COVARIATE MODELS FOR ACCIDENT DATA

GILBERT MacKENZIE*
DEPARTMENT OF COMMUNITY MEDICINE, QUEEN'S UNIVERSITY, BELFAST

(Read before the Society, 12 December 1985)

1 INTRODUCTION

The statistical analysis of accident data has historically relied on the mathematical precepts of classical discrete distribution theory Since the first test of the standard null ("pure chance") hypothesis on accident mortality data, relating to horse-kicks in ten Prussian Army Corps during 1875-1894 (von Bortkiewicz, 1898), various models have been advanced to explain departure from Poisson's (1837) law

The ostensibly separate hypotheses of "proneness" (termed "unequal liability" Greenwood and Yule (1920) and "false contagion" by others Bates, *et al* , (1952), and "contagion" have received extensive attention in the literature (Newbold, 1927, Anscombe, 1950, Bliss and Fisher, 1953, Neyman, 1939, Cresswell and Froggatt, 1963, Kemp and Kemp, 1965 and Kemp, 1967) However, despite considerable development of the associated statistical models, formidable problems of interpretation remain (Froggatt, 1968a)

One complication of the classical approach to the analysis of the accident data arises from the observation

---

* I would like to thank Sir Peter Froggatt, Vice-Chancellor of Queen's University of Belfast and formerly Professor of Epidemiology in the Department of Community Medicine, Queen's University Belfast, for many interesting discussions on the analysis of accident data, and Dr Richard Kay, Department of Statistics, University of Sheffield, for a first introduction to Cox model and much valuable help thereafter

that apparently fundamentally different sets of defining assumptions may generate the same statistical model Such duality of interpretation (for example, the Neyman Type A, Short, and Negative Binomial may each be derived on either of the hypotheses referenced above) appears characteristic of compound Poisson models which are generalised Poisson distributions (Kemp, 1968)

Consequently, it is methodologically difficult (if not impossible) to discriminate between models on the evidence of the data above

Another difficulty relates to the compounding process *per se* The result of modifying the basic Poisson distribution by permitting "liability" to vary over individuals is well known - we arrive at the Compound Poisson

This procedure makes allowance for the hypothesised inter-individual variation in λ but, unfortunately, provides no direct measure of a particular individual's liability The absence of such a measure hampers the twin processes of hypothesis formulation and testing in relation to the possible influence of measurable social and biological characteristics In this respect the contemporary analysis of accident data seems peculiarly unique

A further difficulty arises in relation to the design of the accident studies Adelstein (1952) commented "The statistical approach is to observe a group of persons equal, as far as possible in all relevant respects and exposed to the same essential conditions for a period of time" These requirements would appear unnecessarily rigorous and are, in any case, unlikely to be fulfilled in practice

The purpose of the present paper is to attempt to obviate some of these constraints by modelling the underlying transition probabilities intra-individual on the real-time axis By adopting a proportional hazards parameterisation (Cox, 1972) the effects of covariates on the accident-incidence can be assessed directly The resulting class of stochastic models assumes less structure, *a priori*, than the point processes discussed by Cox and Isham 1980

2 MODEL FORMULATION

In the development which follows accident-incidence is modelled intra-individual on the real-time axis Let A = $(a_o, a_1, \quad , a_r)$ denote the finite collection of ordered transient states traversed by an individual on study for time T_r The state a_o corresponds to an initial, accident-free, state while states $a_k (k=1, \quad r)$ represent accidents Associated with these latter states are times T_k at which transitions occur from states a_{k-1} to a_k so that the underlying stochastic process may be represented by $(a_o, T_1, a_1, T_2 \quad)$

Following Cox (1972) the instantaneous incidence rate for the kth accident may be modelled as

$$\lambda_k(t_k, x_k) = \lambda_{ok}(t_k) \exp (x_k \ \beta_k) \tag{1}$$

where the vector of unknown parameters $\beta_k = (\beta_{k1}, \quad , \beta_{kp})$ measures the influence of the covariates $x_k = (x_{k1}, \quad , x_{kp})$, and $1 \leqslant k \leqslant r$ Equation (1) specifies a proportional hazard regression model of considerable generality in which the underlying hazard and regression component is allowed to vary with k

In practice, however, some additional restrictions among the instantaneous incidence rates themselves, or, in the underlying state space, may be necessary for reliable parametric estimation One of many possible simplifying assumptions is that

$$\lambda_{ok}(t_k) = \lambda_o(t_k) \quad \exp(\beta_k) \tag{2}$$

from which it is clear that the intra-individual event incidence rates are not proportional in the usual sense

From (1) and the nature of the process, the contribution to the likelihood from a single individual on study for time T_r can be written as

$$\prod_k [\lambda_k(t_k, x_k)]^{\delta_k} \quad S_k(t_k, x_k) \tag{3}$$

where the indicators take the following values

$$\delta_k = \begin{cases} 1 & (k \neq r) \\ 0 & \text{otherwise} \end{cases} \qquad \delta_r = \begin{cases} 1 & \text{uncensored} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

whence exactly r events are observed in the uncensored case and $r-1$ otherwise

The conditional event specific survivor function is

$$S_k(t_k, x_k) = \exp[-\Lambda_k(t_k, x_k)] \tag{5}$$

and the conditional event specific cumulative hazard is defined as

$$\Lambda_k(t_k, x_k) = \int_{t_{k-1}}^{t_k} \lambda_k(t, x_k)\,dt \tag{6}$$

where for convenience $t_o = 0$

Kalbfliesch and Prentice (1980a) discuss the method of construction of likelihoods similar to (3) in the context of multivariate failure time data For a sample of n individuals the full likelihood is

$$\prod_i \prod_k [\lambda_k(t_{ki}, x_{ki})]^{\delta_{ki}} \; S_k(t_{ki}, x_{ki}) \tag{7}$$

where $1 \leqslant i \leqslant n$ and $1 \leqslant k \leqslant r_i$

Now suppose that $f_j (j = 0,1, \quad ,r^*)$ individuals experience exactly j events such that $n = \Sigma f_j$ where r^* is $\max_i [r_i]$ the maximum number of events experienced by individuals in the sample and let $R_k(t)$ denote the set of labels for individuals at risk of experiencing the transition from state a_{k-1} to a_k, then the numbers at risk are $n_1 = n$ and $n_k = f_{k-1}, 2 \leqslant k \leqslant r^*$ Furthermore, if $T^*_k(t)$ denotes those individuals who undertake this kth transition then $d[T^*_k(t)] = f_k$ for $1 \leqslant k \leqslant r^*$ and $R_k(t) = T^*_{k-1}(t)$ for $2 \leqslant k \leqslant r^*$ (The use of a subscripted t in parenthesis in the risk set notation will imply a time ordered risk set as in Cox (1972))

For the purposes of function and parameter estimation it follows that Equation (7) may be factorised into r^* separate components Thus (7) may be re-written as

$$\prod_k \prod_u [\lambda_k(t_{ku}, x_{ku})]^{\delta_{ku}} S_k(t_{ku}, x_{ku}) \qquad (8)$$

for $1 \leq k \leq r^*$ and $u \in R_k(t)$, a form which depends on r^* unknown functions $\lambda_{ok}(t_{ku})$, and r^* unknown parameter vectors β_k by virtue of (1) Although (8) permits the instantaneous event incidence function to depend on different covariate vectors throughout the time trajectory, such generality may not be required, especially in randomised controlled trials when the covariates may be the treatment group indicators on entry to the study

3 *ESTIMATION OF* β_k

Let $t_{k1} < t_{k2} < \quad < t_{tn_k}$ be the distinct ordered times associated with the n_k individuals at risk of making the transitions from states a_{k-1} to a_k Furthermore let $T^*(t)$ comprise $t^*_{k1} < t^*_{k2} < \quad < t^*_{kf_k}$, a subset of the times given above at which the f_k transitions actually occur The remaining $n_k - f_k$ times are censored Cox (1972) writes the conditional probability that the sth individual is incident at time t_{ks} as

$$\frac{\lambda_{ok}(t^*_{ks}) \quad \exp(x^*_{ks} \quad \beta_k)}{\sum_\ell \lambda_{ok}(t^*_{ks}) \quad \exp(x_{k\ell} \; \beta_k)} \qquad (9)$$

for $s = 1, \quad , f_k$ and $\ell \varepsilon R_k(t^*_{ks})$ The nuisance function $\lambda_{ok}(t^*_{ks})$ cancels throughout and the partial likelihood (Cox, 1975) is the product of f_k terms like (9)

$$L_k(\beta_k) = \prod_S [\exp(x^*_{ks} \quad \beta_k) / \sum_\ell \exp(x_{k\ell} \quad \beta_k)] \qquad (10)$$

The estimator $\hat{\beta}_k$ of β_k which maximises (10) is asymptotically $N(\beta_k, V(\beta_k))$ where $V(\beta_k)$ may be consistently estimated in finite samples by $I^{-1}(\hat{\beta}_k)$, the inverse matrix of second partial derivatives evaluated at $\hat{\beta}_k$ (Cox, 1975) Partial likelihood estimators may be obtained for each set of parameters $\beta_k, 1 \leq k \leq r^*$ in turn Equation (7) is different from Cox's (1972) model only in the construction of the risk set indicators

4 ESTIMATION OF $\lambda_{ok}(t_k)$

Function estimation for $\lambda_{ok}(t)$ is usually based on a two stage procedure, analogous to the method of relative maximum likelihood (Kalbfliesch and Sprott, 1970), in which $\lambda_{ok}(t_k)$ is estimated at the partial likelihood estimate $\hat{\beta}_k$ Some precision is inevitably lost at both stages first $\hat{\beta}_k$ is not in general fully efficient and, especially if of high dimension, may induce considerable uncertainty in the $\lambda_{OK}(t_K)$ estimates However despite these theoretical disadvantages the methods are known to work well in practice

Perhaps the most direct method in relation to fitted proportional hazard models is based on Breslow's (1974) modification to the cumulative hazard estimator given by Nelson (1969) In terms of the present ordered multi-state framework

$$\hat{\Lambda}_k(t_{ks}, x_{ks}) = \sum_u d_{ku} / \sum_\ell \exp(x_{k\ell} \hat{\beta}_k) \qquad (11)$$

for $s = 1, \quad , f_k$, $\ell \varepsilon R_k(t^*_{ks})$ The right-hand member of (11) takes non-zero values at the distinct incidence times $t^*_{K1} < \quad < t^*_{kf_K}$ in which cases $d_{ks} = 1$ When $k = 1$ and $\beta_1 = 0$ Equation (11) is a relative of the usual Kaplan-Meier estimator for the survivor function (5) Plots of (11) against t for $t = \min[t^*_s]$, where the t'_s are the entry times in $T'_{k-1}(t)$ associated with the f_k individuals in $T'_k(t)$ provide information on the incidence intensity over the appropriate portion of the time trajectory

An alternative method which is computationally more taxing is based on an arbitrary division of the time scale Given originally by Kalbfliesch and Prentice (1973) it has been adapted recently by Kay (1982) to the multi-state framework Let

$$\lambda_{ok}(t) = \lambda_{ok}^{(\ell)} \text{ for } t \ \varepsilon (b_{\ell-1}, b_\ell) = I_\ell \qquad (12)$$

for $1 \leqslant \ell \leqslant J$ and where $b_o = 0 < b_1 < \quad < b_{J-1} < b_J = \infty$ is some suitable sub-division of the real-time axis. Let $w_i^{(\ell)}$ be time spent in state a_{k-1} by the ith individual during I_ℓ and let $\delta_{i\ell}=1$ if this individual undertakes the transition to a_k in I_ℓ and $\delta_{i\ell} = 0$ otherwise for $i \in R_k(t)$. Then given the partial likelihood estimator $\hat{\beta}_k$ of β_k, the likelihood for the $\lambda_{0k}^{(\ell)}$ becomes

$$L(\lambda_{0k}) = \prod_i \prod_\ell [\lambda_{0k}^{(\ell)} \exp(x_{ki} \hat{\beta}_k)]^{\delta_{i\ell}} \exp[-\lambda_{0k}^{(\ell)} w_i^{(\ell)} \exp(x_{ki} \hat{\beta}_k)] \qquad (13)$$

for $1<i<n_k$ whence it is easy to show that

$$\hat{\lambda}_{0k}^{(\ell)} = \sum_i \delta_{i\ell} / \sum_i w^{(\ell)} \exp(x_{ki} \hat{\beta}_k) \qquad (14)$$

where $\sum_i \delta_{i\ell} = d_k^{(\ell)}$

Furthermore, $I(u,v)$ the uth, vth element of the information matrix with respect to $\lambda_{0k}^{(u)}$, $\lambda_{0k}^{(v)}$ is

$$I(u,v) = n_k^{(\ell)} \theta_k^{(\ell)} / [\lambda_{0k}^{(\ell)}]^2 \quad (u = v) \qquad (15)$$
$$= 0 \quad (u \neq v)$$

Where, of course $n_k^{(\ell)}$. $\theta_k^{(\ell)}$ is $E(\Sigma\, \delta_{i\ell})$ and $\theta_k^{(\ell)}$ is the conditional probability of incidence in I_ℓ. With these arrangements $\hat{\lambda}_{0k} = (\hat{\lambda}_{0k}^{(1)}, \ldots, \hat{\lambda}_{0k}^{(j)})$, is approximately asymptotically $N(\lambda_{0k}, V(\lambda_{0k}))$ where the dispersion matrix $V(\lambda_{0k})$ is a $j \times j$ diagonal with typical element reciprocal to (15). The likelihood (13) together with the results above provide a convenient test of the following hypothesis of constant underlying hazard

$$H_0 \quad \lambda_{0k} = \mu_{0k}$$
$$H_1 \quad \lambda_{0k} \neq \mu_{0k} \qquad (16)$$

where $\mu_{0k} = (\mu_{0k}^{(1)}, \ldots, \mu_{0k}^{(j)})$.
When H_0 is true the maximum likelihood estimator of μ_{0k} at $\beta_k = \hat{\beta}_k$ is, from (13)

$$\hat{\mu}_{0k} = \sum_i \sum_\ell \delta_{i\ell} \Big/ \sum_i t_{ki} \exp(x_{ki} \cdot \hat{\beta}_k) \qquad (17)$$

where $\sum_i \sum_\ell \delta_{i\ell} = f_k$

Accordingly, an approximate asymptotic test of H_0 may be based on

$$X_\nu^2 = (\hat{\lambda}_{0k} - \hat{\mu}_{0k}) \quad \hat{V}(\hat{\mu}_{0k})^{-1} \quad (\hat{\lambda}_{0k} - \hat{\mu}_{0k})$$

$$= \hat{\mu}_{0k}^{-2} \quad \sum_j \left[d_k^{(j)} \cdot (\hat{\lambda}_{0k}^{(j)} - \hat{\mu}_{0k}^{(j)})^2\right] \tag{18}$$

where $\nu = j-1$ and

$$\hat{V}(\hat{\mu}_{0k}) = \mathrm{diag}\left[\hat{\mu}_{0k}^2 \quad (1/d_k^{(1)}, \quad ,1/d_k^{(j)})\right] \tag{19}$$

Using either procedure outlined above the cumulative hazard plotting methods of Kay (1977) may be employed to further explore the structure of the underlying hazard

5 REDUCED MODELS

Since the dimension of the risk set is a decreasing function of k, efficiency considerations will usually dictate some reduction of the parameter space associated with the full likelihood (7), especially in studies of rare incidence in which case only the estimation of β_1 and β_2 may be feasible

However, reduced models are of interest for other reasons While, in general, the types of hypothesis to be tested will be governed by the objectives of the study and its particular design, for some accident studies the covariate vectors may be of the same type and dimension Such data structures arise naturally when the covariates are sensory or perception measurements made initially and after each subsequent incident Then, an obvious hypothesis of interest is

$$\text{(A)} \quad \begin{array}{lll} H_0 & \beta_k = \beta & \forall_k \\ H_1 & \beta_k \neq \beta & \forall_k \end{array} \tag{20}$$

which specifies arbitrary underlying hazards and a common regression relationship throughout the time trajectory

Alternatively, from the structure of (1), it may be reasonable to postulate

$$\text{(B)} \quad \begin{array}{lll} H_0 & \lambda_{0k}(t) = \lambda_0(t) & \forall_k \\ H_1 & \lambda_{0k}(t) \neq \lambda_0(t) & \forall_k \end{array} \tag{21}$$

so that the common underlying hazard $\lambda_o(t)$ is modulated by arbitrary regression components β_k

A more restrictive hypothesis than either considered above is of course

$$(C) \quad H_0 \quad \lambda_k(t,x_k) = \lambda(t,x_k) \quad \forall_k \qquad (22)$$
$$H_1 \quad \lambda_k(t,x_k) \neq \lambda(t,x_t) \quad \forall_k$$

When (22) holds the hazard depends on k through event times and the covariate structure the resulting model is closely related to the non-homogeneous Poisson process

The consequence of the hypothesis outlined above are developed in more detail below With regard to testing, an obvious general strategy is to compare particular reduced models with the full model in terms of maximum log likelihoods achieved

In order to derive suitable test procedures for hypothesis (A) above, the arguments leading to (10) may be exploited again The full likelihood (7) depends on r^* nuisance functions, $\lambda_{0_k}(t)$, when (A) holds Accordingly, a partial likelihood for the parameters $\beta_1, \ldots, \beta_{r^*}$ is given by

$$L(\beta_1, \quad, \beta_{r*}) = \prod_k L_k(\beta_k) \qquad (23)$$

in which the nuisance functions have been eliminated When H_0 is true (23) reduces simply to

$$L(\beta) = \prod_k L_k(\beta) \qquad (24)$$

If the vector of partial derivatives with respect to β_k is denoted by $U_k(\beta_k)$ and the matrix of second partial derivatives by $I_k(\beta_k)$, numerical estimation of the common value proceeds on an iterative basis using

$$U(\beta) = \sum_k U_k(\beta) \qquad (25)$$

and

$$I(\beta) = \sum_k I_k(\beta)$$

In addition Equations (25) provide a convenient framework for testing the hypothesis that $\beta = \beta_0$(an assigned value)

Following Cox (1972) the test may be based on the score statistic

$$\chi_{\nu}^{2} = U'(\beta_0) \quad I^{-1}(\beta_0) \quad U(\beta_0) \tag{26}$$

where the degrees of freedom are given by $\nu = \sum_{k} d\,[\beta_{k}]$

Equation (26) has wider applicability than the framework invoked in its derivation since in general there are no restrictions placed on the dimension of the parameter vectors β_k

In the case of hypothesis (B), specification of a common underlying $\lambda_0(t)$ leads to no important simplification of the likelihood (7) The arbitrary regression parameters β_k coupled with the non-homogeneity of the process together imply that partial likelihood arguments must be based on the same risk sets as before Thus, since only individuals who have experienced the k-1th event are at risk of experiencing the kth event, the partial likelihood given by (23) is appropriate irrespective of the condition imposed by hypothesis (B).

Although similar considerations lead to the same conclusion in relation to hypothesis (C) it is clear that tests of these latter hypotheses can be constructed when the underlying hazard function is suitably parameterised

6 *GENERALISATIONS*

The model formulated in Section 2 assumed that the accident incidence was of a single homogenous type However, competing types based on some suitable categorisation may be incorporated into the proportional hazards model (1) This approach may be warranted when some of the events are of a qualitatively different nature - for example, with respect to severity, or when the events occur in materially different environmental conditions, as with the operation of a variety of equipment types Then the instantaneous incidence rates may be written as

$$\lambda_{km}(t_{km}, x_{km}) = \lambda_{okm}(t_{km}) \quad \exp(x_{km}\beta_{km}) \tag{27}$$

for $1 \leq m \leq c$ competing accident types

The resulting likelihoods arise naturally in the theory of competing risks associated with multi-variate failure time data and have been considered briefly by Kalbfleisch and Prentice (1980a) and Kay (1983) Partial likelihood estimation of the parameters β_{km} proceeds along lines similar to these described above An important feature of these models is that the estimates $\hat{\beta}_{km}$ are obtained in the absence of any independence assumptions about the underlying mechanism generating the different event types (Kalbfleisch and Prentice, 1980b) Dependencies between transition types however may be explored using the risk indicator approach suggested by Prentice, *et al* , (1978)

A further generalisation of the proportional hazards model (1) involves the incorporation of internal covariates in the overall covariate structure Neglecting, momentarily, simultaneous incidences, it has been assumed in (1) that the covariate vectors x_k are measured either at the start of the study $(k=1)$ or immediately following the $k-1$th event $(k>1)$ Thus while the covariate vectors x_k are clearly the output of the stochastic process generated by a single individual they are conditional on the number of previous incidences experienced However, this structure permits the inclusion of, more usual, internal covariates which are observed on individuals in $R_k(t)$ only so long as the individuals are not in $T_k'(t)$ Since this type of internal covariate, $x_k(t)$, carries information on the incidence time of the t_k event it is clear that the interpretation of the effects of such time dependent covariates requires additional care (Kay, 1982)

Nevertheless, the availability of these twin devices within the partially parametric proportional hazards framework offers the statistician considerable flexibility in the choice of model

7 DISCUSSION

By focusing principally on the more general concept of accident-incidence, *per se*, this paper attempts to avoid much of the extensive terminology surrounding the analysis of accident data which exists in the literature (Froggatt, 1968b, Kemp, 1968)

A major advantage of the methods discussed lies in their ability to model covariate information collected throughout an individual's period of observation in a consistent and very much general manner This has been achieved using a log-linear parameterisation of the individual time-dependent transition probabilities Estimation of, and test of, hypothesis concerning the effects of factors measured during the study are thus facilitated without the need for stratification of the data by observed covariate values In addition, the components of the covariate vector may be freely chosen to include factors specific to the individuals under study ("proneness") or factors relating to their environment (external) or both ("liability") The methodological advantages of this approach which cases the analysis of accident data in a standard statistical mould would have been appreciated equally by Newbold (1927) and Froggatt (1968b)

Alternatively, other aspects of the stochastic process, for example, $E[N(t)]$ might have been modelled Such methods are, however, less fundamental as they require, *a priori*, the choice of an appropriate statistical distribution for the number of events in $(o,t]$ Furthermore, different distributions, such as the time-dependent Poisson or the time-dependent negative binomial in which identical assumptions were made about the form of $E[N(t)]$ would lead in general to transition probabilities with different parametric structures

Equation (1) and its generalisation (27) define a family of conditional transitional probabilities of great generality A transition from state a_{k-1} to a_k at time t_k^* is made conditional on the whole history of the process prior to t_k^*, say $H(t_k^*)$, which includes information on all previous incidence times and types together with information on censoring (Kalbfleisch and Prentice, 1980c) Likelihood construction is then so straightforward that the method would have been considered an advance even in the fully parametric case ($\lambda_o(t)$ specified) However, the availability of partial likelihood methods (Cox, 1975) permitting the estimation of β irrespective of the functional form of $\lambda_o(t)$ completely relieves the investigator of this awkward problem of specification

What then are the implications for data analysis? First, longitudinal study designs will best facilitate the use of the model, both in relation to (a) the accurate recording of events and their incidence times, and (b) the testing of well-formulated hypotheses within the standard framework of randomised controlled trials [see Kay (1983) for a related paper] It would appear unlikely that the model could be fully utilised in a purely observational study since the choice of covariates following incidence may not be clear cut, although a simple expedient here may be to repeat some of the original covariates and condition on the absolute or relative change as appropriate

Second, since the risk set $(R_k(t)$ is a decreasing function of k, inevitably the estimation of β_k will become infeasible In order therefore to maximise the numbers at risk, studies could be organised on a multi-centre basis The efficiency of β_k for the specialised single failure situation $(k \equiv 1)$ has been investigated by Efron (1977) and Oakes (1977) However, further work comparing the information in (8) with that available from (23) is required

Nevertheless, the proportional hazards regression model should prove more informative than existing procedures and its application to the analysis of accident data will inevitably lead to a reduction in Greenwood and Woods' (1919) "motley host'

REFERENCES

ADELSTEIN, A M , 1952 "Accident proneness a criticism of the concept based upon an analysis of shunters' accidents," *Journal of the Royal Statistical Society*, A, 115 354-410

ANSCOMBE, F J , 1950 "Sampling theory of the negative binomial and logarithmic series distributions," *Biometrika*, 37 358-382

BATES, G E and J NEYMAN, 1952 "Contributions to the theory of accident proneness I An optimistic model of the correlation between light and severe accidents," *University of California Publications in Statistics*, 1 215-254

BLISS, C I and R A FISHER, 1953 "Fitting the negative binomial distribution of biological data," *Biometrics*, 9 176-200

BRESLOW, N E , 1974 "Covariance analysis of censored survival data," *Biometrics* 30 89-99

COX, D R , 1972 "Regression models and life tables" (with discussion), *Journal of the Royal Statistical Society*, B 34 187-220

COX, D R , 1975 "Partial Likelihood," *Biometrika*, 62 269-276

COX, D R and V ISHAM, 1980 *Point Processes*, London Chapman and Hall

CRESSWELL, W L and P FROGGATT, 1963 *The Causation of Bus Driver Accidents An Epidemiology Study*, London OUP, 46-60

EFRON, B , 1977 "Efficiency of Cox's likelihood function for censored data," *Journal of the American Statistical Association*, 72 557-565

FROGGATT, P , 1968a/b "Application of discrete distribution theory to the study of non-communicable events in medical epidemiology " In Patil, E P (ed) *Random Counts in the Biomedical and Social Sciences*, London Pennsylvania State University Press, 2 28-29, *loc cit* , 2 22-27

GREENWOOD, M and G U YULE, 1920 "An enquiry into the nature of frequency distributions representative of multiple happenings, with particular reference to the occurrence of multiple attacks of disease of repeated accidents," *Journal of the Royal Statistical Society*, 83 255-279

GREENWOOD, M , and H M WOODS, 1919 "A report on the incidence of industrial accidents upon individuals with special reference to multiple accidents," Med R Council Industrial Fatigue Research Board, *Report* No 4, London HMSO

KALBFLEISCH, J D , and R L PRENTICE, 1980a/b/c *The Statistical Analysis of Failure Time Data*, New York Wiley, 1980a 181, *loc cit* ,1980b 172, *loc cit* , 1980c 182

KALBFLEISCH, J D , and R L PRENTICE, 19973 "Marginal likelihoods based on Cox's regression and life model," *Biometrics,* 60 267-278

KALBFLEISCH, J D , and D A SPROTT, 1970 "Application of Likelihoods methods involving a large number of parametrics," (with discussion), *Journal of the Royal Statistical Society,* B, 32 175-208

KAY, R , 1977 "Proportional hazard regression models and the analysis of censored survival data," *Journal of the Royal Statistical Society,* C, 26, 3 227-237

KAY, R , 1982 "The analysis of transition time in multistate stochastic processes using proportional hazard regression models," *Communications in Statistical Theory and Methods,* 11(15) 1743-1756

KAY, R , 1983 "Unbiased assessment and treatment effects on survival and disease recurrence in clinical trials," *Statistics in Medicine,* 190, 2 41-48

KEMP, C D , 1967 "On a contagious distribution suggested for accident data," *Biometrics,* 23 241-255

KEMP, C D , 1968 "Accident proneness and discrete distribution theory," in Patil, E P (ed) *Random counts in the biomedical and social sciences,* London Pennsylvania University Press, 2 41-65

KEMP, C D , and A W KEMP, 1965 "Some properties of the 'Hermite' distribution," *Biometrika,* 52 381-394

NELSON, W , 1969, "Hazard plotting for incomplete failure data," *Journal of Quality Technology,* 1 27-52

NEWBOLD, E , 1927 "Practical applications of the statistics of repeated events, particularly to industrial accidents," *Journal of the Royal Statistical Society,* 90 487-535

NEYMAN J , 1939 "On a new class of 'contagious' distribution applicable to entomology and bacteriology", *Annals of Mathematical Statistics,* 10 35-57

OAKES, D , 1977 "The asymptotic information in censored survival data", *Biometrika,* 64 441-448

POISSON, S D , 1837 *Recherches sur la probabilite des judgements* Paris Bachelier

PRENTICE, R L , I D KALBFLEISCH, A V PETERSON, N FLOURNOY, V T FAREWELL and N E BRESLOW, 1978 "The analysis of failure time data in the presence of competing risks," *Biometrics,* 34 541-554

Von BORTKIEWICZ, L 1898 *Das Gesetz der Kleinen Zahlen,* Leipsig Teubner

DISCUSSION

M A Moran I am delighted to be associated with the welcome to Dr MacKenzie here tonight and to have been invited to propose the vote of thanks on behalf of the Society

In his introduction, Dr MacKenzie gave a brief review of the classical treatment of accident data which primarily involved the fitting of the Negative Binomial or more complicated discrete distribution He then proceeded directly to an alternative model formulation based on the proportional hazards model The former makes no provision for covariates, the latter does However, there are intermediate models which do allow for covariates and it might be useful to put the two discussed in tonight's paper in some overall context

The original Greenwood and Yule model envisaged that the number of accidents Y in a period t for a given individual had a Poisson distribution with mean and variance

$$E(Y) = \lambda t$$
$$V(Y) = \lambda t$$

This model will not fit a group of individuals unless they all have the same accident rate λ Individual variation in λ, whether from sources internal or external to individual causes an increase in variance over that indicated by a Poisson model If this variation in λ is modelled by a Gamma, a Negative Binomial distribution for Y results with

$$E(Y) = \lambda t$$
$$V(Y) = \lambda t + \frac{(\lambda t)^2}{k}$$

but a similar inflation in variance can result from alternative models

Disaggregation of groups into homogenous groups on the basis of factors logically associated with variation in the accident rate would be expected to produce a within group variation more in conformance with the Poisson model Thus Weber (1971) described data from a California Driver Study on the accident record of 148,000 motorists over the period 1961-63 in which the overall accident distribution was fitted extremely well by a Negative Binomial When the individuals were partitioned, however, into 2,880 groups according to sex, age, marital status, area of residence, conviction history and accident history, 87 per cent of cells containing 100 observations or more were found to be fitted very well by Poisson distributions Weber went on to express the accident rate λ as a linear function

$$\lambda = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \quad \beta_p x_p$$

of the Covariates which he estimated by iterative weighted least squares Such a model would not be considered as a particular case of the class of generalised linear models introduced by Nelder and Wedderburn (1972) It and plausible alternatives such as

$$\log \lambda = \beta_o + \beta_1 x_1 + \quad + \beta_p x_p$$

are easily fitted using the GLIM program

It is possible, of course, that even with disaggregation and the introduction of covariates that variation in Y in excess of that indicated by a Poisson would still obtain A case in point is the analysis of mortality rates from lung cancer described by Manton, Woodbury and Stallard (1981) They assumed that, after disaggregation, within cell variation in Y, the number of deaths, has a Negative Binomial in which both λ, the death rate and k the measure of excess variation within a cell, are functions of the covariates Thus there are many possible covariate models for accident data apart from the proportional hazards model The latter incidentally assume that all variation between individuals can be attributed to the covariates

Now we come to the substance of tonight's paper which is the modelling of the pattern of accidents over time to allow for repeated incidents At first sight the basic formulation resembles that of Prentice, Williams and

Peterson (1981) who consider two forms for the hazard function

$$\lambda[t \mid N(t), X(t)] = \begin{cases} \lambda_{ok}(t) \exp(x_k, \beta_k) \\ \text{or} \\ \lambda_{ok}(t - t_{(n)} \exp(x_k, \beta_k) \end{cases}$$

The first of these, which is very similar to Equation (1) of Dr MacKenzie's paper, implies that the chance of having an accident at time t, given N(t) = n(t) accidents to date and covariate values X(t) = x(t) is dependent on the total time that the individual has been under observation The second form implies a dependence on time elapsed since the last accident It seems plausible that an accident might introduce a change in the baseline hazard and the dependence on the covariates, but that otherwise one starts from scratch Dr MacKenzie might like to comment on how his formulation related to that of Prentice, *et al*

As a statistician with an interest in applications, I wondered how one would in practice fit the models described tonight It seems to me looking at Equation (7) that the ordered transient states defined in the paper may be treated as distinct strata, note being taken of the conditional nature of the survivorship functions for the second and subsequent strata when it comes to interpretation Thus the usual log minus log plots of the estimated survivor functions in (5) would establish whether the functions for different accident states were proportional, in which case differences between states could be accommodated as a covariate on the lines of Equation (2) A test of equality of such parameters would correspond to test B of Equation (21) For accidents of a single type the fitting and comparison of separate strata can be done using a program such as BMDP2L Alternatively, the approach of Whitehead (1980) could be used in which a set of independent Poisson variables is defined for each risk set with expectations dependent on the covariates in such a way as to produce estimates $\hat{\beta}_K$ and likelihood ratios identical to the proportional hazards model This auxiliary Poisson model can be fitted using GLIM If strata or accident states are incorporated as a factor S, then the main hypotheses of interest can be tested using the sequence of models

$$R$$
$$R + S$$
$$R + S + X$$
$$R + S + X + S\,X$$

where the levels of R correspond to the distinct risk sets and effectively fit the baseline hazard Thus, for example, hypothesis A in (20) is tested by comparing the fit of $R + S + X$ with $R + S + X + S\,X$ The underlying hazard functions can be recovered from the parameter estimates for R and S

I think it a pity that Dr Mackenzie chose not to include any applications or to discuss to what extent accident data in practice are detailed enough to support estimation of the models he has described tonight It is an important topic and there is obviously much to be done I look forward to further contributions from Dr Mackenzie and congratulate him on his initiative in submitting this paper and on his presentation tonight I formally propose tne vote of thanks

References

MANTON, K G , M A WOODBURY, and E STALLARD, 1981 "A variance components approach to categorical data models with heterogeneous cell populations analysis of spatial gradients in lung cancer mortality rates in North Carolina counties,' *Biometrics* 37, 259-69

NELDER, J A and R W M WEDDERBURN, 1972 "Generalized linear models,' *Journal of the Royal Statistical Society,* A, 135, 370-84

PRENTICE, R L , B J WILLIAMS and A V PETERSON, 1981 "On the regression analysis of multivariate failure time data," *Biometrika* 68,2

WEBER D C , 1971 Accident Rate Potential Application of Multiple Regression Analysis of a Poisson Process,' *Journal of the American Statistical Association,* 66, 334, 285-288

WHITEHEAD, J , 1980 "Fitting Cox's regression model to the survival data using GLIM," *Applied Statistics,* 29, 268-275

P J Boland I have found over the past few years that one aspect of the Statistical and Social Inquiry Society of Ireland which I particularly enjoy is the broad diversity of papers that are presented at its meetings Recently I have enjoyed presentations on new plans for the CSO, statistics in the secondary schools, agricultural statistics and tonight a paper on proportional hazard models for accident data I am also glad to see, as is evidenced by tonight's presentation, that papers of a theoretical nature are welcome from time to time at meetings of the Society

Dr MacKenzie has presented a model for analysis of accident data which generalises the log linear proportional hazard model introduced by Cox in 1972 Use of proportional hazards as a basis for modelling has become very popular ever since I note that according to the 1984 Current Index to Statistics (which lists most papers in statistics published in 1984) there were at least 20 papers dealing principally with proportional hazard models

Although I feel strongly that theoretical papers should have a role in the society, I was hoping that Dr MacKenzie would have presented a case study or data analysis illustrating his theoretical considerations It would be interesting to know if say "traffic accidents' or "health-related accidents (such as heart attacks) can be well modelled using the proposed models Some discussion of the specific types of covariates that would be used in accident modelling could also prove very interesting I suspect that time dependent covariates such as the "accidents incurred/ time elapsed at the last previous accident' are of prime importance

I have a few comments related to some technical aspects

(a) I was wondering if the incidence rate functional form defined in Equation (1) precludes the possibility of a kth accident being fatal for some but not all individuals Also it seems that r^* , the maximum accidents experienced, is a random variable This seems to imply that theoretically at least there are an infinite number of initially unknown parameters

(b) I had some difficulty with the terminology for time ordered Risk sets $R_k(t)$ is defined as the set of labels for individuals at risk of experiencing the transition a_{k-1} to a_k But does this refer to those at risk of making this transition *at time* t, or at some time during the interval (0,t)? If it is the former, then the size of the risk set $R_k(t)$ need not be decreasing (as claimed) If it is the latter, then the notation concerning n_k and f_{k-1} should indicate that they are functions of time as well The notation of Cox and Oakes (Analysis of Survival Data) seems to suggest that $R_k(t)$ = those individuals who up until time t have not experienced a k^{th} accident but have experienced $k-1$ accidents

(c) In the estimation of the baseline hazard functions $\lambda_{ok}(t)$ in Section (4) it is stated that a two stage procedure can be used in which $\lambda_{ok}(t)$ are estimated at the partial likelihood estimates $\hat{\beta}_k$ This seems to be intriguing since the defining equations for the baseline hazards seem to indicate that the $\lambda_{ok}(t)$ are independent of the β_k

(d) Some practical examples where the test of hypothesis given in (16) (where the null hypothesis is that the k^{th} hazard function is a step function) might be useful I suspect that in many cases $\mu_{ok}^{(i)}$ would be increasing in i

I have mentioned some technical criticisms I have of the paper However I have enjoyed the paper and would like to firmly second the vote of thanks proposed by Professor Moran Dr MacKenzie has introduced a new class of models which will hopefully prove useful and I wish him luck in promoting their use in practical situations

Reply by Dr MacKenzie I am very grateful to both contributors for their constructive and helpful comments on my paper

The development of models for the analysis of accident data shares a common history with developments in classical discrete distribution theory The close synergism which has existed between these two areas is well illustrated in recent work by Xekalaki (1984) However, it is now clear that a number of useful innovations have been overlooked by practitioners and theorists alike In particular, a rather obvious deficiency has been the absence of general models which account for the effect of explanatory variables Accordingly, the main thrust of tonight's paper has been to draw attention to this problem and offer one possible solution in the form of a semi-Markov proportional hazards model

Professor Moran and Dr Boland raise a number of cogent points The proposed model is only one of a large class of possible alternative formulations Nevertheless there are some compelling reasons for its consideration

First, the model,

$$\lambda_k(t_k,x_k) = \lambda_{0k}(t_k) \cdot \exp(x_k \beta_k) \tag{1}$$

for $k=1$ provides a direct test of the hypothesis of no *ab initio* unequal "liability" (and hence "proneness") allowing for different exposure times among individuals at risk This hypothesis is rejected when $\beta_1 \neq 0$ and then the non-zero components can be identified readily

Second, one can model, conditionally, the incidence of subsequent events $(k>1)$ intra-individual in a similar fashion In this way the influence of particular covariates on each event can be studied and hypothesis A (Equation 20) tested

This approach seems more fundamental than that adopted by Weber (1971) and Manton, Woodbury and Stallard (1981) since the kth step transition probability [Equation (1)] associated with the process is parameterised directly in a systematic and very general manner By contrast, the *post-hoc* parameterisation of function of the first two moments by different linear combinations of the covariates appears, at least, rather inelegant One disadvantage of the semi-Markov model, noted by Dr Boland, is that r' is a random variable so that estimation and inference must be made conditionally on the value observed In practice, however, the number of parameters need not be large as it is often sensible to group-up the tail of the accident distribution Thus k may run from 1 up to m where $m \gg r^*$

Dr Boland comments on the respective roles of $\lambda_{0k}(t_k)$, the underlying hazard, and $\exp(x_k \beta_k)$, the regression component, in Equation (91) Usually the principal problem is the estimation of $\beta_k (k=1, \ldots, m)$ and this is elegantly solved by the method of partial likelihood irrespective of the functional form of $\lambda_{0k}(t_k)$ Estimators for this latter function (Section 4) utilises the partial likelihood estimates, $\hat{\beta}_k$ Such two-stage procedures are not particularly unusual

The hypothesis of "contagion" may be appropriate when $\beta_k = 0$ (all k) and $\lambda_{0k}(t_k)$ is an increasing (step) function in k. It is important to note that, contrary to Professor Moran's conjecture, the exposure times also contribute to the inter-individual variation This fact may be exploited to construct a test for "local" exponentialty of the underlying hazard [Equation (16)]

The risk sets are constructed in the usual way (Cox, 1972) as the table below shows However, it is notationally more cumbersome to cope with the index k

Transition type	k	No at risk	No undertaking transition	No of censorings
a_o, a_1	1	n_1	f_1	$n_1 - f_1$
a_1, a_2	2	n_2	f_2	$n_2 - f_2$
.	.			
	.	.	.	
	.	.	.	.
a_{r^*-1}, a_{r^*}	r^*	n_{r^*}	f_{r^*}	$n_{r^*} - f_{r^*}$

NB . $n_1 = n$ and $n_k = f_{k-1}$ $(2 \leq k \leq r^*)$

Thus Dr Boland's remark on the dimension of the risk set is correct when all the individuals at risk of the kth event actually experience it Perusal of the literature on accident data indicates, however, that this phenomenon is exceptionally rare

Both discussants have commented on the absence of example data and I am bound to agree with their comments Part of the problem relates to the non-availability of accident data collected prospectively in the format required by the model It is important to note that the *a priori* adoption of the proportional hazards model will lead to major improvements in the design of future studies in this field A variant of the model has been used to analyse survival data (Kay, 1983), and the full model is currently in use in Belfast in a study of the incidence of subsequent mitral valvotomies

In these studies specially adapted FORTRAN programs have been employed to estimate the unknown parameters Professor Moran's suggestion that the model might be fitted using GLIM is a useful one and requires further investigation In this connection it may be observed that if N(t) is the random variable denoting the number of events in (0,t) then N(t) is Poisson with parameter

$$\Lambda(t,x) = \int_0^t \lambda(u,x)du$$

where $\lambda(t,x) = \lambda_0(t) \exp(x\ \beta)$ from hypothesis (C)

This formulation by Prentice, *et al*, (1981) is identical to that proposed tonight (my reference Kalbfleisch and Prentice 1980a) and I am grateful to Professor Moran for pointing this out Prentice's gap (time from last accident) model may indeed be appropriate and this hypothesis may be the inclusion of the appropriate time-dependent covariate in Equation (1) The use of total time on study merely provides a general ordering relation which establishes a framework within which many interesting hypotheses may be tested Hypotheses (A), (B) and (C) form one natural group and to my mind the reduction to gap models appears to form another It must be emphasised, however, that the choice of time origin clearly affects the estimation process because of the method of constructing the risk sets and re-estimation of the parameters may be required

In conclusion, I would like to thank the discussants again for their kind contributions

References

MANTON, K G , M M WOODBURY, and E STALLARD, 1981 "A variance components approach to categorical data models with heterogeneous cell populations analysis of spatial gradients in lung cancer mortality rates in North Carolina counties," *Biometrics* 37, 259-69

NELDER, J A and R W M WEDDERBURN, 1972 "Generalized linear models," *Journal of the Royal Statistical Society*, A, 135, 370-84

PRENTICE, R L , B J WILLIAMS and A V PETERSON, 1981 "On the regression analysis of multivariate failure time data," *Biometrika*, 68, 2, 373-79

WEBER, D C , 1971 "Accident Rate Potential Application of Multiple Regression Analysis of a Poisson Process,' *Journal of the American Statistical Association*, 66, 334, 285-288

WHITEHEAD, J , 1980 "Fitting Cox's regression model to survival data using GLIM," *Applied Statistics*, 29, 268-275

XEKALAKI, E , 1984 "The bivariate generalised Waring distribution and its application to accident theory," *Journal of the Royal Statistical Society*, 147, 3, 488-498