JIA 115 (1988) 299-315

MODELLING EXCESS MORTALITY USING GLIM

BY ARTHUR E. RENSHAW, B.SC., PH.D.

(*of the City University, London*)

1. INTRODUCTION

IT is now some sixteen years since Sir David Cox (1972) published his epoch making paper in which he incorporated regression type arguments into life-table analysis. Central to the method was the introduction of the multiplicative hazard

$$\lambda(t, z) = \lambda^*(t) \exp(\beta' \cdot z)$$

with vector of covariates z, unknown regression parameters β, and so-called base-line hazard $\lambda^*(t) = \lambda(t, 0)$. Applications of the method, based on the conditional likelihood argument expounded by Cox in which the base-line hazard λ^* is unknown, have proliferated in the intervening years, largely in the field of medical statistics. There have been relatively few applications in which the base-line hazard is assumed known at the outset. Specific cases include Breslow *et al.* (1983), Berry (1983) and Hill *et al.* (1985).

Attempts to incorporate regression-like models into life-table analysis would appear to have gone largely untried by the British actuarial profession. Essentially this is because in life insurance, data bases are large and the establishment view is that such models are inappropriate when sampling variation is small. Also mortality is of less significance as a factor than economic variables like inflation and investment earnings. All the actuary needs to do is not understate the level of mortality rather than get it exactly right.

The purpose of this paper is two-fold. Firstly we seek to draw attention to the considerable potential of these methods for actuaries when the data bases are relatively small and it is important to get the mortality factor right. Secondly we seek to illustrate this potential by reanalysing part of the Prudential impaired lives data previously reported and discussed at length in the pages of this Journal (Preston & Clarke (1965), Clarke (1978)). Somewhat prophetically, Professor Bernard Benjamin predicted the potential of these methods for investigating the mortality of special groups, such as those with impairments, in the discussion on Cox's 1972 paper.

By way of motivation § 2 contains a brief description of relevant aspects of the impaired-lives data set. The underlying methodology described in § 3 is supported by a Technical Appendix. Sections 4 and 5 deal with peripheral necessities to set the analysis in motion. The results obtained from a reanalysis of aspects of the impaired-lives data set are presented in § 6.

2. THE IMPAIRED LIVES DATA SET

The data are extensive, comprising information derived from well in excess of half a million life insurance policies effected on impaired lives during the period January 1947 to December 1981. In fact the study is on-going with data extending beyond 1981 to the present day. The information on each impaired life includes details of medical status at entry, age next birthday at entry and date at entry, date and mode of exit. Classification by medical status involves nine broad categories, each of which is further subdivided. Full documentation is given in Preston & Clarke (1965) and need not be reproduced here. Entry and exit dates are known to the nearest month.

Typically Figure 2.1 displays the information available for each medical status (here—hypertensive, overweight, specified blood pressure category). Calendar year (January 1947 to December 1981) is represented on the x-axis and age (upwards of 15 years) on the y-axis; with each oblique line or stroke within these bounds representing a single policy experience. The starting coordinates for each stroke, depicted by + are date of entry and age next birthday at entry (minus six months); while the length of each stroke is determined by the duration over which the policy is seen to be operative. The mode of exit may be due to death, depicted by the black spot •, or through censorship for whatever reason, depicted by ○. Two distinctive features present are the anticipated heavy censoring unavoidably induced at the study boundary (December 1981) and the lighter censoring at age 65 years (calendar period 1967 onwards) which presumably is induced by occupational retirement. Indeed many of the policies are endowment assurances maturing at age 65 for this reason.

Insight into the strength of mortality present is provided, in part, by the entry and exit times to and from study of a batch of individuals. Entry times, τ, are defined when policies are issued on individual lives, while exit times, t, are brought about through either death ($\delta = 1$) or censorship ($\delta = 0$). Obviously $\tau < t$ while δ is a zero-one indicator variable. Censorship may be due either to a policy surrender or maturity, or may be induced by the outer temporal limits (on age and calendar year) of the study.

The heterogeneous nature of the collective data set, due to the different (combined) levels of covariates such as medical status, age at entry, policy duration, calendar year of entry is accommodated by partitioning the study data into relatively homogeneous batches or cohorts indexed by $j \varepsilon \mathcal{J}$. Often j will take the form of a vector suffix with one component for each covariate. Collectively such suffices form a so-called product set. Obviously, it is necessary to ensure that the amount of experience for each such cohort is sufficient to make the construction of the traditional actuarial mortality ratios meaningful.

Denoting the age next birthday at entry by the letter a, the information offered by each policy is

$$(\tau_{jk}, t_{jk}, a_{jk}, \delta_{jk})$$

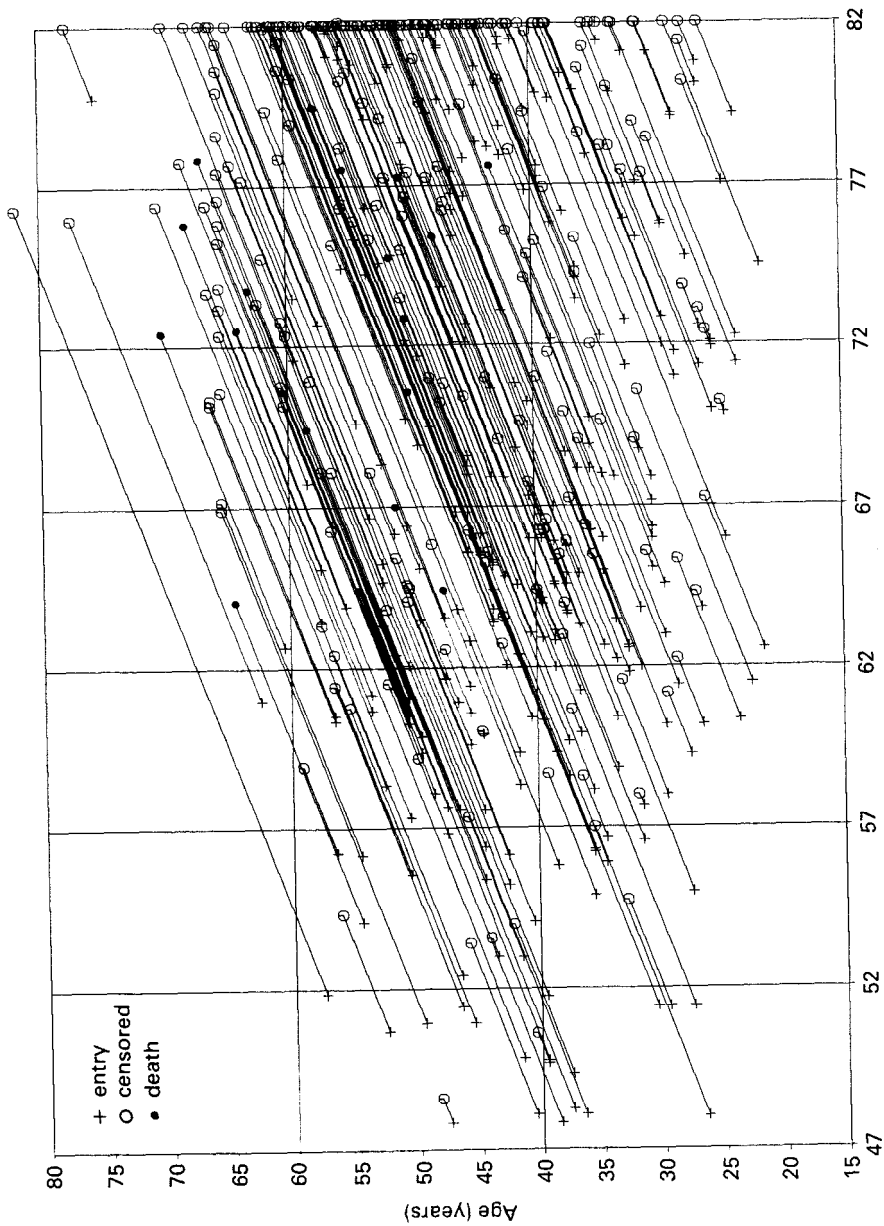

Figure 2.1. Typical set of policy experiences

in which k ranges over the members within a particular cohort j. Thus, we have available

τ_{jk}—entry time
t_{jk}—exit time
a_{jk}—age next birthday at entry
δ_{jk}—mode of exit;

together with information on medical status.

3. METHOD OF ANALYSIS

We focus attention on mortality through the age-specific force of mortality (μ_x in actuarial literature) or hazard function $\lambda(t, z)$ where t denotes time and z the appropriate vector of covariates. A standard or base-line hazard $\lambda^*(t) = \lambda(t, \mathbf{0})$ is constructed for normal life insurance business ($z = \mathbf{0}$), by suitably transforming the representative A1967–70 life-table, adjusted to allow for secular trend in mortality (see § 4). This is used as a yard-stick against which the excess mortality of impaired groups (with covariates z_j) is assessed. It is achieved through Cox's identity

$$\lambda(t, z_j) = \lambda^*(t) \exp(\boldsymbol{\beta}' z_j) \tag{3.1}$$

in which the multiplicative factor, $\exp(\boldsymbol{\beta}' z_j)$, with unknown regression parameters $\boldsymbol{\beta}'$, may be perceived as adjusting the standard hazard by the amount of excess mortality attributable to the specific impairments present in the study groups or cohorts $j \in \mathcal{J}$. It is proposed therefore to investigate the potential for using the $\exp(\boldsymbol{\beta}' z_j)$'s as measures of excess mortality.

Three major questions arise: (1) How are the mortality factors to be computed? (2) What relationship, if any, do they bear to the traditional actuarial mortality ratios computed by Clarke? (3) What advantages, if any, accrue from this approach? We address each of these questions in turn.

3.1 To compute the $\exp(\boldsymbol{\beta}' z_j)$'s we require estimators for the unknown regression parameters $\boldsymbol{\beta}'$. These are obtained by maximum likelihood methods, the theory for which is developed in the Technical Appendix.

Recall the classic linear model (LM) in which the response variables Y_j are assumed to be independent normal variables $y_j \sim \text{IN}(\mu_j, \sigma^2)$ with means $\mu_j = \boldsymbol{\beta}' z_j$ and constant variance σ^2. Specifically if $\boldsymbol{\beta}' = (\alpha, \beta)$ and $z'_j = (1, x_j)$ so that $\mu_j = \alpha + \beta x_j$ we get straight line regression and so on. Such models are generalized by allowing other distributions, typically the Poisson distribution, so that $Y_j \sim \text{IPoi}(\mu_j)$; and also by linking the means μ_j to the linear predictor $\boldsymbol{\beta}' z_j$ through the introduction of a monotonic link-function, g, where

$$g(\mu_j) = \boldsymbol{\beta}' z_j \tag{3.2}$$

We establish in the Technical Appendix that the likelihood function, based solely on the Cox identity (3.1) and data of the type described in § 2, is identical to that for the generalized linear model (GLM) in which the numbers of deaths per cohort, d_j, are independent Poisson response variables $d_j \sim \text{IPoi}(\mu_j)$ with means

$$\mu_j = m_j \exp(\boldsymbol{\beta}' \, z_j),$$

log-link function

$$\log \mu_j = \log m_j + \boldsymbol{\beta}' \, z_j \tag{3.3}$$

and where

$$m_j = \sum_k \int_{\tau_{jk}}^{t_{jk}} \lambda^*(u) \, du. \tag{3.4}$$

m_j is the 'accumulated integrated base-line hazard' and we discuss how to compute the m_j terms in § 5. Notice that they are closely related to the expected number of deaths used to form the denominator in the construction of traditional actuarial mortality ratios as discussed by Haberman (1988). Indeed, the two quantities become identical when we integrate over the standard life-table rather than the standard hazard λ^*.

The practical aspect of generalized linear interactive modelling (GLIM) is conducted using the specially designed computer package bearing this name. It offers the user a wide choice of modelling distributions, link functions and covariate model structures. See Haberman & Renshaw (1988). In particular, model fitting is by maximum likelihood providing the necessary parameter estimates $\hat{\boldsymbol{\beta}}$ with which to compute the exp $(\hat{\boldsymbol{\beta}}' \, z)$'s.

Comparison of expressions (3.2) and (3.3) reveals the extra log m_j terms which have to be subtracted from the log μ_j terms when fitting the model. Such terms are called offsets. They may be perceived as part of the linear predictor $\boldsymbol{\beta}' \, z$, being an extra term with known regression coefficient, having value equal to unity. The GLIM computer package offers this facility.

3.2 The relationship between mortality factors and the traditional mortality ratios, as well as the method of computation, is illustrated by results taken from Table 1 on page 33 of Preston and Clarke (1965). These refer to hypertensives with two covariates, A, denoting weight status with two levels (1—standard + 19%, 2—standard + 20% or over) and, B, denoting age at entry with three levels (1—16 to 39 years, 2—40 to 59 years, 3—60 or more years). The results are reproduced in the first five columns of Table 3.1 in which two suffices (i, j) are required to accommodate the combined crossed levels of the two covariates A and B.

As already indicated, the m_{ij}'s here are based on a standard life-table rather than on the corresponding standard hazard, a minimal change which, for all practical purposes, can be ignored. Fitting the GLM with independent Poisson

Table 3.1. *Mortality data for hypertensives, 1947–63*

A(i)	$B(j)$	d_{ij}	m_{ij}	d_{ij}/m_{ij}	$\boldsymbol{\beta}' z_{ij}$
1	1	64	25·13	2·55	·9348
1	2	480	251·11	1·91	·6473
1	3	264	217·22	1·22	·1950
2	1	21	7·66	2·74	1·0085
2	2	86	43·63	1·97	·6786
2	3	29	23·46	1·24	·2120

responses, log-link, offsets $\log m_{ij}$ and fully interactive structure, denoted by $A*B$, and with parametric representation

$$\boldsymbol{\beta}' z_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \tag{3.5}$$

yields the following parametric estimates

	$j=1$	$j=2$	$j=3$	$\hat{\alpha}_i$
$i=1$	0	0	0	0
$i=2$	0	−·0430	−·0567	·0737
$\hat{\beta}_j$	0	−·2869	−·7398	·9348 $=\hat{\mu}$

The $(\widehat{\alpha\beta})_{ij}$ terms are presented in the body of the table. The linear predictors $\boldsymbol{\beta}' z_{ij}$ are computed using (3.5) and are tabulated in the final column of Table 3.1. Exponentiation yields the mortality factors $\exp(\hat{\boldsymbol{\beta}}' \cdot z_{ij})$. These are identical to the traditional mortality ratios of the preceding column which the reader may readily verify.

Suppose next we decide to ignore one of the two factors, say B, and just fit A. Columns 3 and 4 of Table 3.1 are modified by summation over j to yield the first three columns of Table 3.2, from which column 4 is constructed.

Fitting model A only with parametric form

$$\boldsymbol{\beta}' z_i = \mu + \alpha_i$$

yields estimates

$$\hat{\mu} = \cdot 4931,\ \hat{\alpha}_1 = 0,\ \hat{\alpha}_2 = \cdot 1050$$

from which column 5 of Table 3.2 is constructed. Again exponentiation reveals that the mortality factors are identical to the traditional mortality ratios of the preceding column. An identical situation occurs on fitting B and ignoring A.

Table 3.2. *Condensed mortality data for hypertensives, 1947–63*

A(i)	d_{i+}	m_{i+}	d_{i+}/m_{i+}	$\boldsymbol{\beta}' z_i$
1	808	493·46	1·64	·4931
2	136	74·75	1·82	·5981

Subject to the marginal switch from standard life-table to standard hazard function, we have demonstrated that the proposed method of analysis reproduces traditional actuarial mortality ratios, as a special case, by fitting either single covariates or two (or more) fully interactive covariates.

3.3 In addition to providing measures of excess mortality the method also enables us to conduct a comprehensive statistical analysis of the association between covariates, their interactions, and excess mortality.

The analysis is based on a model goodness of fit statistic called the deviance. This is based in turn on the likelihood ratio principle rather than the possibly more familiar Pearson goodness of fit statistic. It is essential that inferences should be based on differences between model deviances as their absolute values are conditional on the total number of covariates under simultaneous investigation. These differences may be referred to the chi-square distribution with the appropriate degrees of freedom—an approximate result.

The analysis of deviances for the data presented in Table 3.1 is based on the values tabulated in Table 3.3.

The differences in model deviances and corresponding degrees of freedom (in brackets) are displayed on the lattice diagram (Figure 3.1).

Table 3.3. *Model deviances, mortality data for hypertensives, 1947–63*

Model	Linear predictor $\boldsymbol{\beta}' z_{ij}$	Dev.	D.F.
null H_0	μ	57·57	1
weight A	$\mu+\alpha_i$	56·31	2
age B	$\mu+\beta_j$	·16	3
log-additive $A+B$	$\mu+\alpha_i+\beta_j$	·03	4
fully interactive $A*B$	$\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}$	0	6

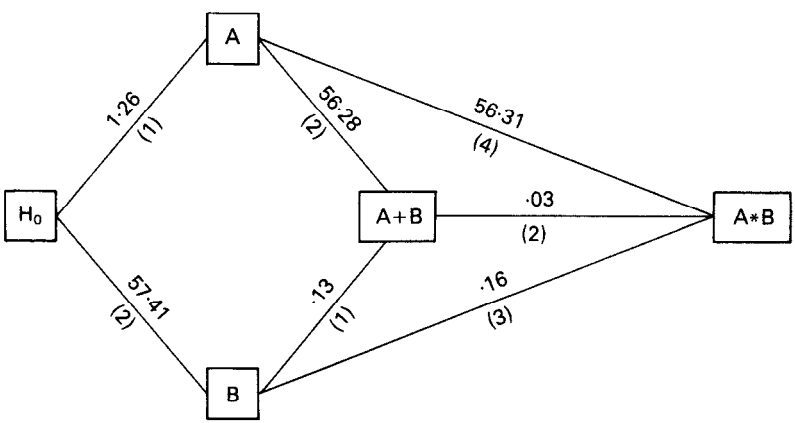

Figure 3.1. Lattice of hypotheses, mortality data for hypertensive, 1947–63

Notice that the lattice deviances (and degrees of freedom) are additive in the sense of moving from one model or hypothesis to another by two different paths, where this is possible (e.g. $A * B$ to A, $A * \mathrm{B}$ to H_o etc.) The GLIM computer package also offers the Pearson goodness of fit statistics as alternatives to the likelihood ratio deviances (same degrees of freedom), but these do not possess this appealing additive property. Examination of the branches connecting the null hypothesis H_o to A, B to $A * \mathrm{B}$, B to $A + B$, all indicate that factor A, weight status, is statistically non-significant; while factor B, age at entry, is highly significant. The branch from $A * \mathrm{B}$ to $A + B$ tests for interaction.

4. THE BASE-LINE HAZARD MATRIX

The A1967–70 table with its four year calendar year span and two-year select period was employed to construct a base-line hazard matrix λ^* as follows. Papaconstantinou (1987) using linear techniques has extrapolated the A1967–70 Table both forward and backward in time in four yearly steps. Then, using linear interpolation, he has computed probabilities of death $q(x, d, c)$ for calendar years $c = 47, 48, \ldots, 81$; durations $d = 1, 2, >2$; and ages $x = 16, 17, \ldots, 79$ ($d = 1, 2$), $x = 16, 17, \ldots, 100$ ($d > 2$). The matrix $q(x, d, c)$ was felt to be unnecessarily detailed and consequently condensed into seven synchronized calendar year periods by averaging over consecutive five yearly intervals, commencing with 1947–51 ($c = 1$) and ending in 1977–81 ($c = 7$). Then the base-line hazard matrix $\lambda^*(x, d, c)$ was computed using

$$\lambda^*(x, d, c) = -\log(1 - q(x, d, c)).$$

5. COMPUTING THE M_j'S

A Fortran 77 program was written to compute the accumulated integrated base-line hazard terms

$$m_j = \sum_k \int_{\tau_{jk}}^{t_{jk}} \lambda^*(u)\, du.$$

Briefly the essence of the program is as follows.

First define the cohorts $j \in \mathcal{J}$. A product set $A \times C \times D \times M$ with at least four component sets is needed to accommodate the traditional actuarial factors of age at entry (A), calendar year of entry (C) and duration (D) along with medical status (M) as model covariates; the set of medical states, M, being sometimes a product set itself. The factors A with either 3 or 4 levels (typically 16–39, 40–49, 50–59, 60–79), C with 7 levels (47–, 52–, 57–, . . . , 77–) and D with 6 levels (0–2, 2–5, 5–10, 10–15, 15–20, 20–) were used in all analyses in conjunction with medical status M (often with the original codes modified).

Next condition in turn on each of the combined levels of $A \times C \times M$ (but not D notice) and compute the contribution of each datum to the different levels of D by

integrating over the appropriate length of time for which the policy is operative. Thus, if $C = 7$ say (1977–81) then the datum contributes only to the first two levels of D at most, and so on.

Recall that each datum comprises

$$(\tau_{jk},\ t_{jk},\ a_{jk},\ \delta_{jk})$$

where:

τ_{jk}—entry time
t_{jk}—exit time
} in months (origin 1/47)

a_{jk}—age next birthday at entry (in years)

δ_{jk}—mode of exit

together with information on medical status. This latter information along with the values of τ_{jk} and a_{jk} is used to sort and classify each datum according to the different levels of $A \times C \times M$. Then the information $(\tau_{jk},\ a_{jk})$ locates the starting element $\lambda^*(a_{jk},\ 1,\ c_{jk})$ in the base-line hazard matrix of the previous section, while it is necessary to assume an exact entry age (exact, that is, to the nearest month and taken to be age next birthday at entry minus 6 months for each datum) in order to commence integrating the step-like base-line hazard matrix. This proceeds in unit steps of one month. The number of deaths per cohort, $d_j = \delta_{j+}$, are accumulated as a by product.

The SPSS-X statistical package was used initially to select, sort and edit the relevant data into a form suitable for feeding into the Fortran 77 program. It was deemed expedient to run the latter program for each medical state M separately because of potential problems with the limitations on array size. A typical piece of output comprises six (or more) columns listing factor levels A, C, D and M together with the values of d_j and m_j. This is now ideal for inputting into the GLIM computer package for statistical analysis.

6. ANALYSIS OF HYPERTENSIVES

Some 25,220 policies were issued on lives diagnosed as hypertensives at entry in the study period 1947–81 (codes 110–168: Preston & Clarke (1965)). The effect on mortality of medical factors:

B—Blood pressure at entry, 7 levels, with codes

		DAP		
		below	average	above
	below		- - - 7 - - -	
S				
A	average	1	3	5
P				
	above	2	4	6

and

E—Weight with 2 levels (1—standard + 19%, 2—standard + 20%, or over) along with factors

A—Age at entry, 4 levels
C—Calendar year of entry, 7 levels
D—Duration, 6 levels

each defined in § 5 above was investigated. So, in the set notation of § 5, medical status $M = B \times E$.

Ignoring all factors and treating the data collectively as a single cohort, the null model or hypothesis H_0, gives rise to the mortality ratio 1·54 implying an excess mortality of +54% for all hypertensives. Fitting each of the five factors separately gives rise to Table 6.1 in which the tabulated model deviance differences give an indication of the strength of association between excess mortality and each factor in turn. On this basis, the weight factor E is not significant statistically and is subsequently dropped from further analysis of these data. This is in agreement with Clarke's findings in his analysis for the period 1964–73 and would appear to be consistent with Preston & Clarke's somewhat limited analysis for the period 1947–63. Parameter estimates for model E (Table 6.2) indicate a slight increase in mortality for the overweights (+64% as opposed to +53%).

The mortality ratios for all five factors fitted separately are given in Table 6.2 together with the observed numbers of deaths. Notice first the pattern of excess mortality with blood pressure, with excess mortality increasing with increasing blood pressure ratings. Clarke observed a similar pattern in his restricted analysis for the period 1964–73, of which more shortly. Secondly, there is a marked decline in excess mortality at the higher ages of entry. Because high blood pressure is very unusual among the young, this higher excess mortality for hypertensives is to be expected at young ages at entry. Conversely, at the oldest ages at entry, all lives are likely to have some level of raised blood pressure so hypertension is not that exceptional. Hence the excess mortality is lower.

Next, examine the variation in excess mortality with duration, an effect which Clarke felt unable to investigate by his methods. Excess mortality is lowest for

Table 6.1. *Testing for main effects, hypertensives 1947–81*

Model	dev.	d.f.	difference	
H_0	1867	1664	dev.	d.f.
A	1723	1661	144	3
B	1754	1658	113	6
C	1831	1658	36	6
D	1843	1659	24	5
E	1865	1663	2	1

Table 6.2. *Excess mortality factors, covariates fitted separately. Hypertensives 1947–81. (Numbers of deaths in parentheses)*

Model							
			Mortality ratio				
H_0			1·52 (3019)				
A	16–39	40–49	50–59	60–79			
	1·80 (353)	2·13 (882)	1·42 (992)	1·22 (792)			
B			DAP				
			below	average	above		
	SAP	below	—	1·42 1(508)			
		average	1·32 (1087)	1·68 (726)	1·77 (104)		
		above	1·63 (153)	2·01 (243)	2·84 (198)		
C	47–	52–	57–	62–	67–	72–	77–
	1·40 (681)	1·45 (803)	1·55 (615)	1·74 (555)	2·05 (217)	1·73 (96)	1·41 (52)
D	0–2	2–5	5–10	10–15	15–20	20–	
	1·32 (223)	1·36 (480)	1·60 (1010)	1·72 (713)	1·55 (365)	1·49 (228)	
E	Standard + 19%		Standard + 20% or over				
	1·53 (2508)		1·64 (511)				

short durations, perhaps reassuringly vindicating company underwriting procedures before accepting business. It rises appreciably for medium duration before declining again at higher duration.

Perhaps the most intriguing feature is the variation in excess mortality with calendar year at entry, with excess mortality peaking significantly for the calendar entry 1967–71. This naturally raises the spectre of whether it is a real effect, and, if so, what possible explanation exists. On the other hand, it is possible that the calendar time variation may be an artefact arising from the particular choice of base line hazard λ^* and its construction, especially as the peak coincides exactly with the calendar period of the A1967–70 standard table used in the construction of λ^*. Clearly further research, which is in progress, is needed here.

The simplest model catering for all four factors A, B, C, D simultaneously has the GLIM additive structure $A+B+C+D$. This translates into multiplicative structure with exponentiation; having the parametric form

$$\exp(\mu + \alpha_i + \beta_j + \gamma_k + \delta_l)$$

where the parameters α, β, γ, δ relates to factors A, B, C, D respectively. The mortality ratios obtained on fitting this model may be deduced from the information displayed in Table 6.3, on forming the product of the relevant entries. Clearly the entries are supportive of the features discussed above.

A more detailed analysis is achieved with the inclusion of interaction terms whose relative significance may be assessed by referring to the differences in model deviances tabulated in Table 6.4. A detailed analysis of each model implies that the interaction terms $B*D$ and $A*D$ are noteworthy.

Mortality ratios for the model $B*D+A+C$ with parametric representation

$$\exp(\mu + \beta_j + \delta_{jl} + (\beta\delta)_{jl}) \exp(\alpha_i) \exp(\gamma_k)$$

may be computed from the entries in Table 6.5 in which the blood pressure categories (j) are presented in the format outlined at the commencement of this section. Clarke (1978) quotes mortality ratios

		DAP		
		below	average	above
	below		1·43	
S				
A	average	1·26	1·60	2·03
P				
	above	2·06	2·36	2·93

Table 6.3. *Excess mortality, multiplicative main effects model. Hypertensives 1947–81*

$\exp \mu = 1{\cdot}32$

i	1	2	3	4			
$\exp \alpha_i$	1	1·14	·75	·67			
j	1	2	3	4	5	6	7
$\exp \beta_j$	1	1·28	1·18	1·50	1·14	1·91	·93
k	1	2	3	4	5	6	7
$\exp \gamma_k$	1	1·03	1·07	1·14	1·31	1·14	1·06
l	1	2	3	4	5	6	
$\exp \delta_l$	1	1·06	1·26	1·28	1·14	1·20	

Table 6.4. *Testing for 1st order interactions. Hypertensives 1947–81. (Tail areas = observed significance levels)*

Model	dev.	d.f.	differences		Tail
$A+B+C+D$	920·7	852	dev.	d.f.	area
$A*B+C+D$	899·5	834	21·2	18	27%
$A*C+B+D$	901·0	834	19·7	18	35%
$A*D+B+C$	900·2	837	20·5	15	14%
$B*C+A+D$	892·7	816	28·0	36	81%
$B*D+A+C$	878·8	822	41·9	30	7%
$C*D+A+B$	912·1	832	8·6	20	99%

for the calendar period 1964–73 with entry age 40–59 years and a duration of 2 or more years. While exact comparisons are not possible here, the new corresponding mortality ratios derived from Table 6.5 compare well.

Rather than reproduce detailed tables for the other noteworthy interactive model $A*D+B+C$, the mortality ratios for the related model $A*D$ are presented graphically in Figure 6.1 which highlights the main source of the interaction.

Table 6.5. *Excess mortality factors, main effects plus major interaction term. Hypertensive 1947–81*

$\exp(\mu+\beta_j+\delta_{jl}+(\beta\delta)_{jl})$

$(j1)$			$(j2)$			$(j3)$		
—	1·13		—	1·34		—	1·39	
1·15	1·65	·90	1·51	1·46	1·70	1·74	1·98	2·19
2·67	2·06	2·97	1·07	2·58	2·75	2·12	2·39	3·04

$(j4)$			$(j5)$			$(j6)$		
—	1·79		—	1·55		—	1·20	
1·52	1·98	2·09	1·48	1·86	1·30	1·72	2·09	0·80
2·69	2·43	3·45	1·92	2·01	1·75	1·75	1·37	4·91

i	1	2	3	4
$\exp \alpha_i$	1	1·15	·75	·67

k	1	2	3	4	5	6	7
$\exp \gamma_k$	1	1·04	1·07	1·14	1·30	1·14	1·02

Clearly more detailed models still involving complex interaction terms are open to scrutiny by this method subject to the upper limit on the number of model parameters within the GLIM environment and provided the data are sufficiently numerous to render the exercise meaningful.

7. CONCLUSIONS

I would suggest that the GLIM approach outlined here could pave the way for a completely new, scientifically sound approach to life insurance underwriting. It offers a more dynamic means of model building that has hitherto been attempted in this field in which the relationship between individual factors and their interactions on excess mortality may be assessed. I would highlight the meagre assumptions on which the models are based, the comparative ease with which they can be fitted and compared using modern statistical packages, and the appealing connexion which these models have with the traditional actuarial standard mortality ratios. I envisage further work on these lines for other impairments, and am investigating further the influence of the specific base-line hazard function used in the analysis and its construction.

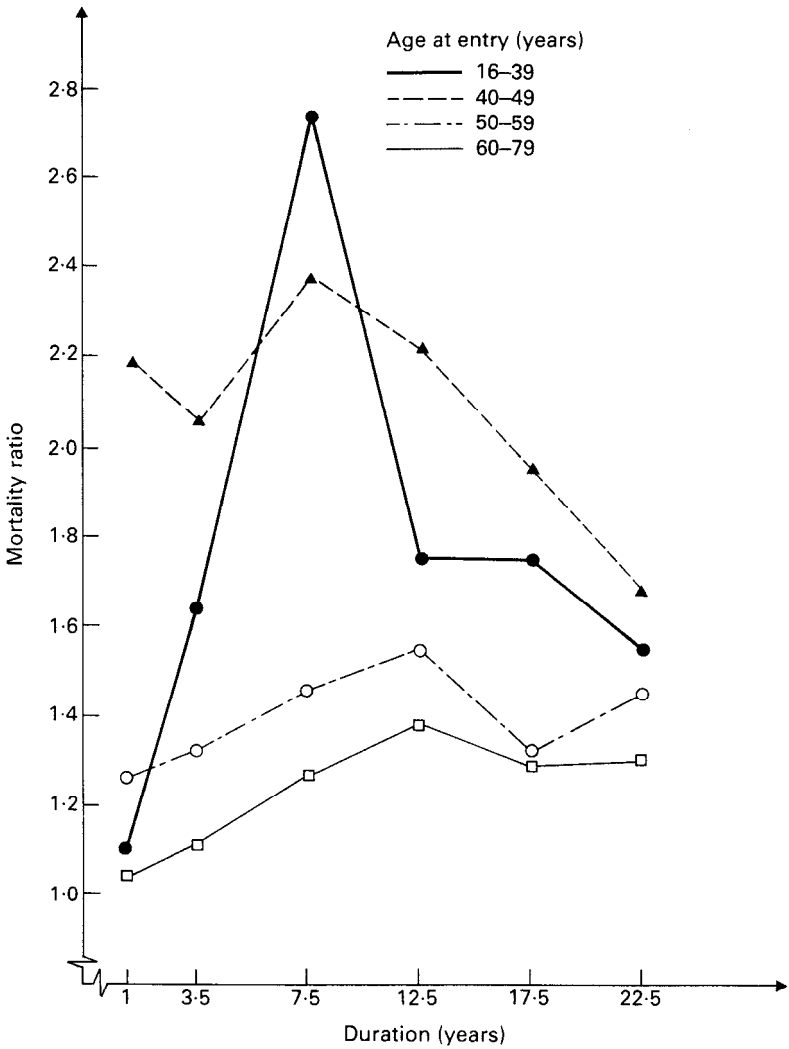

Figure 6.1. Mortality ratios *vs.* duration, by age at entry. Hypertensives 1947–81

8. ACKNOWLEDGEMENT

I am heavily indebted to the Prudential Assurance Company for generously allowing the use of the impaired-lives data set and to my colleague Professor Steven Haberman, not only for drawing my attention to these data, but also for offering unstinted advice on aspects of the analysis and for suggesting improvements to the original draft of this paper.

TECHNICAL APPENDIX

Let the survival time for members of a target population be a non-negative continuous random variable T with density f, hazard λ, survival function $\mathscr{F}$ and integrated hazard Λ; so that

$$\lambda(t) = f(t)/\mathscr{F}(t),\ \mathscr{F}(t) = \exp -\Lambda(t),\ \Lambda(t) = \int_0^t \lambda(u)du. \qquad \text{(A.1)}$$

Recall that the entry and exit times to and from study are denoted by τ and t respectively, while δ denotes the mode of exit.

A study batch of N individuals (i) partitioned according to the mode of exit into a set, $\mathscr{D}$, of observed fatalities ($\delta_i = 1$) and a complementary set, $\bar{\mathscr{D}}$, of non-fatalities and therefore censored individuals ($\delta_i = 0$) gives rise to the following likelihood function

$$L = \prod_{i\in} \frac{f(t_i, z_i)}{\mathscr{F}(\tau_i, z_i)} \prod_{i\in\bar{\mathscr{D}}} \frac{\mathscr{F}(t_i, z_i)}{\mathscr{F}(\tau_i, z_i)}$$

$$= \prod_{i=1}^{N} \lambda(t_i, z_i)^{\delta_i} \frac{\mathscr{F}(t_i, z_i)}{\mathscr{F}(\tau_i, z_i)}$$

(see e.g. Elandt-Johnson and Johnson (Chapter 13)). The vector of covariates, z, is needed to provide additional insight into the nature of T by catering for any real or suspected heterogeneity in the target population. Each observed fatality ($\delta = 1$) contributes an amount

$$\lim_{\delta t \to 0} \frac{P(t < T < T + \delta t | T > \tau)}{\delta t} = \frac{f(t, z)}{\mathscr{F}(\tau, z)}$$

$$= \lambda(t, z) \frac{\mathscr{F}(t, z)}{\mathscr{F}(\tau, z)}$$

on using (A.1), with each censored datum ($\delta = 0$) contributing an amount

$$P(T > t \mid T > \tau) = \frac{\mathscr{F}(t, z)}{\mathscr{F}(\tau, z)}.$$

Again making use of identities (A.1), the log likelihood becomes

$$\log L = \sum_{i=1}^{N} (\delta_i \log \lambda(t_i, z_i) - \int_{\tau_i}^{t_i} \lambda(u, z_i)\, du) \tag{A.2}$$

which is true under very general conditions.

Trivially, the effect of introducing Cox's multiplicative hazard function

$$\lambda(t, z) = \lambda^*(t) \exp(\boldsymbol{\beta}' z)$$

into (A.2) is to give the following result

$$\log L(\boldsymbol{\beta}) = \text{const} + \sum_{i=1}^{N} (\delta_i \boldsymbol{\beta}' z_i - \exp(\boldsymbol{\beta}' z_i) \int_{\tau_i}^{t_i} \lambda^*(u)\, du) \tag{A.3}$$

where we identify specifically the dependence of $\log L$ on the unknown regression parameters, $\boldsymbol{\beta}$. Partitioning individuals i into cohorts j by writing $i = (j, k)$ implies that (A.3) can be written as

$$\log L(\boldsymbol{\beta}) = c + \sum_{j \in \mathscr{J}} (d_j (\log m_j + \boldsymbol{\beta}' z_j) - \exp(\log m_j + \boldsymbol{\beta}' z_j)) \tag{A.4}$$

where

$$d_j = \delta_{j+} = \sum_k \delta_{jk}$$

the number of observed fatalities in the jth cohort, and

$$m_j = \sum_k \int_{\tau_{jk}}^{t_{jk}} \lambda^*(u)\, \mathrm{du}$$

the accumulated integrated base-line hazard; the vector of covariates $z_i = z_{jk} = z_j$ being constant within cohorts. We remark that the first term

$$\sum_{j \in \mathscr{J}} (d_j \log m_j$$

in (A.4), being independent of the unknown regression parameters β, is conjured out of the constant of proportionality in order to match the argument of the exponential terms in (A.4). Writing this argument as

$$\log \mu_j = \log m_j + \beta' z_j$$

so that

$$\mu_j = m_j \exp(\beta' z_j), \tag{A.5}$$

expression (A.4) reduces to

$$\log L(\boldsymbol{\beta}) = c + \sum_{j \in \mathcal{J}} (d_j \log \mu_j - \mu_j);$$

the kernel of the log likelihood function of independent Poisson variables $d_j \sim \text{IPoi}(\mu_j)$ where μ_j is given by (A.5).

REFERENCES

BERRY, G. (1983) The Analysis of Mortality by the Subject-years Method. *Biometrics*, **39,** 173–184.

BRESLOW, N.E., LUBIN, J.H., MAREK, P. & LANGHOLZ B. (1983) Multiplicative Models and Cohort Analysis *J. Am. Statist. Ass.* **78,** 1–12.

CLARKE, R.D. (1978) Mortality of Impaired Lives 1964–73, with discussion *J.I.A.* **105,** 15–46.

COX, D.R. (1972) Regression Models and Life Tables, with discussion *J.R. Statist. Soc. B.* **34,** 187–220.

ELANDT-JOHNSON, R.C. & JOHNSON, N.L. (1980) *Survival Models and Data Analysis*. John Wiley & Sons.

HABERMAN, S. (1988) Measuring Relative Mortality Experience. J.I.A., **115,** 271.

HABERMAN, S. & RENSHAW, A.E. (1988) *The Use of Generalized Linear Models in Actuarial Work*. Presented to a Joint Meeting of the Staple Inn Actuarial Society and Royal Statistical Society General Applications Section.

HILL, C., LAPLANCHE, A. & RAZVANI, A. (1985) *Statist in Medicine*, **4,** 295–302

PAPACONSTANTINOU, I. (1987) *Statistical Analysis of Impaired Assured Lives*, PhD. Thesis. City University.

PRESTON, T.W. & CLARKE, R.D. (1965) An Investigation Into the Mortality of Impaired Lives during the period 1947–63, with discussion. *J.I.A.* **92,** 27–74