       10.88008  P-value =   0.00097


PARAMETER          ESTIMATE       NAIVE S.E.     EST/N.S.E.     ROBUST S.E.    EST/R.S.E.
---------          --------       ----------     ----------     -----------    ----------
Treatment           -0.85939        0.28021       -3.06691        0.29200       -2.94311


NAIVE COVARIANCE MATRIX FOR BETA
--------------------------------

  0.785E-01


ROBUST COVARIANCE MATRIX FOR BETA
---------------------------------

  0.853E-01
```

Fig. 14. Output for the Prentice et al. total time model on the CGD study: *cgd.o4*.

```
174054  1    0  293  0    1
174077  1    0  255  0    0
174109  1    0  213  0    0
174111  1    0  203  0    0
204001  1    0  219  1    1
204001  2    0  154  1    1
204001  3    0   41  0    1
204002  1    0    8  1    0
204002  2    0   18  1    0
204002  3    0  126  1    0
.
.
.
336025  1    0  146  1    0
336025  2    0  170  0    0
336026  1    0  316  0    0
336027  1    0  315  0    1
```

Fig. 15. Data input for the Prentice et al. gap time model on the CGD study.

as the single covariate. The data file for this analysis (see Fig. 8) was created from the original file (see Fig. 4) by inserting a new column (i.e. column 3 in Fig. 8) to indicate the left end-point of the at-risk interval (which equals 0 for $k = 1$, equals the first infection time for $k = 2$ and equals the second infection time for $k = 3$).

As an attempt to account for the influence of prior infections on the risk for a new infection, we add to the preceding model a time-dependent covariate, which indicates by the values 1 vs. 0 whether or not the patient had infections within the past sixty days. As in the previous analysis, we have one record on the first infection for each patient. However, information for a second or third infection needs to be represented by two records if the gap time is more than sixty days, the time-dependent covariate then taking the value 1 in the first time interval and the value 0 in the second interval. The data file for this analysis is shown in Fig. 11. The input of the control parameters and the output are presented in Figs. 12 and 13, respectively.

Prentice et al. [5] proposed two models for recurrent events, the first of which deals with the total time and the second with the gap time (see Lin [4] for more explanations of the Prentice et al. models). To analyze their total time model with the treatment indicator as the single covariate, we use the data file shown in Fig. 8 but specify model (2.2) instead of model (2.1). Then we obtain the results given in Fig. 14. The data file for the Prentice et al. gap time model is shown in Fig. 15. This file differs from the one used for the total time model (see Fig. 8) in that (for the second and third infections) the gap times replace the total times and the left end-points of the at-risk intervals are 0 rather than the times of the previous infections. The estimate for the common treatment effect on the gap times is found to be −0.872 with standard error estimate of 0.279.

5. Availability

The MULCOX2 program can be obtained from the author at no charge. Those interested in obtaining a copy of the program should contact the author by electronic mail at DANYU@BIOSTAT.WASHINGTON.EDU or send him by normal mail a blank diskette.

6. Acknowledgements

The development of MULCOX2 was supported by the National Institutes of Health grants R29 GM-47845 and R01 AI-291968. The author thanks Drs. Kung-Yee Liang and David Harrington for providing the data-sets used here.

7. References

[1] P.K. Andersen and R.D. Gill, Cox's regression model for counting processes: a large sample study, Ann. Statist. 10 (1982) 1100–1120.

[2] E.W. Lee, L.J. Wei and D.A. Amato, Cox-type regression analysis for large numbers of small groups of correlated failure time observations, in Survival Analysis: State of the Art, Ed. J.P. Klein and P.K. Goel, pp. 237–247 (Kluwer Academic Publishers, Dordrecht, 1992).

[3] D.Y. Lin, MULCOX: a computer program for the Cox regression analysis of multiple failure time variables, Comput. Methods Programs Biomed. 32 (1990) 125–135.

[4] D.Y. Lin, Cox regression analysis of multivariate failure time data: the marginal approach, Stat. Med. (submitted).

[5] R.L. Prentice, B.J. Williams and A.V. Peterson, On the regression analysis of multivariate failure time data, Biometrika 68 (1981) 373–379.

[6] L.J. Wei, D.Y. Lin and L. Weissfeld, Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, J. Am. Stat. Assoc. 84 (1989) 1065–1073.