Cox, D. R. (1981). Theory and general principles in statistics. J. Roy. Statist. Soc. A, 144, 289 - 297.

Feinstein, A. R. (1977). Clinical Biostatistics, Philadelphia, C. V. Mosby.

COMMUN. STATIST.-THEOR. METH., 11(19), 2129-2155 (1982)

THE ANALYSIS OF PROSPECTIVE STUDIES OF DISEASE AETIOLOGY

D. G. Clayton

Department of Community Health
Leicester University
Leicester, England

Key Words and Phrases: incidence rates; log-linear models; time-dependent risks; GLIM.

ABSTRACT

This paper reviews the analysis of prospective epidemiological studies using general linear models to describe disease incidence. It is shown that, apart from problems arising from the large size of most studies of this type, these models may be fitted by maximum likelihood (using GLIM, for example) assuming a Poisson likelihood. Alternative methods for dealing with large-scale data are discussed, and some simple procedures for dealing with common problems are outlined. The relationship of the approach to multiple logistic analyses is indicated.

1. INTRODUCTION

Aetiological studies are concerned to elucidate the factors relating to the onset of a disease in individuals. There are two types of such studies and prospective studies. For reasons of

economy and feasibility, the former are more commonly used and many recent papers have clarified the issues involved in their statistical analysis. This paper aims to review the methods of analysis of prospective studies.

It is a central tenet of this paper that prospective aetiological studies yield censored observations of failure times. The risk of contracting one of the diseases with which most epidemiological research is presently concerned, such as cancers and cardiovascular disease, increases with time within an individual (or at least does not decrease). Thus, it is not unreasonable to imagine that, given sufficient time, any studied individual would succumb to any particular disease considered. That such observations are not made may be attributed firstly to the finite duration of observation imposed by study feasibility, and secondly by the actions of other diseases, which eliminate individuals from further consideration by rather more drastic means. With this view, it might be regarded as more relevant to study not whether or not disease onset is observed in a given individual, but when it occurs.

In recent years, a considerable literature has accumulated concerning the analysis of censored survival time data. Most of this work concerns prognostic studies which seek to measure the time to death (or relapse) of patients after the commencement of treatment upon first diagnosis. Prospective epidemiological studies are superficially very similar to these, but pose their own particular problems. These are rarely addressed specifically, and arise mainly because:

(a) it is not always clear from which origin survival time of an individual is to be measured, and

(b) prospective studies are usually (although not always) of very large scale and pose special data-processing problems; although methods which require iterative solution involving many passes through the entire data are, these days, feasible, they could not be embarked upon lightly.

To this list of problems could be added the further requirement that such analyses must be convincing and readily explainable to non-statisticians. Although, perhaps, this requirement is not unique to epidemiological studies, it is particularly important in that area. The long term aim of aetiological studies is the prevention of disease, and the implementation of their findings involves the exertion of considerable political and social pressure. Such campaigns cannot readily be launched on the basis of analyses, understandable only to statisticians. For this reason, this paper will endeavour to relate proposed analyses to traditional epidemiological techniques, notably the 'standardisation' of rates.

2. THE MEASUREMENT OF INCIDENCE RATES

Consider first a disease, the risk of onset of which remains approximately constant with time within an individual. It is well known that the distribution of time-to-onset of such a disease in a homogeneous population is an exponential distribution. If, for an individual who is healthy at t, the probability of onset of disease during the interval $t \to t + \delta t$ is $\lambda\delta t$, then the distribution of time-to-onset is given by

$$f(t) = \lambda e^{-\lambda t} \qquad (1)$$

It is useful, also, to introduce the 'survivor function'

$$F(t) = \int_t^\infty f(u)du = e^{-\lambda t}$$

In general failure-time theory, λ is usually called the 'hazard' or 'failure rate'. In epidemiological terminology, when t represents time to first onset of clinically recognisable disease, then λ is the 'incidence rate' of the disease. The estimation of such an incidence rate from a prospective study is relatively straight forward. Let the i-th individual studied have been

studied from time to t_{oi} up to t_{li} at which time, either onset of the studied disease occurred or observation of the individual had to be terminated for some other reason. Let d_i indicate which of these possibilities occurred with values 1 and 0 respectively.

Let us assume, for the present, that the mechanism of censoring is unrelated to the disease process, (although we shall discuss this assumption later). Then, the log-likelihood for a group of N such individuals, all subject to the same incidence rate, is given by

$$L(\lambda) = \sum_{i=1}^{N} \log \frac{\{f(t_{li})\}^{d_i}\{F(t_{li})\}^{1-d_i}}{F(t_{oi})} \qquad (2)$$

$$= \sum_{i-1}^{N} \{d_i \log\lambda - \lambda(t_{li} - t_{oi})\} \qquad (3)$$

Now this likelihood is exactly the same as would be obtained by assuming $\{d_i\}$ to be independently distributed Poisson variates with expectations $\lambda(t_{li} - t_{oi})$. This assumption is clearly false since d_i cannot exceed 1 and $(t_{li} - t_{oi})$ are sometimes random variables rather than known constants. However, the likelihood is exactly the same as if the Poisson assumption held true, and this fact can be used to considerably simplify estimation. In particular, the asymptotic properties of the maximum likelihood estimate $\hat{\lambda}$ of the incidence rate, λ, holds good. It is given by;

$$\hat{\lambda} = \frac{\sum_{i-1}^{N} d_i}{\sum_{i=1}^{N} (t_{li} - t_{oi})} \qquad (4)$$

the total number of new cases observed, divided by the total of the periods of observation. This expression occurs in several

epidemiological texts as the definition of the incidence rate. Confidence intervals may be obtained for λ by noting that, asymptotically, $\log\hat{\lambda}$ is distributed approximately normally with variance $1/\Sigma d_i$. 'Exact' interval estimation is not possible without knowledge of the values the periods of observation $(t_{li} - t_{oi})$ would have been in those individuals who suffered onset of the disease. These cannot be known exactly for real studies.

Alternatively, a Bayesian approach may be adopted. A gamma distribution provides the most convenient prior for λ, yielding another gamma distribution as posterior. For large samples this approach yields the same interval estimates as the asymptotic approach described above.

It should be noticed that, since inference depends only upon the intervals $(t_{li} - t_{oi})$, and not upon either time alone, the choice of the origin for the time scale is, for the moment, irrelevant. This remains so until we consider time-dependent risks in section 7.

3. MODELLING INDIVIDUAL DIFFERENCES

It would, of course, be a very unambitious study which sought only to estimate the incidence rate of a disease in a population. In general, prospective studies involve measurements on the individuals studied at the time of their admission, $\{t_{oi}\}$. Thus, each individual will be characterised by vector, $\underline{x}_i$ say, of observations of social, geographical and behavioural variables (or factors) which may or may not be related to disease incidence. We shall call this the 'risk factor' vector in line with the usual epidemiological terminology.

In general, there will be more than one risk factor. Some factors may be of great interest to the investigators, and some may be well known but be so important that they cannot be ignored in the analysis. Risk factors may be categorical variables (such as social class, occupation, area of residence), or may be continuous measurements (such as blood pressure and serum cholesterol).

The way of investigating the effects of such risk factors is to assume some parametric form for the dependence of incidence rate upon risk factor, $\lambda(\underline{x}|\underline{\beta})$ say, where $\underline{\beta}$ is a vector of parameters. The most convenient functional form is the general linear model (Nelder and Wedderburn, 1972)

$$g(\lambda) = \mu + \underline{\beta}.\underline{x} \qquad (5)$$

The natural choice of 'link' function, $g(\lambda)$ is the logarithm, since this yields sufficient statistics for μ and $\underline{\beta}$ (as is the case for a simple Poisson dependent variable). With this link, the model is

$$\log(\lambda) = \mu + \underline{\beta}.\underline{x} \qquad (5a)$$

then it is clear that models specifying 'no interaction' between risk factors predict that factors act together multiplicatively in their joint effect upon the incidence rate. This model seems to perform well in a number of situations, at least as a first approximation. The best known examples are the multiple factors related in the incidence of coronary heart disease, as established by the Framingham study and elsewhere, and the joint effect of cigarette smoking and asbestos exposure in lung cancer.

It is important to notice, however, that two factors operating through two different and independent mechanisms would affect incidence additively rather than multiplicatively. It should also be noted that the implications of additive or multiplicative effects of risk factors may be quite different for preventive strategies. The fact that the logarithmic link is more natural statistically should not mean that its choice does not require some investigation.

It has been pointed out elsewhere (Aitkin and Clayton, 1980) that such a model for incidence rates may be fitted using the GLIM program (Baker and Nelder, 1978). The program may be tricked into constructing the correct likelihood by declaring the binary

variable $\{d_i\}$ as the YVAR (with POISSON errors), the link as LOG, and by using the quantities $\log(t_{1i} - t_{oi})$ as OFFSET. However, since most epidemiological prospective studies involve thousands rather than hundreds of individuals and since GLIM holds a data matrix of only limited size, this solution is usually not feasible.

If the risk factors are categorical, this difficulty may be circumvented by first forming a data summary consisting of a multi-way table containing in each cell (a) the number of new cases observed and (b) the total 'person-time' observation for the specified combination of categories of risk factors. The data in this table may then be entered into GLIM, with each cell representing one UNIT. The cell subscripts represent FACTORS in the analysis, and the two cell entries are treated exactly as described in the previous paragraph for the ungrouped data.

Even for risk factors, which are continuous measurements, a perfectly satisfactory approximate solution may be obtained by binning the data into strata with respect to the factor value, and assigning some central value (such as the within-stratum mean) to each stratum.

An example of these grouping methods is illustrated below. Table I shows some data taken from a study described by Morris et

TABLE I

Calorie Intake Range	(Mean)	Bus drivers	Bus conductors	Bank workers	All three occupations
-2249	(2145)	1 (6053)	2 (5230)	3 (6671)	6 (17954)
2250-2499	(2145)	4 (11011)	4 (4495)	3 (16126)	11 (31632)
2500-2749	(2672)	4 (14646)	3 (17497)	5 (22643)	12 (54786)
2750-2999	(2895)	3 (15516)	2 (10921)	5 (30712)	10 (57149)
3000-	(3369)	1 (23481)	3 (19410)	3 (52574)	7 (95465)
		13 (70707)	14 (57553)	19(128726)	46 (256986)

TABLE II

Model	DF	Deviance (Chi-squared)	Change
Overall mean only	14	21.342	
+ Occupation	12	19.386	1.956
+ Calorie intake	8	5.771	13.615

TABLE III

Factor Level	Effect: Additive (Log scale)	Effect: Multiplicative S.E.	Effect: Multiplicative %
Calorie Intake			
-2249	0.0	(By definition)	100
2250-2499	0.1118	0.5101	112
2500-2749	-0.4288	0.4996	65
2750-2999	-0.5916	0.5182	55
3000-	-1.4650	0.5580	23
Occupation			
Bus drivers	0.0	(By definition)	100
Bus conductors	0.3384	0.3888	140
Bank staff	-0.1131	0.3611	89

TABLE IV

	First-step Estimate of θ	Maximum likelihood Estimate of θ
Drivers/Conductors	0.678	0.667
Drivers/Bankers	1.111	1.112
Conductors/Bankers	1.532	1.573

al elsewhere (Morris et al., 1979). The data relates dietary measurements (one week weighed survey) in three different occupational groups, to subsequent incidence of coronary disease. The table shows the number of new cases of disease observed, together with (in parentheses) the man-months of observation, according to occupation and to total daily calorie intake. Tables 2 and 3 show the results of log-linear model-fitting, concentrating upon the effect of total calorie intake, expressed as a 5-level categorical variable. Table II shows the likelihood-ratio chi-squared analysis, and Table III shows the fitted effects in the 'main effects' model.

There is a strong gradient of decreasing risk with increasing calorie intake, but more modest effects of occupation. Indeed, the occupation effects are not statistically significant, but are retained in the model because of the a priori importance of this factor in the design of the investigation.

If the calorie intake effect is modelled by log-linear regression on the stratum means of calorie intake (2145 up to 3369), rather than treating the variable as a categorical variable, the final chi-squared test of fit of the model becomes 7.023 on 11 degrees of freedom. Thus the change in chi-squared for inclusion of calorie intake is 12.363 rather than 13.615, but on only 1 rather than 4 degrees of freedom. The estimated regression coefficient is -0.001310 which compares very closely with the value of -0.001309 obtained by repeating the same GLIM analysis upon the ungrouped data.

4. COMPUTING PROBLEMS

Two practical difficulties are encountered with the methods described above. The first arises out of the now rather surprising lack of generally available survey analysis programs which will form the multiway-tables required in directly machine-readable form. A notable exception is the Rothamsted General Survey Program (RGSP), which has interfaces to both GLIM and GENSTAT (Beasley et al., 1980).

The second difficulty arises out of the fact that, although highly convenient and almost universally available, GLIM is not particularly efficient for larger problems. For categorical risk factors, and providing the multiplicative model (Logarithmic link) only is required, a much more efficient algorithm is provided by the 'iterative scaling procedure'. Slight modification of the classical algorithm is required to ensure that the fitted values are correctly offset with the person-time observation (an A.N.S.I. Fortran algorithm is available from the author).

5. RELATIONSHIP TO THE 'MULTIPLE LOGISTIC' METHOD

Those familiar with the epidemiological literature will have recognised that the method proposed above is different from that usually employed in the analysis of prospective studies, the 'multiple logistic' method, although it resembles it in many ways. The multiple logistic fits the model

$$\text{logit } \pi = \log \frac{\pi}{1-\pi} = \mu + \underline{\beta}'.\underline{x} \qquad (6)$$

where π is the probability of an individual suffering disease onset during the period of observation. This model first obtained wide recognition in epidemiological research with the work of Truett et al (1967), who fitted the model using the assumption of multivariate normality of the risk factor variables within disease groups (this is equivalent to classical discriminant analysis). Later workers, following Walker and Duncan (1967), dropped the multivariate normal assumption and fitted the model directly by maximum likelihood (but at the cost of an iterative, rather than single-pass solution). A full discussion of the model is given by Cox (1970).

This method has several serious disadvantages, however. Firstly, and perhaps most seriously, it relies upon each individual being observed for the same period of time, so that the

probabilities, π, are comparable. This is not achievable in real studies, (a) because recruitment to the study often extends over years so that, at the time of analysis, some individuals have been observed for longer than others; (b) because of migration from the study population, and (c) because of deaths from other causes. This problem can be circumvented to some extent, but only by discarding data.

The second disadvantage is that the logistic model does not allow the analysis to extend to variables which change <u>during</u> the observation period. Thus, although it is possible to take account of age-at-entry to the study upon subsequent risk, it is not possible to take account of any ageing occurring during the study. This difficulty is particularly serious for studies of long-duration. Its resolution with the present approach will be discussed in a later section.

These difficulties apart, the models are very similar. With the assumption of no change in risk with time, the relationship between the incidence rate, λ, and the probability of disease onset during $t_0 \to t_1$, π, may easily be shown, from (1), to be

$$\pi = 1 - \exp\{-(t_1 - t_0).\lambda\}$$

so that

$$\log\{-\log(1 - \pi)\} = \log\lambda + \log(t_1 - t_0)$$

Thus, the analysis of probabilities of onset using the complementary log-log transformation is equivalent to analysing incidence rates using the simple logarithmic link. It is well known that, for small π, the logit and complementary log-log transformations are nearly identical.

Incidentally, a side effect of the widespread use of the multiple logistic method has been to refer to the probability of onset π, quite incorrectly, as an incidence rate. Thus, we have the appearance of the terms '5-year incidence rate' and '10-year

incidence rate'. Such a confusion is to be deplored, and is a recent phenomenon. William Farr, for example, in his writings a century ago drew a clear distinction between the two measures. The same point was made by Elandt-Johnson (1975) in the American Journal of Epidemiology, and provoked some controversy in later issues of that Journal.

6. SOME SIMPLE METHODS

Before moving on to more difficult matters, it is worth noting that, arising out of the general theory described in earlier sections, there are some methods requiring only very simple calculations.

The first of these is for the analysis of the k x 2 summary table of cases and person-time observation, where the prime interest is in the 2-level risk factor. Obviously, the k-level factor is included so that its effect may be discounted in the analysis, and it might be generated by crossing several 'nuisance' factors.

Let $\lambda_{ij}(i = 1 \ . \ . \ k;\ j = 1, 2)$ be the incidence rates corresponding to each cell of the table, and let d_{ij}, T_{ij} represent the corresponding observed number of new cases and person-time observation. The 'main-effects' multiplicative model may be written

$$\lambda_{i1} = \theta.\lambda_{i2} \quad \text{for all } i,$$

or

$$\theta = \lambda_{i1}/\lambda_{i2} \quad \text{for all } i.$$

Thus, θ represents the relative incidence rate between columns within rows of the table. It is easily shown that the maximum likelihood estimate of θ may be obtained by solving

$$\theta = \frac{\sum_i W_i \, d_{i1} \, T_{i2}}{\sum_i W_i \, d_{i2} \, T_{i1}} \qquad (7)$$

by iterative refinements of the weights

$$W_i = 1/(\theta.T_{i1} + T_{i2}), \text{ starting from } \theta = 1.$$

The first step of this iteration itself provides a fully consistent estimate, although it is only fully efficient in the neighbourhood of $\theta = 1$. This first-step estimator is closely related to the Mantel-Haenszel estimator of the common odds-ratio in the 2 x 2 x k table, and the fact that the first stage of the procedure leads to quite a good estimate means that convergence is very rapid; a single refinement stage is all that will be required, except when θ is very far from 1.

The standard error of $\log \theta$ is given by the expression

$$\text{S.E. } (\log \hat{\theta}) = \{\sum_i W_i^2 \, \theta \, T_{i1} \, T_{i2} \, (d_{i1} + d_{i2})\}^{-\frac{1}{2}}$$

where W_i are the final weights. Of course, if the first step estimate is used, the first value of this expression gives the null s.e. of the log of the estimate, so that a test of $H_0: \theta = 1$ may be constructed. An alternative test will be discussed immediately below.

This case seems to be the only one in which relatively simple, yet efficient, estimates of at least some of the parameters of the general multiplicative model, (5A) may be obtained. However, some simple methods remain, based upon tests of hypotheses. In epidemiology, the analysis of prospective studies often involves the examination of a large number of potential risk factors 'in search of hypotheses'. Typically, there will be one or two known factors related to both risk of disease, and to the level of the new factor(s) to be screened. It is not desirable to embark upon a full model-fitting exercise (involving, as it does, some considerable computation) for each such factor. Although, of course, nominal significance levels must be treated with considerable caution, hypothesis tests can be most useful, particularly tests based upon 'score statistics' (Cox and Hinkley, 1974).

We shall consider, first, the score test for column effects in the k x m table of incidence rates (the equivalent test for row effects is obvious). This test involves calculating the vector, $\underline{r}$, of discrepancies between the observed number of cases of disease in each column and the number 'expected' on the basis of the distribution of the row variable; i.e.

$$r_j = d_{.j} - \sum_i (d_{i.} T_{ij}/T_{i.})$$

The score test is a chi-squared test on (m-1) degrees of freedom, and is given by the quadratic form

$$X^2_{(m-1)\ d.f.} = \underline{r}' . \underline{R}^{\theta} . \underline{r} \qquad (8)$$

where θ indicates a generalised inverse, and $\underline{R}$ is the matrix

$$R_{jk} = \delta_{jk} \sum_i (d_i T_{ij}/T_{i.}) - \sum_i \{d_{i.} T_{ij} T_{ik}/(T_{i.})^2\}$$

(δ_{jk} is the Kronecker delta). This test is, however, rather cumbersome to calculate and an extremely close approximation is provided by the Pearsonian $(O-E)^2/E$ formula;

$$X^2 \simeq \sum_j \{(r_j)^2 / \sum_i (d_{i.} T_{ij}/T_{i.})\}\ ,\ \text{a conservative approximation.}$$

This test could be used on the data of Table 1 to test whether there is a difference between the calorie intake groups, over and above any occupational differences. In this example the true score test and the approximate Pearsonian test agree closely, yielding chi-squared values of 13.647 and 13.558 respectively, and also agree closely with the likelihood ratio chi-squared value of 13.615 (table II).

However, these tests are not ideal for examining the effect of calorie intake, since they are on 4 degrees of freedom and ignore the fact that the calorie breakdown reflects an underlying continuum. They could not, therefore, be expected to be sensitive

against an alternative hypothesis of a steady trend in risk with increasing total calorie intake. The remainder of this section is concerned with some more sensitive methods for detecting relationships between incidence rates and risk factors measured on a continuous interval scale.

When the continuous factor, x, is the only one under consideration, the log-linear model (5(a)) becomes

$$\log \lambda = \mu + \beta . \underline{x}$$

Although the maximum likelihood estimate of β requires iterative computation, the score test for $\beta = 0$ is remarkably simple and deserves to be more widely known in epidemiology. If X_i is the value of the factor for the i-th individual, d_i indicates whether or not (with values of 1 and 0 respectively) onset of disease was observed in that individual, and T_i represents the time for which he or she was observed, then it may easily be shown that considerations leading to the log likelihood of the form of (2) and (3) lead to the score statistic

$$\left.\frac{\partial L(\mu,\beta)}{\partial \beta}\right|_{\substack{\mu=\mu \\ \beta=0}} = \bar{X}_1 - \bar{X}_0$$

$$\text{where } \bar{X}_1 = \sum_{i=1}^{N} d_i X_i / d.,$$

the 'expected' value which is the mean of all measurements using the observation times, T_i, as weights. The sampling variance of this statistic is simply

$$(S_0^2 - \bar{X}_0^2)/d. \qquad (10)$$

where

$$S_0^2 = \sum_{i=1}^{N} T_i X_i^2 / T.,$$

the mean square of X, using observation times as weights. Thus, from (9) and (10) a simple asymptotic standard normal deviate test may easily be constructed.

This method is convenient, simple and can provide convincing evidence of an effect of a continuous factor. It may also be readily adapted to examine an association within strata of the study population. Such an analysis can indicate, firstly, whether the stratifying variable, by confounding, has spuriously exaggerated the effect of the factor considered and, secondly, whether the effect is consistent throughout the data. Using the superscript (j) to indicate strata, the procedure is to compare the observed stratum means for 'cases', $\bar{X}_1^{(1)}$, with their 'expected' values $\bar{X}_0^{(j)}$. The overall test for the effect of X, after adjustment for the stratifying variable is given by comparing the overall mean in cases,

$$\bar{X}_1 = \sum_j d_.^{(j)} \bar{X}_1^{(j)} / \sum_j d_.^{(j)}$$

with

$$\bar{X}_0^* = \sum_j d_.^{(j)} \bar{X}_0^{(j)} / \sum_j d_.^{(j)}$$

which is the 'expected' value which takes account of the relationship between disease incidence and the stratifying variable.

The sampling variance of the difference $(\bar{X}_1 - \bar{X}_0^*)$ is

$$\sum_j d_.^{(j)} \{(S_0^{(j)})^2 - (\bar{X}_0^{(j)})^2\} / \{\Sigma\, d_.^{(j)\ 2}\}$$

This method is simple, convincing and computationally tractable. An example of its use is discussed at the end of the next section.

7. TIME DEPENDENCE OF RISKS

So far, the discussion has been concerned solely with problems in which it is to be assumed that incidence rates do not vary within the study period. This will usually only be a legitimate assumption for studies with a very short duration of follow-up. In general, risk of onset of disease will vary during the observation period for several reasons, notably

(a) increase of risk with increasing age of the subject,

(b) increase of risk with increase in duration of the length of time exposed to causal agents, and

(c) dependency of risk upon the duration of time in the study.

(d) Secular effects operating at the time of disease onset; i.e. dependency of risk upon chronological time.

Most of the literature concerning survival data is concerned with determining effect of various factors upon prognosis. Usually, in such studies, patients are admitted to the study on diagnosis of the disease. Thus, it is survival time, measured from the start of study, which is of prime interest. The variation of risk throughout the study is dominated by the progression of the disease, so that the relationship of risk to time under study is of most importance. In prospective studies of previously healthy individuals this is no longer the case and, a priori, it is more natural to expect the first two reasons for time dependency of risk to be dominant. In general, however, these variables must be expected to be heavily confounded so that it might be difficult to disentangle their effects. Only in certain cases will duration of exposure to a causal agent be known, so that the first choice of time scale will be age, although it may be a surrogate for some other scale.

The multiple logistic analyses described earlier cannot allow for such dependencies. For example, age at entry into a study may be incorporated into the model as a risk factor, but any effect of ageing during the study is ignored. The present approach is, however, easily adapted. The contribution to the log-likelihood of an individual observed from t until t and subject to incidence rates $\lambda(t)$ and corresponding survivor function, $F(t)$ may easily be shown to be

$$d \log \lambda(t_1) - \Lambda(t_0, t_1)$$

where

$$\Lambda(t_0, t_1) = \int_{t_0}^{t_1} \lambda(u)\, du$$

and, as before, d indicates whether or not onset of the disease was observed. In its simplest form, the log-linear model would specify that the incidence rates for individuals with risk factor vector $\underline{X}$ as being

$$\log \lambda(t/\underline{x}) = \log \lambda_0(t) + \underline{\beta}'.\underline{x} \quad \text{(Cox, 1972)}$$

In this form the model specifies that the effect of the factors is not time dependent, but it is perfectly straightforward to incorporate 'interaction' terms to allow the factor effects to vary with time.

There are four possible approaches to the problem of what to do about $\lambda_0(t)$. We shall consider them, briefly, one by one.

7.1 KNOWN $\lambda_0(t)$

This provides the most simple analysis. Aitkin and Clayton (op. cit.) show that the analysis of this problem in GLIM is straightforward; as before, the indicator variable, d, is declared as the YVAR with POISSON errors, but, now, $\log \Lambda_0(t_0, t_1)$ must be used as OFFSET. Note that $\Lambda_0(t_0, t_1)$ is the 'expected' value of d if the known rates, $\lambda_0(t)$ had applied throughout the interval:

$$\lambda_0(t_0,t_1) = \int_{t_0}^{t_1} \lambda_0(u)\,du$$

Many readers will recognise that this analysis is essentially the traditional method of indirect standardisation, instead of analysing incidence rates, one analyses the ratio of observed new events to those expected had $\lambda_0(t)$ applied. Thus, if $\lambda_0(t)$ is known, all the analyses described in earlier sections may be applied with this small modification. However, this argument also shows that the method of indirect standardisation depends upon knowledge of $\lambda_0(t)$, and only rarely is this the case. In traditional epidemiological analyses, $\lambda_0(t)$ is provided by a suitable

set of 'index' rates. In the absence of such a suitable set of rates, it is necessary to adopt one of the remaining three strategies.

7.2 ASSUME A PARAMETRIC FAMILY FOR $\lambda_0(t)$

If $\lambda_0(t)$ is not known (as is usually the case), it might be that, at least, we may assume it to be a member of a parametric family of functions, $\lambda_0(t;\underline{\gamma})$ say. Aitkin and Clayton (op. cit.) have shown how certain families might be fitted using GLIM. Essentially the approach is to use the method described above for estimation of the regression equation conditional upon the current estimate of $\underline{\gamma}$, then to update the estimate of $\underline{\gamma}$, to re-estimate the regression equation and so on. There are two difficulties with this approach. Firstly, the choice of parametric function may be difficult. Where t represents either duration of exposure to some factor or age (which may or may not be a surrogate for duration of exposure to, as yet, unknown factors), there is empirical evidence to suggest that a power-law relationship of the form

$$\lambda_0(t) = \gamma_1(t - \gamma_2)^{\gamma_3}$$

provides an appropriate model in a wide range of settings (see, for example, Doll. 1972). When, however, t represents the time elapsed since admission into a prospective study, it is more difficult to suggest an appropriate function. Two processes may operate. Heterogeneity of risk within a selected cohort will yield an apparent fall in incidence with time as the high risk groups succumb early (this model underlies the Pareto distribution, a common failure-time distribution). On the other hand, there is a well known tendency in prospective studies for a selection bias towards healthy individuals, so that initial incidence might be low. The combined effects of these two processes might produce a variety of shapes of curve.

A second difficulty with this approach is, again, the computation involved with the scale of data generally encountered in prospective studies. Several passes through the raw data would be required for each model to be fitted.

7.3 ASSUME A STEP-FUNCTION FOR $\lambda(t)$

An alternative approach is to stratify the time axis, and to assume constant incidence rates within time bands. For example, if t represents age, an individual might be assumed to be subject to incidence rate λ_1 between the ages of 30 and 39, λ_2 between the ages of 40 and 49 and so on. The log-linear model might then be written, for the i-th age band,

$$\log \lambda_i = \alpha_i + \underline{\beta}'.\underline{X}$$

It may easily be shown that the likelihood for this model is identical to that which would be obtained if each individual studied were to be treated as several 'pseudo individuals' - are for each age band within which he is observed. Thus, an individual observed from age 35 until he suffers onset of disease at age 55 may be treated as if he were three different people: (a) one person of age 30-39, five years observation, no onset of disease; (b) one person of age 40-49, 10 years observation, no onset of disease; (c) one person of age 50-59, 5 years observation, suffering disease onset.

With this device, the methods described in earlier sections for the constant incidence rate model may be used. In particular, for categorical factors, the data may be summarised in multiway tables prior to analysis; the time axis becomes simply another dimension of the table. The formation of such tables is fairly straightforward but cannot usually be carried out directly by a survey analysis program, since one individual is required to contribute to several cells of the table. Where the survey analysis system allows a user-supplied input routine, the problem is

TABLE V

Bank staff

Age at attack of CHD	No. of cases	Energy intake (kcal) Observed	Expected
40-49	4	2769	3015
50-59	8	2514	2894
60-69	7	2725	2846
Total	19	2645	2902

Bus drivers

Age at attack of CHD	No. of cases	Energy intake (kcal) Observed	Expected
40-49	2	2918	2853
50-59	4	2808	2838
60-69	6	2458	2833
Total	12	2651	2838

Bus conductors

Age at attack of CHD	No. of cases	Energy intake (kcal) Observed	Expected
50-59	5	2515	2845
60-69	9	2718	2828
Total	14	2646	2834

All occupations

Age at attack of CHD	No. of cases	Energy intake (kcal) Observed	Expected
40-49	6	2819	2961
50-59	17	2583	2867
60-69	22	2649	2835
Total	45	2647	2864

easily solved by writing a routine which reads the data for an individual and, on successive calls, passes the data for 'pseudo individuals' to the main program. If this facility is not available, then a preprocessor program must be written to expand the data to a disc file of 'pseudo individuals'.

Table V is taken from Morris et al (op. cit.) and shows one such analysis using the observed and 'expected' mean method described in section 6. The experience of the three occupational cohorts in three age bands is examined; association between risk and total calorie intake is remarkably consistent. The 'expected' means in 'all ages' now takes account of differences in age structure of the different occupations and, likewise, those in the 'all occupations' column standardises for occupation. The overall 'expected' mean adjusts for both age and occupation, and differs

TABLE VI

Age	Calorie Intake				
	-2249	2250-2499	2500-2749	2750-2999	3000-
40-49	0 (2298)	1 (4460)	1 (9461)	3 (10260)	1 (21349)
50-59	3 (7359)	5 (14096)	4 (24203)	2 (24438)	3 (41713)
60-69	3 (6675)	4 (10597)	7 (17439)	5 (18902)	3 (27321)

highly significantly from the observed mean calorie intake in the coronary disease cases (s. n. d. = 3.36, P<0.001).

Table VI shows, for all occupations, observed number of cases and (in parentheses) the person-weeks observation in the calories x age classification. Again, age refers not to the age at entry to the study, but the age at which exposure to risk is observed. Fitting the multiplicative model to this table yields, for the age-adjusted multiplicative effects of the calorie-intake groups; 100% (by definition), 95.9%, 66.7%, 52.6% and 22.6%. These multiplicative effects may be thought of as age-standardised incidence ratios, and Mantel and Stark (1968) have referred to their calculation as 'internal' indirect standardisation - indirect standardisation without an external reference set of rates. The numerical method given by these authors is the weighted form of the iterative scaling procedure mentioned in section 4.

The score test, (8), for the column effect in table 6 yields a chi-squared of 11.991 on 4 degrees of freedom for the effects of calorie intake after allowing for age effects, and the approximate form of the test gives 11.914. In this context, this test is closely related to the log rank test of Peto and Peto (1972).

The device of 'discretising' the time scale can be useful when fitting smooth parametric families to $\lambda_0(t)$ to avoid some of the computational difficulties mentioned in 7.2. All disease experience within a discrete band is treated as if it were experienced at some central time. This method is approximate, but allows the data to be grouped before model fitting. Gehan and

Siddiqui (1973) have discussed essentially this method for fitting the Weibull distribution (power law for $\lambda(t)$).

7.4 ARBITRARY $\lambda_0(t)$

Finally, it is possible, using the method of Cox (1972) to consider arbitrary $\lambda_0(t)$. The likelihood is based upon construction, for each 't' at which a case of disease occurs, the set of individuals at risk. Thus, when 't' represents age, the risk set is made up of all those individuals under observation at the age at which a case of disease occurred. Clearly, the computational problems of this procedure are considerable with the scale of data commonly encountered in prospective studies. However, little efficiency is lost by the replacement of each risk set by a much smaller one made up of the index case and several 'controls', randomly sampled from the disease-free members of the risk set. Mantel has described this method as a 'synthetic retrospective study' (Mantel, 1973). The procedure seems to have little to recommend it except for studies which involve very laborious coding of records; such coding may then be restricted to the cases and relatively few controls (see, for example, Morris et al 1973).

8. MULTIPLE TIME AXES

The methods of section 7 apply regardless of which time axis is to be considered. Often, however, more than one time variable will need to be considered simultaneously. Ultimately, it may be desirable to attempt to disentangle, say, age, time since entry into the study, and duration of exposure to pathogen. The methods described above may readily be adapted to such an analysis, though methods 7.1 and 7.2 will probably not be practicable. Method 7.4 reduces simply to the choice of controls 'matched' with respect to all three time variables. Method 7.3 simply requires further proliferation of 'pseudo individuals'. Each individual may contribute to any cell in the three-way grid formed by stratifying the time

variables; if he suffers disease onset, this is ascribed to the cell in which it occurred, and this total observation time is partitioned between all the cells in the grid.

The algorithm for partitioning the observation time is simple and requires only routines for choosing the earliest and latest of a set of dates and for calculating the elapsed time between two dates. For example, if we wish to determine the observation time of one individual (a) during the age range 40-49, (b) within 2-4 years of entry to the study and (c) with 5-10 years of first exposure to some risk factor, then the procedure is as follows.

i. Choose the LATEST of the three dates:
   Date of birth + 40 years,
   Date of entry into study + two years, and
   Date of first exposure + five years.

ii. Choose the EARLIEST of the four dates:
   Date of birth + 50 years
   Date of entry into study + four years,
   Date of first exposure + 10 years, and
   Date of exit from study.

iii. If (ii) precedes (i), then the individual makes no contribution to this cell.
   Otherwise, the observation time contributed is the time interval from date (i) until date (ii).

Analysis of the resultant multiway tables may procede as described in the remainder of this paper. It is interesting at this stage to mention 'birth cohort effects', i.e. effects attributable to the chronological date of <u>birth</u> of an individual. This time variable is not a time axis in the sense considered here, since it does not vary within individuals, but, birth cohort effects may be manifested as a particular form of interaction between age and chronological time.

9. MULTIVARIATE PROBLEMS

A casual reader might be forgiven for thinking that multivariate problems had already been discussed! Earlier sections

discussed analyses involving multiple risk factors, and section 8 discussed analyses involving 'multivariate time', but in all these problems only a single disease process is involved and it is modelled by a single stochastic process. Thus, the theory is essentially that of a univariate problem. However, aetiological studies present problems in which more than one disease process is involved. These problems have received little attention, and present some considerable difficulties, and, although it would not be appropriate to deal with them in detail here, they should be pointed out.

The first problem is that of 'family history', relating the disease experience of an individual to the disease experience of his parents, his siblings and other relatives. Adequate models of this problem must involve more than one stochastic process, these being to some extent interdependent. Elsewhere, I have suggested a class of models which seem to have highly desirable characteristics (Clayton, 1978). This paper also discusses some methods of inference from prospective studies. Unfortunately, Oakes (1981) has pointed out that the method of estimation proposed overstates the precision of the key parameter estimate.

The second important problem involves multiple disease processes <u>within the same individual</u>. This is usually referred to as the problem of 'competing risks' and has been discussed in detail by Prentice et al (1978). Here we have followed conventional epidemiological practice in concentrating upon incidence of one particular disease. Death from (and usually even incidence of) other diseases preclude further observation of the individual and has, therefore, been treated as simply a mechanism of censoring. This is legitimate only if the different disease processes are independent of one another. Unfortunately, the nature of the censoring is such that no information is available for testing the truth of this assumption. The only sensible way out of this impass would seem to be the use of computer simulation to investigate the importance of the difficulty in any particular case. The models mentioned above for the family history problem (Clayton, 1978), would seem very suitable for this purpose.

BIBLIOGRAPHY

Aitkin, M. & Clayton, D., (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. Applied Statistics, 29, 156.

Baker, R. J. and Nelder, J. A., (1978). General linear interactive modelling (GLIM) release 3. Numerical Algorithms Group, Oxford.

Beasley, J. D., Church, B. M. & Yates, F. (1980). The Rothampsted general survey program, fourth edition, Rothampsted Experimental Station, Harpenden, Hants, 1980.

Clayton, D., (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.

Cox, D. R., (1970). The analysis of binary data. London: Methuen.

Cox, D. R., (1972). Regression models and life tables. J. R. Statist. Soc. B, 34, 187-202

Cox, D. R. & Hinkley, D. V., (1974). Theoretical statistics. London: Chapman & Hall.

Doll, R., (1971). The age distribution of cancer: implications for models of carcinogenesis. J. R. Statist. Soc. A, 134, 133 -155.

Elandt-Johnson, R. C., (1975). Definition of rates: some remarks on their use and misuse. Am. J. Epidemiology, 102, 267-271.

Mantel, N., (1973). Synthetic retrospective studies and related topics. Biometrics, 29, 479.

Mantel, N. & Stark, C. R., (1968). Internal indirect standardisation of rates. Biometrics, 24, 997.

Morris, J. N., Adam, C., Chave, S. P., Sirey, C., Epstein, L. & Sheehan, D. J., (1973). Vigorous exercise in leisure-time and the incidence of coronary heart disease. Lancet, Feb. 17, 333-339.

Morris, J. N., Marr, J. W. & Clayton, D. G., (1977). Diet and heart: a postcript. Brit. Med. J., 1977, 2, 1301-1368.

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

Oakes, D., (1981). A model for association in bivariate survival data. Unpublished manuscript.

Peto, R. & Peto, J., (1972). Asymptotically efficient rank invariant test procedures. J. R. Statist. Soc., A, 135, 185-198.

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

Truett, J., Cornfield, J. & Kannel, W., (1967). Multivariate analysis of the risk of coronary heart disease in Framingham. J. Chronic Dis., 20, 511.

Walker, S. H. & Duncan, D. B., (1967). Estimation of the probability of an event as a function of several independent variables. Biometrika, 54, 167-179.