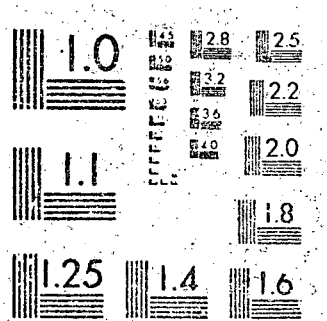
If you have issues viewing or accessing this file contact us at NCJRS.gov.

1 OF 1

AD

A046209




MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD A046209

ACCESSION for
NTIS White Section ☒
DDC Buff Section ☐
UNANNOUNCED ☐
JUSTIFICATION
BY
DISTRIBUTION/AVAILABILITY CODES
Dist. AVAIL. and/or SPECIAL
A

12

SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK

9 TECHNICAL REPORT NO. 333

11 May 1977

12 25p.

14 TR-333

6

# A SURVEY OF COVARIANCE MODELS FOR CENSORED LIFE DATA WITH AN APPLICATION TO RECIDIVISM ANALYSIS

by

10 Russell R. Barton and Bruce W. Turnbull

15 Prepared under contracts

DAAG29-77-C-0003, U.S. Army Research Office - Durham,

and

N00014-75-C-0586, Office of Naval Research

Approved for Public Release; Distribution Unlimited.

DDC
RECEIVED
NOV 9 1977
D

143

409 869

THE FINDINGS IN THIS REPORT ARE NOT TO BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, UNLESS SO DESIGNATED BY OTHER AUTHORIZED DOCUMENTS.

# A SURVEY OF COVARIANCE MODELS FOR CENSORED LIFE DATA WITH AN APPLICATION TO RECIDIVISM ANALYSIS

Russell R. Barton and Bruce W. Turnbull

Cornell University

## ABSTRACT

A survey is given of techniques for covariance analysis of censored life data. Both parametric and nonparametric approaches are reviewed. An application is given to the evaluation of parolee followup data. ~~We examine~~ The effects of covariates, such as age, income, and drug use, on time to rearrest were examined. One of these covariates varies with time. The records of two correctional institutions are compared after adjusting for non-homogeneity of covariate values.

## 1. INTRODUCTION

Regression techniques applied to survival or failure rate data have been the subject of much recent interest in the areas of medical followup and industrial life testing studies. In this paper we review the various models and methods that have been proposed and present an example from a new area of application, namely the study of the recidivism rate of exoffenders released from

*This research supported by DAAG29-77-C-0003, U.S. Army Research Office - Durham and N00014-75-C-0580, Office of Naval Research.

correctional institutions. The data come from an extensive followup study of the post-release behavior of 108 parolees in the State of Connecticut. In Section 2 we examine various parametric models that have been proposed. Section 3 describes the conditional likelihood approach in Cox's "semi-parametric" model, together with some of the difficulties. Alternate models, developed subsequently, are described in Section 4 and in the section following we discuss some general problems in analysis of covariance and how they relate to the treatment of censored life data. Finally we describe the Connecticut recidivism data which is used to illustrate some of the techniques surveyed in the earlier sections.

We will use the following notation: Let $Y$ denote the random variable representing response time (time to failure, death, rearrest, etc.) and let $F(t) = 1 - S(t) = P(Y \leq t)$. If $F$ has a density $f$, the hazard (failure) rate $\lambda(t)$ is defined to be the density at $t$ conditioned on survival to $t$, i.e.

$$\lambda(t) = f(t)/S(t) \tag{1}$$

The cumulative hazard is defined by $\Lambda(t) = \int_0^t \lambda(u)du$, and so by (1) we have the relation $S(t) = \exp(-\Lambda(t))$. Thus the hazard function can be used to characterize any continuous distribution on $(0,\infty)$. A good introduction to survival distributions can be found in Gross and Clark (1975).

We suppose that the sample consists of $n$ items (subjects). Let the response times be denoted by $Y_1, Y_2, \ldots, Y_n$. Because of right censoring, not all $Y_i$ are observed exactly; some are known only to exceed some censoring value which may vary from item to item. This censoring may come about because the data are analyzed while some of the subjects are still "at risk", or because losses to followup occur during the course of the study. All covariance techniques assume that the censoring mechanism operates independently of response time, an assumption not always valid in practice.

We assume that associated with the i'th item $(1 \leq i \leq n)$ are $p$ $(\geq 1)$ covariates represented by the vector $z_i = (z_{1i}, z_{2i}, \ldots, z_{pi})$. All the covariance models discussed below describe the covariance

relationship as:

$$\lambda(t,z) = g(t,z\beta,\gamma) \tag{2}$$

Here $\beta^T = (\beta_1,\beta_2,\ldots,\beta_p)$ is a p-vector of unknown regression coefficients and $\gamma$ represents a vector of other (possibly nuisance) parameters. It is possible that some or all components of $z$ vary with time. The statistical problem at hand is the estimation of $\beta$ in (2) and tests of hypotheses concerning $\beta$, although inference concerning $\gamma$ is also sometimes of importance.

## 2. PARAMETRIC MODELS

In an early paper, Feigl and Zelen (1965) proposed an exponential covariance model, where $Y_i$ has density:

$$f_{Y_i}(t) = \lambda_i \exp(-\lambda_i t)$$

and

$$\mu_i = 1/\lambda_i = \alpha + z_i\beta, \qquad z_i \in \mathbb{R}^1 \tag{3}$$

where $\mu_i$ is the mean lifetime. An extension to include simple Type I right censoring was made by Zippin and Armitage (1966). Note that improper estimates of $\beta$ in (3) can lead to negative values for some $\mu_i$.

Feigl and Zelen (1965) also proposed a loglinear model:

$$\mu_i = 1/\lambda_i = \alpha \exp(z_i\beta), \qquad z_i \in \mathbb{R}^1 \tag{4}$$

This model, independently proposed by Glasser (1967), was extended to the censored case by Zippin and Lamborn (1969). Lamborn (1969) showed that the goodness of fit statistic for both of the above models was distributed as a linear combination of squared standard normals, and not as $\chi^2$ as had been conjectured. Both models are easily extended for $z_i \in \mathbb{R}^p$ $(p \geq 2)$. Sprott and Kalbfleisch (1970) examined likelihood methods for the estimation of $\beta$ in (4). Mantel and Myers (1971) noted convergence problems of Newton-Raphson techniques in (3), and suggested the use of expected rather than observed Hessian values.

Peto and Lee (1973) considered the Weibull model:

$$f_{Y_i}(t) = \lambda_i k t^{k-1} \exp(-\lambda_i t^k) \qquad (5)$$

$$\lambda_i = \exp(z_i \beta), \qquad z_i \in R^p$$

with $\lambda_i$ estimated from the (possibly right censored) data via maximum likelihood.

Prentice (1973) provided significance tests for (4) and (5), and inferences on $\beta$ and the hazard corresponding to $\bar{z}$, the mean covariate level. Prentice (1974) and Farewell and Prentice (1977) go on to develop a very flexible "generalized gamma" model that includes exponential, Weibull, gamma, and lognormal as special cases. In the papers above, Prentice develops a marginal likelihood measure utilizing Fraser's (1968) structural inference model. This approach is also applied to Cox's (1972) model which we discuss in the next section.

Different parametric models are supported by different sources of survival data. Carcinogenesis data may follow lognormal or Weibull survival patterns. Several authors have proposed these two models - see Hoel, et al, (1975), Pike (1966) and others. Whittemore and Altschuler (1976) applied graphical techniques to fit both models to the Doll and Hill lung cancer data. The exponential model was suggested by Stollmack and Harris (1974) as appropriate for exoffender recidivism analysis.

The analysis of accelerated life tests prompted the consideration of stress as a covariate. This concept is discussed by Lee and Thompson (1976) and Cox (1972). Nelson and Hahn (1972, 1973) discuss a general parametric model with unknown location parameter $\mu(z) = \beta_0 + z\beta_1$ unknown scale parameter $\sigma$. The authors go on to describe graphical and linear unbiased techniques for estimation of $\beta$ with Type II censored data. In part II of the paper the BLUE is given.

Nelson and Kielpinski (1975, 1977) describe the optimal choice of stress levels for accelerated life test plans for normal and lognormal distributions. Nelson and Meeker (1975) provide similar

results for Weibull and extreme value distributions. Hahn and Nelson (1975) examine MLE's as well as LUE's and graphical estimates. Nelson (1975) considers the competing risks problem and plots failure by mode. Applications are made in the above papers to motor winding failures and wire-bond failure on semiconductor chips.

## 3. "CONDITIONAL" MODELS

Mantel (1966) has considered the problem of testing for homogeneity of $k$ ($\geq 2$) treatment groups. He obtained a $\chi^2_{k-1}$ statistic based on the Mantel-Haenszel procedure for combining a set of $2 \times k$ contingency tables -- one table is constructed for each distinct observed time of death ($t_1, t_2, \ldots$, say). This statistic is given by

$$\chi^2_{k-1} = (0 - E)^T V^-(0 - E) \qquad (6)$$

where $0 = \sum(r_{1j}, r_{2j}, \ldots, r_{kj})$, $r_{ij}$ = no. of deaths in group $i$ at time $t_j$, $E = E(0)$, $V = \mathrm{Var}(0)$, and $V^-$ is a generalized inverse. This is appropriate for testing against Lehmann (proportional hazard) alternatives. Peto and Peto (1972) show it is an asymptotically efficient rank invariant test. (See also Peto (1972a), Crowley (1974a).) A conservative approximation to (6) is given by Peto and Pike (1973), but if the $k$ groups have similar censoring patterns, there is little loss in power over (6) (Crowley and Breslow, 1975). Mantel (1966) suggests generalizing the model to include covariates by dividing the sample into subgroups depending on covariate values. This approach is used by Hankey and Myers (1971). The method above is inefficient as it does not take into account trends in the quantitative covariates (Tarone, 1975). The model discussed next does not suffer this shortcoming.

A proportional hazards approach to nonparametric analysis of covariance was taken by Cox (1972), based on the relation:

$$\lambda_i(t) = \lambda_0(t) \exp(z_i \beta), \qquad z_i \in \mathbb{R}^p \qquad (7)$$

Assuming $\lambda_0(t)$ arbitrary, Cox derived a maximum 'likelihood'

estimate of $\beta$ by conditioning on the death times and individuals at risk (i.e. known to be alive) at those times. Under the assumption of no ties, the contribution to the 'likelihood' at the i'th distinct observed death given the risk set $R_i$ (i.e. those known alive) at $t_i - 0$, is:

$$\exp(z_i\beta)/\sum_{j\in R_i}\exp(z_j\beta)$$

Taking the product over all failure times $\{t_i\}$ as the 'likelihood' $L^*$, we have:

$$L = \log L^* = \sum_i (z_i\beta - \log \sum_{R_i} \exp(z_j\beta)) \qquad (8)$$

The model is attractive in that neither right censoring nor time dependent $z_j$'s cause problems in evaluating (8). Unfortunately, tied deaths, which arise commonly in grouped data, do cause a problem. In this case, Cox proposed the use of a logistic model that reduced to the original model in the limit. The proper likelihood calculation would require the evaluation of $m_i!$ orderings of $m_i$ tied deaths at $t_i$. Peto (1972b) suggested a simple approximation to the likelihood contribution at the i'th death time:

$$\exp(v_i\beta)/\binom{N_i}{m_i}\left(\frac{\sum_{R_i}\exp(z_j\beta)}{N_i}\right)^{m_i}$$

and

$N_i = \|R_i\|$, $v_i = \sum_{D_i} z_j$, $D_i$ = {index set of all dying at the i'th death time}

The rough average given above then yields the 'log likelihood' function:

$$L(\beta) = \sum_i (v_i - m_i \log \sum_{R_i} \exp(z_j\beta)) + \text{constant} \qquad (9)$$

Cox defined $U(\beta) = \nabla L(\beta)$ and $I(\beta) = \nabla^2 L(\beta)$, and argued that $I(\hat{\beta})$ can be used to estimate the variance of the MLE $\hat{\beta}$ directly. As a test of the global null hypothesis $H_0$: $\beta = 0$, Cox suggested the asymptotically $\chi^2_p$ statistic:

$$U(0)^T(I(0))^{-1}U(0) \tag{10}$$

In the simple two sample case this reduces to the (asymptotically) standard normal statistic: $U(0)/\sqrt{I(0)}$. Cox indicated the equivalence of (10) with (6) in the two (or multi-) sample case.

Oakes (1972) suggested a reasonable approach to the estimation of the failure cdf, $F(t)$. He considered $\lambda_0(t)$ to be constant between failures, which leads to the MLE at the i'th failure (Breslow, 1972, 1974, 1975):

$$\log(1 - F(t_i)) = -\sum_{k=1}^{i} m_k / \sum_{j \in R_k} \exp(z_j\hat{\beta})$$

The Cox model was applied to remission times for leukemia patients, and an alternative likelihood ratio test for significance was used:

$$2(L(\hat{\beta}) - L(0)) \sim \chi^2_p \quad \text{for} \quad \hat{\beta} \in \mathbb{R}^p \tag{11}$$

Cox also addresses the problem of accelerated life testing in the light of his conditional model.

The discussion following Cox(1972) contains many interesting remarks. Kalbfleisch and Prentice point out, disturbingly, that the "conditional likelihood" proposed by Cox is not a conditional likelihood at all, which places in doubt the asymptotic MLE properties claimed above. However several followup papers have shed more light on this situation. Kalbfleisch and Prentice (1973) show that, for censored but untied data with time-constant covariates, the class of models (7) is invariant under the group of differentiable monotone increasing transformations on the time scale. This enables the derivation of (8) as a marginal likelihood. For censored tied data and fixed covariates, the authors derive a marginal likelihood different from that in Cox (1972). They point our that, for substantial grouping, the regression parameters in the logistic model of Cox may be considerably different from those in (8). Simulation results are presented supporting this contention and showing the improved performance of the marginal likelihood estimate. The authors note that the group invariance property is lost when time dependent covariates are allowed, and thus the

Cox results for this case could not be supported by marginal likelihood arguments.

Breslow (1972), in the discussion following Cox's paper, assumes a constant hazard between observed failures, and assumes censoring occurs at the start of each interval. Joint maximization of $\beta$ and $\lambda_0(t)$ lead to Peto's approximation and Oakes' estimate, respectively, providing further justification forthe use of (8) and (9). Breslow (1974) applies this model to leukemia data and compares the covariate estimates with models (3) and (4). A clear description of the relationship of these models and further applications are presented in a later paper (Breslow, 1975).

Cox (1975) generalized the ideas of conditional and marginal likelihood, and argued that, under mild assumptions, the usual asymptotic properties hold for "partial likelihoods". He showed that (8) is a partial likelihood, allowing time dependent covariates but not ties.

Kalbfleisch (1974) examines the efficiency of the Cox estimate of $\beta$ relative to the usual MLE based on $\lambda_0(t) = \lambda$, the exponential model. With uncensoreddata, the Cox procedure is found asymptotically fully efficient at $\beta = 0$. This generalizes known results for the special cases of the exponential scores of Savage (1956) and Cox (1964), and the relative efficiency of .75 for twin studies (Holt and Prentice, 1974). Kalbfleisch shows an efficiency (relative to the exponential model) of .94 at $n = 20$ deaths. Efficiency at $\beta \neq 0$ is approximated at greater than .75 over a reasonable range of $\beta$ values. Kalbfleisch and McIntosh (1977) examine corresponding efficiencies in a two sample Weibull shape shift problem $(\lambda_0(t) = \lambda k t^{k-1})$.

Efron (1975), states that under reasonable assumptions about the "average hazard rate" and with no ties, the Cox estimate $\hat{\beta}$ is asymptotically fully efficient, compared with a $\hat{\beta}^*$ based on the full likelihood function.

There have been many applications of the Mantel and Cox models to the analysis of medical followup data. Analysis of heart transplant data provides an interesting example of a covariate (group membership) that changes over time -- see Turnbull, Brown and Hu(1974), Mantel and Byar(1974). Crowley(1974b) demonstrates the asymptotic normality of the resulting statistic (10). More recently, Crowley and Hu(1977) have performed a more extensive analysis of heart transplant data including prognostic factors as additional covariates. Cangir et al. (1975) and Gehan and Smith (1976) apply Cox's model in a forward stepwise manner to determine prognostic factors in leukemia. Tarone (1975) provides tests for trends (and departures from trends) in hazard functions.

## 4. RELATED MODELS

Thompson(1977) and Holford(1977) consider life table applications of the proportional hazards model. Both models allow covariates to change with time.

Thompson assumes the following form for the grouped data hazard (cf. (7)) for the j'th individual in the interval $(t_i, t_{i+1})$:

$$\frac{N_i - N_{i+1}}{N_{i+1}} = \lambda_i \exp(z_j \beta)$$

yielding the log likelihood (c.f.(9)):

$$L = \sum_i [v_i \beta + m_i \log \lambda_i - \sum_{j \epsilon R_i} \log(1 + \lambda_i \exp(z_j \beta)] \qquad (12)$$

Here we are using the notation of Section 3.

Thompson considers the loglikelihood ratio test described above (11). He also shows that as the grouping becomes fine, (12) goes to (8) when there are no ties, and to (9) when there are ties.

Holford uses (7) as the covariate relation, but assumes the exact time of failure is known for the individual that fails in an interval. However, this assumption may be unreasonable when dealing with grouped data. Holford acknowledges this and proposes

estimates for the each time of failure if only the interval is known. If the (possibly interval dependent) adjustment rule is the same over all persons, then the Holford likelihood is proportional to (9).

Thus we find that all three of the above models should yield results that are the same or similar to those yielded by (9).

In a recent paper, Miller (1976) extends the techniques of standard least squares to the case of censored data through the use of the Kaplan-Meier estimate of a distribution function. Unfortunately there are some computational difficulties with this approach.

## 5. OTHER TOPICS

We now examine some topics relevant to all of the models discussed above. Cox and Snell(1968) discuss a generalized residual concept that could be useful in examining goodness of fit; Peduzzi et al. (1976) examine model fit using this technique. The question of sample size determination has not been thoroughly addressed, although some work by Nelson (with coauthors) has been mentioned in Section 2. Missing covariate values can be handled perhaps by regression techniques (Rubin, 1976).

Several approaches have been suggested for selecting appropriate covariates for inclusion in a model. Gehan and Smith (1976) use a forward stepwise procedure, choosing the entering variable based on the largest increase in the likelihood. Greenberg et al. (1974) use a "pseudo log likelihood" backward regression procedure, for a model of the form:

$$\lambda_i = z_i\beta = \sum_j \beta_j z_{ij},$$

the authors eliminate the covariate corresponding to $\beta_k$ yielding the largest ratio $L(\lambda^k)/L(\lambda)$ where

$$\lambda_i^k = \sum_{j \neq k} \beta_j z_{ij}$$

Byar and Corle(1974) discuss the general backward elimination rule for maximum likelihood regression, and suggest the

computationally simpler rule of choosing $k$ to minimize $\hat{\beta}_k/\sigma(\hat{\beta}_k)$.

Unfortunately, this rule cannot guarantee that the selected variable will reduce the likelihood least. Here we propose an alternate rule that is exact in the case of a quadratic likelihood function: choose $k$ to maximize

$$-\frac{1}{2}\, g_k^T\, (I_k)^{-1} g_k$$

where $g_k$ is the gradient of the overall likelihood, excluding the kth coordinate, and $I_k$ is Hessian of the overall likelihood, excluding the kth row and column. In the quadratic case, this yields the maximum likelihood for each submodel. We apply this method in the next section.

Finally, the interpretation of the results of covariance analysis of censored survival data must be considered. Hartley and Sielken(1977) and Hoel et al. (1975) examine the problems in extrapolating results from animal dose-response experiments to low dose carcinogenic effects in man. Byar and Corle(1974) describe how to detect treatment-patient type interactions based on regressions results.

## 6. APPLICATION TO RECIDIVISM DATA

The sample consisted of 37 and 71 maximum security offenders paroled from the Cheshire and Somers correctional institutions respectively, between November 1974 and March 1975. Cheshire is primarily for youthful offenders, but the age ranges of the two institutions overlap. Followup data were collected monthly until January 31, 1976 (Christie, et al. 1976). The dependent variable measured was time from release until first arrest. Of the 37 Cheshire parolees, 15 (41%) were rearrested, while 23 of 71 (32%) Somers parolees were rearrested. We strongly emphasize that the data came from a nonrandomized observational study, and that care should be taken in assessing significance levels (P-values) of the statistical tests that follow; see McKinlay (1975) for a further discussion of this problem.

The covariates measured on each offender were:

$z_1$ - Institution, a 0 - 1 indicator variable for the offenders institution. (1 = Somers).

$z_2$ - Previous major offense, coded '1' for burglary and larceny, '-1' for murder and rape, and '0' for all others.

$z_3$ - Age at release.

$z_4$ - Drug use, a 0 - 1 indicator with '1' indicating use.

$z_5$ - Monthly income, a time varying covariate.

These covariates are among those considered by Glaser (1969) and others to be correlated with recidivism. Summary statistics for these data are shown in Table 1.

As a first step, we compare the rearrest experience of the two institutions. A two sample test can be performed for $\omega_A$: $\beta_i = 0$, $1 \leq i \leq 5$ against $\Omega_B$: $\beta_i = 0$, $2 \leq i \leq 5$. Stollmack and Harris(1974) analyze a similar problem, assuming exponential failure times and no covariates. (See also Turnbull, 1977.) Using their goodness of fit test for exponentiality, the constant hazard rate assumption could not be rejected at any reasonable level of significance, and so their F test is of interest. The results are presented in Table 2, together with the Mantel-Haenszel test and the loglikelihood ratio test (see (10) and (11)). The tests all demonstrate a considerable difference between the arrest rates for the two institutions. The Cox likelihood test may exaggerate this difference because the censoring patterns are different for the two groups. The estimate $\hat{\beta}_1$, in the Cox model (see Table 3, Model B) suggests a proportional hazard of exp (.633) = 1.88, which compares with 1.68 for the exponential model (Table 2).

The important question is whether this difference can be attributed to the effectiveness of the two institutions. To do this, adjustments should be made for non-homogeneity of covariates between groups. In the analyses that follow, we use Cox's approach, with Peto's (1972b) "rough probability" to handle ties (equivalently Breslow, 1974). Approximate maximum likelihood

solutions are found using a function maximization routine due to Fletcher and Powell (1963).

Before further analyses, we examine the income covariate. We consider income as a single explanatory variable and test $\omega_A$: $\beta_i = 0$, $1 \leq i \leq 5$ against $\Omega_C$: $\beta_i = 0$, $1 \leq i \leq 4$, obtaining $2(L_C - L_A) = 11.31$ for a $\chi^2_1$ P-value of $<.005$. This high significance is not surprising, since nearly half the arrests resulted in reincarceration and consequent loss of income. As a crude compensation, we considered lagging the income covariate by one full month. Now the statistic $2(L_D - L_A) = .932$, and the significance of the income effect is greatly reduced. Finer grouping would permit one to lag by one or two weeks, perhaps a more reasonable amount. In the analyses below, income has been replaced by "lagged income".

Returning to the question of differences between Cheshire and Somers rearrest rates, we test $\omega_F$: $\beta_1 = 0$ against the full model $\Omega_E$, yielding $2(L_E - L_F) = .94$ with a $\chi^2_1$ P-value of .33. The regression parameters are given in Table 3, models E and F. Thus, after adjustment for the four other covariates, the difference between institutions is now far from significant. Indeed, age alone can explain most of the difference. Testing $\omega_I$: $\beta_1 = \beta_2 = \beta_4 = \beta_5 = 0$ against $\Omega_J$: $\beta_2 = \beta_4 = \beta_5 = 0$, we have $2(L_J - L_I) = 1.76$ with a P-value of .19; an appropriate transformation of the age covariate could reduce the significance still further. Testing the overall significance of $\omega_A$: $\beta_i = 0$, $1 \leq i \leq 5$ against $\Omega_E$ yields (Table 3) $2(L_E - L_A) = 6.02$ for a $\chi^2_5$ P-value of .32, and so the joint effect of the covariates does not appear significant.

A backward elimination stepwise regression procedure was carried out on the data, yielding successively models F through I, using the method we described in Section 5. The five variables are eliminated in the order: institution, offense, drug use, lagged income, and age. The method of Byar and Corle (1974) yields different choices for the first two models, suggesting drug use

as the first exiting variable. Although no step exhibits a higher level of significance than age alone, the backwards procedure is still of some value as a rough ranking of the importance of factors assumed to affect recidivism. We also see, through this procedure, that the regression coefficients are reasonably stable as one moves to reduced models.

## 7. SUMMARY

The intent of the above analysis was to illustrate the potential application of recent techniques for censored life data to recidivism data and to observational studies in general. A more thorough analysis with a larger sample size could include consideration of nonlinear effects, either in a direct manner or by appropriate stratification of the independent variables (age, income). Other covariates, such as age at first arrest, number of previous convictions, and type of prison industry training, should be considered. No attempt is made to extrapolate behavior beyond the 15 month followup period. This short period is roughly half the mean life (under an exponential assumption) so we have little information on the right tail of the distribution. As further nonparametric studies of recidivism data are performed, one may be able to choose an appropriate parametric model to increase the small sample power of testing procedures. Recording of failures by day, or at least by week, would reduce the problem of ties, and increase the usefulness of time dependent covariates.

The modification (9) of Cox's model was used above because it was simple, and because several other recent models yield similar likelihoods. Given the severe grouping of our sample, Thompson's approach (12) might be an appropriate alternative.

## BIBLIOGRAPHY

Breslow, N. E. (1972). Contribution to the discussion on the paper by D.R. Cox. J.R. Statist. Soc. B. 34, 216-17.

Breslow, N.E. (1974). Covariance analysis of censored survival data. Biometrics, 30, 89-100.

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

Byar, D.P. and Corle, D. (1974). Selecting optimum treatment in clinical trials using covariate information. Paper read at the Meeting of Biometric Society, St. Louis, Missouri, August, 1974. Abstract #2263 in Biometrics, (1975) 31, 591.

Cangir, A., George, S. and Sullivan, M. (1975). Unfavorable prognosis of acute leukemia in infancy. Cancer, 36, 1973-78.

Christie, R.J. et al. (1976). Study of the Economic and Rehabilitative Aspects of Prison Industries, by ECON, Inc. for U.S. Dept. Justice under contract J-LEAA-033-75.

Cox, D.R. (1964). Some applications of exponential ordered scores J.R. Statist. Soc. B., 26, 103-10.

Cox., D.R. (1972). Regression models and life tables (with discussion). J.R. Statist. Soc. B., 34, 187-220.

Cox, D.R. (1975). Partial likelihood. Biometrika, 62, 269-76.

Cox, D.R. and Snell, E.J.(1968). A general definition of residuals. J.R. Statist. Soc. B., 30, 248-75.

Crowley, J. (1974a). A note on some recent likelihoods leading to the log rank test, Biometrika, 61, 533-38.

Crowley, J. (1974b). Asymptotic normality of a new nonparametric statistic for use in organ transplant studies. J. Am. Statist. Assoc., 69, 1006-11.

Crowley, J. and Breslow, N.E. (1975). Remarks on the conservatism of $\sum(O - E)^2/E$ in survival data. Biometrics, 31, 957-61.

Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data. J. Am. Statist. Assoc., 72, 27-36.

Efron, B. (1975). The efficiency of Cox's likelihood function for censored data. Technical Report No. 15, (PHS 1 R01 GM21215-01), Division of Biostatistics, Stanford University.

Farewell, V.T., and Prentice, R.L. (1977). A study of distributional shape in life testing. Technometrics, 19, 69-75.

Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. Biometrics, 21, 826-38.

Fletcher, R., and Powell, M.J.D. (1963). A rapidly convergent descent method for minimization. Computer Journal, 6, 163-68.

Fraser, D.A.S. (1968). The Structure of Inference, New York: Wiley.

Gehan, E.A., and Smith, T.L. et al (1976). Prognostic factors in acute leukemia. Seminars in Oncology, 3, 271-82.

Glaser, D. (1969). The Effectiveness of a Prison and Parole System. Abridged Ed. Indianapolis: Bobs-Merrill.

Glasser, M. (1967). Exponential survival with covariance. J.Am. Statist. Assoc., 62, 561-8.

Greenberg, R.A., Bayard, S., and Byar, D.P. (1974). Selecting concomitant variables using a likelihood ratio step-down procedure and a method of testing goodness of fit in an exponential survival model. Biometrics, 30, 601-608.

Gross, A.J. and Clark, V.A. (1975). Survival Distributions: Reliability Applications in the Biomedical Sciences, New York Wiley.

Hahn, G. and Nelson, W.B. (1975). A comparison of methods for analyzing censored life data to estimate relationships between stress and product life. IEEE Transact.Reliability, 23, 2-11.

Hankey, B.F. and Myers, M.H. (1971). Evaluating differences in survival between two groups of patients. J. Chron. Diseases, 24, 523-31.

Hartley, H.O. and Sielken, R.L. (1977). Estimation of "safe doses" in carcinogenic experiments. Biometrics, 33, 1-30.

Hoel, D.G., Gaylor, D.W., Kirschstein, R.L. Saffiotti, U. and Schneiderman, M.A. (1975). Estimation of risks of irreversible delayed toxicity. J. Toxicology and Envir. Health, 1, 133-51.

Holford, T.R. (1976). Life tables with concomitant information. Biometrics, 32, 587-97.

Holt, J.D. and Prentice, R.L. (1974). Survival analysis in twin studies and matched pair experiments. Biometrika, 61, 17-30.

Kalbfleisch, J.D. (1974). Some efficiency calculations for survival distributions. Biometrika, 61, 31-37.

Kalbfleisch, J.D. and McIntosh, A.A. (1977). Efficiency in survival distributions with time-dependent covariables. Biometrika, 64, No. 1, 47-50.

Kalbfleisch, J.D. and Prentice, R.L. (1973). Marginal likelihoods based on Cox's regression and life model. Biometrika, 60, 267-78.

Lamborn, K. (1969). On the chi-square goodness of fit test for sampling from more than one population with possibly censored data. Technical Report No. 21, Dept. of Statistics, Stanford University. (PHS 2T01GM00025-11).

Lee, L. and Thompson, W.A. (1976). Failure rate - a unified approach. J. Appl. Prob., 13, 176-82.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports, 50, 163-70.

Mantel, N. and Byar D.P. (1974). Evaluation of response-time data involving transient states: an illustration using heart transplant data. J. Am. Statist. Assoc., 69, 81-86.

Mantel, N. and Myers, M. (1971). Problems of convergence of maximum likelihood iterative procedures in multiparameter situations. J. Am. Statist. Assoc., 66, 484-91.

McKinley, S.M. (1975). The design and analysis of the observational study - a review. J. Am. Statist. Assoc., 70, 503-20.

Miller, R.G. (1976). Least squares regression with censored data. Biometrika, 63, 449-64.

Nelson, W.B. (1975). Graphical analysis of accelerated life data with mixed failure modes. IEEE Transact. Reliability, 24, 230-237.

Nelson, W.B. and Hahn, G.J. (1972-3). Linear estimation of a regression relationship from censored data I,II. Technometrics, 14, 247-269 and 15, 697-715.

Nelson, W.B. and Kielpinski, T.J. (1975). Optimum accelerated life tests for the normal and lognormal life distributions. IEEE Transact. Reliability, 24, 310-20.

Nelson, W.B. and Kielpinski, T.J. (1976). Theory for optimum accelerated life tests for normal and lognormal life distributions. Technometrics, 18, 105-14.

Nelson, W.B. and Meeker, W.Q. (1975). Optimum accelerated life tests for the Weibull undertone value distributions. IEEE Transact. Reliability, 24, 321-32.

Oakes, D., (1972). Contribution to the discussion on the paper of D.R. Cox. J.R. Statist. Soc. B., 34, 208.

Peto, R. (1972a). Rank tests of maximum power against Lehmann type alternatives. Biometrika, 59, 472-4.

Peto, R. (1972b). Contribution to the discussion on the paper of D.R. Cox, J.R. Statist. Soc. B., 34, 205-7.

Peto, R. and Lee, P. (1973). Weibull distributions for continuous carcinogenesis experiments. Biometrics, 29, 457-70.

Peto, R., and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. J. R. Statist. Soc. A., 135, 185-206.

Peto, R. and Pike, M.C. (1973). Conservatism of the approximation $\Sigma(O-E)^2/E$ in the logrank test for survival data or tumor incidence data. Biometrics, 29, 579-83.

Peduzzi, P.N., Holford, T.R., and Hardy, R.J. (1976). Regression methods in life table analysis with time dependent covariates. Yale University School of Medicine, Department of Epidemiology and Public Health. Unpublished.

Pike, M.C. (1966). A method of analysis of a certain class of experiments in carcinogenesis. Biometrics, 22, 142-61.

Prentice, R.L. (1973). Exponential survivals with censoring and explanatory variables. Biometrika, 60, 279-88.

Prentice, R.L. (1974). A log gamma model and its maximum likelihood estimation. Biometrika, 61, 539-44.

Rubin, D.B. (1976). Comparing regressions when some predictor values are missing. Technometrics, 18, 201-6.

Savage, I.R. (1956). Contributions to the theory of rank order statistics: the two sample case. Ann. Math. Stat., 27, 590-615.

Sprott, D.A., and Kalbfleisch, J.D. (1969). Examples of likelihoods and comparison with point estimates and large sample approximations. J. Am. Statist. Assoc., 64, 468-84.

Stollmack, S. and Harris, C.M. (1974). Failure rate analysis applied to recidivism data. Operations Research, 22, 1192-1205.

Tarone, R.E. (1975). Tests for trend in life table analysis. Biometrika, 62, 679-82.

Thompson, W.A. (1977). On the treatment of grouped observations in life studies. Biometrics, 33.

Turnbull, B.W. (1977). A note on the nonparametric analysis of the Stollmack-Harris recidivism data, Operations Research, 25.

Turnbull, B.W., Brown, B.W. and Hu, M. (1973). Survivorship analysis of heart transplant data. J. Am. Statist. Assoc., 69, 74-80.

Whittemore, A., and Altshuler, B. (1976). Lung cancer incidence in cigarette smokers: further analysis of Doll and Hill's data for British physicians. Biometrics, 32, 805-16.

Zippin, C. and Armitage, P. (1966). Use of concomitant variables and incomplete survival information with estimation of an exponential survival parameter. Biometrics, 22, 665-72.

Zippin, C., and Lamborn, K. (1969). Concomitant variables and censored survival data in estimation of an exponential survival parameter, part II. Technical Report No. 20, (PHS 2 T01 GM00025-11). Dept. of Statistics, Stanford University.

TABLE I

Parolee Sample Description by Institution

| | SOMERS | CHESHIRE |
|---|---|---|
| Failure (censoring) Times | 1,2,2,3,3,4,4,(5),6,6,6, 6,6,6,(6),(6),7,7,(7),(7), 8,8,9,9,(9),10,(10),(11), (12),(12),(12),(12)13, (13),(13),(13),(13),(13), (13), and thirty at (14) or more. | 2,2,3,3,3,3,4,4,4,(4), (4),5,5,5,(5),6,(6),7, (8),9,10,(10),(10),(10), (10),(10),(10),(10),11, 11, and seven at (12) or more. |
| Rearrest | 23 | 15 |
| Months at Risk | 724 | 281 |

Summary Statistics on Covariates

| | Mean | Range | Mean | Range |
|---|---|---|---|---|
| Offense | .043 | -1,0,1 | .297 | -1,0,1 |
| Age | 32.63 | 22-71 | 20.57 | 17-45* |
| Drug Use | .46 | 0,1 | [illegible] | 0,1 |
| Monthly Income | $261 | $0-$800 | $130 | $0-$600 |

*Only one individual -- next oldest was 24.

TABLE 2

Two Sample Tests for Difference Between Institutions with No Adjustments for Other Covariates

| Test: | Stollmack-Harris* | Mantel-Haenszel | Log likelihood ratio |
|---|---|---|---|
| Test Statistic | $\lambda_0/\lambda_1$ | $U(0)/\sqrt{I(0)}$ | $2(L_\Omega - L_\omega)$ |
| Null Distribution | $F_{46,30}$ | $N(0,1)$ | $\chi^2_1$ |
| Observed Value | 1.68 | 2.04 | 3.87 |
| P-Value | .05 | .02 | .05 |

*Assumes constant hazard.

TABLE 3

Coefficients (with Standard Errors)

for the Regression Models

| Models | Log-Likelihood | Coefficient Estimates ($\hat{\beta}$) Institution | Offense | Age | Drug Use | Income | Lagged Income |
|---|---|---|---|---|---|---|---|
| A | -185.20 | . | . | . | . | . | . |
| B | -183.27 | .633 (.315) | . | . | . | . | . |
| C | -179.55 | . | . | . | . | -.0027 (.00094) | . |
| D | -184.74 | . | . | . | . | . | -.00068 (.00072) |
| E | -182.19 | -.380 (.388) | .298 (.307) | -.0109 (.0190) | -.156 (.319) | . | -.00050 (.00074) |
| F | -182.66 | . | .344 (.303) | -.0183 (.0170) | -.189 (.317) | . | -.00065 (.00072) |
| G | -183.31 | . | . | -.0245 (.0166) | -.200 (.317) | . | -.00068 (.00073) |
| H | -183.52 | . | . | -.0241 (.0166) | . | . | -.00063 (.00073) |
| I | -183.91 | . | . | -.0253 (.0169) | . | . | . |
| J | -183.03 | -.495 (.372) | . | -.0123 (.0185) | . | . | . |

| | | |
|---|---|---|
| 1. REPORT NUMBER<br>Technical Report No. 333 | 2. GOVT ACCESSION NO. | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle)<br>A SURVEY OF COVARIANCE MODELS FOR CENSORED LIFE DATA WITH AN APPLICATION TO RECIDIVISM ANALYSIS | | 5. TYPE OF REPORT & PERIOD COVERED<br>Technical Report<br>6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s)<br>Russell R. Barton and Bruce W. Turnbull | | 8. CONTRACT OR GRANT NUMBER(s)<br>DAAG29-77-C-0003<br>N00014-75-C-0586 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>School of Operations Research & Industrial Engineering, College of Engineering<br>Cornell University, Ithaca, NY 14853 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br>Sponsoring Military Activity<br>U.S. Army Research Office<br>Durham, N.C. 27706 | | 12. REPORT DATE<br>May 1977<br>13. NUMBER OF PAGES<br>21 |
| 14. CONTROLLING OFFICE NAME AND ADDRESS<br>Sponsoring Military Activity<br>Statistics and Probability Program<br>Office of Naval Research<br>Arlington, Virginia 22217 | | 15. SECURITY CLASS. (of this report)<br>Unclassified<br>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)

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

A survey is given of techniques for covariance analysis of censored life data. Both parametric and nonparametric approaches are reviewed. An application is given to the evaluation of parolee followup data. We examine the effects of covariates, such as age, income, and drug use, on

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

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

time to rearrest. One of these covariates varies with time. The records of two correctional institutions are compared after adjusting for non-homogeneity of covariate values.

END

DATE
FILMED

12-77

DDC

SHIPPED FROM NTIS

OCT 31 1980