Monographs
on Statistics and
Applied Probability

Analysis of Survival Data

D.R. Cox and D. Oakes

Chapman and Hall

Analysis of Survival Data

D. R. COX

Department of Mathematics,
Imperial College, University of London, UK

D. OAKES

Department of Statistics,
University of Rochester, USA
formerly TUC Centenary Institute of Occupational Health
London School of Hygiene and Tropical Medicine, UK

LONDON NEW YORK
CHAPMAN AND HALL

First published 1984 by
Chapman and Hall Ltd
11 New Fetter Lane, London EC4P 4EE
Published in the USA by
Chapman and Hall
733 Third Avenue, New York NY10017

Printed in Great Britain at the
University Printing House, Cambridge

ISBN 0 412 24490 X

British Library Cataloguing in Publication Data

Cox, D. R.
Analysis of survival data.—(Monographs on statistics and applied probability)
1. Medical statistics 2. Life expectancy
I. Title II. Oakes, D. III. Series
616.6′8′01519532 RA409

ISBN 0-412-24490-X

Library of Congress Cataloging in Publication Data

Cox, D. R. (David Roxbee)
Analysis of survival data.

(Monographs on statistics and applied probability)
Bibliography: p.
Includes indexes.
1. Failure time data analysis. I. Oakes, D.
II. Title. III. Series.
QA276.C665 1984 519.5 83-20882
ISBN 0-412-24490-X

Preface

The statistical analysis of the duration of life has a long history. The recent surge of interest in the topic, with its emphasis on the examination of the effect of explanatory variables, stems mainly from medical statistics but also to some extent from industrial life-testing. In fact the applications range much more widely, certainly from physics to econometrics. The essential element is the presence of a nonnegative response with appreciable dispersion and often with right censoring.

The object of the present book is to give a concise account of the analysis of survival data. We have written both for the applied statistician encountering problems of this type and also for a wider statistical audience wanting an introduction to the field.

To keep the book reasonably short we have omitted both some of the very special methods associated with the fitting of particular distributions and also the mathematically interesting topic of the application of martingale theory and weak convergence to the rigorous development of asymptotic theory. We have also firmly resisted the temptation to extend the discussion to the statistical analysis of point processes, i.e. systems in which several point events may be experienced by each individual.

We thank warmly Ms P. J. Solomon for comments on a preliminary version.

London, March 1983

D. R. Cox
D. Oakes

Contents

CHAPTER 1

The scope of survival analysis

1.1 Introduction

In survival analysis, interest centres on a group or groups of individuals for each of whom (or which) there is defined a point event, often called failure, occurring after a length of time called the failure time. Failure can occur at most once on any individual.

Examples of failure times include the lifetimes of machine components in industrial reliability, the durations of strikes or periods of unemployment in economics, the times taken by subjects to complete specified tasks in psychological experimentation, the lengths of tracks on a photographic plate in particle physics and the survival times of patients in a clinical trial.

To determine failure time precisely, there are three requirements: a time origin must be unambiguously defined, a scale for measuring the passage of time must be agreed and finally the meaning of failure must be entirely clear. We discuss these requirements in a little more detail in Section 1.2.

Sometimes we are concerned solely with the distribution of failure times in a single group. More often, we may wish to compare the failure times in two or more groups to see, for example, whether the failure times of individuals are systematically longer in the second group than in the first. Alternatively, values may be available for each individual of explanatory variables, thought to be related to survival. The lifetime of a machine component may be influenced by the stress exerted on it, or by the working temperature. White blood count is known to influence prognosis in leukaemia. In clinical practice, it is quite common for information on 100 or more variables to be routinely collected on each patient, giving the statistician the unenviable task of summarizing the joint effect of these variables on survival.

Survival analysis is properly thought of as a univariate rather than a multivariate technique because there is only a single response variable, failure time, even though there may be many explanatory variables. Some special problems involving a multivariate response are, however, discussed in Chapter 10.

1.2 The definition of failure times

We now comment briefly on the requirements for measuring failure time.

The time origin should be precisely defined for each individual. It is also desirable that, subject to any known differences on explanatory variables, all individuals should be as comparable as possible at their time origin. In a randomized clinical trial, the date of randomization satisfies both criteria, and would be the normal choice. While it might be more biologically meaningful to measure time from the first instant at which the patient's symptoms met certain criteria of severity, the difficulty of determining and the possibility of bias in such values would normally exclude their use as time origin. Such information might, however, be useful as an explanatory variable.

The time origin need not be and usually is not at the same calendar time for each individual. Most clinical trials have staggered entry, so that patients enter over a substantial time period. Each patient's failure time is usually measured from his own date of entry. Fig. 1.1 illustrates the calculation.

The evaluation of screening programmes for the detection of breast cancer provides an instructive example of the difficulties in the choice of origin. The aim of screening, of course, is to detect the disease at an earlier stage in its development than would otherwise be possible. Even in the absence of effective treatment, patients with disease detected at screening would be expected to survive longer after diagnosis than patients whose disease is detected without the aid of screening. This bias seriously complicates any comparison of the failure times of the two groups. Perhaps the only satisfactory way to evaluate the effect of screening in reducing mortality is to compare the total mortality rate in a population offered screening with that in a population where no screening programme is available.

The time origin need not always be at the point at which an individual enters the study, but if it is not, special methods are needed. For example, in epidemiological studies of the effects on mortality of occupational exposure to agents such as asbestos, the natural measure of time is age, since this is such a strong determinant of

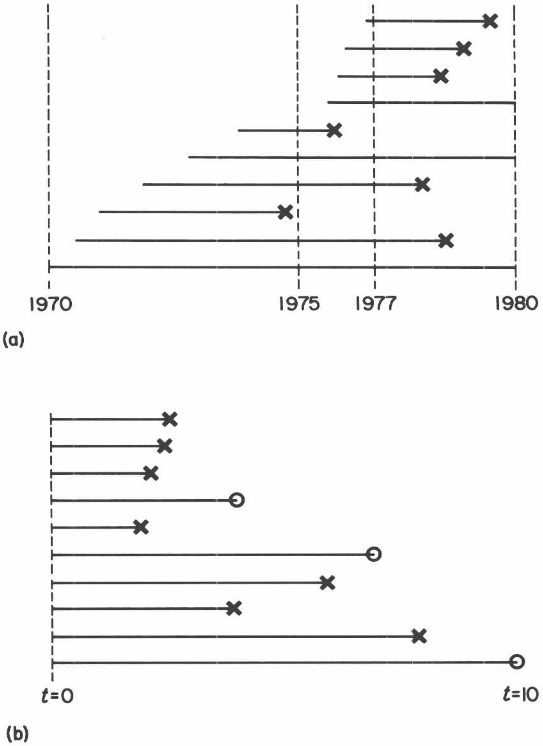

Fig. 1.1. Experience of ten individuals with staggered entry and follow-up until 1980: ×, death; ○, censoring. (a) Real time; (b) time, t, from entry into study.

mortality. However, observation on each individual commences only when he starts work in a job which involves exposure to asbestos. Likewise, in industrial reliability studies, some components may already have been in use for some period before observation begins. We refer to such data as 'left-truncated' and the appropriate methods are discussed in Chapter 11.

Often the 'scale' for measuring time is clock time (real time), although other possibilities certainly arise, such as the use of operating time of a system, mileage of a car, or some measure of cumulative load encountered. Indeed, in many industrial reliability applications, time is most appropriately measured by cumulative usage, in some sense. Or failures may consist of flaws in textile yarn, when failure 'time' would be the length measured up to the first flaw. There are interesting applications in geometrical probability, where the failure time denotes the length of a line segment contained in a convex body. About the only universal requirement for failure times is that they are nonnegative.

One reason for the choice of a timescale is direct meaningfulness for the individual concerned, justifying the use of real time in investigating survival in a medical context. Another consideration is that two individuals treated identically should, other things being equal, be in a similar state after the lapse of equal 'times'; this is the basis for the use of cumulative load encountered in an engineering context. If two or more different ways of measuring time are available, it may be possible, having selected the most appropriate timescale, to use the other 'times' as explanatory variables.

Finally, the meaning of the point event of failure must be defined precisely. In medical work, failure could mean death, death from a specific cause (e.g. lung cancer), the first recurrence of a disease after treatment, or the incidence of a new disease. In some applications there is little or no arbitrariness in the definition of failure. In others, for example in some industrial contexts, failure is defined as the first instance at which performance, measured in some quantitative way, falls below an acceptable level, defined perhaps by a specification. Then there will be some arbitrariness in the definition of failure and it will be for consideration whether to concentrate on failure time or whether to analyse the whole performance measure as a function of time.

1.3 Censoring

A special source of difficulty in the analysis of survival data is the possibility that some individuals may not be observed for the full time to failure. At the close of a life-testing experiment in industrial reliability, not all components may have failed. Some patients (many, it is to be hoped) will survive to the end of a clinical trial. A patient who has died from heart disease cannot go on to die from lung cancer.

An individual who is observed, failure-free, for 10 days and then withdrawn from study has a failure time which must exceed 10 days. Such incomplete observation of the failure time is called censoring. Note that, like failure, censoring is a point event and that the period of observation for censored individuals must be recorded.

We suppose that, in the absence of censoring, the ith individual in a sample of n has failure time T_i, a random variable. We suppose also that there is a period of observation c_i such that observation on that individual ceases at c_i if failure has not occurred by then. Then the observations consist of $X_i = \min(T_i, c_i)$, together with the indicator variable $V_i = 1$ if $T_i \leqslant c_i$ (uncensored), $V_i = 0$ if $T_i > c_i$ (censored). We refer to the c_i of individuals who in fact are observed to fail as unrealized censoring times, as contrasted with the realized censoring times of the censored individuals. The term potential censoring time is usual when c_i is considered without regard to whether censoring or failure occurs.

In some applications, all the c_i will be known, as for example if the only cause of censoring is the planned ending of follow-up at a predetermined time. Another example is so-called Type I censoring, in which all the c_i are equal, $c_i = c$, a constant under the control of the investigator. In Type II censoring, observation ceases after a predetermined number d of failures, so that c becomes a random variable. Type II censoring is a useful technique for economical use of effort in industrial life-testing. Other forms of so-called random censorship are possible. A crucial condition is that, conditionally on the values of any explanatory variables, the prognosis for any individual who has survived to c_i should not be affected if the individual is censored at c_i. That is, an individual who is censored at c should be representative of all those subjects with the same values of the explanatory variables who survive to c.

The simplest way to ensure this is to take the c_i to be in principle predetermined constants, and this viewpoint will be adopted throughout most of this book. Note, however, that often the c_i will not be known to the investigator in advance, and that the unrealized c_i corresponding to observed failures may never become known. The above condition is also satisfied if the potential censoring times are random variables c_i, which are independent of the T_i. Type II censoring is an example of a more general scheme in which, loosely speaking, censoring can depend on the past history, but not the future, of the whole process. We may call this evolutionary censoring.

1.4 Other methods of analysis

Besides the techniques to be discussed in this book, a number of other approaches have been used to analyse survival data. Perhaps the simplest method, much used by clinicians, is to dichotomize according to survival or nonsurvival at a critical period such as five years. Comparisons of the five-year survival rates of subjects in various groups can be made using techniques for binary data. Although this approach is often quite satisfactory, it has two major disadvantages. Concentration on a single point of the survival curve necessarily wastes some information. More seriously, calculation of survival rates as simple proportions is directly possible only when no individuals are censored during the critical period. This restriction can lead to some absurdities; see Exercise 1.1.

With survival dichotomized as above, and with quantitative explanatory variables, discriminant analysis has sometimes been used to identify variables that are related to survival, although such use of discriminant analysis is better regarded as an approach to binary logistic regression. Discriminant analysis, can, however, be a useful way of sifting through a large set of variables to determine a few variables or combinations of variables which can then be considered in more detailed analyses. By itself, discriminant analysis provides little insight into the way the explanatory variables affect survival.

Reduction to a binary response is most useful when the survival of each individual is easily classified as either very short or very long. When the potential censoring times are related to the explanatory variables, discriminant analysis will give biased results. Note also that the inclusion of the actual failure time as an explanatory variable in a discriminant analysis would be a serious error, as the failure time is part of the response, not part of the factors influencing response.

In the absence of censoring, the dependence of failure time on the explanatory variables can be explored through multiple regression. Because failure times are never negative and often have highly skewed distributions, preliminary transformations of the data such as the logarithm or reciprocal are often used. The log transformation is closely related to the accelerated life model, discussed in Chapter 5. Either transformation may give undue weight to very short failure times, which will have high negative logarithms and high positive reciprocals.

1.5 Some examples

We now describe in outline three examples that will be referred to a number of times throughout the book. Other examples will be introduced at the appropriate point in the development. Some of the examples, especially the first, have been widely used in the literature to illustrate alternative techniques.

Example 1.1 Leukaemia: comparison of two groups

Table 1.1 (Gehan, 1965, after Freireich *et al.*) shows times of remission (i.e. freedom from symptoms in a precisely defined sense) of leukaemia patients, some patients being treated with the drug 6-mercaptopurine (6-MP), the others serving as a control. Treatment allocation was randomized. Note the great dispersion and also that censoring is common in the treated group and absent in the control group. It is important to have methods of analysis that are effective in the presence of such unbalanced censoring. In fact, the trial was designed in matched pairs with one member of the pair being withdrawn from study when, or soon after, the other member comes out of remission. This is an aspect we shall ignore.

Example 1.2 Failure times and white blood count, WBC

Table 1.2 shows, for two groups of leukaemia patients, failure time (time to death) in weeks and white blood count, WBC (Feigl and Zelen, 1965). The formal difference from Example 1.1 lies partly in the presence of a continuous explanatory variable, WBC, and partly in that the division into groups is based on an (uncontrolled) measurement for each individual rather than on a randomized treatment allocation.

Example 1.3 Failure times of springs

Table 1.3 illustrates an application from industrial life-testing kindly supplied by Mr W. Armstrong. Springs are tested under cycles of repeated loading and failure time is the number of cycles to failure, it being convenient to take 10^3 cycles as the unit of 'time'. Here 60 springs were allocated, 10 to each of six different stress levels. At the lower stress levels, where failure time is long, some springs are censored, i.e. testing is abandoned before failure has occurred.

Table 1.1 *Times of remission (weeks) of leukaemia patients* (*Gehan*, 1965, *from Freireich et al.*)

Sample 0 (drug 6-MP)	6*, 6, 6, 6, 7, 9*, 10*, 10, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*
Sample 1 (control)	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

* Censored

Table 1.3 *Cycles to failure* (*in units of* 10^3 *cycles*) *of springs*

Stress (N/mm^2)										
950	225	171	198	189	189	135	162	135	117	162
900	216	162	153	216	225	216	306	225	243	189
850	324	321	432	252	279	414	396	379	351	333
800	627	1 051	1 434	2 020	525	402	463	431	365	715
750	3 402	9 417	1 802	4 326	11 520*	7 152	2 969	3 012	1 550	11 211
700	12 510*	12 505*	3 027	12 505*	6 253	8 011	7 795	11 604*	11 604*	12 470*

* Censored

Table 1.2 *Failure time and white blood count* (*Feigl and Zelen*, 1965)

(*AG positive*), $N = 17$		(*AG negative*), $N = 16$	
White blood count, WBC	*Failure time (weeks)*	*White blood count, WBC*	*Failure time (weeks)*
2 300	65	4 400	56
750	156	3 000	65
4 300	100	4 000	17
2 600	134	1 500	7
6 000	16	9 000	16
10 500	108	5 300	22
10 000	121	10 000	3
17 000	4	19 000	4
5 400	39	27 000	2
7 000	143	28 000	3
9 400	56	31 000	8
32 000	26	26 000	4
35 000	22	21 000	3
100 000	1	79 000	30
100 000	1	100 000	4
52 000	5	100 000	43
100 000	65		

1.6 Computing

Some of the simpler techniques to be described in this book can be applied to modest sets of data using a programmable (or even nonprogrammable) pocket calculator. If large amounts of data are involved or if some of the more elaborate methods of analysis are contemplated, use of the computer is essential and, under the working conditions of most statisticians, the writing of special programs is impossible on other than a very small scale. Therefore, the availability of packaged programs is crucial.

All aspects of computing change so rapidly that a very detailed discussion is not appropriate in a book like this. There follow a few notes on the position at the time of writing, 1983.

The packages GLIM (Release 4), BMDP and SAS contain programs for many of the analyses described in this book. Points to watch in the choice of program include the facilities available for checking the