National Child Development Study
User Support Group

Working Paper
No 14

DROP OUT FROM APPRENTICESHIP: AN APPLICATION OF SURVIVAL METHODS TO GROUPED DATA

by

D. Hutchison

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

Social Statistics Research Unit
City University
Northampton Square
LONDON EC1V 0HB

AUGUST 1986

NCDSUSGWP14:DH;080586

National Child Development Study User Support Group Working Paper Series

This Working Paper is one of a number, available from the National Child Development Study User Support Group, which report on the background to the Study and the research that has been based on the information collected over the years. Other Working Papers in the series are listed below.

No.	Title	Author(s)	Date
1.	The National Child Development Study: an introduction to the origins of the Study and the methods of data collection	P. Shepherd	October 1985
2.	Publications arising from the National Child Development Study	NCDS User Support Group and Librarian, National Children's Bureau	October 1985
3.	After School: the education and training experiences of the 1958 cohort	K. Fogelman	October 1985
4.	A Longitudinal Study of Alcohol Consumption Amongst Young Adults In Britain: I Alcohol consumption and associated factors in young adults in Britain	C. Power	December 1985
5.	A Longitudinal Study of Alcohol Consumption Amongst Young Adults In Britain: II A national longitudinal study of Alcohol consumption between the ages of 16 and 23	M. Ghodsian and C. Power	December 1985
6.	A Longitudinal Study of Alcohol Consumption Amongst Young Adults In Britain: III Childhood and adolescent characteristics associated with drinking behaviour in early adulthood	M. Ghodsian	December 1985
7.	Report on the longitudinal exploitation of the National Child Development Study in areas of interest to DHSS	Mildred Blaxter	April 1986

DROP OUT FROM APPRENTICESHIP

NOTES

1. The National Child Development Study (NCDS) is a longitudinal study which takes as its subjects all those living in Great Britain who were born between 3 and 9 March, 1958. Since the original birth survey in 1958, the National Children's Bureau has sought to monitor the social, economic, educational and health circumstances of the surviving subjects. To this end, major surveys were carried out in 1965 (NCDS1), 1969 (NCDS2), 1974 (NCDS3) and 1981 (NCDS4). For the purposes of the first 3 surveys, the birth cohort was augmented by including those new immigrants born in the relevant week and information was obtained with the active cooperation of parents, teachers and the schools' health service as well as members of the NCDS cohort. The 1981 survey differs in that no attempt was made to include new immigrants since 1974 and information was obtained from the subject only.

The target sample for the 1981 survey was a total of 16,450 individuals - all those who had participated in NCDS1, NCDS2 and NCDS3, excluding those known to have emigrated or died. Following initial tracing by the Bureau, details of names and addresses were passed to NOP Market Research Limited and Social and Community Planning Research who carried out further tracing and subsequent interviews. The 12,538 interviews obtained represented 76 per cent of the orginal target sample and 93 per cent of those traced and contacted by interviewers.

2. Strictly speaking these times are grouped rather than tied, and we discuss this point later.

3. I am indebted to my colleague, Dr. Tony Ades, for acquiring this program.

METHOD OF DEALING WITH GROUPED DATA: AN APPLICATION TO DROP-OUT FROM APPRENTICESHIP

I. INTRODUCTION

This paper describes the application of survival-type methods to a large data set where there is a substantial amount of grouping in the reporting of time. It has two main objectives in view; firstly to provide an example of the process of screening independent variables and carrying out the multivariate analysis that follows; and secondly to compare various methods of handling the grouped data involved. The data came from the Fourth Sweep of the National Child Development Study (1) and the particular application is to examine drop out from apprenticeship, that is leaving an apprenticeship without completing it and without being sacked or made redundant, though here the substantive interest is secondary to the methodological. It is hoped to report more fully on the former elsewhere (Cook & Hutchison, in preparation).

A total of 2568 members of the cohort, 22 per cent of the ever employed, had started an apprenticeship at some time. Of these, 884, about a third, failed to complete it.

The group considered in this paper is those who started their apprenticeship, whether completed or not, in their first job when they were under 18.

(Table 1 about here)

Table 1 shows, for males and females, outcome of first apprenticeship. We see that two thirds of apprentices completed their apprenticeships, but fully one third did not. The great majority of these, about a quarter of all apprentices, dropped out of their own accord but small proportions (between 1 and 3 per cent for each cause of termination) were either sacked or made redundant, or alternatively their firm closed, or they left for other unspecified reasons. Because of the relatively small numbers of female apprentices, and of those leaving other than of their own volition, we shall confine attention to male apprentices and the focus of interest will be those leaving of their own volition.

II UNIVARIATE ANALYSES

A number of factors might be expected to affect whether an apprentice gave up his apprenticeship. We shall briefly outline the ones we consider here:-

(a) Trade of apprenticeship
(b) Age at entry (16 or 17)
(c) Region
(d) Size of firm
(e) Whether the apprentice had signed articles
(f) Education level
(g) Entry (method of learning about job)
(h) Marital status
(i) Leaving home

(a) Trades are grouped as follows for the purpose of the analysis

(1) Professional, managerial, education, health and welfare, scientific (64)
(2) Clerical and sales
(3) Security and personnel services (53)
(4) Manufacturing and repairing not metal or electrical (308)
(5) Engineering (1252)

(6) Painting, construction and mining (203)
(7) Miscellaneous (farming and transport) (35)

Numbers in each group ar

at risk at the start of period, t less half of those censored during period t. Other conventions have been proposed (Thomson 1977, Hull & Nie 1981) which take account of elements which terminate during the period and are thus not available for the entire period. We have replicated some of the analyses using this alternative convention and results are virtually identical.

Figure 1(a) shows the cumulative survival of drop-out month by month comparing these 7 'trade' groups. Cumulative survival curves are shown, rather than hazard curves, because the cumulation damps down sampling fluctuations and patterns are more easily seen.

The graph shows the estimated cumulative proportions surviving graphed against time in months. The method used allows for the fact that some of the population were censored, that is they left the at-risk population before experiencing the terminal event. In this analysis censoring takes the form of the majority of apprentices completing their apprenticeship successfully after serving their time and hence being no longer at risk of dropping out. If there were no censoring, the graph would represent the cumulative proportion surviving against time.

The graph shows, for example, that those still available, just over .4 of group (security and personal services) have not dropped out by 36 months and that this proportion remains fairly steady after that. Numbers are rather small by this point, because of attrition, and one should be wary of assuming this represents any sort of a plateau.

In general it seems that apprentices in groups labelled 2 and 3 (clerical and sales, and security and personal services, respectively) are the most likely to quit, and those in group 4 (manufacturing group 5 (engineering) group 6 (painting and building) and group 7 (miscellaneous) the least. Table 2(a) gives a summary for this. The group differences are highly statistically significant ($X^2 = 52.2$, 6 d.f.). This shows that, in addition to security and personal services (particularly high drop-out rates) and clerical and sales apprentices in professional and managerial trades apprentices are more liable to drop out. Many of the multivariate analyses reported later in this paper use the (constant) proportional hazards model. It is desirable to test this assumption and

this can be done by comparing log (-log (proportion surviving)) between groups, as indicated by Kay (1977). Figure 1(b) presents the graph: it can be seen that intergroup differences are substantially constant, confirming that the constant hazard assumption is a reasonable one for comparing the 7 trade groups.

(b) Age at Entry

Another possible influence on abandoning an apprenticeship is the age at which it was started. Note that this analysis is confined to those starting before 18 so we only have two categories, those starting before their 17th birthday, and the rest. The NCDS cohort was 'ROSLA' (Raising of the School Leaving Age) year, so they could not legally leave school before age 16.

(Table 3 about here)

Table 3 shows that the difference between age of starting is relatively small and not statistically significant. For this reason we do not show the graph of drop-out rates.

Table 4 compares hazard rates between areas grouped into 5 'regions'.

(1)	South	(384)
(2)	North	(204)
(3)	Midlands	(399)
(4)	Wales	(62)
(5)	Scotland	(179)

It will be seen that these differences are not statistically significant. For this reason, as before, we do not show the graphed results.

(d) Size of Firm

We group firms as follows:-

(1) Less than 25 employees' (420)
(2) 25-100 employees' (617)
or less than 25 but a branch of a larger firm
(3) Over 100 (876)

(Table 5 about here)

Category (b) includes branches of larger firms since these might be expected to have an apprenticeship training scheme covering the whole firm.

Figure 2(a) compares (1) and (2) and (3) above. It can be seen that the relatively small firms show considerably higher drop out rates than medium firms and these in turn considerably higher than large firms. Table 5 confirms this impression. Figure 2(b) shows that the proportional hazards assumption is not violated in comparing these groups, since the lines are effectively parallel.

(e) Formal apprenticeships

The data tell us whether or not an apprentice signed articles, but if he had more than one apprenticeship we do not know whether he signed articles in the first or subsequent apprenticeship (or both). However the number of individuals for whom there is any possible confusion is small (less than 5 per cent) so we ignore this potential complication.

(Table 6 about here)

Figure 5 shows drop-out rates among those who had, or had not, signed articles. Group 2 is those who had not, and it may be seen that those who had not signed articles were much more likely to drop out. This is confirmed in summary Table 6(a) which shows that the difference between the two groups is significant at the .001 level. Figure shows that the proportional hazards assumption is not violated in comparing these groups.

(f) Educational level

Education level may be connected with apprenticeship drop-out.

The three categories we consider, together with the numbers involved, are:

(1)	None	458
(2)	CSE only	720
(3)	O-level	536

(Table 7 about here)

Figure 3(a) shows that the drop-out rate amongst those with one or more O-levels is consistently lower than amongst those with CSE only, which is in turn lower than that for those with no qualifications. Table 7 confirms this. Figure 3(b) shows that the proportional hazards assumption is not violated for qualification groups.

(g) Entry

The next factor we have investigated was method of entry. It is of interest to know whether young people who were found apprenticeships by official channels were as satisfied with them as those who found an apprenticeship by their own efforts, or those who were introduced into a firm by a relative. Entry has been classified as:-

(1)	Own initiative	(594)
(2)	Via statutory services (careers, service, school careers teachers etc.)	(514)
(3)	Via Relative	(689)
(4)	Other	(160)

Table 8 about here)

Table 8 compares the drop-out rates of these 4 groups. We see that there is no statistically significant increase or decrease in the probability of drop-out amongst those who found their own job or had it found for them by relatives compared with those who had found their apprenticeship through the statutory services).

(h,i) Marital Status and leaving Home

Two further possible influences on whether or not an apprentice drops out before completion are marital status and leaving home. In the event, only 16 apprentices got married before completing their apprenticeships and of these only one dropped out. For this reason we do not consider this possible influence here. A second way in which the balance of income and responsibilities is whether or not the apprentice is living at home. Since leaving home is likely to take place during the apprenticeship this is a time-varying covariate and has to be treated in a different way from such fixed status variables as whether the apprenticehad signed articles, or the qualification obtained at school. In particular we do not show graphs comparing drop-out patterns between the two groups.

Summary of Univariate effects

So far we have investigated relation between drop-out (of young men in their first job and first apprenticeship starting before 18) and various factors, taken singly. We have made and checked the assumption that while overall rates of drop-out may fluctuate quite widely over time, the ratios of drop-out rates between groups remain substantially constant.

We found that age at entry, region, marital status did not appear to affect the probability of drop-out. On the other hand, we found that trade of apprenticeship, size of firm, whether the apprentice (or his parents) had signed articles and educational level had quite strong relationships with drop-out.

Thus we found that the traditional industries, such as engineering or other manufacturing were more successful in keeping their apprentices. Similarly we found that those in large firms, those who had signed articles and those who had taken O-levels were more likely to stay the course. However these are just univariate comparisons and it is likely that these relationships are more complicated. For example, it seems likely that engineering firms are in general larger ones, and it is also possible that large firms can offer more organised training and thus attract abler recruits. Only a multivariate analysis, such as we describe below, can start to unravel these effects, and see whether size of firm has an effect beyond that which can be due to selectivity of intake to different sizes of firms. For this reason we decided to carry out multi-variate survival analyses to look at these questions. However as a preliminary we looked at the two-way, and in some instances 3-way, relationships between the independent factors.

III RELATIONSHIPS BETWEEN INDEPENDENT FACTORS

Table 9 shows the distribution of sizes of firms in different trades. We see that a substantially higher proportion of engineering firms fall into the 'large' (100+) category.

It is possible also that engineering firms might be more traditional, or more organised and thus more likely to expect their apprentices to sign articles. Table 10 shows the relationship between trade and proportion signing articles. While the proportion of those with articles in engineering is high, there is little or no difference between it and the other large groups namely manufacturing non-metal and painting and building.

Table 11 shows that there does seem to be a fairly strong relationship between trade and qualifications with O-levels.

Professional' apprenticeships, not surprisingly, have the highest qualifications, but even ignoring these, there is a strong relationship between trade and proportion with O-levels, with engineering apprentices being the most qualified.

Table 12 shows the levels of qualifications in small, medium and large firms. The last, larger firms with 36 per cent are more likely to have recruits with O-levels than small or medium (26 or 27 per cent).

(Table 13 about here)

Following this question further, Table 13 shows the proportion of apprentices with O-levels for small medium and large firms in the different trades. In general there is little pattern though there does seem to be a relationship between proportion with O-levels and size within engineering.

(Table 14 about here)

Size of firm also seems to be related to the proportion signing articles, as is shown in Table 14. Comparison with Table 10 shows that size of firm is important in itself rather than a proxy for other characteristics such as trade.

(Table 15 about here)

Table 15, which shows the proportions with articles by size for trades, is in keeping with this.

(Table 16 about here)

Finally, Table 16, showing the proportion signing articles among those with various qualification levels shows that the main difference is between those with no qualifications and those with some (O-level or CSE).

These crosstabulations shows the existence of interrelationships among the independent variables which require multivariate survival analyses to interpret.

IV MULTIVARIATE SURVIVAL ANALYSES

To carry out such analyses we use Survival Analysis (see, for example Cox and Oakes 1984), alias Event History Analysis (Tuma and Hannan 1984). We employ the former term here. We do not propose to give an in-depth description of this type of technique, but the following brief outline will help to make it clearer to those unfamiliar with it.

The basic concept is the hazard function h(t). This is the probability that an individual will experience a teminal event: in this context drop out of his apprenticeship, given that he has survived so far.

Where an individual does not survive until time t, for example drops out, is sacked or successfully completes before time t then he is not included in the analysis for that or subsequent time. If this loss is for other reasons than that under investigation, then the individual is described as censored. For example, in this paper voluntary drop-outs are the terminal event in question and those who are made redundant or who complete their apprenticeship are treated as censored for all subsequent times.

The form of the basic hazard function h (t) is of relevance to the conduct of the analysis.

Figure 5 shows a graph of hazard function against time in months. It may be seen that the hazard function is a combination of a gradual downward trend and a distinct 'lumpiness' with clear peaks at whole years and a suggestion of smaller peaks of half-years. This may be a

combination of a real effect whereby apprentices actually leave preferentially at such time (for example at the beginning or end of formal technical college courses) together with an artefactual one whereby dates of relatively distant events are rounded, for example, to whole years from the date of the interview.

In any event it is clear that neither a constant hazard, nor one which is a continuously increasing or decreasing hazard, such as the Weibull, will fit satisfactorily. The 'lumpy' and decreasing form of the hazard function as seen in Figure 5 suggests that it may be appropriate to fit a year factor for the decreasing trend in the hazard, and a month factor to deal with the peaks and troughs within the year. We could also use a more general approach which does not assume a fixed functional form for the hazard. One model we use is the proportional hazards which assumes that the hazard in one group at any given time may vary but is always a constant number of times the hazard in another group at the same time. In mathematical notation, where $h_i(t)$ is the hazard in group i at time t, $h_i(t) = k_i * h_o(t)$

This gives $h_i(t)/h_j(t) = k_i/k_j$ (constant)

A model with an arbitrary h(t) is known as a Cox model: in this the arbitrary $h_o(t)$ is conditioned out at each terminal event by constructing the conditional likelihood that the event occurs, to the individual that it did given that some event occurred.

A complication arises here. The Cox model assumes that all survival times are distinct, an assumption which does not hold with NCDS data where times are recorded to then nearest month. Cox (1972) proposed methods for dealing with a small number of tied times, but these are potentially cumbersome, particularly if the number of ties is at all large. Other approaches and approximations have been proposed (see Allison 1982, Prentice and Gloeckler 1978, Thompson 1977).

Models

Possible models may be summarized as follows:

(a) Discrete time models
(b) Tied continuous time models
(c) Grouped continuous time models
(d) Grouped discrete time models

We outline these briefly below.

(a) - (c) are fairly standard. (d) is, we believe, original to this paper. We start by assuming that one can ignore censoring, to make the exposition simpler. This is obviously not a realistic assumption here so we follow this by outlining how it is treated.

(a) -Tied discrete times. Where times are discrete Cox (1972) defined the hazard as:

$$h(t) = (T \leq t + \mid T>t) \qquad \ldots.(9)$$

and proposed a discrete logistic model

$$\frac{h(t,\underline{Z})}{1-h\ (t,\underline{Z})} = \exp\ (\underline{Z}\ \underline{B})\ \frac{h_o\ (t)}{1-h_o(t)} \qquad \ldots.(10)$$

(i) Conditional analysis

The expression for one time period for the likelihood of d elements out for r simultaneously failing is

$$\frac{P(i_1) * P\ (i_2) * \ldots\ldots\ldots\ldots P\ (i_d)}{\sum_{k\ E\ S\ (j;\ d)} P\ (k\ ) * P\ (k_2) * \ldots\ldots\ldots P\ (k_d)}$$

Where S (j;d) denotes the set of all selections of d = d_j items from the risk set of size r. The denominator contains $\binom{n}{d}$ elements, and this rapidly becomes extremely large as n and d increase. This assumes that all censorings occur after failures at the same (discrete) time t: it is of course possible to modify the likelihood to relax this assumption.

For computing purposes, simpler approximations have been suggested by Peto (1972) Efron (1977) Oakes (1977) and Farewell & Prentice (1980). We replicated some of the analyses using BMDP 2L which uses an approximation attributed to Breslow (1974).

However we expected that the paper by Gail et al (1981) setting out a simple algorithm for the exact likelihood would reduce the computational load of carrying out exact analyses for tied times, (2). Accordingly we replicated the analyses further using PECAN (Storer et al 1983) (3)

(ii) Unconditional analyses

We again assume that $h(t;\underline{Z}_i)$ = is related to $h_o(t)$ by the logistic equation.

$$\frac{h(t;\underline{z}_i)}{1 - h(t;\underline{Z}_i)} = \exp(\underline{z}\,\underline{\beta}) \frac{h_o(t)}{1 - h_o(t)}$$

If the number of failures per interval is relatively large then one can devote a separate nuisance parameter to estimate $h_o(t)$ in each interval, and this procedure is described as an unconditional analysis (Breslow & Day 1980).

(b) Tied continuous time models

Again, examples in which tied observations occur in continuous time are rare and in general, apparent ties could have been resolved by a finer recording of the timing events. Where there are genuinely tied events then the discussion in the previous section substantially applies.

(c) Grouped continuous data methods

This is frequently the most realistic situation in the social sciences.

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

$$1-(1-\lambda_i)\exp(\underline{Z}\,\underline{\beta}) \qquad \ldots\ldots(11)$$

where $\lambda_i = \exp\int_{t-1}^{t} \lambda(u)\,du$

$\underline{\beta}$ can be estimated by putting

$$C[h(t)] = \alpha + \underline{\beta}\,\underline{X}$$

where C is the complementary log-log transformation. Both this and the logistic approximation discussed below can expressed in the general linear model form (see McCullagh & Nelder 1983). The unconditional logistic model for discrete data is frequently used for such analyses since 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 β is independent of the length of the time interval, and consequently the β thus estimated is unbiased for the continuous time model. This property is not shared by the logistic link (Singer & Spilerman 1976) in which not only the precise values of the β coefficient, but even the form of the relationship can change (Myers et al 1973). However, where the hazards in each interval are small one should expect the methods to give very similar results (McCullagh & Nelder 1983). We investigate this question in this paper.

Grouped discrete-time model

We also propose a model for the process which is an attempt to produce a more 'realistic' description of the actual process. This postulates the existence of a (dis) satisfaction function Dt. Satisfaction could be a combination of the factors above which may be considered as relating to general level of suitability (S) of the young person for the training and vice versa; the level of formal and informal commitment (C) of both parties; and the varying unobservables, such as short term fluctuations (E) - e.g. the Monday feeling as evidenced by absenteeism rates and relationships at home.

$$D(t) = f(S, C(t) + E(t)$$

where $C(t) = g(A(t), I(t))$ A formal commitment

I human capital investment

The unit of measurement of t is arbitrary with only the condition that it should be such that the number of intervals should be relatively large within each period considered. Arbitrarily we could set this to a day as this seems a natural unit in considering employment.

For parismony we postulate the linear relationship

$$D(t) = aI + bA + cQ + dT + eF + E(t)$$

where D(t) is the level of dissatisfaction on day t.

I = time (in months) already spent on the apprenticeship
This we take as a measure of human capital already invested by the apprentice.

A = whether or not apprenticeship has already signed articles.

Q = Qualifications

T = Trade of apprenticeship

F = Size of firm

E(t) =is a 'random' fluctuation, specific to the day.

The experience of apprenticeship is considered as a series of trials of a process with a latent variable D(t) which is a combination of a fixed level of (dis) satisfaction for that month and a random disturbance term.

When the dissatisfaction level for a particular day passes beyond a fixed arbitrary level, then the apprentice quits. The chance of an apprentice quitting during a given month can thus be approximated by the extreme value distribution, the probability of the largest value of D(t) in a month going over the threshold tolerance level.

Then the procedure can be modelled using GLIM with binominal error and complementary log-log link.

Some comments on the model are appropriate here.

1. The disturbance term will not in fact be purely random since it will probably vary fairly systematically for example by day of the week. However in this context it seems acceptable to treat it as though it were random.

2. The precise distribution function of the disturbance term is unimportant. It can be one of many functional forms. Exponential, normal or logistic are examples.

3. It is obviously an oversimplification to assume that an individual's basic dissatisfaction level can be determined completely by the measured factors defined above: individual apprentice and individual firm factors must have an impact to give rise to variation within the group defined by a particular set of covariates. The question of allowing for such unobserved heterogeneity is discussed elsewhere in this volume.

Censoring in grouped or discrete data. So far in this discussion we have ignored censoring in general ties between failures, and censorings are handled by deeming that the censorings occur instantaneously later than their tied failures (Cox and Oakes 1984, 8.5). Obviously this does not hold here where data are recorded to the nearest month, and within a month censorings can occur either before or after failures.

We follow Kalbfleisch & Prentice (1980) by taking as the hazard h(t) taking the proportion terminating in period t

$$h(t) = \frac{T(t)}{R(t) - 1/2\ C(t)}$$

where:

h(t) is the proportion terminating (hazard) in period T

T(t) is the number terminating in period T

R(t) is the number at risk at the start of period t.
the individuals 'at risk' are those who have not already dropped out or been censored (for example those sacked before time t).

C(t) is the number censored in period t.

In this approach the number astrisk over period t is taken as the number

at risk at the start of period, t less half of those censored during period t. Other conventions have been proposed (Thomson 1977, Hull & Nie 1981) which take account of elements which terminate during the period and are thus not available for the entire period. We have replicated some of the analyses using this alternative convention and results are virtually identical.

V RESULTS

Basic Hazard Function

The 'lumpy' and decreasing form of the hazard function, as seen in figure 5, suggest that it may be appropriate to fit a year factor for the decreasing trend in the hazard. and a month factor to deal with peaks and troughs within the year.

Table 17 shows the results of such a multivariate analysis of apprenticeship drop-out. The second and third columns are used to assess the statistical significance of the effect of each variable except the time variables as 'last one in' that is, after having allowed for the possible effect of all other independent variables. Comparing 17(a) and 17(b) we see that allowing for time as 61 months is significantly better than as year and month. However the intergroup comparisons are very little affected indeed, in magnitude and significance, and in addition the latter shows up the pattern of time variation more clearly at the cost of a relatively small worsening of the fit. Note of course that the value of chi-squared for overall fit are biased, but we should expect the significance of individual effects to be unbiased (see for example McCullagh & Nelder 1983).

We therefore concentrate our attention on 17(b). We notice that the coefficient of the year variable decreases, indicating that by and large, the longer an individual has been in an apprenticeship, the less likely he is to drop out.

The coefficient of the month variable is clearly highest at month 1 and then at month 7 confirming our impression (see above) from figure 5 of 'peaks' in recorded leaving, whether actual or artefactual, at whole and half-years.

Turning to the independent variables, we see that all four, whether signed articles, whether taken O-levels size of firm and trade of apprenticeship are all still highly statistically significant, after allowing for the effect of all the others. Thus we see that apprentices who have not signed articles (or whose parents have not signed articles) are considerably more likely to drop out than those who have. The actual size of the difference does not immediately have an obvious interpretation, though this will be made clearer below. It seems that the formal procedure of signing articles does have an effect in encouraging commitment to an apprenticeship, since it comes out as an extremely strong difference even after allowing for other important factors. Similarly those with higher qualifications are considerably less likely to dropout.

Since this effect also holds strongly after allowing for trade and size of firm it shows that the more academically successful recruit finds it easier to cope with an apprenticeship, perhaps because of the formal education element in apprenticeships, or perhaps because of a more positive attitude to authority, as exemplified by his making the effort to sit public exams.

Large firms are more successful in hanging on to their apprentices. After allowing for the effects of the other independent variables, drop out rates in firms with more than 100 employees are roughly half those in firms with less than 25. Finally, those in more traditional trades: engineering, manufacturing and painting and building, are less than half as likely to drop-out as those in clerical and sales or security and personal services.

However, there is a potential complication here because the differences between engineering apprenticeships and the rest. We discussed earlier interactions between the independent variables, such as

the larger mean size of engineering firms, and their tendency to have more highly qualified recruits. This is potentially complicated by the generally greater length of completed apprenticeships in engineering than in other trades. Table 18 shows for each trade the length of apprenticeship (grouped) by which 50 per cent had completed. It can be seen that the great majority of engineering apprenticeships finish between 3.5 and 4 years, whereas the others tend to be shorter.

Table 19 confirms the effects of articles, qualifications and firm size independent of trade by carrying out the analysis purely on engineers. As may be seen, patterns of fitted constants are highly similar in general import to those of Table 16 through two of the three Chi-squareds are smaller, (but still highly significant) because of the smaller sample created by excluding non-engineers. The effect of size of firm is increased by confirming the analysis to engineers, as would be predicted by comparison with Table 9.

A further possibility described earlier was that leaving home might precipitate a decision to quit an apprenticeship . Table 20 shows the results of an approximate conditional analysis carried out using BMDP 2L including the time varying covariate leaving home. It shows that leaving home is very strongly associated with the decision to quit an apprenticeship, those who have done so being nearly 3 times as likely to quit as those who remain without their parents.

The analysis so far, though relatively simple, has been described in some detail as an example of an approach to a survival analysis. The remainder of the paper compares a variety of ways carrying out the multivariate analysis, and in particular method of handling the question of non-distinct failure times.

VI COMPARISON OF ANALYSIS PACKAGES

This section deals with the comparison of results and costs of different analysis techniques and packages. There are two main reasons for this.

(a) Many of these packages use different approaches, particularly to the question of non-distinct failure times (See Section IV, above). It would be of interest to see whether they all give different results.

(b) The unconditional analysis carried out by GLIM reported earlier on a month-by-month basis was rather expensive in computing time, and we were interested to see whether other packages were less heavy, and also whether a broader grouping, giving less cells, would reduce computer time without materially affecting results.

We first investigate the effect on GLIM of reducing the dimensions of the problem by considering drop-out over 6-month, rather than one-month periods, (Table 21a). We also compare the more widely available logistic regression with the complementary log-log link - only available in GLIM and GENSTAT (Table 21b). We take care of the time factor in these analysis by allowing one parameter per 6-month period. We see that the fitted constants and significance levels for articles, qualifications, size of firm and trade are similar for clog-log and logistic links. We also observe that the results differ very little from those for the month by month analysis, and so we conclude that logistic regression gives basically the same results as the complementary log-log link and that 6-month grouping gives a satisfactory method of analysis. With hindsight this closeness of the results using logistic and clog-log links is eminently predictable if one compares the Taylor series expansions of the two functions for small P.

Table 22 and 23 show the results of the same analysis carried out using BMDP-2L which uses the Peto/Breslow (Peto 1972, Breslow 1974) approximation for tied times, and using PECAN (Storer et al 1983) which carries out exact analysis of tied times, using the Gail/Howard algorithm (Gail et al 1981, Howard 1972) Comparison of the results with each other shows no consistent differences in fitted effects. Nor moreover do consistent differences emerge in probability levels between either the BMDP or the PECAN and the unconditional GLIM analysis of Table 17.

The conclusion emerges that in a situation where as here, effects are sizeable, sample sizes large and rates small that it matters little to the estimation of beta values which analysis techniques is used. Where one is investigating the effects of covariates BMDP-2L and PECAN are of course conditional probability approaches and do not directly estimate the underlying hazard.

For this reason it is valuable to look at computing costs. Table 24 shows approximate c.p.u. times on the ULCC Amdahl 4/75. These comparisons are only approximate as not all packages were performing exactly this same tasks, as some of the GLIM analyses in particular were done taking more than one bite at the cherry. However they do give order-of-magnitude estimates of computing times. It can be seen that there are very drastic differences between computing resources used, a factor or over 300 between highest and lowest. These comparisons are perhaps rather unfair to GLIM which is intended as an interactive package: GENSTAT might well be much cheaper.

It seems from inspection of the table that BMDP-2L is the 'best buy'.

It is also seen, from Comparing analysis 19 and 21, that the introduction of a time-varying covariable increases computing times drastically (by a factor of over 20.) This of course would not hold with GLIM where computing times should not change once the time x covariate input table was prepared.

Summary and Discussion

Non-distinct failure times potentially present problems in the analysis of failure time data particularly in semi-parametric methods, (Cox 1972) where the computational load can become extremely high if there are large numbers of failures. A wide variety of approaches have been suggested for coping with the situation but the literature is sometimes confusing and indeed some authorities (Tuma & Hannan 1984) consider that their treatment is still a matter for debate.

Such approaches divide into those where the times are genuinely tied (generally discrete time situations, and those where the apparent equality arises mainly or entirely as result of grouping. Tied and grouped time are typically handled by assuming a logistic or complementary log-log link between predicted hazard and the covariates (McCullagh & Nelder 1983).

The logistic link was first proposed (Cox 1972) and it is here that computational overdrafts arise. Methods suggested of reducing with this include a variety of approximations (Peto, 1972; Breslow 1974; Efron 1977 and Oakes 1977). An algorithm for calculating the exact likelihood has also been developed (Howard 1972, Gail et al 1979). This is quite computationally thrifty, particularly when combined with sampling from the risk set (Breslow & Patten 1979). An alternative approach where the number of ties is substantial, and the number of time points is relatively small is described as unconditional anlaysis, in contrast with the conditional analyses noted so far, and involves fitting a separate parameter for each interval.

The complementary log-log analysis is of course only an unconditional approach. We considered these approaches in investigating drop-out from apprenticeship in the National Child Development Study (Payne 1985). Both a grouped continuous and a grouped discrete time model led to a complementary log-log analysis. In general, in social science approaches events will only be recorded as occurring within a period, so this type of analysis will be appropriate.

However logistic regression is more familiar to social scientists and computer packages are more widely available. We considered that if logistic regression gave results which differed little and were less expensive in computer time, it might well be worth using them, particularly in exploratory analyses. We compared complementary log-log and unconditional logistic analyses using GLIM, approximate conditional analysis (BMDP2L) and 'exact' analyses using PECAN (Storer et al 1983). There were no consistent differences between the results using any of the analysis. The results showed that unconditional complementary log-log analyses using GLIM were drastically more expensive than approximate conditional analyses using BMDP2L, and indeed substantially more expensive than 'exact' analyses using PECAN. The computing costs of complementary log-log and logistic regression were identical in GLIM. These results are in accord with other published comments (Storer et al 1983, Morris 1985), above the relative expense of GLIM. In fairness to GLIM, it must be observed that it was not designed for this kind of work and it is a tribute to its versatility that it can be used here.

In fine, it is clear that on large data sets with a small hazard but substantial ties there is nothing to chose between the packages as to accuracy but that in terms of computer usage, and to some extent convenience, BMDP2L is a clear winner.

Table 1 Outcome of first apprenticeship by sex

	Female %	Male %	Total %
Completed	53	67	66
Left	41	24	26
Redundant	1	3	3
Sacked	2	3	3
Firm closed down	2	2	2
Other	2	2	2
N = (100%)	315	2253	2568

Table 2 Comparison of hazard rate by trade

(Lee Desure Statistic)

			N
(a)	Professional and Managerial	-111.3	64
	Clerical and Sales	-331.4	25
	Security and Personal Service	-662.3	53
	Manufacturing non metal	6.0	308
	Engineering	51.3	1252
	Painting and Building	- 73.6	203
	Miscellaneous	- 16.9	35

Overall X^2 52.2*** (6.d.f)

Table 3 Comparison of hazard rate by Age Started

(Lee Desu Statistic)

Under 17	1.1	1787
Under 18	-10.9	181
Overall X^2	.0 (NS)	(1df)

Table 4 Comparison of hazard rate

(Lee Desu Statistic)

Region

(2)	South	38.0	384
	North	- 8.1	204
	Midlands	-36.4	399
	Wales	44.0	62
	Scotland	- 6.3	179
	Overall X	4.8 (NS)	4df.

Table 5 Comparison of hazard rate by firm size

(a) (Lee Desu Statistic)

		N
⅛25	-176.1	420
⅛25-100 or 25 branch	-36.4	617
100+	110.1	876
Overall X^2	41.1***	(2 df)

Table 6 Comparison of hazard rate by signing article

(a) (Lee Desu Statistic)

		N
Yes	120.8	1306
No	-283.2	557
Overall X^2	117.2***	(1 df)

Table 7 Comparison of hazard rate by Qualification

(Lee Desu Statistic)

		N
None	-127.6	458
CSE	- 4.1	720
O-level	114.5	536
Overall X^2	30.6***	(2 df)

(b) Proportional hazards test X^2 .3NS (1df)

Table 8 Comparison of hazard rate by entry

(Lee Desu Statistic)

		N
Own initiative	-8.5	594
Statutory services	1.6	514
Relatives	9.1	689
Others	-12.8	160
Overall X^2	0.2 (NS)	(3df)

Table 9 Trade by Size of Firm

	⅛25	25-100 or branch	100+	Total (100%)
	%	%	%	
Professional	19	38	44	64
Clerical and Sales	(5)	(8)	48	25
Security and personal services	22	37	41	54
Manufacturing non-metal	32	37	31	300
Engineering	18	31	52	1239
Painting and building	35	35	30	198
Miscellaneous	30	33	36	33

X2	78.5***	12 df
	75.0***	4 df Engineering, Building, Manufacturing

Table 10

Trade and proportion signing articles

	%	N=100%
Professional	58	64
Clerical and Sales	60	25
Security and personal services	57	51
Manufacturing non-metal	68	292
Engineering	72	1200
Painting and building	71	197
Miscellaneous	69	35

X2	12.4 P .05	6 df
X2	1.6 - NS	2 df Engineering, Building Manufacturing

Table 11 Trade and Qualification

	CSE only %	O-levels %	Total (=100%)
Professional	35	50	58
Clerical and Sales	(7)	(7)	25
Security & Personal Services	47	(8)	45
Manufacturing non-mental	42	(30)	272
Engineering	42	33	1102
Painting and Building	49	20	180
Miscellaneous	32	38	34
X^2	35.9***		12 df.
X^2 (Engineering, Manufacturing, Painting & Building)	17.0***		4 df.

Table 12 Size of firm by Qualifications

	CSE only %	O-levels %	Total (=100%)
⅛25	42	26	375
25-100 or branch	46	27	555
100+	40	36	759
X2	18.3**		4 df.

Table 13 Proportion of apprentices with qualifications by size of firm trade

	Size of firm					
	⅛25		25-100 or branch		100+	
	CSE	O-levels	CSE	O	CSE	O-level
Manufacturing non-mental	46	32	41	29	40	30
Engineering	41	26	47	28	40	38
Painting & Building	45	16	49	18	54	17

Significance tests for fit	X^2	Df
Overall	39.7***	16
Size	24.5*	12
Trade	20.2 P .05	12
Size + Trade	8.7 NS 8	8

Table 14 Percentage of apprentices signing articles by size of firm

	%	N (=100%)
⅛25	54	409
25 - 100 (or branch)	69	600
100+	78	829
X^2 76.2***	2 df	

Table 15 Percentage of apprentices with articles by size of firm by trade

	≤25	25-100	100+
Professional	(6)	70	54
Clerical and Sales	(2)	(5)	(7)
Security and personal services	83	32	65
Manufacturing non-mental	54	71	78
Engineering	52	70	79
Painting and Building	55	67	91
Miscellaneous	52	40	9

Significance tests of fit	X^2	
Overall	87.4	8 df
Size only	6.1	6 df
Trade only	85.9	6 df
Size + Trade	5.1	4 df

Table 16 Qualifications and Signing articles

Qualifications	% Signing Articles	N(=100%)
None	57	430
CSE only	75	695
O-level	72	520
X^2	41.4*** 2 df.	

Table 17 Apprenticeship Drop-out (1 month intervals)

Complementary Log - Log Model

			(a) Month (0-60) Fitted const.	X^2	Df	(b) Year + Month Fitted const	X^2	Df
Overall								
Year	1					0		
	2					- .36		
	3		Intervals not			- .54		
	4		shown			-1.14		
	5					-1.44		
	6					-3.53		
Month	1					0		
	2					- .72		
	3					- .40		
	4					- .65		
	5					- .76		
	6					- .90		
	7					- .26		
	8					- .73		
	9					- .52		
	10					- .72		
	11					- 1.17		
	12					- .92		
Articles		Yes	0	61***	1	0	60***	1
		No	.87			.87		
Quals		O Level	0					
		CSE	- .17	21***	2	-.16	19***	2
		None	- .61			-.60		
		25	0			0		
Size		25-100	- .22	22***	2	-.22	22***	2
		100+	- .63			-.63		
		Prof	.43			.44		
Trade		Clerical & Sales	.73			.72		
		Security & Personal	1.08			1.08		
	Mfg. non mental		- .13	25***	6	- .13	25***	6
		Engineering	0			0		
	Paint & Build		.19			.20		
	Misc.		.13			.14		
Overall			1374	5174			1474	5218

Interactions tested		
Articles	*Quals	
Quals	*Size	All NS
Size	*Trade	
Size	*Articles	

Table 18 50th and 80th Percentile Suruvival Time by Trade: 6 month periods

	50th Percentile	N = 100%
Professional	30 - 35	64
Clerical and Sales	24 - 29	25
Security and Personal Services	18 - 23	53
Manfacturing non-metal	36 - 41	308
Engineering	42 - 47	1252
Painting and Building	30 - 35	203
Miscellaneous	30 - 35	35

Table 19 Apprenticeship Drop-out (6 months intervals
M Engineers only)

		Logistic model Fitted constant (exponentiated)		
Overall		.083		
Interval (months)	- 5	1		
	6 - 11	.74		
	12 - 17	.60		
	18 - 23	.67		
	24 - 29	.62		
	30 - 35	.41		
	36 - 41	.49		
	42 - 47	.11		
	48 - 53	.14		
	54 - 59	.19		
	60+	.002		
Articles	Yes	1	X^2 29.7**	df 1
	No	2.40		
Quals	O-level	1		
	CSE	.82	12.5***	2
	None	.53		
Size	25	1		
	25 - 100/branch	.76	24.2	2
	100+	.42		
Overall X^2			170.6	8 df

Table 20 Apprenticeship Drop Out (1 month intervals)

BMDP-2L Approximate conditional logistic Analysis with Time - Varying Covariate

Factor		Fitted constant (Exponentiated)	X^2	DF
Articles	Yes	1		
	No	2.53	65.5***	1
Quals	None	1		
	CSE	.88	16.3***	2
	O-level	.56		
Size	⅛25	1		
	25-100 or branch	.85	18.8***	2
	100+	.55		
Trade	Prof.	1.35		
	Clerical and Sales	2.08		
	Security and Personnel	2.44		
	Mfg Non-mental	.86	21.9	6
	Engineering	1	P .001	
	Painting and Building	1.07		
	Misc.	.73		
Left home	No	1	46.2***	1
	Yes	2.87		

Interactions tested As Table 17 (all NS)

Table 21 Apprenticeship Drop-Out (6 Month Intervals: all M)

		(a) Clog-log Fitted const.	X^2	Df	(b) Logistic Fitted const.	Exp ( )	X2	D
Overall		-2.77			-2.74	.06		
Interval	-5	0			0	1		
(6 months)	6 - 11	.05			.05	1.06		
	12 - 17	- .34			- .36	.70		
	18 - 23	- .36			- .37	.69		
	24 - 29	- .30			- .30	.74		
	30 - 35	- .84			- .87	.42		
	36 - 41	- .78			- .80	.45		
	42 - 47	-1.90			-1.94	.14		
	48 - 53	-1.83			1.87	.15		
	54 - 59	1.24			-1.27	.28		
	60+	-4.66			-4.81	.01		
Articles	Yes	0			0	1		
	No	.86	59.1***	1	.89	2.44	593***	1
Quals	None	0			0	1		
	CSE	-.15	18.2***	2	- .16	.85	18.1***	2
	O-level	-.59			- .61	.55		
	<25	0			0	1		1
Size	25-100 or branch	-.22	21.4***	2	- .22	.80	21.5***	2
	100+	-.62			- 65	.65		
Trade	Professional	. 42			.43	1.54		
	Clerical & Sales	. 61			.64	1.91		
	Security & Personal	1. 12			1.18	3.27		
	Mfg non-metal	- .13	24.6***	6	- .13	.88	24.7***	6
	Engineering	0			0	1		
	Painting & Building	.19			.20	1.23		
	Mis.	.12			.12	1.13		
Overall X^2		597.1	924			596.		924

Interactions tested (as Table 17 - all NS)

Table 22 Apprenticeship Drop Out (1 month intervals)

BMDP - 2L approximate conditional logistic Analysis

		'Exact' Fitted Constant (Exponentiated)	Sampled Fitted Const	X^2	D.f.
Factors					
Articles	Yes	1			
	No	2.51	68.8***	1	
	None	1			
Quals	CSE only	.83	20.2***	2	
	0-level	.53	.53		
	25	1			
Size	25-100 or branch	.82			
	100+	.55	18.8***	2	
	Prof	1.52			
Trade	Clerical and Sales	2.06			
	Security and Personal	2.89			
	Mfg Non-metal	.87	29.5***	6	
	Engineering	1			
	Painting and Building	1.13			
	Misc.	1.13			
Interactions tested		As Table 17 (all NS)			

Table 23 Apprenticeship Drop Out (1 month intervals)

PECAN (Gail et al) Program for Exact Conditional Likelihood Survival

		'Exact' Fitted Const. (Exponentiated)	X2	Sampled Fitted Const.	X2	D.f.
Factors						
Articles	Yes	1	65.8***	1	53.6**	1
	No	2.51		2.43		
Quals	CSE only	1				
	O-level	.83	20.2***	1		
		.53		.86	15.6***	
				.56		
	25	1		1		
Size	25-100 or branch	.82	19.4***	.83	16.6***	
	100+	.55		.56		
Trade	Professional	1.53		1.50		
	Clerical & Sales	2.08		2.60		
	Security & Personal	2.97		3.11	22.7***	
	Mfg. non-mental	.88	24.0***	.90		
	Engineering	1		1		
	Painting & Building	1.14		1.25		
	Miscellaneous	1.13		1.13		

Table 24 — Analysis Described: Techniques and costs

Type of analysis	Results in Table	'Exact' (E) or Approx(A)	Population	Size of time interval	Programe	No. of input units	Time element in model	Approx. cpu time (Sec) AMDAHL 4/75
Clog-log	17(a)	-	All M	1 month	GLIM	52	Months	2000
	17(b)	-	All M	1 month	GLIM	52	Year + months	900
	21(a)	-	All M	6 months	GLIM	946	6 months	50
Logistic unconditional	21(b)	-	All M	6 months	GLIM	946	6 months	50
	19	-	Engineers M	6 months	GLIM	197	6 months	13
Logistic conditional	21	A	All M	1 month	BMDP-2L	1014	Conditioned out	6
	20	A	All M	1 month	BMDP-2L	1014	Conditioned out+time-varying	150
'Exact' conditional	23	E	All M	1 month	PECAN	1042	Conditioned out	900
Sampled exact condition	23	E	All M	1 month	PECAN	3204	Conditioned out	780

Note: These analyses are not completely comparable, since the programs tested do not all all carry out exactly the same tasks. Additionally some of the GLIM analysis were broken up into a number of runs thus incurring additional costs for reading the data. However they give order-of-magnitude estimates for computing costs.

REFERENCES

Allison, P. D. (1982) 'Discrete-time methods for the analysis of event histories'. Sociological Methodology 61-98.

BAKER, R.G. & NELDER, J.A.(1978) The GLIM System, Release 3, Generalised Linear Interactive Modelling Numerical Algorthms Group, Oxford.

BRESLOW, N.E. (1974) 'Covariance analysis of censored survival data' Biometrics 89-99.

BresLOW, N.E. & PATTEN, J. (1979) 'Case-control analysis of cohort studies' in Breslow, N.E. & Whittemore, A.S. (eds) Energy and Health SIAM.

Cook, L. & Hutchison, D. (in preparation) Factors affecting drop-out from apprenticeship

Cox, D.R. (1972) 'Regression models and life tables' Jr. Roy. Statist. Soc. B. 187-202.

Cox, D. R. (1975) 'Partial Likelihood' Biometrika 599-607

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

Efron, B. (1977) 'The efficiency of Cox's likelihood function for censored data 'Jr. Amer. Statist. Assoc. 557-65

Farewell, V.T.& Prentice, R.L. (1980) 'The approximation of failure time with emphasis on case-control studies'. Biometrika 273-8.

Gail M.H. Lubin J.H.& Rubinstein, L.V. (1981)
'Likelihood calculations for matched case-control studies with tied death times ' Biometrika, 703-7

Howard, S. V. (1972) Contribution to discussion of paper by D. R. Cox. Jr Roy Statist Soc. B., 210-1.

Hull, C.H. & Nie, N.H. (1981) SPSS Update 7-9 McCraw-Hill.

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

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

Morris, R. W. (1985) 'On the application of Cox's Proportional Hazards Model in GLIM' GLIM Newsletter 9, 35-36

Myers, M. H., Hankey, B.F. & Mantel, N. (1973) 'A logistic-exponential model for use with response-time transformation involving regressor variables ' Biometrics 257-267.

Oakes, D. (1977) 'The asymptotic information in censored survival data' Biometrika 441-8.

Peto, R. (1972) Contribution to discussion of paper by D.R. Cox. Jr. Roy Statist. Soc. 34, 205-7.

Prentice, R.L.&. Gloeckler, L.A. (1978) 'Regression analysis of grouor calculating the exact likelihood has also been developed (Howard 1972, Gail et al 1979). This is quite computationally thrifty, particularly when combined with sampling from the risk set (Breslow & Patten 1979). An alternative approach where the number of ties is substantial, and the number of time points is relatively small is described as unconditional anlaysis, in contrast with the conditional analyses noted so far, and involves fitting a separate parameter for each interval.

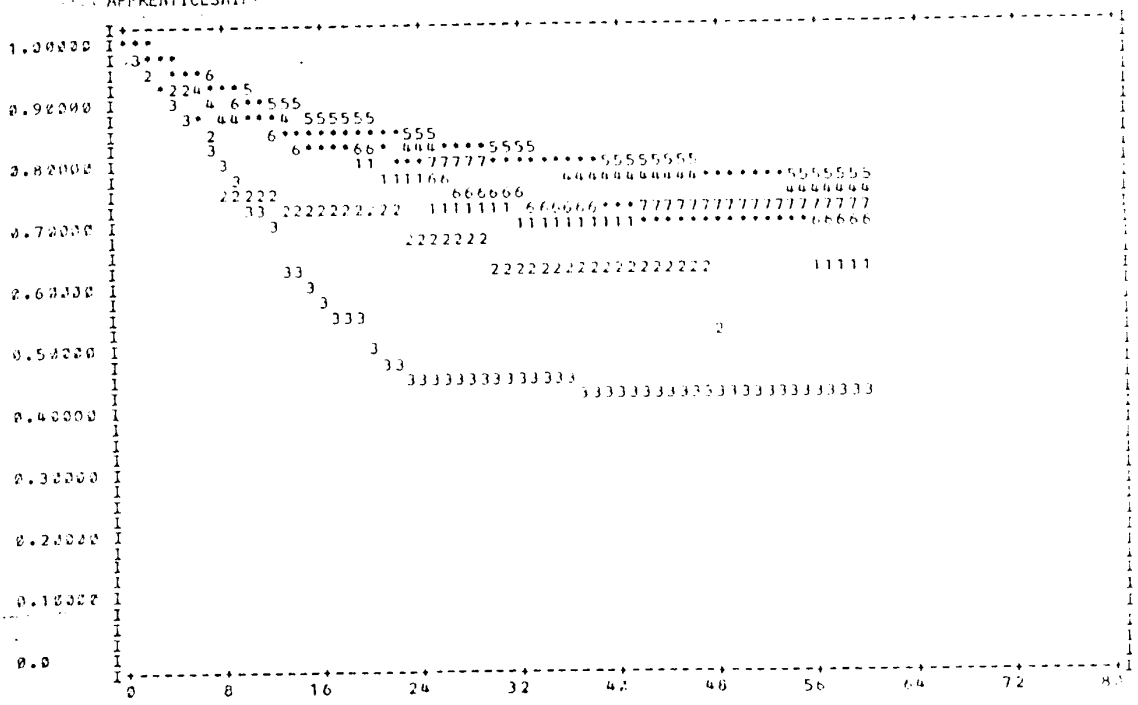
Storer, B.E. (1980) Pecan Users' Notes: Version 2.1 - 10/25/83
Mimeo. Dept. of Biostatistics, University of Washington.

Storer, B.E., Wacholder, S. & Breslow, N.E. (1983).
'Maximum Likelihood Fitting of General Risk Models to Stratified Data'
Applied Statistics 172-181.

Thompson, W.R. (1977) 'On the Treatment of Grouped Observations in life
Studies' Biometrics. 463 -470.

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

FIGURE 1(a) (CUMULATIVE) SURVIVAL FUNCTION BY TRADE OF APPRENTICESHIP



VALUES OF RTRADE AND THEIR GRAPH SYMBOLS

VALUE	GRAPH SYMBOL	VALUE LABEL
1	1	PROF ZTC
2	2	Clerical & sales
3	3	Security & pers serv
4	4	Mtg non-metal
5	5	Engineering
6	6	Printing & building
7	7	Misc

Figure 1(b) GRAPH OF LOG (-LOG SURVIVAL FUNCTION) VS LOG-TIME BY TRADE

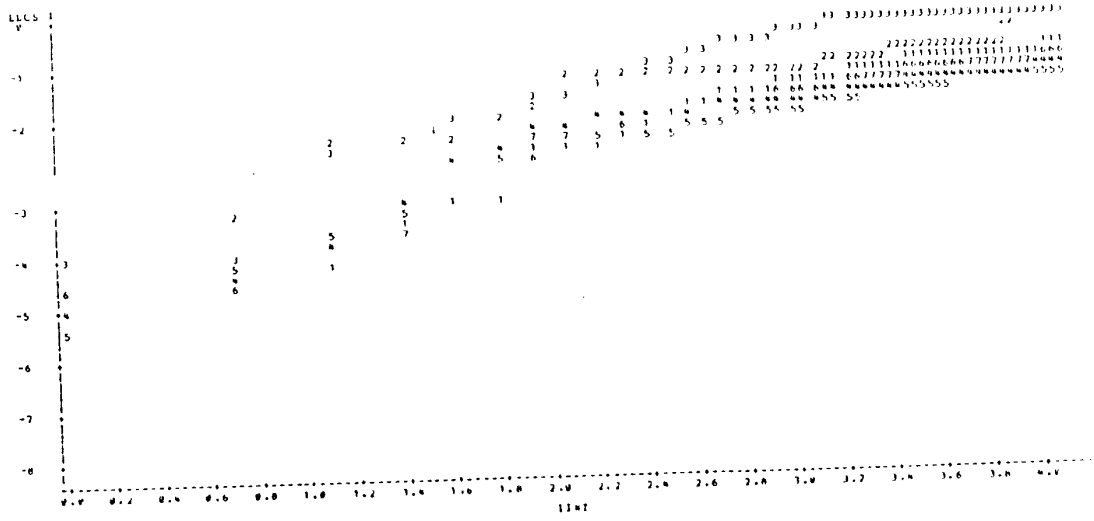


FIGURE 2(a) (CUMULATIVE) SURVIVAL FUNCTION BY SIZE OF FIRM

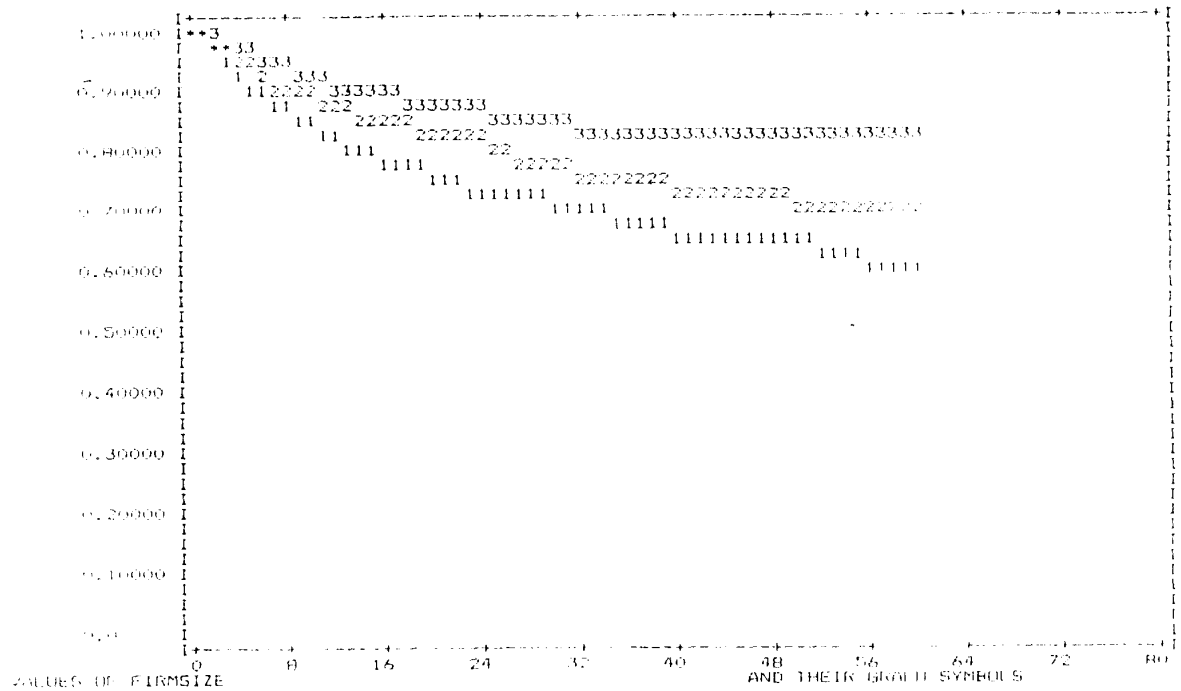


Figure 2(b) GRAPH OF LOG (-LOG SURVIVAL FUNCTION) VS LOG-TIME BY SIZE OF FIRM

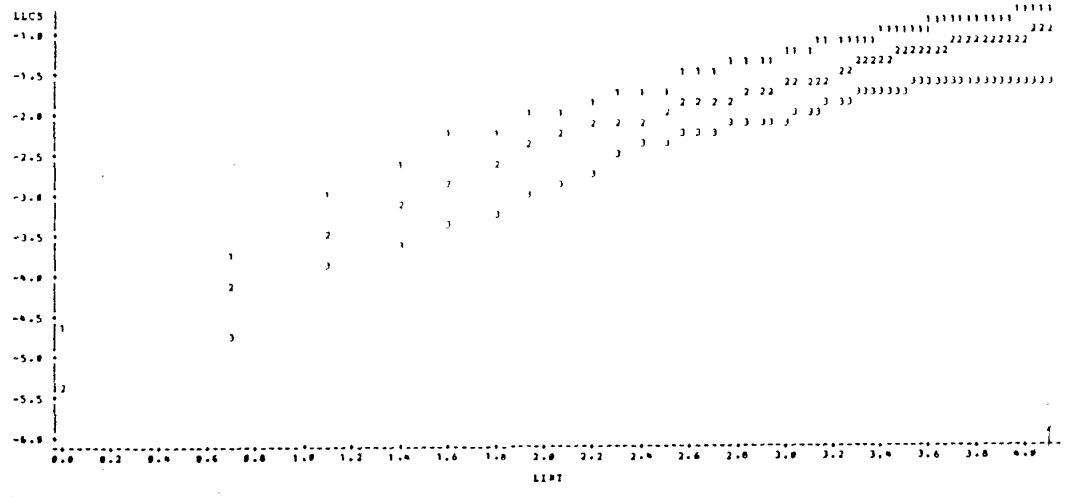
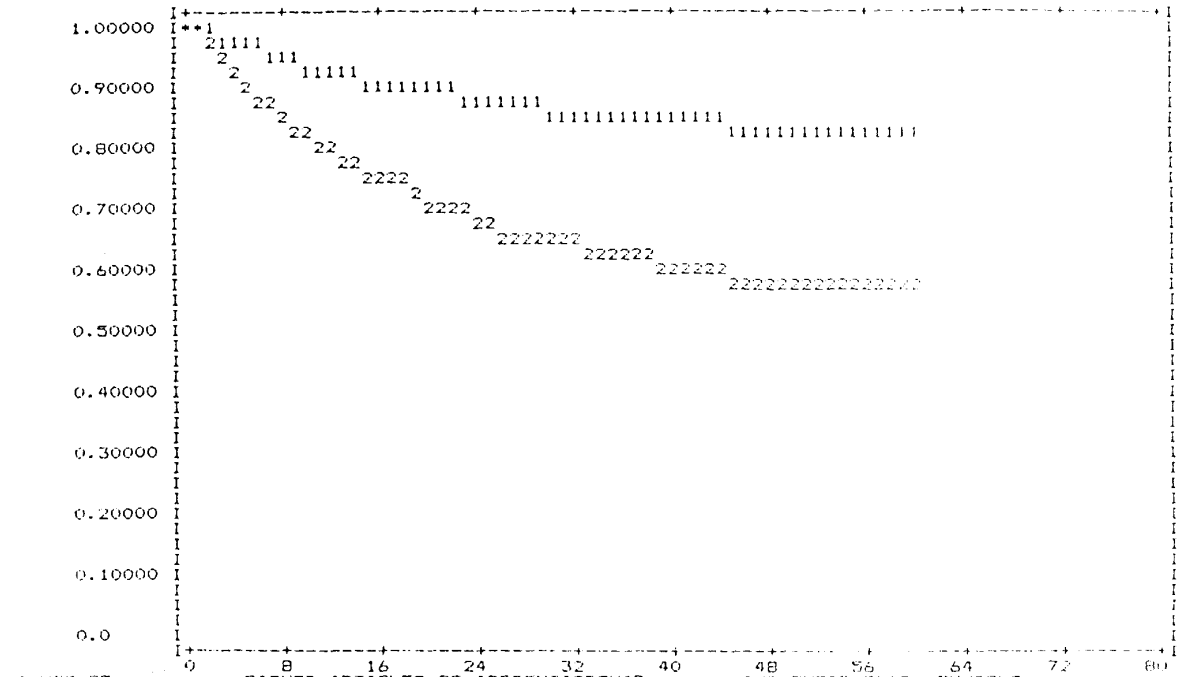


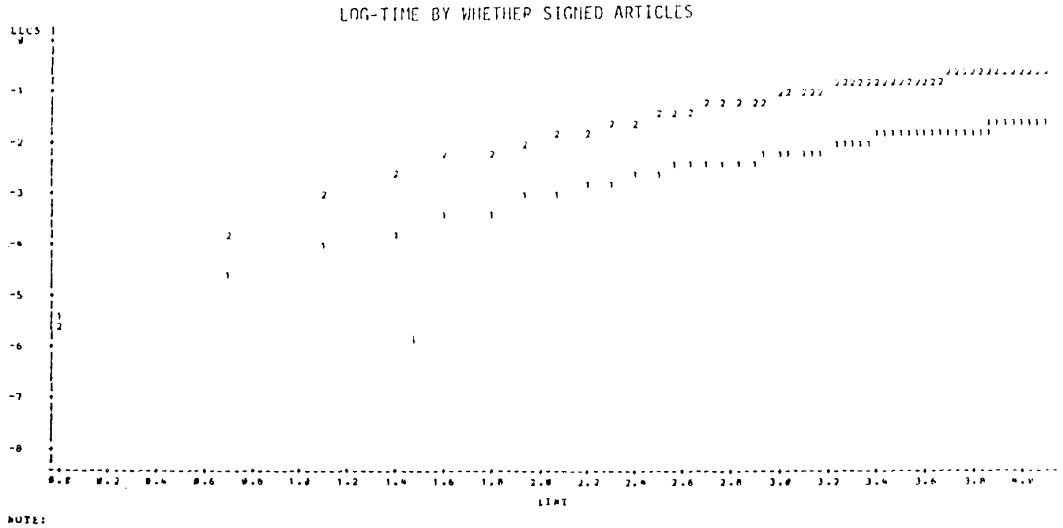
FIGURE 3(a) (CUMULATIVE) SURVIVAL FUNCTION BY WHETHER SIGNED ARTICLES



VALUES OF SIGNED ARTICLES OF APPRENTICESHIP AND THEIR GRAPH SYMBOLS

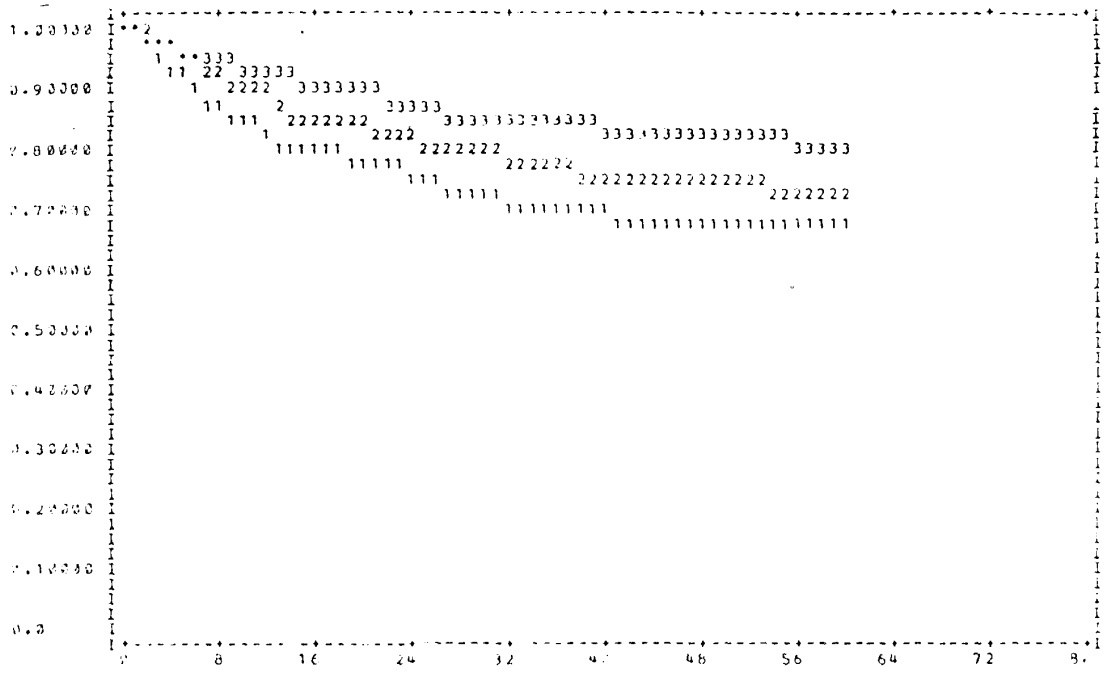
VALUE	GRAPH SYMBOL	VALUE LABEL
1	1	YES
2	2	NO

Figure 3(b) GRAPH OF LOG (-LOG SURVIVAL FUNCTION) VS LOG-TIME BY WHETHER SIGNED ARTICLES



NOTE:

FIGURE 4(a) (CUMULATIVE) SURVIVAL FUNCTION BY QUALIFICATIONS



VALUES OF QUALS AND THEIR GRAPH SYMBOLS

VALUE	GRAPH SYMBOL	VALUE LABEL
1	1	None
2	2	CSE only
3	3	O-le

Figure 4(b) GRAPH OF LOG (-LOG SURVIVAL FUNCTION) VS LOG-TIME BY QUALIFICATIONS

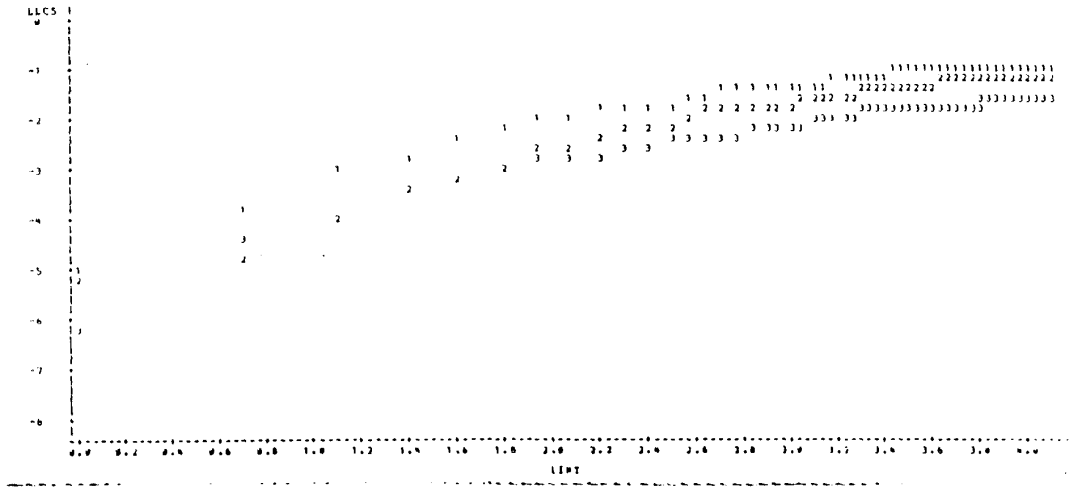
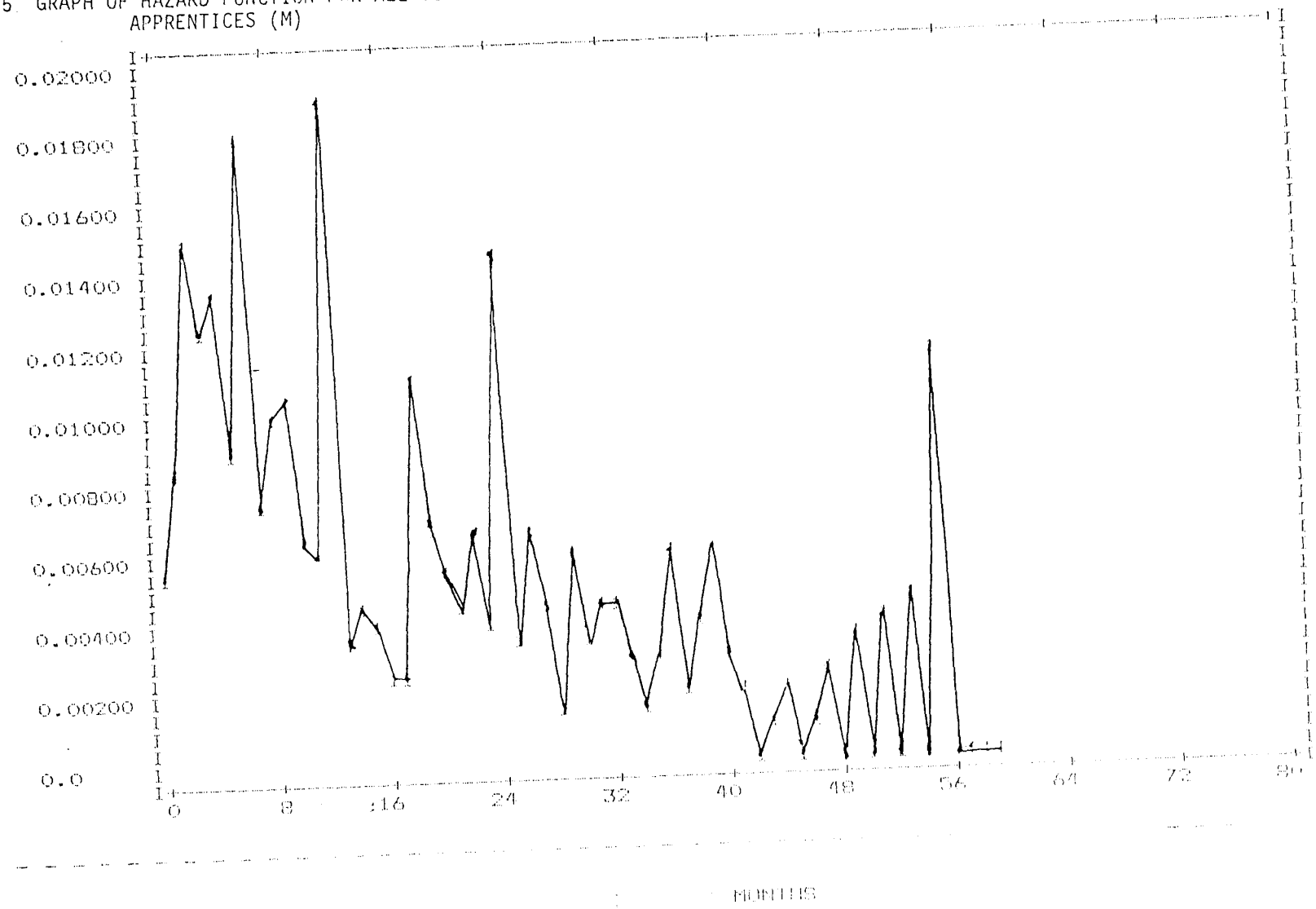


FIGURE 5 GRAPH OF HAZARD FUNCTION FOR ALL 'YOUNG' APPRENTICES (M)



Appendix

Sampling from the risk set where there are tied failure times

A number of writers (e.g. Liddell et al 1977, Breslow et al 1983, Cox and Oakes 1984) have discussed sampling from the risk set in a survival analysis to reduce the computational complexity of the maximum likelihood calculations without seriously effecting statistical efficiency. Breslow and Patton (1977) have shown that for case-control studies in the two sample problems, where the ratio of hazards is exp(B) to exp(o), the local (B=0) efficiency of 1:R matching compared with complete knowledge of the control population is R/(R+1). Because of the well-known equivalence of the likelihood function between case-control and survival studies, the same arguments hold for sampling from the risk set in survival analysis and Cox and Oakes (1984, p129) recommend that 8 to 10 survivors per failure usually suffice to recover most of the information from the full partial likelihood. Does the same ration of 1:8 hold when more than one failure occurs at a time? Common sense would suggest that it does, and this in fact does hold, as we now prove.

Following Breslow and Patton (1977) we consider for simplicity the two-sample problem where the explanatory variable has just one component which takes the values 1 and 0 for groups 1 and 0 respectively. Let

$$\tau_1 < \tau_2 < \ldots \tau_i \ldots < \tau_g$$

denote the g distinct failure times and D_k denotes the multiplicity of failure times at τ_k The partial likelihood is equal to the product over i of the conditional probabilities at each failure time.

We start by considering the situation at one time point. Dropping time subscripts for clarity, we see that each stratum corresponds to a 2 x 2 table of the form

	Group 1	Group 0	
Failure	X	D-X	D
Survivor	W	S-W	S
	u	D+S-u	D+S

where X = 0, 1, ...D denotes the number of failures in group 1.

W = 0, 1,.....S denotes the number of survivors in group 1

m = X + W denotes the total number in group 1 at that time

If S is large compared with D, and we wish to sample from the survivors to reduce the computation involved, how large a sample do we require to obtain an acceptably efficient estimate? Sampling from S gives the following 2 x 2 table.

	Group 1	Group 0	
Failure	X	D-X	D
Survivor	Y	R-Y	R
	t	D+R-t	D+R

X elements failing in group 1 contribute a factor proportional to

$$\frac{\exp(XB)}{R-t+D+t\exp(B)}$$

to the likelihood.

To find the efficiency of estimation based on sampling from the risk set we first find the contribution to the information conditional on the precise value of t obtained, and then find its expectation over the possible samples from the survivors. Cox and Oakes (1984, p104) show that the Information I(B) is the variance of X under weighted sampling without replacement from the risk set, the probability of selection of each element being proportional to exp (BZ) . For simplicity we find the value of the information I(0) at B=0 in order to allow us to test the Null Hypothesis. This is equal to the variance of a hypergeometric random variable so $I(0) = \frac{D.t.\ (D+R-t).R}{(D+R)^2\ (D+R-1)}$

If the probability of being in group 1 (Z = 1) is given by p = 1 -q under the null hypothesis, then the probability of a sample of (R + D) elements containing t elements from group 1 is given by

$$\binom{R+D}{t} p^t q^{R+d-t}$$

Then $E\{I(0)\} = \frac{DR}{R+D}\, pq$

and thus the local efficiency of D:R sampling from the risk set compared with complete knowledge D:∞ is given by $\frac{R}{R+D}$

In other words, as might be expected, provided the sample represents a relatively small proportion of the original risk set, the sampling fraction for a given efficiency depends on the ratio of failures to sampled survivors. The relationship does not depend on the proportion p of the total number at risk, so it is applicable at each time point, Consequently sampling throughout to give a ratio of 1 failure to 8 survivors will mean that a sufficiently efficient analysis can be performed.

REFERENCES

BRESLOW, N.E, LUBIN, J.H, MAREK, P & LANGHOLTZ, B(1983)
'Multiplicative Models and cohort analysis' in Jr.Amer Statist Assoc. 78,1-12

BRESLOW,N.E & PATTON, J (1977) 'Case-Control Analysis of Cohort Studies' in Energy and Health BRESLOW & WHITTEMORE (eds) Philadelphia , SIAM.

COX, D.R. & OAKES,D. (1984) Analysis of Survival Data
Chapman and Hall, London.

LIDDELL, F.D.K., McDONALD,J.C. & THOMAS, D.C. (1977)
'Methods of cohort analysis: appraisal by application to asbestos mining' Jr. Roy Statist Soc A, 140, 469-91.

National Child Development Study User Support Group Working Paper Series

No.	Title	Author(s)	Date
8.	Health and social mobility during the early years of life	C. Power K. Fogelman A.J. Fox	May 1986
9.	Effects of ability grouping in secondary schools in Great Britain	A.C. Kerckhoff	June 1986
10.	Leaving the parental home: an anlysis of early housing careers	G. Jones	July 1986
11.	Stratification in youth	G. Jones	July 1986
12.	Social class changes in weight-for-height between childhood and early adulthood	C. Power and C. Moynihan	July 1986
13.	Response to a national longitudial study: policy and academic implications for the study of change	D. Hutchison	August 1986
14.	Drop out from apprenticeship: an application of survival methods to grouped data	D. Hutchison	August 1986
15.	Event history and survival analysis in the social sciences: review paper and introduction	D. Hutchison	August 1986

National Child Development Study User Support Group, Working Paper No 14, August 1986. Social Statistics Research Unit, City university, Northampton Square, LONDON EC1V 0HB. Telephone (01) 253 4399

N A T I O N A L C H I L D D E V E L O P M E N T S T U D Y

The National Child Development Study (NCDS) is a continuing longitudinal study which is seeking to follow the lives of all those living in Great Britain who were born between 3 and 9 March, 1958.

It has its origins in the Perinatal Mortality Survey (PMS). This was sponsored by the National Birthday Trust Fund and designed to examine the social and obstetric factors associated with the early death or abnormality among the 17,000 children born in England, Scotland and Wales in that one week.

To date there have been four attempts to trace all members of the birth cohort in order to monitor their physical, educational and social development. These were carried out by the National Children's Bureau in 1965 (when they were aged 7), in 1969 (when they were aged 11), in 1974 (when they were aged 16) and in 1981 (when they were aged 23). In addition, in 1978, details of public examination entry and performance were obtained from the schools, sixth-form colleges and FE colleges.

For the birth survey information was obtained from the mother and from medical records by the midwife. For the purposes of the first three NCDS surveys, information was obtained from parents (who were interviewed by health visitors), head teachers and class teachers (who completed questionnaires), the schools health service (who carried out medical examinations) and the subjects themselves (who completed tests of ability and, latterly, questionnaires). In addition the birth cohort was augmented by including immigrants born in the relevant week in the target sample for NCDS1-3.

The 1981 survey differs in that information was obtained from the subject (who was interviewed by a professional survey research interviewer) and from the 1971 and 1981 Censuses (from which variables describing area of residence were taken). Similarly, during the collection of exam data in 1978 information was obtained (by post) only from the schools attended at the time of the third follow-up in 1974 (and from sixth-form and FE colleges, when these were identified by schools). On these last two occasions case no attempt was made to include new immigrants in the survey.

All NCDS data from the surveys identified above are held by the ESRC Data Archive at the University of Essex and are available for secondary analysis by researchers in universities and elsewhere. The Archive also holds a number of NCDS-related files (for example, of data collected in the course of a special study of handicapped school-leavers, at age 18; and the data from the 5% feasibility study, conducted at age 20, which preceded the 1981 follow-up), which are similarly available for secondary analysis.

Further details about the National Child Development Study can be obtained from the NCDS User Support Group.