WORKING PAPER NO. 15

EVENT HISTORY AND SURVIVAL ANALYSIS IN THE SOCIAL SCIENCES

DOUGAL HUTCHISON

PART 1 - Background and introduction
PART 2 – Advanced applications and recent developments

Published in, Quality and Quantity, 22: part 1, p.p. 203 - 219 (1988)
part 2, p.p. 255 – 278 (1988)

Quality & Quantity 22: 203–219 (1988)

Event history and survival analysis in the social sciences
I. Background and introduction

DOUGAL HUTCHISON
National Foundation for Educational Research in England and Wales, The Mere, Upton Park, Slough, Berks, SL1 2DQ, UK

Table of contents

I. Introduction

An increasing number of large longitudinal and retrospective studies – the National Child Development Study (Shepherd 1985), the Women and Employment Survey (Martin & Roberts 1984), the (US) Panel Study of Income Dynamics (Duncan & Mathieovitz 1984), the (US) National Longitudinal Study (Palmore et al. 1981) are just a few examples – are attempting to analyse life histories.

All these studies have as a major focus the occurrence of events either singly or in sequence, which involve the transition between a discrete number of states, and the time at which such transitions occur. One example would be the study of duration of unemployment, in which we would investigate the timing of the transition from the state of being unemployed to the state of being employed. Other examples would be age of marriage or childbirth; when people become owner occupiers; the effect that exposure to radiation has on cancer incidence.

II. Panel vs. survival methods – discrete or continuous time?

Such a focus represents a major departure from the panel techniques which were the original methods developed to handle longitudinal studies and are

of course still used (Plewis 1985). Longitudinal studies of children and young people in Britain (The National Study of Health and Development, NSHD, Atkins et al. 1982; The National Child Development Study, NCDS, Fogelman 1983; Child Health and Education in the Seventies, CHES, Osborne et al. 1984) have typically consisted of a series of sweeps, each of which is a cross-sectional snapshot of the situation, recording such information as test scores, health, membership of minority groups and so on. Multivariate analyses of these differ according to whether the response variable is considered continuous or categorical. Whether an individual had had contact with welfare services since the last sweep (a categorical dependent variable) could be analysed using logistic regression, whereas in the analysis of progress in mathematics ability between 11 and 16, ordinary least squares regression may be used. Such techniques may be described as *discrete time* or *panel* models.

However, in analysing discrete events occuring over time, particularly if there is more precise information on the time of events, there are a number of arguments, some more telling than others, favouring the use of *continuous time* models.

(1) Frequently there is no natural time period within which respondents make their decisions or attain certain statuses.

(2) Data collection frequently occurs at time to suit funding or administrative convenience. Even where the timing of data collection is planned for crucial points on the life cycle, this may only hold for a small proportion of processes in a multipurpose study, or only for one part of the geographical research area. To give an example, the National Child Development Study planned its 11 year sweep (Fogleman 1983) for the time when the children first started in secondary school: this of course only holds for England and Wales, Scottish children transferring a year later, at 12.

(3) Where a panel study collects information on the state of an individual at well-spaced discrete time points it can by its very nature contain only partial information about the path of a stochastic process. Thus a child who is at care when surveyed at 11 and 16 may have been in care all that time, or these may represent two isolated and short-term incidents. Even when we can be sure that an individual has undergone at most one such change, or where we are only considering the first occurrence of an event, biases arise when discrete time models, such as logistic regression, are used to estimate continuous time processes. This problem is discussed below, under Non-distinct failure times, in Section V, Covariance Analyses of Censored Survival Data.

(4) It has been said that if the only tool you have is a hammer, then you tend to treat everything as if it were a nail. In the past, quantitative analysis

in the social sciences has been very much restricted by the availability of particular statistical techniques. All our readers must be familiar with examples where a theoretical analysis has had Procrustean treatment to fit it into the framework of (say) multiple regression or factor analysis. Yet, with the advent of new, powerful and readily accessible software, particularly SAS (SAS Institute Inc. 1982) and GLIM (Baker & Nelder 1978) as well as the advent of hardware sufficiently powerful to permit the maximum likelihood calculations involved, we are approaching a situation where the majority of properly specified hypothesis can be tested statistically.

Survival-type approaches permit a relatively realistic modelling of many histories where discrete transitions are of interest. For this reason they will be of considerable value to research workers constructing causal models in sociology, economics, education, psychology and biomedical sciences.

(5) Tuma and Hannan (1984) refer to panel analysis as "equilibrium" analysis and the survival approach as "dynamic" analysis, and advocate the claims of the latter by drawing a parallel with a change in sociological thinking from dominance by a structural-functional perspective to a more diverse spectrum with a greater emphasis in many of the new directions on disequilibrium and change. In our view such a gloss is somewhat suspect. The labels "equilibrium" and "dynamic" analysis are at least misleading, and at worst downright biased against the panel approach. While it is undeniable that the latter does consist of a series of cross-sectional snapshots, it does not follow that these snapshots represent equilibrium positions, and indeed such methods are explicitly focussed towards change and development. The charge of being merely (?) an "equilibrium" approach would be more accurately levelled at cross-sectional studies. It is of interest to note that a similar contrast was made between panel and cross-sectional studies (in favour of panel studies!) in Wall and Williams (1970).

For these reasons, where an individual is in one of a number of mutually exclusive states which are constantly at risk of changing, it may be preferable to view such process not from a panel perspective but from a discrete state, continuous time (event or survival) perspective.

Since the seminal paper by Cox on the covariance analysis of censored survival data (1972) there has been a veritable explosion of publications on survival and event history, the majority in econometrics, engineering and biomedical science.

The purpose of this paper is to inform social sciences about the applicability of such techniques in their own fields. Where appropriate, we illustrate characteristics of survival-type analysis by contrast with the more traditional panel approach.

III. Nomenclature

Approaches which consider the occurrence and timing of events and variously described as *survival analyses* (Cox and Oakes 1984), *event history analyses* (Tuma and Hannan 1984) and *reliability theory* or *renewal theory* (Barlow and Proschan 1975). For simplicity we shall use the term *survival analysis* to subsume the other two. We shall use the term *event history analysis* but confine it to multispell survival techniques.

Survival analysis as a rule is a continuous time approach. Discrete time survival analysis is possible (see Section V below), but more usually apparently discrete time processes are actually continuous time processes where the recorded time has been rounded to the nearest whole unit which may be relatively large – for example to the nearest month (Hutchison 1987) or even year (Davies 1983).

Discrete-time approaches and where the focus is on a change of state between a number of sweeps, we shall continue to describe, as above, as *panel-type* techniques. Survival analysis, in some senses, may be regarded as a dual of the panel-type approach, in that the panel approach assumes a relatively small number of time points and may assume a continuous outcome space, whereas survival analysis typically assumes a relatively small number of outcome states and generally a continuously-measured time variable.

The paradigm of these approaches is summarised most clearly in the table below:

Discrete and continuous time and outcome: example of approaches

Time	Outcome	
	Discrete	Continuous
Discrete (Panel)	Logistic and log linear regression (Discrete time Survival methods)	OLS regression
Continuous	Survival methods	Deterministic and Stochastic Differential Equations (Tuma and Hannan 1984) Integral equations

Deterministic and stochastic differential equation and integral equations are not covered in this paper but included for completeness. Those interested will find an outline in Tuma and Hannan (1984).

IV. Basic concepts

Terminal event

A Terminal Event (TE) occurs when the transition under investigation occurs, for example an individual gets married, leaves a job, has an accident or dies of a particular illness. The somewhat gloomy terminology is an acknowledgement that much of the theory for this kind of approach developed in biomedical science where it was utilized to investigate survival, particularly in cancer studies. However, the terminal event does not need to be actually terminal and the approach can be used on repeatable events such as non-fatal accidents, or spells of unemployment.

Censoring

It is not generally possible to obtain complete histories for all elements in a study, either because the study is terminated before the event being investigated has occurred, or because the individual concerned experiences some other type of Terminal Event first: for example if one is studying marital breakdown two marriages in three will be terminated by death rather than breakdown. This is known as (right) censoring, and is the reason why special maximum likelihood methods are necessary for analysing survival-type problems. Otherwise survival times could be analysed using ordinary regression methods, perhaps after a transformation (e.g., logarithmic) to compensate for the skewness of the distribution (Kalbfleisch & Prentice 1980).

However since censored observations tend to be longer, either treating these observations as missing, or taking them as equal to the last recorded occasion, will bias results (Tuma and Sorensen 1979). In any event it is inefficient to treat a censored observation as "missing", that is saying we know nothing about, since we *do* have some information about it: namely that it is at least as long as the time of censoring. In maximum likelihood methods, the likelihood of a given set of observations is equal to the product of

(a) the likelihood of cases with TEs (the product of the likelihood of surviving until the time of the TE and the likelihood of having a TE at that time) and
(b) the likelihood of the cases censored before reaching TEs (the likelihood of survival until the time of censoring).

The development above has assumed that the mechanism for censoring may be ignored on investigating the T.E. This is obviously reasonable when the

censoring process is independent of the process under investigation, for example where censoring is random or where it takes place at a set time. It is also sound on a wider class of situations, provided one is able to say that, at any time the items withdrawn from risk are representative of the items at risk, and in particular items cannot be censored because they appear to be at an unusually high or low risk of failure. Somewht confusingly this wider class is known as *independent* censoring, though it concludes censoring processes which do depend on the failure process, such as type II censoring (Cox & Oakes 1984) where a study is concluded after observing a pre-specified number of events. This topic is discussed in Williams and Lagakos (1977), Lagakos and Williams (1978) and Kalbfleisch and Mackay (1979). This discussion has referred to *right* censoring in which it has proved impossible to follow the individual under study up to the TE being studied. Equally, also it frequently occurs that the process starts before observation does. In right censoring, all that is known about the timing of a event is that it had not occurred by a given time, while in left censoring all that we know is that it has occurred before a particular time. An example in social sciences would be where one collected employment status data at intervals, and a recording of "unemployed" for the first time at a data collection point would mean that the respondent had become unemployed before that time (and after the last collection time). Grouped data are thus in effect treated as if they were right and left-censored at the ends of the time interval.

An allied concept is that of left *truncation* where events occurring before a given time point are not known to the researcher. Thus if one considered the life-expectancy a cohort of people of a given age, those who had died before that age would be ignored. Not all authors (see Tuma and Hannan 1984) make the distinction between left censoring and left truncation. There has been less attention paid to left censoring and truncation in the literature, though see Turnbull (1974), Tuma & Hannan (1984, 5.4.) Cox and Oakes (1984, 11.6).

To compare different populations in their propensity to change state we need to define some measures. Two basic definitions are the *hazard function* and the *survivor function*: in many ways the central concept of all survival analysis is the former.

Hazard function

We confine ourselves here to the sitution where there is only one type of Terminal event and where only one such event can occur or where we only consider the first occurrence (first passage time-see Flinn and Heckman 1982a): we generalise this later.

Drop-out and the occurrence of terminal events will mean that the numbers of those still left in the study will decrease over time. The hazard function at any given time is the rate of change of state, given that the individual in question has survived this far. Mathematically the hazard function $h(t)$ is defined as

$$h_T(t) = \lim_{\Delta \to 0+} \frac{P(t \leqslant T < t + \Delta t \leqslant T)}{\Delta}. \quad (1)$$

In a real-life situation, rates are hypothetical quantities which cannot be observed. At any instant an event either occurs or it does not: we can only compare rates by observing a greater or lesser number of events over unit of time. This is very simply conceptualized as (number of events) divided by the (number at risk) divided by (length of the time period). Complications arise in estimating the number at risk since this varies over the interval as individuals die or drop out. In the absence of the precise information one convention is to assume that those who are censored or experienced TEs during the period are at risk for half of it. This leads to what is known as the actuarial method (Berkson and Gage 1950, Elandt-Johnson and Johnson 1980) in which the hazard rate over the period (t_i, t_{i+1}) is given by

$$\frac{2q_i}{h_i(1 + p_i)}, \quad (2)$$

where q_i is the proportion of terminal events in the interval, p_i is the proportion surviving, h_i is the width of the interval.

Other definitions are possible, e.g., Allison (1982a), where all censorings are assumed to occur at the end of the interval. Cox (1972) proposes a special form of the hazard for discrete time points. While truly discrete *time* points are rare this approach can be used on quasi-time series, such as the length of a football team's unbeaten run, and can also be a useful approximation to grouped continuous data.

When the time unit becomes small, all these conventions tend to the same limit, namely that of the continuous time definition (1) above.

The hazard function is also referred to as mortality rate, force of mortality or failure rate. Barlow & Proschan (1975) confusingly call it hazard *rate* and reserve the term hazard function for another quantity.

Survivor function

When no change in the risk set occurs except due to terminal events then the survivor function $S(t)$ is simply the proportion of the original population

remaining in the study at time t. In practice the population will also lose members over the time due to censoring. The actuarial estimate of the survivor function is simply the product of (one minus the actuarial hazard rates (2)) above for each time period. Thus the number at risk at the beginning of each time period is adjusted for terminal events and dropouts which occurred beforehand. A more recent and statistically preferable method which does not require the data to be grouped into time periods is the Product Limit or Kaplan–Meier method (Kaplan & Meier 1958).

While it may be of interest to examine the behavior over time of a phenomenon in a population which is considered to be uniform, so many imponderables influence human behaviour that in practice precise description of social processes does not arise, and social scientists tend to confine their interest to more-less comparisons between broad groups. This is performed using some form of covariance analysis.

V. Covariance analysis of censored survival data

If we wish to compare two groups, A and B, on how long it is before an event occurs, one approach would be to compare the average or expected lifetimes. We could posit, for example, that on average lifetimes in Group A would be a constant number of times that in Group B or a constant plus that in Group B. An alternative perspective would be to view the comparison as one of hazard rates, with the possibility that the hazard rate in Group A could be a constant number of times that in group B, or that group A's rate was a constant plus that of group B. We deal with the various approaches under the following headings:

(a) Additive models for survival time or hazard.
(b) Multiplicative models of survival time – Accelerated Failure Time (AFT) models
(c) Multiplicative models of hazard – Proportion Hazards (PH) models.
(d) Other models.

(*a*) *Additive models of survival time or hazard*

Zippin and Armitage (1966) modelled the predicted survival time as a linear function of risk and confounding variables. They assumed no censoring, though the method could readily be adapted to deal with censored times. In the case of constant hazard then this procedure is equivalent to modelling $1/\lambda$ additively.

The additive risks model

$$\lambda = a + bX_1 + cX_2 + \ldots \tag{3}$$

has the attractive property that each sub-population makes an additive contribution to the risk. However, additive models generally are unsatisfactory in that they can give rise to inadmissible estimates, the Zippin and Armitage model possibly giving rise to negative hazard rates. The program RATE (Tuma 1980) can estimate additive models constraining results to be non-negative. However, such a procedure is rather *ad hoc* and statisticians generally prefer to use a multiplicative model which guarantees that estimated survival times or hazard rates are positive.

(*b*) *Accelerated failure time models*

The multiplicative model for survival time can be extended to posit that the expected proportion surviving at t_1 in group A is equal to the proportion surviving in group B at t_2 where $t_1 = kt_2$. Put more colloquially it is as if one group were living faster than the other. This is the Accelerated Failure Time (AFT) Model. In statistical terms this states that:

$$\log T = a + bX_1 + cX_2 + \cdots \quad \cdots + \varepsilon, \tag{4}$$

where T = survival time, a, b, c are constants to be estimated, X_1, X_2 are values of independent variable, ε is an error term of specified distribution.

This model has been used with an error distributed log-normally to study lung cancer in uranium miners (Lundin et al. 1979) and with a log-gamma error in the assessment of risk due to toxic environment substances (Rai & van Ryzin 1979).

If all failure times are recorded, ordinary regression can be used, but censoring requires the use of special methods. This model has been used in the biomedical sciences, though seldom in the social sciences. However, it could usefully be employed to model a situation in which different groups accumulate some quality at different rates to reach the same threshold, for example, saving for the deposit on a house or in learning theory. Cox (1972) has shown that in a "random shocks" model under certain circumstances a cumulative shock gives an AFT model while a single shock going over a threshold corresponds to the PH model.

One drawback, however, is that the AFT model requires that the distribution form of the error term be known, or at least estimable. This has tended to draw researchers towards the proportional hazards model, which

is less dependent on the precise distribution involved. This is not as much of a restriction as one might imagine because of the extreme flexibility of the PH model, particularly with time-varying covariates. In principle, some non-parametric approach, comparable to the mass points techniques of Heckman & Singer (1982) for the proportional hazards (see Section VIII below) would seem to be possible.

(*c*) *Proportional hazards*

An alternative model, known as Proportional Hazards (PH), models the hazard rather than the survival time as a log–linear function of independent variables.

$$\lambda_1(t) = \lambda_0(t)^*\psi(\mathbf{X}_1). \tag{5}$$

If the $\mathbf{X}$ are not related to T then $\lambda_1(t)$ and $\lambda_2(t)$, the hazards in group 1 and 2 respectively are related thus:

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \frac{\psi(\mathbf{X}_1)}{\psi(\mathbf{X}_2)} = K \text{ constant over time for constant covariates.} \tag{6}$$

This property that the ratio of hazard functions remains constant over time is known as proportional hazards. The adequacy of the proportional hazards model can be checked by graphical methods (Kay 1977, Kalbfleisch & Prentice 1980). This model may be estimated by:

Maximum likelihood ML (fully parametric)
Partial likelihood PL (semi parametric)

(1) *Maximum likelihood (fully parametric) approach for proportional hazards*

In some instances it may prove reasonable to assume a functional form for the element $\lambda(t)$ in the formula $\lambda_0(t)$ exp $(\boldsymbol{\beta}\mathbf{Z})$. Widely used for the hazard are

$\lambda_0(t) = \lambda$ Exponential

$\lambda_0(t) = \lambda_i \alpha_i t_i^{\alpha_i - 1}$ Weibull.

Nelson (1972) and Kay (1977) describe graphical procedures for deciding whether a number of particular functional forms are approprite. Nelson's

procedures involve the use of specialised types of graph paper with functional scales, such as logarithmic, on either or both axes. The availability of packages for data presentation means that it is simpler to ask the computer plot these directly, for example to plot the Survival function $S(t)$ against log (t) rather than to plot $S(t)$ against t on semi-log graph paper.

However, it may turn out that the hazard is too irregular to fit with a regular function, and in this case it may be desirable to assume a quite arbitrary hazard function. This approach is the basis of the partial likelihood or semi-parametric approach (Cox 1972, 1975).

(2) *Semi-parametric approach for proportional hazards*

Where no regular function fits readily the Cox regression approach (Cox 1972) gives a proportional hazards model

$$\lambda(t) = \lambda_0(t) \exp(\boldsymbol{\beta}\mathbf{Z}), \qquad (7)$$

$$\lambda_0(t) \text{ arbitrary.}$$

The elements of $\boldsymbol{\beta}$ give an estimate of log (relative risk). A clever argument due to Cox (1972) means that the form of the basic hazard function λ_0 is irrelevant. If the ith element in the risk set fails at time t then the conditional probability of this, given that one element fails is

$$\frac{\lambda_0 \exp(\boldsymbol{\beta}\mathbf{Z}_i)}{\sum_j \lambda_0 \exp(\boldsymbol{\beta}\mathbf{Z}_j)}, \qquad (8)$$

where the summation is taken over all elements of the risk set, the arbitrary element λ_0 cancelling out. The full likelihood of all the data may be written

$$Lik_f = \Pr[E_1, E_2, \ldots, E_d; C_2, \ldots C_d], \qquad (9)$$

where E_i contains the information about the ith terminal event at t_i and C_i contains all information about censoring in the time interval $[t_{i-1}, t_i)$. This can be factored to give:

$$Lik_f = \Pr[E_1|C_1] \prod_{i=2}^{d} \Pr[E_i|E_1, \ldots, E_{i-1}, C_1, \ldots C_i]$$

$$* \Pr[C_1] \prod_{i=2}^{d} \Pr[C_i|E_1, \ldots, E_{i-1}, C_1, \ldots C_{i-1}]. \qquad (10)$$

The second half of this expression contains information about censorings. Cox proposed that this should be discarded, and the resultant expression he called partial likelihood (Cox 1972, 1975), for obvious reasons.

Note that this equal to the product of the conditional probabilities (8) above.

This semi-parametric partial likelihood approach is of particular value in the social sciences, where there is barely any reason to assume a particular functional form. If $\mathbf{Z}$ does not depend on time, the only information needed on times in this approach is their ordering. For large samples the partial likelihood method is unbiased and nearly as efficient as using the full likelihood provided that:

(a) β is not far from zero
(b) censoring is not strongly dependent on $\mathbf{Z}(t)$ &
(c) the covariates do not exhibit strong time trends.

For small samples the loss in practice from using the partial likelihood (PL) is rather greater (Oakes 1977, Efron 1977) though the quality of PL estimates is very high in samples of moderate size even when a high proportion of the sample observations are censored (Caroll et al. 1978, Tuma 1982). In fact such comparisons underestimate the efficiency of PL, since they assume that the full maximum likelihood estimator is correctly specified. Where the full maximum likelihood estimator is incorrectly specified then the PL estimator is likely to perform better.

Non-distinct failure times

However, the use of this semi-parametric approach can give rise to technical problems. Information on the timing of events is rarely collected with any degree of precision. Much data on job and fertility histories is collected retrospectively and one is doing well to get data accurate to the nearest month. Unfortunately where times are tied Cox's likelihood very rapidly becomes exceedingly complex with increasing numbers of ties. Note that the problem does not arise where the hazard function $\lambda_j(t)$ is specified up to a finite number of constants (i.e., is parametric).

The question of non-distinct failure times can be approached in a variety of ways, depending on whether the times are viewed as (i) tied discrete, (ii) tied continuous (iii) grouped continuous, or (iv) grouped discrete.

(*i*) *Tied discrete times.* Where times are discrete Cox (1972) defined the hazard as:

$$h(t) = P(T \leqslant t + 1 | T > t) \tag{11}$$

and proposed a discrete logistic model

$$\frac{h(t, \mathbf{Z})}{1 - h(t, \mathbf{Z})} = \exp(\beta \mathbf{Z}) \frac{h_0(t)}{1 - h_0(t)}. \tag{12}$$

At time t, the conditional probability that items $i_1 i_2 \ldots i_d$ fail from the risk set R is then

$$\frac{\psi(i_1)\psi(i_2) \ldots \psi(i_d)}{\sum_{k \in s(j,d)} \psi(k_1)\psi(k_2) \ldots \psi(k_d)},$$

where $s(j, d)$ denotes the set of all samples of d items from the set R.

This type of approach which involves conditioning out the unknown hazard is known as *conditional analysis*. In practice of course applications involving discrete *times* are rare, though it could be argued that a day represented a natural quantum in the study of (say) employment or education. In this case, there are structural factors (common school leaving dates, jobs terminating at the end of a week or a month) which mean that tied times are inevitable. However, this approach is frequently more relevant to discrete quasi-time scales whose sole function is to order sequences of discrete events, such a number of football matches constituting an unbeaten run or even number of repeat purchase of a brand commodity (Amemiya 1981).

Conceptualising process as occurring in discrete time with a logistic relationship between hazards however provides a convenient and easily computed approximation to the grouped continuous model described under (iii).

(*ii*) *Tied continuous failure times.* In practice of course, tied *continuous* times are rare, and could generally, at least in principle, be separated by a finer recording of times. Medical investigations frequently record times to the nearest day for this reason. The most realistic situation is generally to view times as *grouped continuous.*

(*iii*) *Grouped continuous failure times.* In analysing events which are recorded as having occured simultaneously, as in analysing anything else, it is important to be clear how these can have arisen. The exigencies of data collection generally mean that times are recorded to the nearest day, month or even year. For example the (GB) National Child Development Study measures employment history and family formation events by the month, in

its fourth 23-year-old sweep, so that, with an attained sample size of 12 538, there are inevitably many individuals whose recorded times are tied.

If the proportional hazards assumption holds and discounting censoring, the hazard for the interval and for the group with covariates $\mathbf{Z}$ is given by

$$1 - (1 - \lambda_i) \exp(\boldsymbol{\beta}\mathbf{Z}), \tag{14}$$

where $\lambda_i = \int_{t-1}^{t} \lambda_0(u)\mathrm{d}u$, $\boldsymbol{\beta}$ can be estimated by setting $C\{\hat{h}(t)\} = \alpha + \boldsymbol{\beta}\mathbf{Z}$, and where C is the complementary log–log transformation. Both this and the logistic approximation discussed earlier can be expressed in the general linear model form

$$Link\ \{h(t)\} = \alpha + \boldsymbol{\beta}\mathbf{Z} + \varepsilon, \tag{15}$$

where *link* is either the logit function or the complementary log–log function and ε is an error distribution (see McCullagh & Nelder 1983, Nelder & Wedderburn 1972.)

The logit link is perhaps more familiar and comprehensible to social science users, but the complementary log–log has the important advantage that the coefficient $\boldsymbol{\beta}$ is independent of the length of the time interval, and consequently the $\hat{\boldsymbol{\beta}}$ thus estimated is unbiased for the continuous time model. This property is not shared by the logistic link (Singer & Spilerman 1976) as not only the precise values of the $\boldsymbol{\beta}$ coefficient, but even the form of the relationship can change as the length of the interval changes (Myers et al. 1973). However, where the hazards in each interval are small the two methods give very similar results (McCullagh & Nelder 1983, Hutchison 1987).

(*iv*) *Grouped discrete failure times.* These have not been widely treated, though one example has been provided by Hutchison (1987). He postulates a discrete time process with a function $D(t)$ which he calls a dissatisfaction function

$$D(t) = \alpha + \boldsymbol{\beta}\mathbf{Z} + \varepsilon, \tag{16}$$

and that a Terminal Event occurs if $D(t) \geqslant$ some threshold. If this relation holds over a period containing a relatively large number of discrete time units, then an extreme value distribution (Johnson & Kotz 1970) holds and the parameters may be estimated by use of the complementary log–log as in the grouped continuous model above. Note that (16) is Linear rather than log–linear as compared to (12).

Censoring in grouped or discrete data. So far in this discussion we have ignored censoring. The majority of derivations involving non-distinct times effectively still ignore censorings by arbitrarily deciding that censorings occur uniformly before (Breslow 1972, Pierce et al. 1979) or uniformly after (Cox & Oakes 1984) terminal events with the same recorded failure times. If this is not considered satisfactory because of a large number of censorings, some adjustment may be made, generally by subtracting 0.5 of the number of censored observations from the risk set. (Thompson 1977, Holford 1976). Cox & Oakes (1984) give a precise likelihood for the situation where censorings and failures occure throughout the interval, and show that this is equivalent, for small hazards, to the actuarial estimator referred to earlier (Berkson & Gage 1950).

Underlying hazard function. Whereas the ML approach estimates both the dependence of β on covariates and the underlying hazard h_0, the PL approach conditions out h_0, so that it is not estimated. Various procedures are available to estimate h_0, the most widely used at present being that of Breslow (1974).

(*d*) *Other models*

Diekman & Mitter (1984) have proposed that they describe as a Sickle model to provide a realistic model of a process which starts at a low rate, climbs rapidly and tails off more gradually.

This paper has covered the more straightforward applications of survival analysis methodology. A companion paper, detailing with more elaborate applications, and outlining some possible future lines of development, will be published in the next issue of *Quality and Quantity*.

Acknowledgements

Particular thanks are due to Tony Ades and Andrew Pickles for encouragement and valuable comments on earlier drafts. Much of this work was done while the author was working with the National Child Development Study and the author is grateful to former colleagues at the National Children's Bureau for encouragement.

References

Allison, P.D. (1982). "Discrete-Time Methods for the analysis of event histories", *Sociological Methodology 1982*: 61–98.

Atkins, E., Cherry, N., Douglas, J.W.B., Kiernan, K.E. & Wadsworth, M.E.J. (1981). "The 1946 British birth cohort: an account of the origins, progress nd results of the National Survey of Health and Development", in S.A. Mednick & A.E. Baert (eds.), *Prospective Longitudinal Research*. OUP for WHO.

Baker, R.J. & Nelder, J.A. (1978). *The GLIM System*. Release 3. Oxford: Numerical Algorithms Group.

Barlow, R.E. & Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing*. New York: Holt Rinehart, Winsten.

Berkson, J. & Gage, R. (1950). "Calculation of Survival Rates for Cancer", Proceeding of the Mayo Clinic 25: 270.

Breslow, N.E. (1972). Contribution to discussion of paper by D.R. Cox. *J. Roy. Statistics Soc.* B34: 216–7.

Breslow, N.E. (1974). "Covariance analysis of censored survival data", *Biometrics* 30: 89–100.

Carroll, G.R., Hannan, M.T., Tuma, N.B. & Warsavage, B. (1978). "Alternative estimation procedures for event history analysis: a Monte Carlo Study", Technical Report No. 70, Laboratory for Social Research, Stanford University.

Cox, D.R. (1972). "Regression models and life tables", *J. Roy. Statist. Soc.* B34: 187–220.

Cox, D.R. (1975). "Partial likelihood", *Biometrika* 62: 269–276.

Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman-Hall.

Davies, R.B. (1983). "Duration dependence: a re-evaluation of the competing risk approach", *Environment and Planning A* 1057–1065.

Diekmann, A. & Mitter, P. (1984). "A comparison of the "sickle function" with alternative stochastic model of divorce rates", in A.R. Diekman & P. Mitter (eds.), *Stochastic Modelling of Social Processes*. Academic Press.

Duncan, G.J. & Mathiowetz, N.A. (1984). "A Validation Study of Economic Survey Data" Mimeo. Survey Research Centre, Institute for Social Research, University of Michigan.

Efron, B. (1977). "The efficiency of Cox's likelihood function for censored data", *Jr. Amer. Statist. Ass.* 72: 557–575.

Elandt-Johnson, R.C. & Johnson, N.L. (1980). *Survival Methods and Data Analysis*. New York: Wiley.

Flinn, C.J. & Heckman, J. (1982). "New methods for analysing individual event histories", *Sociological Methodology* 1982: 99–144.

Fogelman, K. (ed.) (1983). *Growing up in Great Britain*. Macmillan for the National Childrens Bureau.

Heckman, J. & Singer, B. (1982). "The identification problem in econometric models for duration data", in W. Hildenbrand (ed.), *Advances in Econometrics; Proceedings of World Meeting of the Econometric Society* 1980.

Heckman, J. & Singer, B. (1984). "Econometric Duration Analysis", *Jr. Econometric 63: 132.*

Holford, T. R. (1976). "Life tables with concomitant information", *Biometrics* 32: 587–597.

Hutchison, D.A. (1987). "Methods of dealing with grouped data: an application to drop out from apprenticeship", in R. Crouchley (ed.) *Longitudinal Data Analysis*. Aldershot: Avebury.

Johnson, N.L. & Kotz, S. (1970). *Distributions in Statistics, Continuous Univariate Distributions*. Boston: Houghton Mifflin.

Kalbfleisch, J.D. & Mackay, R.J. (1979). "On constant-sum models for censored survival data", *Biometrika* 66: 87–90.

Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Kaplan, E.L. & Meter, P. (1958). "Nonparametric estimation from incomplete observations", *Jr. Amer. Statist. Ass.* 53: 457–481.

Kay, R. (1977). "Proportional hazard regression models and the analysis of censored survival data", *Jr. Roy Statist. Soc.* C: 227–237.

Lagakos, S.W. & Williams, J.S. (1978). "Models for censored survival analysis: a cone class of variable-sum models", *Biometrika* 65: 181–189.

Lundin, F.E., Archer, V.E. & Wagoner, J.K. (1979). An exposure-time response model for lung cancer mortality in uranium miners: effects of radiation exposure, age and cigarette smoking", pp. 243–264 in Breslow & Whittemore (eds), *Environment and Health*. SIAM.

McCullagh, P., Nelder, J.A. (1983). *Generalised Linear Models*. Chapman & Hall.

Martin, J. & Roberts, C. (1984). *Women and Employment: a Lifetime Perspective*. HMSO.

Myers, M., Hankey, B.F. & Mantel, N. (1973). "A logistic-exponential model for use with response-time data involving regressor variables", *Biometrics* 29: 257–269.

Nelder, J.A. & Wedderburn, R.W.M. (1972). "Generalised linear models", J. Roy. Statist. Soc. A 135: 370–384.

Nelson, W. (1972). "Theory and applications of hazard plotting for censored failure data", *Technometrics* 13: 201–201.

Oakes, D. (1977). "The Asymptotic information in censored survival data", *Biometrika* 64: 441–448.

Osborne, A.F., Butler, N.R. & Morris, A.C. (1984). *The Social Life of Britain's Five Year Olds*. Routledge & Kegan Paul.

Palmore, E.B., Fillenbaum, G.G. & George, L.K. (1984). "Consequences of retirement", *J. Gerontol.* 109–116.

Pierce, D.A., Stuart, W.H. & Kopecky, K.J. (1979). "Distribution free regression analysis of grouped survival data", *Biometrics* 34: 785–793.

Plewis, I.F. (1985). *Analysing Change*. Wiley.

Rai, K. & van Ryzin, J. (1979). "Risk assessment of toxic environment section using a generalised multi-hit response model", pp. 99–117 in Breslow & Whittemore (eds.), *Environment and health*. SIAM.

SAS Institute Inc. (1982). SAS Users Guide. SAS Institute Inc. Cary N.C.

Shepherd, P. (1985). *The National Child Development Study: an introduction to the origins of the Study and the methods of data collection*. Working Paper No. 1 Mimeo, NCDS User Support Group, City University.

Thompson, W.A. (1977). "On the treatment of grouped observations in life studies", *Biometrics* 33: 467–470.

Tuma, N.B. (1980). *Invoking RATE*. Mannheim: ZUMA.

Tuma, N.B. (1982). "Nonparametric and partially parametric approaches to event history analysis", *Sociological Methodology 1982*: 1–60.

Tuma, N.B. & Hannan, M.T. (1984). *Social Dynamics: Models and Methods*. Academic Press.

Turnbull, B.W. (1974). "Nonparametric estimation of a survivorship function with doubly censored data", *J. Amer. Statist. Ass.* 69: 169–173.

Wall, W.D. & Williams, H.L. (1970). *Longitudinal Studies and the Social Sciences*. Heinemann.

Williams, J.S. & Lagakos, S.W. (1977). "Models for censored survival analysis: constant sum and variable models", *Biometrika* 64: 215–224.

Zippin, C. & Armitage, P. (1966). "Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter", *Biometrics* 22: 665–672.

Quality & Quantity 22: 255–278 (1988)

Event history and survival analysis in the social sciences
II. Advanced applications and recent developments

DOUGAL HUTCHISON
National Foundation for Educational Research in England and Wales, The Mere, Upton Park, Slough, Berks, SL1 2DQ

Abstract. A previous paper (Hutchison, 1988) in this journal has provided an introduction to the basic concepts of survival and event history analysis, originally developed in medical research, econometrics and engineering, and argued the case for their wider application in the social sciences. This paper introduces some further complications that the researcher is likely to meet, and offers some guidelines for handling problems that arise in applying such methods to the highly complex social situations involved.

Contents

I. Time-varying covariates

The description in Hutchison (1988) has envisaged independent variables which remain constant over time. With the exception of a few characteristics of the individual such as date of birth and, generally, sex, this is not a reasonable assumption, and, fortunately, it is not one which is necessary for methods of this kind to work.

In his original paper Cox (1972) assumed that the covariates varied over time in the proportional hazards model. Indeed he allowed the inter-group proportionality also to vary over time, though to make any sense of this, it is necessary for the variation to follow some functional relationship, such as linear, with time. While time-varying covariates differ conceptually depending on how they arise, there is relatively little difference in treatment. In increasing order of complexity, we can consider:

(1) changing environment – external covariates;
(2) changing status – independent internal covariates;
(3) complete history of the process to date. Cox & Oakes (1984, 8.1) put forward a different classification of time dependent covariates, based more on statistical properties, but we feel ours is perhaps more relevant to social science users.

1. *Changing environment.* Frequently the circumstances in which a process takes place will alter in ways that are effectively beyond the control of the individual or unit concerned, or to which he or she is only marginal. For example the government may change, or the unemployment rate could increase. Other such "external" covariates could include the size of the risk set or the number of terminal events in the group so far. These are examples of *evolutionary* covariates – see (3) below.

2. *Changing status.* It can also happen that the status or characteristics of an individual may change in a way that we feel could be a cause or modifier of the terminal event, but not vice versa; or that the change in status may be of a different nature from the TE. For example if we are comparing the health over time of married and single people, or employed and unemployed then not everyone will remain in the same status over any substantial length of time. In this situation, when an individual becomes married, he or she is transferred from the "single" risk set to the "married". Such a procedure of course takes no account of individual continuities in unobserved heterogeneity, that is, an individual's characteristics that we cannot summarise from knowledge of measured covariates such as (say) their sex, age, and social class. This topic is discussed more fully in Section III below.

   Another possibility is that statuses may be combined to form a composite dependent variable: for example in studying economic activity for women one possibility would be to treat marital status as a time-varying covariate, but another possibility would be to decide that working while married involves such different attitudes and commitments that it makes better sense to view the outcome space as consisting of (say) the 4 situations:

   (a) single working
   (b) single non-working
   (c) married working
   (d) married not working.

3. *Evolutionary covariates* (alias the story so far). The two methods outlined so far make a fairly clear distinction between cause and effect. An alternative view comes closer to a perspective in which all of an individual's experience so far is contained in his or her present state and can influence the occurrence of a TE. Such evolutionary covariates can include the external and independent internal covariates discussed earlier and also the entire history of the process, including previous occurrences of the terminal event. Thus in a study of unemployment the size of the risk set, the number of times previously unemployed, or the duration of previous unemployment could be included as covariates.

Handling time-varying covariates

The method of handling time-varying covariates depends on whether they are considered continuous or categorical, and whether semi-parametric or fully parametric approaches are used.

Of these the simplest is the semi-parametric approach with categorical variables. Where an individual changes status, then he or she is simply transferred from one risk set to another at the appropriate time. This is in fact the equivalent of considering the individual as providing two (or more) censored observations, the first right censored at the point of change and the second left censored at the same point. Similarly this is how fully parametric estimation can be treated, with the second (and subsequent) observations representing the probability of surviving from the time of the change in the covariate situation to the time of a terminal event or until censoring by leaving the study or changing risk group. Some slight complications arise with the accelerated failure time model since the effective time of left censoring is dependent on the parameter β to be estimated, and so an iterative procedure, such as the EM algorithm (Dempster *et al.*, 1977), may be required. Where there are a large number of observations of some rapidly changing influence, as could occur where monthly or more frequent unemployment figures are included in an analysis, then technical and other problems may arise. As frequently occurs when maximum likelihood methods are used for any but relatively simple applications, computing rapidly becomes onerous. A number of possible simplications have been suggested to reduce the computational burden. Covariates may fixed as having their values at the start of a spell, or the spell may be broken up into a number of subspells where the value is considered constant, or averaged over the spell. That these short cuts are potentially hazardous is shown by Flinn & Heckman (1982) and Heckman & Singer (1984) who analyse problems using these methods and find that the results may be seriously biased compared with using a

model where the exogenous variables are allowed to vary freely over the spell. It is generally best to fit a model using all the known time variation compatible with efficient use of the computer, or if possible, to represent the time-variation by a functional form, for example $f(t) = at$, or

$$f(t) = at + \sin(bt),$$

the latter possibly being appropriate as an approximation to seasonal fluctuations. Tuma and Hannan (1984) suggest if there are a number of observations of a continuously-changing variable, one may represent the process by a linear function of time between pairs of measurements. They show that a variable which changes in this form gives an exponential form for the corresponding rate.

Time-varying covariates and time dependency of hazard

Many factors which we cannot or do not observe change over time and affect the rates. Thus individuals will mature over time, or cumulative exposure to some toxic substances will build up. It may be that some monotonic function of time will provide a reasonable working approximation. It is well to remember that it is unlikely that time in itself will be a causal factor for many processes, but that it represents a development which is generally closely related to time. Thus change in an underlying covariate may be the "explanation" of a time dependent process.

Multiple time scales

Time can be used as a proxy for a number of processes – physical ageing, progression of disease, work experience, pressure of social norms. We may consider that more than one of these is important in a process, and this can mean that where this occurs they are highly correlated. For example, we may feel that both age and work experience, or age of mother and age of youngest child affect a decision relating to female economic activity, or that age of patient and length of time since contracting a disease will affect mortality.

To handle such questions we can avail ourselves of the fact that time may enter into the specification of the hazard process in two ways, firstly as the ordering quantity t in $h(t)$ and secondly in time varying covariates $\mathbf{Z}(f(t))$. We write $\mathbf{Z}(f(t))$ rather than $\mathbf{Z}(t)$ to indicate that it is not necessary for the two times to be measured on the same or even proportionate scales. Indeed with partial likelihood it is not necessary for time t in $h(t)$ to be measured on a scale at all, merely that the observations be ordered.

If one variable is considered to affect the process on a more rapid time-scale or to be more central to the process then it makes sense to make that the ordering variable and to include the other in some other way. For example moving home is affected both by the time an individual has stayed at that abode and by their age. We could stratify age into fairly broad bands of five or ten years and include age as a covariate, either constant at its value at the start of the period of residence, or as a time-varying covariate changing as an individual changes from one band to another, though see the section on *Handling Time Varying Covariates* above.

Another area investigated could be the time spent out of the labour force by women while having a family. A very important determinant is the age of youngest child and the researcher could include a time-varying covariate to deal with this, restarting the clock with second or subsequent births. Some care may be needed in interpreting time dependency since there may be a close relationship between time scales: for example there is a linear dependency between time out of the labour force and age of youngest child for those women with only one child. In this the common sense of the statistician is probably the best guide.

A different approach is that of Farewell & Cox (1979) who advocate using more than one time scale and trying to interpret the weighted result.

II. Competing risks

The discussion so far has been nominally in terms of only one terminal event destination, for example death or quitting a job, to quote two somewhat disparate phenomena. This may be lumping together events which proceed in quite different ways, such as deaths from different causes. In examining deaths in a population we might wish to combine a factor dropping away relatively quickly over time, for example to represent the effects of an epidemic, together with a more gradually time-trended relationship representing the usual effects over time (Barlett, 1978; Pierce et al., 1979). This would be principally useful if one were unable to distinguish between different causes of death.

However, one may consider it of interest to distinguish between different types of terminal events. A sociologist, investigating the process of leaving a job, would consider that leaving for a promotion elsewhere is very different, at least from the point of view of the worker, from being sacked.

The methods in this section deal with the situation in which an individual is prey to a number of possible types of terminal events, but can only

experience one, or, equivalently, we only observe the first. The case where an individual may be observed to experience more than one type of terminal event is described in other sections.

A destination – specific hazard function $h_v(t)$ can be defined for the vth cause of death similarly to the general hazard $h(t)$

$$h_v(t) = \lim_{\Delta \to 0+} \left\{ \frac{P(t_v \leqslant T \leqslant t_v + \Delta | t_v \leqslant T)}{\Delta} \right\}.$$

$f_v(t)$ can be defined similarly. The destination-specific hazards $h_v(t)$ may be summed to give the general hazard $h(t)$. (Prentice et al., 1978; Gail, 1975).

(a) *Independent risks approach.* The situation is described as one of competing risks (Cox & Oakes 1984). If the risks are independent then the models may be estimated by treating first one destination as TE and including the other destinations with censored observations and then vice versa. We used the expression "nominally" at the start of this section since our treatment has included censored observations. Censoring, where the censoring process is independent of the risk, may be considered as another destination and thus results about dependent or independent censoring are applicable.

(b) *Proportional hazard models.* At the other end of the continuum one can assume that

$$h_v(t) = \alpha_v h(t)$$

for all t. This is somewhat confusingly known as the proportional hazards approach by analogy with the model for different groups on the independent variable. The assumption of proportional hazards in the dependent variable is conceptually the same as that of destination independence (Davies, 1983), widely used in social sciences. The assumption of destination independence may be used to divide modelling of a stochastic process into separately analysed subprocesses of timing and outcome (Wrigley, 1980). Ginsberg (1978) has produced a test for destination independence, refined by Davies (1983), which involves testing the effect of constraining to equality some parameters in a fully parametric representation of the process. Cox & Oakes (1984) show how the assumption of dependent variable proportional hazards may be tested in the semi-parametric model. Graphical methods as outlined in the section of this paper on independent variable proportional hazards are also applicable.

Dependent failure times

Very little has been published on situations in between the two extremes of proportional and completely independent hazards. Davies (1983) has given an example where two hazards are both of the same (Weibull) form $\lambda_i \alpha_i t_i^{\alpha_i - 1}$ and the effect of constraining the constant x to be equal. Other examples are given in Nadas (1970, 1971), David & Moeschberger (1978), Moeschberger (1974) and Lagakos & Williams (1978).

In some instances the information on type of Terminal Event may not be helpful. One possible example could be in a study of marital breakdown, where the researcher has information on which partner petitioned for divorce. It is quite conceivable that in many cases either of these processes, rather than indicating the more dis-satisfied partner, would simply represent an acknowledgement that the marriage was at an end, and that the decision over which party would instigate proceedings would be a matter of convenience. In a situation of this type the soundest analysis strategy might well be to group the two events together.

The discussion in this section generally has been in terms of a limited range of destinations which can be described by a categorical variable equalling 1 or 0 for the possible destination states. Cox and Oakes (1984) propose that this could be extended to encompass real valued destination states (e.g., length of time off sick when an individual is ill) or vector-valued destinations: for example one might look at illnesses in a family with the failure type being described by a vector combination of individual affected, and type of illness.

III. Overdispersion and unobserved heterogeneity

The equations (12) earlier propose that *Link* (h) where h is the hazard, is a linear function of certain background covariates. Obviously this cannot entirely determine the duration till the occurrence of a TE of the individual. The indeterminacy remaining is acknowledged in that the hazard is simply the parameter driving a stochastic process which gives rise to a probability distribution of times. However, one does not expect that everything relevant about an individual would be known from (say) their age, sex and social class. Yet this is effectively what is being said if these are the only covariates used. The situation can be improved by increasing the information available, either by a finer classification of age and or social class for example, or by the inclusion of other covariates, such as years of education, income or

measure of parental social status. However, one rapidly runs into problems with numbers in cells, and still other covariates suggest themselves – personality, qualifications, school attended and so on. The traditional approach to this in social research has been to investigate the statistical significance of such refinements in prediction of the dependent variable and to stop when further refinement brings no statistically significant reduction in some appropriate statistic, most recently the deviance. This of course suffers from the drawback that one can only include data which is available. Tautological though this may sound, it means that one cannot take account of characteristics which have been inadequately measured, or completely omitted. Further, the type of influencing factor that cannot readily be recorded by the data collection instrument will also be omitted. This is potentially a more serious problem, since researchers can tend to lose sight of influences which they cannot assess.

Ignoring within-category differences can lead to misleading conclusions, as a problem in migration theory will make clear. It has been observed that the longer a unit (individual or household) has been staying in the same place, the less likely it is to move during the next unit of time: this is described as cumulative intertia (McGinnis, 1968). An alternative explanation postulates heterogeneity in the population so that some individuals are more likely to move at any time than are others. Consequently as time goes by more of the former have already moved, and the population increasingly consists of the stick-in-the-muds and thus of course moving rates decrease. This argument is summarised somewhat heuristically here but a more mathematical treatment is widely available, for example in Flinn & Heckman (1982). A very much oversimplified version of this forms the basis of the mover/stayer model of Goodman (1961). Davies & Pickles (1985) show that this type of process gives rise to a spurious correlation between exogenous and endogenous variables which biases the estimated coefficients.

It is instructive to treat the topics of this section within the framework of the General Linear Model (see e.g., McCullagh & Nelder, 1983), and departures therefrom.

Applied to the proportional hazards survival model:

t is the time parameter
$\Pi(t)$ is the hazard at t
$Y(t)$ is the observed number of TEs
$n(t)$ is the size of the risk set
$\mathbf{Z} = \{z_j\}$ are the covariates.

Then the GLM may be stated as follows: $y_i \sim B(n_i, \Pi_i)$ the binomial distribution with size n_i and probability Π_i

$$\text{Link } (\Pi_i) = \eta_i = \alpha_i + \sum_j Z_{ij}\beta_j,$$

where α, β_j are coefficients to be estimated and η_i is the linear predictor. Link (Π_i) is some function of Π_i, for example the logistic

$$\log\left(\frac{\Pi_i}{1 - \Pi_i}\right) = \text{logit } (\Pi_i).$$

This model contains a stochastic element already in the form of Y. Note however, that the binomial distribution is the distribution of scores from trials with probability Π_i so it implies that all n_i elements are equal. Thus for a given cell, the probability Π_i, the linear predictor η_i and the covariates Z_i are considered fixed. These constraints may be relaxed in the interests of a more realistic representation of a process.

(i) The variance for binomial proportion Π_i is $n_i\Pi_i(1 - \Pi_i)$ Where the probabilities Π_{ij} of the individual elements vary, but the overall proportion is Π_i then the variance can be written in the form $\sigma^2 n_i\Pi_i(1 - \Pi_i)$ which can be handled using the concept of quasi-likelihood (Wedderburn, 1974).

Generally σ^2 is great than 1 and this sitution is referred to as overdispersion. McCullagh and Nelder (1983) describe how σ^2 may be estimated.

(ii) In an annual panel study of the labour force participation of married women, Heckman & Willis (1977) proposed that individual participation was distributed according to a beta-binomial distribution with population mean Π, and investigated how much participation rates were correlated from year to year. See also Davies et al. (1982). The beta-binomial was employed since it has a wide range of forms which means that many situations can be modelled, and as the authors freely admitted, for mathematical convenience.

(iii) The Heckman & Willis approach involves taking account of heterogeneity by including a random effect in Π_i. A closer representation of the process could involve a person-specific effect in the characteristics of the individual, that is by including a random effect in the linear predictor η_i

$$\eta_i = \sum_j \beta_j Z_{ij} + \varepsilon_i \qquad (19)$$

or

$$\eta_i = \left(\sum_j \beta_j Z_{ij} \right) * \varepsilon_i. \tag{20}$$

Heckman & Willis (op cit) do in fact set up a model of the first type in their economic-theoretic development but this is not continued in the statistical model used.

Allison (1982b) distinguishes between what he refers to as *external* models of the form

$$\Pi_i = \text{link}\,(\eta_i) + \varepsilon_i \tag{21}$$

and *internal* models of the form

$$\Pi_i = \text{link}\,(\eta_i + \varepsilon_i) \tag{22}$$

Flinn & Heckman (1982) develop an "internal" approach which does take account of unobserved heterogeneity in the characteristics of the population. The importance of allowing for unobserved heterogeneity is underlined by three observations: firstly, Heckman & Borjas (1977) argue that most heterogeneity is the result of unobserved components; secondly, that being forced to be aware of the shortcomings of their models by including unmeasured heterogeneity should induce a healthy humility in most social theorists; thirdly, introducing an allowance for heterogeneity can alter quite seriously the coefficient estimates for independent variables, and time dependency of the hazards (Heckman & Singer, 1982). This last is quite surprising given that intuitively one might merely have expected some dilution of the effects (see also Allison, 1982). Davies & Pickles (1985) argue that this may be due to correlation of omitted variables with exogenous or endogenous regressors. It is also likely that the precise form of the error distribution changes over time as the more extreme elements are more likely to experience TEs.

The situation here is in contrast with that in the ordinary least squares models used in panel analysis of continuous outcomes. Unobserved heterogeneity has received less attention here since the model states that

$$y_i = \boldsymbol{\beta}\mathbf{Z}_i + \varepsilon_i, \tag{23}$$

where effects due to population heterogeneity, unspecified causal mechanisms, sampling variation and measurement error in the dependent variable combine to produce ε_i, assumed to be distributed $N(0, \sigma^2)$. The statistical treatment in OLS thus takes heterogeneity into account without distinguishing it from the other elements in the random term.

Given that constant but heterogeneous hazards can mimic the effect of a time dependent hazard, it is natural to ask whether these can be distinguished statistically. While Heckman & Singer (1984) give some conditions for making such decisions, these are at present difficult to apply. One simple rule is that heterogeneity which remains constant over time can only cause an apparently decreasing rate. Thus an increasing observed hazard function is incompatible with a constant rate, and while there may indeed be heterogeneity with an increasing hazard, correcting for this will cause further increase over time in the hazard. The matter is not particularly relevant since the Markov assumption is simply one of mathematical convenience in most social sciences fields and there is no reason to expect a constant hazard rate.

If there is considered to be unobserved heterogeneity the underlying period hazard (probability of an event in the interval $(t_i, t_{i+1}]$) may be considered as varying over the population according to a distribution to be estimated. Heckman & Singer (1984) give as examples the normal, log-normal and the gamma distribution. These distributions are chosen partly because they are flexible enough to cover a wide range of possible underlying probability distributions and partly for mathematical convenience, and it is frequently difficult to justify a precise distributional form. In particular, discrete probability masses, for example, at $P = 0$ (those who never do something) or at $P = 1$ are not well accommodated. This is one reason why the "non-parametric" procedure proposed by Heckman & Singer (1982) which does not assume any distribution for unobserved heterogeneity but simply assumes that it may be satisfactorily approximated by a number of mass points, is attractive.

The following equation gives an extension to the proportional hazard model

$$\text{link}\,(P_{jt}) = f(t) + g(\mathbf{Z}_j) + \gamma_t \varepsilon_j, \tag{24}$$

where ε_j (density $\mu(\varepsilon)$) is the unobserved heterogeneity, an individual specific element, assumed constant over time; t is time, $\mathbf{Z}_j$ covariates, and functions f and g and scalar γ_t are to be estimated. P is the probability of a terminal event at time t. Since there are as many ε_j as there are individuals, further information is needed on ε_j. This could be

(a) by means of restrictions so that ε_j has a specified functional form, or
(b) by assuming that its distribution is concentrated at a finite (and relatively small) number of mass points or can be satisfactorily approximated by a distribution of this form (Heckman & Singer, 1982).

This is conceptually similar on the one hand to an extension of the mover/stayer model to include more than two groups, and on the other hand to the well known process of approximating the effect of a continuous independent variable by grouping it into a number of categories.

Alternatively, (c) more information on the ε_j can come from further episodes for the same individuals. (c) is considered later, in the section on multievent histories.

One drawback to this approach is that in practice $f(t)$ must be a parametric function for it to be possible to estimate from one episode the heterogeneity term using the non-parametric methods outlined (Elbers & Ridder, 1982). This unfortunately rules out the Cox (arbitrary hazard) model unless it is possible to specify a parametric form for the error.

IV. Measurement error and latent variables

Measurement error refers to the situation where the measurement of a variable is fallible and the observed value is considered to be the sum of the true value plus an error term.

Fuller and his co-workers (Warren et al., 1974; Fuller & Hidiroglou, 1978) have provided the theoretical background, and a program (SUPERCARP, Hidiroglou et al., 1979) for handling the situation in panel studies where an independent estimate is available of the measurement error variance, or the reliability, which is the ratio of measurement error variance to total variance. Examples of the use of this approach include Hutchison (1980).

Joreskog and his co-workers (see e.g., Joreskog 1979) have provided theory and program (LISREL-VI, SPSS inc. 1984) which deals with a formally similar situation where one or more indicator variables are considered to be imperfect indicators of underlying or latent variables:

$$\left.\begin{aligned} X_1 &= \lambda_1 L + \varepsilon_1 \\ X_2 &= \lambda_2 L + \varepsilon_2 \end{aligned}\right\}. \tag{25}$$

This approach does not assume an independent estimate of measurement error variance but rather obtains estimates for the ratio of the coefficients λ_1, λ_2 and of Var (Σ_1) and Var(Σ_2), in addition to the panel regression coefficients, from the structure of variance and covariances of the data. Examples of the use of this techniques are given (e.g.) in Joreskog op cit.

While formally similar, these two approaches use different conceptions of error and may well give somewhat different results, in particular where the measurement error on two indicators is correlated.

In survival-type analysis the situation is much less advanced. There is little research which handles measurement of error in the *independent* variable or latent variables (though see Prentice, 1982; Clayton, 1986). This may well be because the basic theory for covariance analysis even with infallable variables in survival analysis is much more recent.

V. Event histories – series of events

In general so-called event history techniques are better described as event analysis techniques. Writers, for example, Tuma (1980) et al., deal with multiple events for an individual by making the unit of analysis the person-spell rather than the person. Thus, if an individual was married, divorced, and then remarried (and still remarried at the time of data collection) then that individual would contribute 3 units to the analysis:

(1) time to divorce;
(2) time to remarriage;
(3) time remarried (censored).

This is somewhat unsatisfactory, as these methods ignore the connection between the three units of analysis. This is not fatal, as it still gives rise to unbiased estimates of beta parameters in (for example) the proportional hazards model, provided that unobserved person-specific affects are uncorrelated with the independent variables, but since the observations are not independent techniques are statistically inefficient and the estimates of variance are biased.

This implies that one assumes that there is no carry-over from one event to the next and one can consequently ignore the multiplicity of an event. This type of approach has been widely used in engineering-type applications (Barlow & Proschan, 1975) under the name of *Renewal theory*. However, while such an approach is all right for the replacement of light bulbs, it is in general not satisfactory for processes involving human beings, since it fails to take account of ageing or learning processes which the individual has experienced in the interim. Second marriages are not the same as first marriages, nor do the births of second children share the same concomitant circumstances as those of first children. To quote from Herodotus: "It is not possible to step twice into the same river".

One possibility might be to elaborate the modelling so that (say) second marriages were analysed completely separately from first, and remarriages

of divorced individuals were considered to be different from that of those who had been widowed. However, such an approach could rapidly produce a large number of small cells if there were a sequence of relevant events, and particularly where there are a number of possible outcomes of any process. Other methods will generally be required and we discuss this below.

Unobserved heterogeneity. Multiple spell data can assist in taking account of unobserved hetergeneity. If one assumes that such heterogeneity remains constant over spells, then one can use repeated event information to partial out unobserved heterogeneity, in much the same way as the Cox Partial Likelihood approach conditioned out the unknown base hazard functions, and gain an estimate of true duration dependence. Of course, in practice, this is unlikely: for example, individuals embarking on their second marriage, after a divorce, would be expected to have either learned from or to become embittered by the experience of breakdown. A more useful approach in many instances might be autocorrelated heterogeneity:

$$a(t + 1) = k^*a(t) + e(t), \tag{26}$$

where $e(t)$ is a random term, and $a(t)$ is the measure of heterogeneity at time t.

No authors use this precise model in event history analysis, though Joreskog and his co-workers use it in panel-type data. Flinn & Heckman (1982) describe a model for error with a certain similarity in the short run:

$$a(t) = k^*C + e(t). \tag{27}$$

The lagged autocorrelation suggests a process of the individual heterogeneity being modified by events experienced. Strictly of course, an individual's entire history is relevant to their action or decisions, and any part of it may be crucial, such as chance meetings in the street, resurgence of long-forgotten childhood fears or ambitions. Nevertheless some simplification is necessary if statistical methods are to be used at all.

Thus the researcher could assume that the probability of an event occurring at a given time of or, more likely, its logit, was a linear combination of a number of factors and that (say) the same combination held for those had and had never been unemployed, except for a dummy variable indicating whether the individual had ever been unemployed. This assumption could of course be tested against the alternative that a quite different mechanism was involved by comparing the overall fit under the alternative hypothesis. One also assumes that the quantities to be included in the model gave a sufficient picture of the process. This is less easy to assess, but some impression can be gained from the overall likelihood ratio.

More generally, one can allow for the history of an individual by including an independent variable relating to the occurrence of duration of previous states (Flinn & Heckman, 1982): these are described, unsurprisingly, as *occurrence dependence* and *lagged duration dependence*.

Taking account of an individual's history can give rise to problems of identification. Unobserved heterogeneity and history dependence can produce similar effects. Thus, if we found that employees who had been made redundant once were more likely to be made redundant again, we could postulate that the least competent workers were those made redundant, or that they worked in a less secure sector of the job market (heterogeneity) or that the experience of being made unemployed was such a blow to their self-esteem that they became less effective workers subsequently (occurrence dependence) or that the time out of a job meant a lowering of intellectual or physical capital, with a corresponding reduction in competence (duration dependence). Similar considerations could apply to accident proneness or to the remarriage of divorced individuals.

Individual heterogeneity at any given time, and the individual's history so far, are to some extent different sides of the same coin. On the one hand, much of what renders an individual different must be due to his or her cumulative experience, while on the other the heterogeneity which is not a product of past experience may well have manifested itself in the events of an individual's history. Thus one would expect that including a number of items of past history (where a large enough number of historical covariates could be introduced) would have an effect virtually the same as formally allowing for unobserved heterogeneity.

Which time scale? Particularly if a process occurs by moving through a series of intermediate events it may be difficult to decide from where the clock starts. Thus when Kay (1984) compares two cancer treatments considering the progression from incidents to remission, local or distant recurrence and death, he considers whether measuring the time from recurrence or first occurrence is the more appropriate time scale.

Other topics. In considering sequences of events it may be important to consider the order of events: taking a deep breath and jumping into the water is not the same as jumping into the water and taking a deep breath. If one found that independent variables affecting (say) having a child were different depending on whether the woman concerned was married beforehand, then it would be reasonable to infer that this was an important influence in investigating such an event. Similarly the order of events may be an important consideration in deciding the effect of such events viewed as historical independent variables: Students who marry before graduation may well be less successful than those who complete the course first.

VI. Associated failure processes

So far we have considered failure processes one at a time: if there is an association between the failure times of two individuals, this is considered to arise because of observed covariates in common. For example, if there was a male–female difference in hazard, then the failure history of two females taken at random could be expected to be more similar than that of similar male–female pair. However, we now extend this to cover association arising from unobserved heterogeneity of proneness or "frailty". This would arise in studies of breast cancer in sisters, or to give a social science example since most of the theorical development has been medical, studies of the occurrence of a particular behaviour where several pupils per class are considered. There is an analogy here with the literature on multi-level component analysis (Aitkin & Longford, 1985; Goldstein, 1986). Alternatively the association could arise between two different processes happening to the same individual as this would give rise to a much closer degree of similarity than would be accounted for by measured covariates, such as age, sex and social class. This type of approach where the association is caused by unmeasured heterogeneity is sometimes described as a correlation-type approach, in contrast to techniques in the previous section which may be described as regression-type approaches where occurrence of one event directly influences the occurrence of a second event (Armitage, 1985). Thus in studies of recidivism, we would expect to find that individuals who had committed one crime were more likely to commit a second and a "correlation" explanation could suggest that the individuals concerned were "criminal types" whereas a "regression" explanation could suggest that an individual once convicted might be more likely to have a second conviction because other opportunities were closed, because prison had hardened them or because police were more likely to suspect an already committed individual.

Linear model

Given the focus in survival analysis on the proportional hazards model, it is natural to allow for association in hazards. If (T_1, T_2) denotes the bivariate survival time with hazard function $(\lambda_1(t), \lambda_2(t))$ then we assume that

$$\lambda_k(t|\omega) = \lambda_k^0(t) \exp(a_k\omega), \quad k = 1, 2, \tag{28}$$

where ω is the unobserved common covariate for example, the effect of the two processes occuring to the same individual, $\lambda_k(t/\omega)$ are the hazard rates

conditional on ω, $\lambda_k^0(t)$ are unknown baseline hazard rates, different for each of the two processes. As usual the covariate factor is in exponential form to ensure that the resultant hazard must be positive. However, while the process is proportional hazard conditional on ω it is not proportional hazards in general unconditionally, thus losing the desirable properties of the PH model.

A convenient approach to analysis of the model (28) above comes from integrating over time and taking logs to give

$$\log\left(\int_0^t \lambda_k(s)\,ds\right) = a_k\omega + \log\left(\int_0^t \lambda_k^0(s)\,ds\right), \tag{29}$$

or, where

$$Y_k = \log\left(\int_0^t \lambda_k^0(s)\,ds\right)$$

have the linear functional relationship

$$Y_1 = a_1\omega + \varepsilon_1$$

$$Y_2 = a_2\omega + \varepsilon_2,$$

where ε_1, ε_2 are independent random variables with (minus) extreme value distributions

$$\Pr(\varepsilon > x) = \exp(-\exp(x)).$$

Note that we are considering different processes so a_1 and a_2 are not equal and the transformations of T_k to Y_k differ for each k. This model can be made rather more general by assuming a different error distribution: for example, a log-normal or log-logistic error distribution where $Y = \log T$ would give an accelerated failure time distribution. This linear model approach was used by Cuzick (1982) and Wu (1982).

An alternative model for correlation was put forward by Clayton (1978) who required his model to fulfil these requirements:

(i) It should be relatively straightforward to estimate the association parameter given censored observations for both covariates.
(ii) The effect of one variable on the other should be expressible as a constant ratio of age specific rates.

(iii) The model should be symmetrical in the two variables. This gives the model

$$\frac{\lambda_1(s_0|t = t_0)}{\lambda_1(s_0|t > t_0)} = \frac{\lambda_2(t_0|s = s_0)}{\lambda_2(t_0|s > s_0)} = \vartheta \tag{31}$$

This model is thus a specialisation of the proportional hazards model. It has another interpretation in that

$$h(s, t) = \vartheta.h_1(s|t > t_0).h_2(t|s > s_0). \tag{32}$$

Clayton & Cuzick (1985) state that this model is not valid for negative association.

This this measure is an appropriate measure of association can be seen by considering either form (30) or (31). For example the second form shows that the probability (density) of the two events is less than the product of the two probability (densities). There appears to be no nomenclature to distinguish the two models but it may be helpful to describe them as the common heterogeneity or linear model and the shared odds or Clayton model, respectively.

Estimation of these models is somewhat more difficult then the individual case, most models for both cases involving the estimation of the integrated base hazard function

$$\Lambda_0(t) = \int_0^t \lambda_0(t).$$

At least theoretically this is fairly straightforward where $\lambda(t)$ is a parametric function. Clayton (1978) gives a likelihood for the Weibull and piecewise exponential cases of the Clayton model while Wu (1982) indicates how estimation may be carried out for the linear model.

When the basic hazard function is completely unknown, estimation of the bivariate model is more difficult than for the univariate case estimated by Cox regression. Rank tests asymptotically efficient for the linear model have been derived by Cuzick (1982) and Wu (1982): these gives a test based on generalised Savage scores. Fully efficient tests for both models based on adjusted generalised Savage (log-rank) scores are given by Clayton & Cuzick (1985) who state that the computations involved would appear to be conveniently carried out in GLIM. Other approaches are based on a generalisation of the partial likelihood approach (Clayton 1978) and Kendall's rank test, but these have drawbacks (Oakes, 1982; Clayton & Cuzick, 1985).

The Clayton and Cuzick approach however, while the most satisfactory statistically, suffers from the drawback that it cannot handle negative association. This means that further work is needed to model a failure process which involves competition for resources, or where termination arises as a result of near-exclusive processes, for example individual leaving a job either because they were sacked or because they were promoted.

In some circumstances (Holt & Prentice, 1974; Woolsen & Lagenbruch, 1980) it is appropriate to assume that the basic hazard function is the same for each member of the pair, and the hazard function for each pair is a constant multiple for a basic hazard function

$$\lambda_{ik}(t|\omega) = a_i\lambda^0(t) \exp(\beta\omega) \tag{33}$$

for the kth member of the ith pair (see Holt & Prentice, 1974; Clayton & Cuzick, 1985). This model could apply in an educational experiment where members of matched pairs are allocated at random to an old or new teaching method and the time taken to master a particular task is compared.

VII. Discussion and conclusions

Recent developments in maximum likelihood theory are enabling greater flexibility and verisimilitude in the application of quantitative methods to social processes. While the basic theory of maximum likelihood dates from 1929, it is only the dramatic increases in power and availability of computer hardware and software that have enabled ML techniques to change from statistical curiosities to central elements of statistical practice. One such set of approaches is variously known as survival analysis, event history analysis or reliability theory, depending on whether it is used in biomedical, social or econometric, or engineering applications. To date it has been most widely used in medicine, engineering and econometrics but there is wide scope for application in the social sciences.

Survival and event history analysis have been developed to investigate how long elements remain in a state, and in particular for use in the situation where a sizeable proportion of the elements had not yet left the state: thus doctors wanted to assess treatment for affecting sick people before they all died, and engineers wanted to be able to estimate how long (say) lightbulbs lasted without burning them all out.

In the social sciences correspondingly observations are *incomplete* if either an individual leaves the state being investigated for some reason other than that being investigated, or if he or she is surveyed before the event has

occurred. These techniques are not confined to use where time is the dimension measuring the extent of a state. They could be used for such quasi-timescales as number of games without defeat or even repeat purchases of a commodity, but they are most generally used for time.

At present these methods are restricted to discrete states, though the extension to continuous studies represents an obvious (and challenging) development.

They enable researchers to exploit their data more fully by making use of more accurate information on the time of occurrence of events than is possible in panel-type approaches which simply record whether or not an event has taken place. There are other possible reasons for prefering survival-type (continuous time discrete event) approaches to panel-type (discrete time continuous or discrete event). These include the lack of a natural time unit within which respondents change status, the statistical complication that data collected by extending the period between observations can frequently give a misleading picture of the underlying process, or indeed not give a picture at all, referred to as the problems of *identifiability* and *embeddability* (Singer & Spilerman, 1976). Also it is possible, by using time-varying covariates, to take more accurate account of changing circumstances over time than is possible in panel-type approaches.

Social and perhaps more especially econometric sciences have extended the approaches and methods of engineering and biomedicine, the former by emphasis on the unobserved heterogeneity of different individuals who cannot to be fully classified by their age, sex and social class and the latter by looking at repeatable events and sequences of different events in which individual heterogeneity continues to exert an effect from one event to the next.

Incomplete observations represent the main reason why it has proved necessary to develop special methods for the analysis of survival data. Maximum likelihood (ML) methods assume a hazard function of the form

$$h(t; Z) = h_0(t) \exp(\beta \mathbf{Z}),$$

where $h_0(t)$ is assumed to be a standard function of t. The likelihood of the sample is the product, over all observations, of the probability of lasting until the time of censoring or termination and the probability of a terminal event occurring at that time for the observations which were incomplete. The maximum likelihood approach estimates the values of β and of the parameters of $h_0(t)$ for which the sample likelihood is maximised. An alternative approach, that of Partial Likelihood (PL), assumes that the investigator is not particularly interested in the precise form of $h_0(t)$, but

rather in comparing subgroups, and estimates β while partialling out h_0. In this approach the likelihood to be maximised is the product over time points of the conditional likelihood of the actual event occurring at each time point, given that *some* event occurs.

There are two main models used, namely the proportional hazards (PH) model and the accelerated failure time (AFT) model, both of which have interpretations in terms of stochastic models of process – the thousand natural shocks to which flesh is heir. AFT is a "last straw breaks the camel's back" model, which be fitted to the cumulation of random shocks, though more purposive inputs such as learning processes are also covered. The other model is the proportional hazards (PH) model. This can be envisaged as a non-cumulative random shocks or "bolt from the blue" model in which the severity of the shock (or the susceptibility of the group) is related to the characteristics of the group. The PH model is far more widely used, despite the apparent relevance of AFT to many real-life situations, but this is less of a drawback than might appear because of the versatility of the PH model.

Developments in survival/event analysis have been extremely rapid since the early 1970's, and it is likely that social science applications will continue to extend. It is likely that more "realistic" models of processes will be developed and it is possible that this may be tied in wich catastrophe theory (Thom, 1975) for predicting sudden state changes. Applications should also include attempts to view series of events as a whole, particularly taking account of autocorrelated unmeasured heterogeneity, possibly introducing the method of variance components and the effect of non-normative historical influences (Baltes, 1979). Finally it may even prove possible to extend the methods to deal with the behavior of dyads (see e.g. Dowdney et al., 1985; Lincoln, 1984.)

Acknowledgements

Particular thanks are due to Tony Ades and Andrews Pickles for encouragement and valuable comments on earlier drafts. Much of this work was done while the author was working with the National Child Development Study and the author is grateful to former colleagues at the National Children's Bureau for encouragement.

References

Aitkin, M. & Longford, N.T. (1986). "Statistical modelling issues in school effectiveness studies", *Jr. Roy Statist. Soc.* A. 14: 1–43.

Allison, P.D. (1982). "Introducing a disturbance into logit and probit regression models", Mimeo. Cornell University.

Armitage, P. (1985). "Contribution to discussion of paper by Clayton & Cuzick", *Jr. Roy Statist. Soc* A 148: 109–110.

Baltes, P. & Nesselroade, J.R. (1979). "History and rationale of longitudinal research", in Nesselroade, J.R. & Baltes, P. (eds.) *Longitudinal Research in The Study of Behaviour and Development*. New York: Academic Press.

Barlow, R.E. & Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing*. New York: Holt Rinehart, Winsten.

Bartlett,N.R. (1978). "A survival model for a wood preservative trial", *Biometrics* 34: 673–679.

Clayton, D.G. (1978). "A model for association in bivariate life-tables and its application in epidemiological studies of familial tendency in chronic disease incidence", *Biometrika* 65: 141–151.

Clayton, D.G. & Cuzick, J. (1985). "Multivariate generalisations of the proportional hazards model", *Jr. Roy Statist. Soc* A 148: 82–117.

Clayton, D.G. (1986) "'Errors in variables' models in epidemiology", paper delivered to Medical Section, Royal Statistical Society, 25th February 1986.

Cox, D.R. (1972). "Regression models and life tables", *Jr. Roy Statist. Soc* B34: 187–220.

Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman-Hall.

Cozick, J. (1982). "Rank tests for association with right censored data", *Biometrika* 69: 304–364.

David, H.A. & Moeschberger, M.L. (1978). *The Theory of Competing Risks*. London: Griffin.

Davies, R.B., Pickles, A. & Crouchley, R. (1982). "Event history testing", *Sociological Methods and Research* 285–302.

Davies, R.B. (1983). "Duration dependence: a re-evaluation of the competing risk approach", *Environment and Planning A* 15: 1057–1065.

Davies, R.B. & Pickles, A. (1985). "Longitudinal vs. cross-sectional methods for behavioural research: a first round knockout", *Environment & Planning A* 17: 1315–1329.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). "Maximum likelihood from incomplete data via the EM Algorithm", *Jr. Roy Statist. Soc.* B. 39: 1–38.

Dowdney, L., Skuse, D., Rutter, M., Quinton, D. & Mrazek, D. (1985). "The nature and qualities of parenting provided by women raised in institutions", *Jr. Child Psychol. Psychiatr* 26: 599–626.

Elbers, C. & Ridder, G. (1982). "True and spurious duration dependence: the identifiability of the proportional hazards model", *Rev. Econ. Studies* 49: 403–410.

Farewell, V.T. & Cox, D.R. (19??). "A note on multiple time scales in life testing", *Appl Statist* 28: 73–70.

Flinn, C.J. & Heckman, J. (1982). "New methods for analysing individual event histories", *Sociological Methodology 1982*: 99–144.

Fuller, W. & Hidiroglou, M. (1978). "Regression estimation after correcting for attenuation", *Jr. Amer Statist. Ass.* 73: 99–104.

Gail, M.H. (1975). "A review and critique of some models used in competing risk analysis", *Biometrika* 31: 209–222.

Ginsberg, R.B. (1978), "The relationship between timing of moves and choice of destination in stochastic models of migration", *Environment and Planning* A 10: 667–679.

Goldstein, H. (1986). "Multilevel mixed linear model analysis using iterative generalised least squares", *Biometrika* 73: 43–56.

Goodman, L.A. (1961). "Statistical methods for the mover-stayer model", *Jr. Amer. Statist. Ass.* 56: 841–868.

Heckman, J. & Borjas, G. (1980). "Does unemployment cause future unemployment?", *Economica* 47: 247–283.

Heckman, J. & Willis (1977). "A beta-logistic model for the analysis of sequential labour force participation by married women", *Jr. Pol. Econ.* 85: 27–58.

Heckman, J. & Singer, B. (1982). "The identification problem in econometric models for duration data", in W. Hildenbrand (ed.), *Advances in Econometrics: Proceedings of World Meeting of the Econometric Society* 1980. Cambridge: Cambridge University Press.

Heckman, J. & Singer, B. (1984). "Econometric duration analysis", *Jr. Econometric* 24: 63–132.

Hidiroglou, M.A., Fuller, W.A. & Hickman, R.D. (1979), SUPER CARP. Mimeo. Statistical Laboratory. Iowa State University.

Holt, T.D. & Prentice, R.L. (1974). "Survival analyses in twin studies and matched pair experiments", *Biometrika* 61: 17–30.

Hutchison, D. (1980). "Statistical appendix", in J. Steedman, *Progress in Secondary Schools*, National Children's Bureau, 8 Wakley Street, London, EC1V 7QE.

Hutchison, D. (1988) "Event history and survival analysis in the social sciences, part I", *Quality & Quantity* 22: 203–219.

Joreskog, K.G. (1979). "Statistical estimation of structural models in longitudinal-development investigation", in J.N. Nesselroade and P.B. Baltes (eds.), *Longitudinal Research in the Study of Behaviour and Development*. Academic Press.

Kay, R. (1984). "Multistate survival analysis: an application in breast cancer", paper delivered to Royal Statistical Society Medical Section, 28th Feb. 1984.

Lagakos, S.W. & Williams, J.S. (1978). "Models for censored survival analysis: a cone class of variable-sum models", *Biometrika* 65: 181–189.

Lincoln, J.R. (1984). "Analysing relations in dyads: problems, models and inter-organisational research", *Sociological Methods and Research* 13: 45–79.

McCullagh, P., Nelder, J.A. (1983). *Generalised Linear Models*, London and New York: Chapman & Hall.

McGinnis, R. (1968). "A stochastic model of social mobility", *Amer. Soc. Rev.* 23: 712–722.

Moeschberger, M.L. (1974). "Life tests under dependent competing causes of failure", *Technometrics* 16: 39–47.

Nadas, A. (1970). "On estimating the distribution of a random vector when only the smallest co-ordinate is visible", *Technometrics* 13: 923–924.

Nadas, A. (1971). "The distribution of the identified minimum identifies the distribution of the pair", *Technometrics*, 14: 201–202.

Oakes, D. (1982). "A model for association in bivariate survival data", *Jr. Roy Statist. Soc.* B 44: 414–422.

Pierce, D.A., Stuart, W.H. & Kopecky, K.J. (1979), "Distribution free regression analysis of grouped survival data", *Biometrics* 34: 785–793.

Prentice, R.L., Kalbfleisch, J.D., Peterson, A.V., Flournoy, N., Farewell, V.T. & Breslow, N.E. (1978). "The analysis of failure times in the presence of competing risks", *Biometrics* 34: 541–554.

Prentice, R.L. (1982). "Covariate measurement errors and parameter estimation in a failure time regression model", *Biometrika* 69: 331–342.

Singer, B. & Spilerman, S. (1976). "The representation of social processes by Markov Models", *Amer. Jr. Sociol.* 82: 1–54.

SPSS-Inc. (1984). "USERPROC LISREL: Using LISREL VI with SPSSX", SPSS Inc.

Thom, R. (1975). *Structural Stability and Morphogenesis* (translated by D.H. Fowler). Reading, Mass.: Benjamin.

Tuma, N.B. (1980). *Invoking RATE*. Mannheim: ZUMA.

Tuma, N.B. & Hannan, M.T. (1984). *Social Dynamics: Models and Methods*. Orlando: Academic Press.

Warren, J., White, J.K. & Fuller, W.A. (1974). "An error-in-variables analysis of managerial rold performance", *Jr. Amer. Statist. Ass.* 69: 886–893.