Presented to the Staple Inn Actuarial Society

on 2\textsuperscript{nd} February 1988

GENERALISED LINEAR MODELS
IN ACTUARIAL WORK

by

S Haberman and A E Renshaw
INTRODUCTION

It is probably not an overstatement to say that generalised linear models together with the statistical package GLIM have revolutionised both the teaching of undergraduate and graduate statistics and the carrying out of data analysis in practice. As an example of this teaching revolution, we refer to changes in our own institution where mainstream courses in the analysis of experimental data (at the BSc and MSc levels) were completely revised in 1980, since when generalised linear models and GLIM have together been an integral part of the academic content.

As will be described in a later section, generalised linear models are a natural generalisation of the familiar classical linear models. The class of generalised linear models includes, as special cases, linear regression, analysis-of-variance models, log-linear models for the analysis of contingency tables, logit models for binary data in the form of proportions and many others.

The use of classical linear models in actuarial work is not new. Thus, such models have become an established part of the description of claim frequency rates and average claim costs in motor insurance – as evidenced by a number of papers in the British actuarial literature, including Johnson & Hey, Grimes, Bennett, Baxter et al and Coutts.

However, the use of generalised linear models in actuarial work is relatively new. Thus, McCullagh and Nelder in their excellent and comprehensive monograph give a number of examples of the fitting of generalised linear models to different types of data. One of these relates to data from Baxter et al on the average claim costs in a motor insurance portfolio (originally modelled by Baxter et al using a weighted least squares approach). In our turn, we have made a small step in the direction of using generalised linear models in life insurance when we modelled the variation of lapse rates with age at entry, duration of policy, type of policy and insurance company. Some of these models are described further in section 5.

Our purpose in this paper is to show that generalised linear models have a wide area of application in actuarial work and are not confined merely to models for motor insurance premiums. We hope to fulfil this purpose by demonstrating three separate practical applications in actuarial work:

(i) Fitting of loss distributions in non-life insurance (section 3);

(ii) Representing the variation in force of mortality in life insurance underwriting (section 4); and
In section 3 we will be concerned with fitting distributions, in section 4 with representing the force of mortality (or hazard rate) as a function of several variables when we have data for a cohort being followed up over time and in section 5 with representing an actuarial rate or probability as a function of several variables.

We hope that the reader will appreciate the wide applicability of the ideas described here. The methodology of section 3 can be used for fitting various types of distribution. Section 4 would apply to any cohort experiment where the effect of several covariates on survival over time is to be investigated. Section 5 would apply to the variability of actuarial rates with different covariates (as in motor insurance claim frequencies or in mortality rates for group life insurance, for example).

Before we describe these three applications we shall use section 2 to introduce the structure of generalised linear models.

(iii) Representing the variation in lapse rates with policy characteristics in life insurance (section 5).

Each of these applications involves different types of data and a different type of model.
2. **GENERALISED LINEAR MODELS. (GLMs)**

Classical linear models comprise a vector of \( n \) independent normally distributed response random variables \( Y \) with means \( \mu = \mathbb{E}(Y) \) and constant variance \( \alpha^2 \). A non-random, systematic structure is incorporated by assuming the existence of vector covariates \( x_1, x_2, \ldots, x_p \) with known values such that

\[
\mathbb{E}(Y) = \mathbb{E}(Y_i) = \mu = \sum_{j=1}^{p} \beta_j x_{ij} = X \beta
\]

where \( x_{ij} \) is an \( n \times 1 \) vector, \( \beta \) is a \( p \times 1 \) vector, \( X \) is an \( n \times p \) matrix and \( j = 1, 2, \ldots, p \).

So letting \( i \) index the observations we have that:

\[
\mathbb{E}(Y_i) = \mu_i = \sum_{j=1}^{p} \beta_j x_{ij}
\]

where \( x_{ij} \) is the value of the \( j \)th covariate for observation \( i \).

The \( \beta_i \)'s are usually unknown parameters which have to be estimated from the response data \( y \) assumed to be a realisation of \( Y \). Specific choices of the design matrix \( X \) lead to a broad class of linear models which includes the familiar regression type model:

\[
Y_i \text{ are independent random variables distributed as } N(\alpha_i + \beta x_i, \alpha^2)
\]

and the respective one factor and two factor non-interactive models which form the basis of the familiar analysis of variance tests, viz:

\[
Y_i \text{ are independent random variables distributed as } N(\alpha_i, \alpha^2)
\]

\[
Y_{ij} \text{ are independent random variables distributed as } N(\alpha_i + \beta_i, \alpha^2).
\]

Such models are traditionally fitted by a least squares method which, because of the independent normal assumption, is equivalent to maximum likelihood. An unbiased estimator based on the residual sum of squares is taken for \( \alpha^2 \). The adequacy of the model, including the strong constant variance assumption, normal assumption and imparted systematic structure are monitored through residual plots. Attempts to simplify the systematic structure by nesting models may be tested statistically using the familiar F-test.

Classical linear models may be generalised in two respects: firstly, through the introduction of a much wider class of distributions, the so-called exponential family of distributions; and, secondly by linking the systematic component or linear predictor

\[
n = X \beta
\]
to the means $\mu$ of the independent response variables through the introduction of a monotonic differentiate function $g$ where

$$n = g(\mu) \quad (2.2)$$

so that

$$\mu = g^{-1}(n) = g^{-1}(X\beta).$$

The function $g$ is called the link function.

The exponential family of distributions includes the normal, binomial, Poisson and gamma distributions amongst its members. Clearly, the normal distribution has to be selected in conjunction with the identity link function in order to retrieve the classical linear model (described above).

The GLIM computer package is specifically designed to enable the user to fit generalised linear models interactively. It offers the choice of modelling distribution, link function and linear predictor. Parameter estimates, fitted values and a goodness of fit measure, called the deviance, all form part of the output as each model is fitted; while a variety of residual plots can be displayed. In addition, users may add their own programs (or macros) to GLIM, providing further versatility.

Estimation of the linear predictor parameters $\beta$ is by maximum likelihood using an iterative weighted least squares algorithm. There exist sufficient statistics for the parameters in the linear predictor (equation 2.1) provided that the modelling distribution is matched with a specific link function called the canonical or natural link function. Alternative link functions are, however, available.

With $n$ observations $y$ available, models with between 1 and $n$ parameters may be contemplated. The extreme cases have an important role to play. The null model, comprising a single parameter has the property that all the $\mu$ components are identical, leaving all of the variation in the data $y$ to be accounted for in the error structure of the model. At the other extreme, with $n$ independent data components and $n$ parameters to estimate, the so-called saturated model is attained in which the fitted values are the data themselves ($\mu = y$). Clearly, we seek an optimum model somewhere between these extremes. Ideally, it should involve as few parameters as possible while accounting for the salient structure present in the response data, leaving a small pattern-free set of residuals.

The examination of residual plots plays an important role in the assessment of the viability of any proposed model.
Goodness of fit is based on the likelihood ratio principle with the saturated model providing the benchmark, rather than on an adaptation of the possibly more familiar Pearson (chi-square) goodness of fit criterion. Consider any intermediate model, with \( p \) parameters, called the current model. Then, a measure of the goodness of fit of the current model is provided by the ratio of the maximum value of the likelihood function under the current model to that under the saturated model. Minus twice the logarithm of this ratio (a monotonic mapping) is defined to be the (current) model deviance. Nested models can then be compared on the basis of the differences between their model deviances. In the case of the classical linear model, the model deviance collapses into the familiar residual sum of squares term.

It is important that inferences should be based on differences between the deviances for different models since their absolute values are conditional on the total number of covariates under simultaneous investigation. Differences between the deviances may be referred to the chi-square distribution with appropriate degrees of freedom for a formal test (this is an approximate result).

The reader is referred to Chapter 2 of the book by McCullagh and Nelder for a more detailed description as well as to the GLIM manual.
3. FITTING LOSS DISTRIBUTIONS

The term "loss distribution" was introduced by Hogg and Klugman as a brief title for their monograph which has since become an established part of the course of reading for the professional examinations of both the Institute of Actuaries and the Society of Actuaries (USA). The term refers to the distribution of single losses that are related to claims made against a variety of insurance policies. So, we do not consider the probabilities of the occurrence of various numbers of claims and we are not concerned here with the aggregate claims experience of a portfolio or company. We are concerned just with the distribution of the size of a loss, given that such a loss has occurred.

Losses in the insurance industry tend to be heavy-tailed and skewed to the right and so the appropriate theoretical distributions have the same features and are "unusual" in the context of conventional distribution theory.

In this section, we explain how GLIM can be used to facilitate the fitting of loss distributions to data and illustrate the methodology using a real data set quoted by Hogg and Klugman. Although GLIM is used in this section, the problem discussed does not strictly fit into the realm of generalised linear models.

For the methodology, we adapt the iterative procedure for fitting censored data suggested by Aitkin and Clayton.

Focus attention on a class of modelling distributions with the density \( f \) and "survivor" function \( F(F(x) = P(X>x)) \) of the type

\[
f(x) = \alpha \lambda(x) \exp(-\alpha \Lambda(x)) \\
F(x) = \exp(-\alpha \Lambda(x))
\]

where both \( \Lambda(X) \) and

\[
\Lambda(x) = \int_{d}^{x} \lambda(u) \, du
\]

involve parameters, \( \theta \), other than the parameter \( \alpha > 0 \). This class of distributions includes the Burr, Pareto and Weibull distributions, specific details of which are listed in Table 3.1. When \( d > 0 \) the distributions are truncated below, a not uncommon feature associated with claims distributions: for a direct insurer this may reflect a policyholder's excess and for a reinsurer this may reflect truncation via the direct insurer's retention level.
When the data comprise independent claims of amounts \((x_1, x_2, \ldots, x_n)\) then the likelihood function is

\[
L = \prod_{i=1}^{n} a \lambda(x_i) \exp(-a\lambda(x_i))
\]

\[
= \prod_{i=1}^{n} \mu_i \exp(-\mu_i \lambda(x_i))
\]

where

\[
\mu_i = a \lambda(x_i)
\]

so that

\[
n_i = \log \mu_i = \log a + \log \lambda(x_i).
\]

The first two terms of \(L\) form the kernel of the Poisson likelihood based on \(n\) independent Poisson responses, \(Y_i \sim \text{IPoi}(\mu_i)\), all of which realise the value one \((y = 1)\). In addition, the remaining term \(\Lambda(x) / a\Lambda(x_i)\) involves the parameters \(\theta\), but not \(a\). Comparison of equation (3.2) with equations (2.1) and (2.2) is suggestive of the logarithmic link function, a linear predictor with \(n \times 2\) design matrix \(X\) satisfying

\[
x_{i1} = \log \lambda(x_i); \quad x_{i2} = \log \lambda(x_i); \quad i = 1, 2, \ldots, n
\]

and parameters \(\beta_1 = \log \alpha, \beta_2 = 1\). This is a little unusual in-so-far as one of the two \(\beta_i\)'s is known in advance and does not have to be estimated. The combined contribution of such terms to the linear predictor \(n, x_\star\) in this instance, is called an "offset". In fitting such a model, the offset is first subtracted from the linear predictor and the result can then be regressed on the remaining covariate(s) \(x\). It follows therefore that, given the values of \(a\) and hence \(\Lambda(x)\), the maximum likelihood estimate for \(\log a\) (and hence \(\alpha\)) is obtained on fitting the generalised linear model with independent Poisson response variables (values all unity) in conjunction with the natural log-link. The \(a\Lambda(x)\) terms are declared as offsets and the null model is selected.

To establish an iterative procedure, we further require starting values for the \(\theta_i\)'s as well as a means of updating the values at each stage. The updating may be done by using any convenient arrangement of the maximum likelihood equations as suggested by Aitkin and Clayton or by using Newton-Raphson methods.

We illustrate the technique by fitting the Pareto distribution to the trend adjusted losses tabulated in Table 3.2. These losses are based on data compiled by the American Insurance Association and
are attributed to hurricane storm damage events from 1949 through to 1980, excluding those hurricanes for which the total adjusted losses were less than $5 million. The reader is referred to Chapter 4 of the book by Hogg and Klugman for further details.

The Pareto distribution has unknown parameters $\alpha, \lambda > 0$ with

$$P(X > x) = \left(\frac{\lambda + d}{\lambda + x}\right)^\alpha, \quad x > d > 0.$$  

Truncation is at the point $x = d$, which is given. A convenient starting value for $x$ can be based on estimation by the method of moments. Denoting the mean and variance of the observed losses by $\bar{x}$ and $s^2$ respectively, this estimator is

$$\hat{x} = \hat{\alpha}(\bar{x} - d) - \bar{x}$$

where

$$\hat{\alpha} = \frac{2s^2}{(s^2 - (\bar{x} - d)^2)}.$$

The specific form for the log-likelihood is

$$\log L = n \log \alpha - (\alpha + 1) \sum \log (\lambda + x_i) + n\alpha \log (\lambda + d).$$

Setting its partial derivative with respect to $x$ equal to zero yields the equation

$$g(\lambda) = \frac{n\alpha}{\lambda + d} - (\alpha + 1) \frac{1}{i} \frac{1}{\lambda + x_i} = 0.$$  

Values of $x$ may then be updated at each stage of the iteration procedure using the Newton-Raphson formula

$$\lambda_{n+1} = \lambda_n - \frac{g(\lambda_n)}{g'(\lambda_n)}.$$

The offsets are

$$\log(\log(\lambda + x_i) - \log(\lambda + d)).$$

The maximum likelihood (m.l.) estimates obtained by this method together with those obtained by Hogg and Klugman, using a different computation routine, are tabulated below:
Similar experiments have been successfully carried out for the Burr and Weibull distributions.

At present we are investigating the possibility of extending this methodology to data which are in a grouped form so that the individual losses themselves are not known.
Loss Distributions

Table 3.1

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Burr</th>
<th>Pareto</th>
<th>Weibull</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha, \lambda, \gamma &gt; 0$</td>
<td>$\alpha, \lambda &gt; 0$</td>
<td>$\alpha, \gamma &gt; 0$</td>
<td></td>
</tr>
<tr>
<td>Domain</td>
<td>$x &gt; d &gt; 0$</td>
<td>$x &gt; d &gt; 0$</td>
<td>$x &gt; d &gt; 0$</td>
</tr>
<tr>
<td>$f(x)$</td>
<td>$\frac{\gamma-1}{\alpha} \frac{(\lambda+d)^\alpha}{(\lambda+x)^{\alpha+1}}$</td>
<td>$\frac{\lambda \alpha}{(\lambda+x)}$</td>
<td>$\alpha \gamma x^{-1} \exp(-\alpha(x^\gamma-d^\gamma))$</td>
</tr>
<tr>
<td>$F(x)=P(X&gt;x)$</td>
<td>$\frac{(\lambda+d)^\alpha}{(\lambda+x)^{\alpha}}$</td>
<td>$\exp(-\alpha(x^\gamma-d^\gamma))$</td>
<td></td>
</tr>
<tr>
<td>$\lambda(x)$</td>
<td>$\frac{\gamma \lambda x^{-1}}{\lambda+x^\gamma}$</td>
<td>$\frac{1}{\lambda+x}$</td>
<td>$\gamma x^{-1}$</td>
</tr>
<tr>
<td>$\Lambda(x)$</td>
<td>$\log\frac{\lambda+x^d}{\lambda+d^\gamma}$</td>
<td>$\log\frac{\lambda+x}{\lambda+d}$</td>
<td>$x^{-d^\gamma}$</td>
</tr>
</tbody>
</table>

Table 3.1

<table>
<thead>
<tr>
<th></th>
<th>6766</th>
<th>19030</th>
<th>47905</th>
<th>103217</th>
<th>361200</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7123</td>
<td>25304</td>
<td>49397</td>
<td>123680</td>
<td>421680</td>
</tr>
<tr>
<td></td>
<td>10562</td>
<td>29112</td>
<td>52600</td>
<td>140136</td>
<td>513586</td>
</tr>
<tr>
<td></td>
<td>14474</td>
<td>30146</td>
<td>59917</td>
<td>192013</td>
<td>545778</td>
</tr>
<tr>
<td></td>
<td>15351</td>
<td>33727</td>
<td>63123</td>
<td>198446</td>
<td>750389</td>
</tr>
<tr>
<td></td>
<td>16983</td>
<td>40596</td>
<td>77809</td>
<td>227339</td>
<td>863881</td>
</tr>
<tr>
<td></td>
<td>18383</td>
<td>41409</td>
<td>102942</td>
<td>329511</td>
<td>1638000</td>
</tr>
</tbody>
</table>

Adjusted losses ($000's$), hurricane storm damage, USA 1949-80.

Table 3.2
4. MODELLING EXCESS MORTALITY FOR IMPAIRED LIVES

4.1 Introduction

Consider the following hypothetical set of data. We have a group of male insured lives whose survival experience has been recorded by the insurance company. The lives are accepted for insurance at different ages and are observed for different periods of time. We are interested in the mortality differences between the LEFT handed (L) and the RIGHT handed (R).

The classical, actuarial approach to such a problem would be to calculate the age-specific mortality rates estimates for the L and R groups using exposed-to-risk theory (so that we allow for censoring). We might then, as a first stage, compare graphically the two derived sets of $\mu_x$, the observed force of mortality or hazard rate (Figure 4.1). How do we measure whether these curves are significantly different? Do we need to be careful about measuring differences at particular ages or durations? Should we be looking for overall measures of differences? The last two questions are written in expectation of the answer "Yes".

Similarly, a comparison of life tables based on these $\mu_x$ values, either in the form of cumulative survival or cumulative mortality rates up to a specific age or duration would have the serious disadvantage that we would be concentrating on a single age or time interval - a process which necessarily wastes information.

Suppose now that we believe, as might be suggested by Figure 4.1, that the gulf between the two groups widens with age. How do we allow for this? Perhaps this gulf is associated with sex, too. We may speculate that there is an interaction between handedness and age and sex. Perhaps we should allow for graphs showing the variation in $\mu_x$ with age: for the male right-handers, male left-handers, female right-handers, female left-handers.

We believe that these unanswered questions suggest that the situation needs a statistical model to be formulated which can accommodate the incorporation of several covariates simultaneously and the tests of such hypotheses.

The key is to bring together the methodology of survival analysis (as represented by the life table) and that of linear regression. This was first proposed by Cox in his epoch-making paper of 1972 which, in statistical terms, has become the progenitor of a new discipline - survival analysis.

Like Cox, we focus attention on the specific hazard rate $\lambda(s)$ where $s$ denotes time. As noted above this usually appears in actuarial literature as the age-specific force of mortality, $\mu_x$. 
Central to the method is the introduction of the multiplicative hazard model

\[ \lambda(s, z) = \lambda^*(s) \exp(\beta'z) \]

where the vector \( z \) represents the different levels of \( p \) covariates (e.g. age or sex or handedness for the earlier hypothetical example), \( \beta \) is a vector of unknown regression parameters to be estimated from data and \( \lambda^*(s) = \lambda(s, 0) \) is the so-called base-line hazard rate. Applications of Cox's method, in which \( \lambda^* \) is unknown, have proliferated in the intervening years since 1972, largely in the field of medical statistics. There have been some, albeit relatively few, applications in which the base-line is assumed known at the outset.

We intend to apply the methodology to the mortality of insured impaired lives (i.e. lives with an identified health impairment present when they are accepted for life insurance) and here there is an obvious choice for \( \lambda^* \) - viz the standard life table in general use for insured unimpaired ("healthy") lives.

The multiplicative mortality factor, \( \exp(\beta'z) \), may be perceived as adjusting the base-line hazard by the amount of excess mortality attributable to the specific impairments defined by the vector covariates. We propose that the \( \exp(\beta'z) \) factor be used as a measure of excess mortality.

Three major questions arise. There is not space to address these fully here but the interested reader is referred to Renshaw\(^1\) for an extensive discussion. The questions are:

1. How are the mortality factors to be computed?

2. What relationship if any do these factors bear to the actual/expected (A/E) mortality ratios traditionally computed (for example, by Preston and Clarke\(^2\))?

3. What advantages, if any, accrue from this more complex approach?

Regarding question 1., to compute and assess the significance of the mortality factors \( \exp(\beta'z) \) for a variety of known covariate structures \( \beta'z \), we require estimates for the regression parameters \( \beta \). These may be based on the method of maximum likelihood and computed using a generalised linear model - full details are provided by Renshaw\(^1\). Some of the formulae are introduced in section 4.3.

Regarding question 2., the relationship is illustrated in section 4.3 of this paper. Further, Renshaw\(^1\) demonstrates 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). Thus, the approach advocated here is a generalisation of the A/E ratios.
Regarding question 3, the approach provides measures of excess mortality and enables a comprehensive statistical analysis of the association between covariates, their interactions and excess mortality to be carried out.

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 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 underestimate the level of mortality rather than get it exactly right. Of course, the situation is different with annuities and pensions in payment but the general point remains. The advent of Acquired Immune Deficiency Syndrome (AIDS) may change this prevailing view.

The purpose of section 4 of this paper is to illustrate the potential of mortality analysis based on GLIM by looking again at part of the impaired lives data set (from one large insurance company) previously reported and discussed by Preston and Clarke\textsuperscript{12}, Clarke\textsuperscript{13} and Leighton\textsuperscript{14}. Somewhat prophetically, Bernard Benjamin predicted the value of these methods for investigating the mortality of special groups, such as those with impairments, in the discussion on Cox's 1972 paper.

4.2 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 and Clarke\textsuperscript{12} and need not be reproduced here. Entry and exit dates are known to the nearest month.

Typically, Figure 4.2 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 co-ordinates 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 o. 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.

4.3 Analysis of Results for Peptic Ulcer Cases

A full analysis of the excess mortality experienced by lives in this study, examining all the major impairment categories in detail, will be provided by Papaconstantinou\textsuperscript{15}.

In this paper we consider only impaired lives identified at entry as suffering from peptic ulcers (medical codes 200 to 207 and 210 to 217). The term peptic ulcer is used by physicians to describe an ulcer occurring anywhere in that part of the alimentary tract which comes into contact with the digestive juices. The most common types are duodenal, gastric and oesophageal ulcers.

34,631 male lives with a diagnosis of peptic ulcer are included in this study. The data allow the incorporation of six covariates as factors into the model structure. Three define medical status at entry according to the format:

- **H** - history, 2 levels
  - 1 - short, apparently acute
  - 2 - long, apparently chronic

- **P** - operation, 2 levels
  - 1 - no
  - 2 - yes
    - 1 - no complication
    - 2 - history of haematemesis (i.e. vomiting of blood)
    - 3 - history of perforation
    - 4 - other complications

- **I** - index, 4 levels
  - 1 - no complication
  - 2 - history of haematemesis (i.e. vomiting of blood)
  - 3 - history of perforation
  - 4 - other complications

and three define age, year and duration as follows:

- **A** - age at entry, 3 levels
  (15-39 yrs; 40-49 yrs; 50 and over)

- **C** - calendar year at entry, 7 levels
  (5 yearly intervals commencing Jan. 1947)

- **D** - policy duration, 6 levels
  (0-1 yrs, 2-4 yrs, 5-9 yrs, 10-14 yrs, 15-19 yrs, 20 yrs and over).
This classification of covariates is undesirable in some respects. Firstly, in cases with a short history (H=1) there may have been little time for complications to develop prior to acceptance for insurance so that the categorisation I=1 may be misleading in terms of measuring severity. Further, it is not clear how reliable is the category I=4 which may comprise a heterogeneous mixture of minor and major complications.

The numbers of cases available for analysis are shown in Table 4.1. The six covariates, with levels as defined in Table 4.1, give rise to $2 \times 2 \times 4 \times 96 = 1536$ cross-classified cells or cohorts which we index by $j \epsilon J$. Each policy experience provides information comprising:

- $\tau_{jk}$ - entry date
- $t_{jk}$ - exit date (nearest month)
- $a_{jk}$ - age next birthday at entry
- $\delta_{jk}$ - mode of exit ($\delta=0$ for censorship, $\delta=1$ for death)

Together with medical status at entry. The index $k$ is a serial index ranging over the members within the particular cohorts. Some of the cross-classified cells may be devoid of data.

Subject only to the assumption that the time, $S_{jk}$, which each policy contributes to the overall experience is a non-negative continuous random variable whose hazard $\lambda(s_{jk}, z_{j})$ satisfies the multiplicative hazard model,

$$\lambda(s_{jk}, z_{j}) = \lambda^*(s_{jk}) \exp(a^*z_{j})$$

it is possible to show (see Renshaw\textsuperscript{11}) that the log likelihood for such data is given by

$$\log L(\beta) = \text{const.} + \sum_{j \epsilon J} \left( d_{j} \left( \log m_{j} + a^*z_{j} \right) - \exp \left( \log m_{j} + a^*z_{j} \right) \right)$$

where

$$d_{j} = \sum_{k} \delta_{jk},$$

the number of observed deaths in the $j$th cohort, and

$$m_{j} = \sum_{k} \int_{\tau_{jk}}^{t_{jk}} \lambda^*(s) \, ds,$$

the "accumulated integrated base-line hazard". This quantity may be shown to be closely related to the expected number of deaths used to form the traditional Actual/Expected mortality ratio (Haberman\textsuperscript{16}).
A Fortran 77 program has been written to compute both \( d_j \) and \( m_j \) for each cohort \( j \). Integration is achieved numerically using a base-line hazard matrix \( \lambda^*(x,d,c) \) with 7 calendar year increments \( c=1(1947-51), \ldots , c=7(1977-81) \); for 3 durations \( d=1(0-1 \text{ yrs.}) \), \( d=2(1-2 \text{ yrs.}) \), \( d=3(>2 \text{ yrs.}) \); and with annual age increments \( x=16, 17, \ldots , 79 \) \( (d=1,2) \), \( x=16, 17, \ldots , 100 \) \( (d=3) \). \( \lambda^*(x,d,c) \) is based on the representative A1967-70 life table suitably transformed according to

\[
\lambda^*(x,d,c) = -\log(1-q(x,d,c))
\]

where \( q \) denotes the probability of death, adjusted for the secular trend in general insured lives mortality using linear interpolation and extrapolation - see Papacontantinou\(^{15}\) for a full discussion of these adjustments. The values \( (r_{jk}, a_{jk}) \) locate the starting element of the base-line matrix while integration proceeds in unit steps of one month. Subject to the minimal switch from standard life-table to standard hazard in the construction of \( m_j \) and the choice of \( j \), readers will immediately identify the ratio \( d_j/m_j \) with the traditionally constructed Actual/Expected (A/E) mortality ratios.

The A1967-70 life table is based on the experience of male lives accepted at standard premium rates by insurance companies transacting life insurance in the United Kingdom. How applicable is this table to the particular insurance company being considered here? To answer this question, we have recourse to the work of Clarke\(^{13}\) who compared the mortality experience of standard (unimpaired) lives of this insurance company with the combined experience of all insurance companies and concluded that the A1967-70 table was an appropriate basis.

To compute maximum likelihood estimators for the regression parameters \( \beta \) and various model structures \( \beta'z \) we use the GLM (generalised linear model) analogue based on independent Poisson variables \( d_j \sim \text{Poi}(\mu_j) \) with means

\[
u_j = m_j \exp(\beta'z_j),
\]

natural log-link

\[
\eta_j = \log \nu_j = \log m_j + \beta'z_j
\]

and offsets \( \log m_j \). It is a trivial matter to verify that the log likelihood function of this GLM is identical to that quoted in expression (4.1).

To establish a connection between the mortality factors \( \exp(\beta'z_j) \) and the traditional actuarial mortality ratios, consider the equation

\[
u_j = E(d_j) = m_j \exp(\beta'z_j)
\]
which we can write as

\[
\frac{d_j}{m_j} = \exp(\hat{\beta}'z_j),
\]

on replacing \(u_j\) and \(\beta\) by their respective estimators \(\hat{u}_j = d_j\) and \(\hat{\beta}\). Then, provided we select the specific covariate structures associated with either single factor models or two (or more) fully interactive factors models, the mortality measures \(\exp(\hat{\beta}'z_j)\) are identical to the traditional (A/E) mortality ratios (subject only to the minimal switch from standard life table to standard hazard function in the construction of the \(m_j\) terms). In addition to this close relationship with existing actuarial practice, the proposed method of analysis offers an additional comprehensive means of assessing the effects of covariates and their interactions on excess mortality.

Fitting the null hypothesis \(H_0\) with model structure \(\beta'z_j = \mu\) for all \(j\) yields the mortality factor \(\exp(\mu) = 1.15\), implying an estimated excess mortality of +15% for peptic ulcer sufferers treated as a homogeneous mass. We stress that this figure (in common with all other mortality ratios quoted) is based on the specific experience of a single company. Consequently great caution must be exercised in inferring that this amount of excess mortality applies to the total population of insured lives with this impairment due to possible bias induced by the company case selection procedure.

The model deviances and mortality factors computed for each covariate fitted separately are given in Table 4.2 and Table 4.3 respectively. Examination of the differences in model deviances (Table 4.2) gives an indication of the strength of association between excess mortality and the covariate concerned. These differences may be tentatively related to the chi-square distribution with appropriate degrees of freedom as quoted. We also quote the observed significance levels - the probability that a more extreme observation than that observed can occur. On this basis we note that each covariate, except for calendar year at entry, has a highly significant effect on excess mortality. The nature of these effects may be gauged by examining the mortality ratios quoted in Table 4.3. Although these were computed directly from the model parameter estimates outputted by GLIM, we also quote the number of cases, observed number of deaths and values of the accumulated integrated base-line hazards for completeness and for comparison with the published tabulations of Preston and Clarke\(^1\), Clarke\(^2\) and Leighton\(^3\).

Table 4.3 shows the estimates of the mortality ratios for the six models where each of the main effects is fitted separately. The ratios indicate that mortality is higher for those cases with a short history rather than a long history (apparently acute \(v\), apparently chronic), that mortality is higher for those cases with
an operation and that mortality is higher as we move from no complication to haematemesis to perforation and then to other complications. These trends in relation to the medical covariates are as we would expect. For age at entry, the trend is as expected with the ratios decreasing with increasing age at entry (peptic ulcers are more "abnormal" among young lives). The trends for calendar year of entry and duration of policy are uneven - for calendar year of entry there is a peak for 1961-66 and then a subsequent fall with a sharp drop during 1977-81 (although here the numbers are small with only 14 deaths in this period). This recent downward trend may be a real effect or it may be caused by interaction with the other covariates or it may be an artefact arising from the method of adjustment used to introduce a secular trend in the underlying standard mortality rates (i.e. $\lambda^*(\cdot)$). The trend with policy duration also peaks (at 15-19 yrs) and then falls. Here, the low levels of the ratios for durations under 5 years may indicate the effectiveness of the underwriting process practised by this insurance company and the rise in the ratios as duration elapses may reflect the success of this underwriting process in eliminating lives that would have contributed deaths to the early durations.

Mortality ratios for each individual medical category are presented in Table 4.4. These ratios indicate that the mortality experience of lives with a long history of peptic ulcer ($H=2$) and either no complication or only a history of haematemesis ($I=1$ or 2) is broadly close to that of the unimpaired lives (the ratios in Table 4.4 lie between 1.01 and 1.17). The long history cases with perforation or with other complications show higher levels of mortality ratio. The short history cases show the same broad features. It is worth noting that the short history cases with no complication or with only a history of haematemesis seem to have higher ratios than their long history counterparts. As noted earlier the no complication category for the short history cases may be misleading.

Comparison with the conventionally computed mortality ratios published by previous authors from this data set is difficult because (a) ratios were only calculated and published for certain of the cells in Table 4.4 and (b) those ratios that were published refer to particular calendar years of death.

As an example of the comparison we present below figures available in relation to the short history cases for four of the eight cells:
Years of death included | Preston and Clarke 1947-63 | Clarke 1964-73 | Leighton 1974-83 | Haberman and Renshaw 1947-81
--- | --- | --- | --- | ---
Short history, no operation, no complication. | 0.92 | 1.22 | 1.19 | 1.25
Short history, no operation, history of haematemesis. | 1.30 | 1.07 | 1.17
Short history, with operation, no complication. | 1.07 | 0.85 | 1.08 | 1.65
Short history, with operation, history of perforation. | 1.49 | 1.20 | 1.50
We now investigate the effect of interactions more fully. To do so, we first condense the number of categories associated with some of the individual covariates mainly to ensure realistic numbers of cases in the resulting crossed classifications. Consider, therefore, modified covariate categories as follows-

medical status

- H - history, 2 levels: short/long (no change)
- P - operation, 2 levels: without/with (no change)
- I - index of complications, 2 levels: without/with (merge 2,3,4)

policy status

- A - age at entry, 3 levels: 15-29, 40-49, 50 and over (no change)
- D - duration, 5 levels: 0-4, 5-9, 10-14, 15-19, 20 and over (merge 1,2).

This will have the obvious effect of masking some of the finer detail already discussed particularly in relation to the index of complications and the appreciable drop in excess mortality recorded in the final five year period in comparison with all other calendar periods, for whatever reason. Earlier analysis suggests that there is little difference between the first two levels of the duration covariate so that their combining into the 0-4 yrs category seems justified. Now, Table 4.4, showing mortality ratios by medical status, telescopes into Table 4.5.

The main features here seem to be the higher ratios when the history is short (except with operation and without complications) and when complications are present. Both features are in accord with commonsense.

To simplify the presentation we define a covariate M (medical status) which captures the different levels of the three covariates H, P and I. Formally, we introduce $M = H \times P \times I$ which now takes eight-levels.

We can judge the significance of the main effects of medical status, age at entry, calendar year at entry and policy duration by referring to Figure 4.3. This diagram plays the same role as Table 4.2 where this time we have presented the differences in model deviances and their degrees of freedom on the branches of a lattice. The only substantive difference in our conclusions is to confirm our suspicion that the merging of the last two calendar year categories is to render the main effects of this covariate even less significant than before, while we still need to retain the main effects of the three (five condensed to three: A,D,M) remaining covariates.

Estimates of the excess mortality multipliers based on the main effects model $M+A+D$ are presented in Table 4.6. (Readers should recall that the model is additive on the log scale). To estimate the excess mortality ratio for a specific medical status, specific
age at entry and specific duration (any calendar year category), multiply the appropriate entries selected from this table. Notice that the pattern of entries in Table 4.6 is similar to that to be found in Table 4.4 and 4.5 and the relevant two parts of Table 4.3.

Specifically, the downward trend with age at entry, the trend with duration which peaks at 15-19 yrs are repeated as are the relationships between the three medical covariates.

Figure 4.4 shows the results of testing for interactions between pairs of covariates. The GLIM notation M*A should be interpreted as representing M+M+A i.e. additive main effects and an interaction term. Interaction between age at entry and duration is the most dominant term (the observed significance level is marginal since $P(X^2 > 15.7) = 5\%$) which we investigate further by quoting estimates for the excess mortality multipliers based on the appropriate interactive model M+A*O in Table 4.7. Again, to compute the estimated excess mortality ratio for a specific combination of medical status, duration and age at entry multiply the relevant entries selected from this table.

Table 4.7 has the same features for medical status as Table 4.6. The principal difference concerns age at entry and duration. The duration profiles for the three ages at entry are now markedly different:

for the youngest age group (15-39), the ratios peak at duration 5-9 yrs;

for age group 40-49, the ratios peak at 15-19 yrs;

for age group 50 and over, the ratios increase with duration.

For each duration group, (except for 20 yrs and over), the three age at entry subgroups show the same features as Table 4.6.

We draw attention to the particularly high mortality ratio of 1.83 for age at entry 15-39 and duration 5-9 yrs.

4.4 Comparison with other Studies

A number of other studies have reported on the mortality experience of life insurance policyholders with peptic ulcer disease and these are reviewed in detail by Singer and Levinson.$^{17}$

The two principal other studies (both US based) are the Society of Actuaries 1951 Impairments Study$^{18}$ (based on experience over the period 1935 to 1950 and now updated by the soon to be published 1983 Impairments Study) and the analysis by Pauley of the Prudential of America experience over the period 1952-1965$^{19}$. The main features from these studies noted by Singer and Levinson are:-
(i) the secular downward trend in mortality;

(ii) the decline in mortality ratios with increasing age at entry;

(iii) the higher relative mortality of gastric ulcer cases than duodenal ulcer cases;

(iv) the contradictory evidence regarding the trend in mortality ratios with policy duration: for example the 1951 Impairments Study\textsuperscript{18} reports a downward trend for duodenal and gastric cases separately (operated and unoperated) while Pauley\textsuperscript{19} reports an upward trend for duodenal cases.

Brackenridge\textsuperscript{20}, in his review, comments that the risk for most medically treated ulcers is low and "decreases with time".

Some of these results are confirmed by our analysis - in particular, (ii) above. The secular downward trend is not apparent from the lack of significance of the calendar year of entry variable but the secular trend is controlled for through our use of a standard hazard rate $\lambda(t)$ which decreases with time. The data we have available do not permit a separate consideration of gastric and duodenal cases. Regarding (iv), we have identified (Table 4.7) different durational trends for three age groups - other studies have not been able to identify this interaction between age at entry and duration.

4.5 Final Word

The principal advantages of this approach to analysing mortality over the traditional approach (used for example by Preston and Clarke, Clarke and Leighton) are that, rather than merely presenting results in the form of Table 4.4, (a) we are able to present a model to represent the variation in excess mortality with all the covariates present simultaneously; (b) we are able to test whether certain components are contributing to the model to a significant extent; and (c) we are able to investigate complex interactions between covariates as well as models with only main effects contributions.
FIGURE 4.1 Hypothetical Example: Mortality and Handedness
Peptic ulcer cases 1947-81 subdivided by H(history), P(operation), I(index of complications), A(age at entry), C(calendar year of entry).

Table 4.1

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>Degrees of Freedom</th>
<th>Differences</th>
<th>Observed significance level</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$: null</td>
<td>1541.5</td>
<td>1714</td>
<td>0</td>
<td>0.5%</td>
</tr>
<tr>
<td>H: history</td>
<td>1533.6</td>
<td>1713</td>
<td>7.9</td>
<td>0.1%</td>
</tr>
<tr>
<td>P: operation</td>
<td>1531.4</td>
<td>1713</td>
<td>10.1</td>
<td>&lt;&lt; 0.1%</td>
</tr>
<tr>
<td>I: index of complications</td>
<td>1518.6</td>
<td>1711</td>
<td>22.9</td>
<td>&lt;&lt;&lt; 0.1%</td>
</tr>
<tr>
<td>A: age at entry</td>
<td>1483.4</td>
<td>1712</td>
<td>58.1</td>
<td>36%</td>
</tr>
<tr>
<td>C: calendar year of entry</td>
<td>1534.9</td>
<td>1708</td>
<td>6.6</td>
<td>2.5%</td>
</tr>
<tr>
<td>D: policy duration</td>
<td>1528.9</td>
<td>1709</td>
<td>12.6</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2
Mortality ratios, main effects fitted separately. Peptic ulcers 1947-81.

Table 4.3
<table>
<thead>
<tr>
<th>Short history</th>
<th>Long history</th>
<th>No operation</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apparently acute</td>
<td></td>
<td>1.25 (330)</td>
<td>1.08 (61)</td>
</tr>
<tr>
<td></td>
<td>perforation</td>
<td>1.17* (89)</td>
<td>1.71* (13)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.40* (5)</td>
<td>1.50 (119)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.01 (472)</td>
<td>1.15 (525)</td>
</tr>
<tr>
<td>Apparently chronic</td>
<td></td>
<td>1.07 (734)</td>
<td>1.17 (525)</td>
</tr>
<tr>
<td></td>
<td>perforation</td>
<td>1.09 (109)</td>
<td>1.65* (11)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.43 (149)</td>
<td>1.50 (119)</td>
</tr>
</tbody>
</table>

Table 4.4


*based on 232 cases or less
(Number of deaths in brackets.)

*haematemesis

index of complications

with operation
<table>
<thead>
<tr>
<th></th>
<th>Complications</th>
<th>without</th>
<th>with</th>
</tr>
</thead>
<tbody>
<tr>
<td>short history</td>
<td>no operation</td>
<td>1.25</td>
<td>1.21</td>
</tr>
<tr>
<td>apparently acute</td>
<td>with operation</td>
<td>1.08</td>
<td>1.51</td>
</tr>
<tr>
<td>long history</td>
<td>no operation</td>
<td>1.01</td>
<td>1.14</td>
</tr>
<tr>
<td>apparently chronic</td>
<td>with operation</td>
<td>1.17</td>
<td>1.37</td>
</tr>
</tbody>
</table>

Mortality ratios by (modified) medical status. Peptic ulcers 1947-81.

Table 4.5
null hypothesis

Lattice of hypotheses, testing for main effects. Peptic ulcers 1947-81.

Figure 4.3
<table>
<thead>
<tr>
<th>age at entry</th>
<th>15-39</th>
<th>40-49</th>
<th>50 and over</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor</td>
<td>1</td>
<td>0.89</td>
<td>0.67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>duration</th>
<th>0-4</th>
<th>5-9</th>
<th>10-14</th>
<th>15-19</th>
<th>20 and over</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor</td>
<td>1</td>
<td>1.12</td>
<td>1.11</td>
<td>1.17</td>
<td>1.07</td>
</tr>
</tbody>
</table>

### Medical Status

<table>
<thead>
<tr>
<th>history</th>
<th>operation</th>
<th>complication</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>1.37</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>1.24</td>
</tr>
<tr>
<td>short</td>
<td>no</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>1.35</td>
</tr>
<tr>
<td>long</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td></td>
</tr>
</tbody>
</table>


Table 4.6
Lattice of hypotheses, testing for interactions.
Peptic ulcers 1947-81.

Figure 4.4
### Table 4.7

<table>
<thead>
<tr>
<th>Age at Entry</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-4</td>
</tr>
<tr>
<td>15-39</td>
<td>1.39</td>
</tr>
<tr>
<td>40-49</td>
<td>1.31</td>
</tr>
<tr>
<td>50 and over</td>
<td>0.85</td>
</tr>
</tbody>
</table>

**Medical Status**

<table>
<thead>
<tr>
<th>History</th>
<th>Operation</th>
<th>Complication</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>short</td>
<td>no</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>0.91</td>
</tr>
<tr>
<td>long</td>
<td>no</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Excess mortality ratio estimates, main interaction effects. 
Model M+A*D.
Peptic ulcers 1947-81.
5. MODELLING INSURANCE LAPSE RATES

5.1 Introduction

We recently published a full, statistical analysis of the lapse data generously supplied by the Faculty of Actuaries Withdrawals Research Group. These data cover the lapse or withdrawal experience for 1976 of seven Scottish life insurance companies. An extensive study had been published by this Research Group although, for reasons given below, we believe that our approach based on linear models is better able to describe the structure of the data than the detailed tabulations and cross-tabulations of this earlier paper.

The data enable the experience of 1976 to be investigated with particular reference to the variation of lapse or withdrawal rates with various policy characteristics. The expression "lapse" is used throughout to denote the removing of a policy from the live file, due to premature termination of the contract, with or without payment of a surrender value. It excludes the conversion of a policy to a paid-up amount, the reduction of premium and/or sum assured or the surrendering of bonuses.

The characteristics are summarised in Table 5.1 together with the categories into which each has been divided. The total exposed to risk is in excess of 750,000. As noted in Table 5.1 there are some missing data.

The Report of the Faculty of Actuaries Withdrawals Research Group published in 1978 presented the data for 1976 in a factual way, without attempting to set up any theoretical models.

Data of this particular type are described by statisticians as categorical. The authors of the 1978 Report identified nine characteristics with which the withdrawal rate may be expected to vary. Each of these characteristics has been divided into a number of discrete categories. Merely to present the data in a complete way would require a nine-dimensional tabulation which, of course, is not practicable. To present the possible two-way marginal tables would require 36 such tables - the authors of the 1978 report have shown results for only eleven of these. More complex interactions were not investigated further.

The use of theoretical models for such a data set has the advantage of providing a structure to the data in order to improve the estimation of underlying parameter values. Further, statistical theory enables different models to be compared and contrasted so that conclusions about the structure of the data may be reached. The fitting of the models described in a subsequent section can be regarded as analogous to parametric graduation in that the crude data are "smoothed" in order to satisfy an assumed relationship.
The analyses published in 1978 indicated that four characteristics contribute "significantly" to the variation in lapse rates viz. office, type of policy, age at entry, duration of policy. As described above, no formal statistical investigations were undertaken to qualify the term "significant". We shall view these results as arising from a preliminary examination of the data and shall take the identification of these four factors as the starting point for our analyses. The categorisation of these four factors is shown in Table 5.1. Regarding policy type, because the open-ended endowments and unit-linked policies in the investigation (1976) were mainly of short duration with little or no data beyond eight years' duration, it was decided to exclude these two types and concentrate on the remaining five. The temporary insurance class includes family income benefits, reducing and level temporary insurances as well as those with conversion options - this is, therefore a heterogeneous group of policies. Where a single policy combined a basic type of insurance and some type of temporary assurance, the Research Group considered the policy as one of the appropriate basic type and ignored the temporary insurance portion. Altered policies were grouped by their current policy class in the investigation.

5.2 The Data

The raw data were edited and the way in which policy lapses, the response, varied with the following covariates was investigated:

A - age at entry, 3 categories
- i=1: early (15 to 29 yrs)
- i=2: medium (30 to 39 yrs)
- i=3: late (40 to 64 yrs)

D - duration of policy, 3 categories
- j=1: short (1 to 3 yrs)
- j=2: medium (4 to 8 yrs)
- j=3: long (9 or more yrs)

F - office, there are 7 denoted by k=1,2, ..., 7

T - type of policy, 5 categories
- l=1: with profit
- l=2: non profit endowment
- l=3: with profit whole-life
- l=4: non profit
- l=5: temporary

The cross-classification of covariates gives rise to a set of cells or units {u = (i,j,k,l)}. The numbers of lapses w_u, out of n_u exposures, for different u, are available for analysis. The data were not quite balanced in the sense that no temporary policies of long duration were recorded by office number seven*, giving rise to

Footnote
[* offices 6 & 7 were renumbered 7 & 6].
a total of $N = 3^2 \times 7 \times 5 - 3 = 312$ units or non-empty cells. The choice of categories for covariates A and D is, to some extent, arbitrary, and could have been adjusted by editing the raw data differently.

5.3 Results

In cell $u$, let $w_u$ be the observed number of lapses and $n_u$ be the observed exposed-to-risk.

Attempts were made to fit a variety of model structures to the following response variables firstly using independent normal homoscedastic errors:

(i) the annual lapse rate, $w_u/n_u$;

(ii) the lapse frequency, $n_u/w_u$;

(iii) the log odds of lapsing, $\log_e \left( \frac{w_u}{n_u - w_u} \right)$.

The first two choices of response variable failed to produce satisfactory residual plots when fitted for a variety of model structures. Residual plots for the log odds response variable (choice (iii)) with an additive, main effects structure and normal homoscedastic error structure led to satisfactory plots, subject to a few outliers. These plots offer reasonable supporting evidence for the model. Summary details of the initially adopted overall model therefore are:

Data: $(w_u, n_u)$

Covariates: $A, D, F, T$ with $u = (i, j, k, l)$

Decomposition: $y_u = m_u + \varepsilon_u$

Response: $y_u = \log \left( \frac{w_u}{n_u - w_u} \right)$ *

Errors: $\varepsilon_u \sim N(0, \sigma^2)$ with independence

Structure: $M: m_u = \alpha_i + \beta_j + \gamma_k + \delta_l$ (parametric)

or $M = A + D + F + T$ (GLIM notation).

For the parametric form of the additive (no-interaction) model structure, parameters $\alpha, \beta, \gamma, \delta$ naturally relate to the covariates $A, D, F$ and $T$ respectively.

Footnote

We remark here that, since the lapse rates encountered are of the general order of 1 in 20, lapse odds are effectively the same as lapse rates i.e.

$$\text{lapse odds} = \frac{w_u}{n_u} \frac{u}{1 - w_u / n_u} = \frac{w_u}{n_u}.$$
The model has the advantage of a particularly simple and readily interpreted structure. Firstly, however, we enquire whether certain (fine tuning) adjustments to the model structure M are in order before interpretation begins.

Specifically, are all four covariates statistically significant or will an even simpler model suffice?

And is there any significant interaction between the covariates which matters?

To answer the first of these questions, each of the four factors was omitted in rotation, and their significance formally assessed using familiar F-tests. These are valid because of the assumed error structure which was not, in turn unsupported by the residual plots. Test details are displayed on a lattice of hypotheses (Figure 5.1) in which the nodes represent the different model structures (written in GLIM notation). Residual sums of squares or deviances and the associated degrees of freedom are displayed at each node. Departure sums of squares and the associated degrees of freedom, obtained by differencing, are displayed on the branches of the lattice. Details of the F-tests are then tabulated along-side the lattice. The tests clearly demonstrate that each of the four factors is highly significant and suggests what might be called a "pecking order of significance".

To answer the second of the above questions, we consider formal significance F-tests for the interaction between the various covariates. These are summarised in the same way (Figure 5.2). Here for example, the model A* D + F + T, which has the parametric representation

\[ m_u = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + (\alpha \beta)_{ij} \]

is composed of the main effects terms plus an interaction term between covariates A and D. (Note: In GLIM notation A*D=A+D+A.D so that the interaction term is A.D). The F-ratios for all three interaction terms involving the covariate F have been calculated and are shown in Figure 5.2 - all three are clearly non-significant. Of the remaining three interaction terms, two have calculated F-ratios which border on the upper 5% level, while the interaction term D.T is clearly the most significant, overwhelmingly so. Thus, we might reasonably conclude that the essential features of the data set are encapsulated within the model structure

\[ A + D + F + T + D.T \]

with its parametric representation

\[ m_u = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + (\beta \delta)_{ij} \].
Fitted values (using GLIM)

\[ \hat{m}_u = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_k + \hat{\delta}_l + (\hat{\beta}\hat{\delta})_{lj} \]

are based on the maximum likelihood (or least squares) estimators shown in Table 5.2.

Examination of the differences (or contrasts) in estimated covariate levels leads us to draw the following conclusions:

1. As already stated, the office effects, while statistically highly significant, are additive by nature. The evidence for this lies with the appropriate F-tests of Figure 5.2. Thus, we conclude that, essentially, all offices experience a similar pattern of lapses across the different combined levels of the other factors under investigation, but to varying degrees of intensity. Contrasting the (non-unique) estimators \( (\hat{\gamma}_k) \) indicates that offices 1, 2 and 6 experience (near) identical intensities of lapses across the board, with the remaining four offices experiencing somewhat lower intensities of lapses, to varying degrees. A similar conclusion was reached by the Faculty Research Group. This finding raises the issue of whether these systematic differences are "real" or are perhaps rather a function of the way in which the data were selected and recorded from office to office. "Real" reasons for variations between offices might be, for example, the varying quality of after-sales service, the results of different marketing strategies or the varying generosity in the level of surrender values.

2. The "pecking order of significance" mentioned above and displayed, for example, in Figure 5.1 indicates that, of the four factors being considered in these models, Office is the least significant and Duration is the most significant. These comments contrast with the Faculty Research Group who concluded that Policy Type was the most significant factor (page 277 of reference 21), albeit with no formal scientific validation of this statement. Our findings do, however, agree with the earlier paper on the role of office. Indeed the Faculty Research Group amalgamated the data for all offices to produce lapse rates by policy type. It is true that the combined data are likely to give a better picture of the market place created by the various types of intermediary. But here we emphasise the inter-office differences in order to indicate the extent of variation that might be anticipated between offices. This is pursued further in Renshaw and Haberman where a more detailed analysis is carried out for one office.

3. The interaction between age at entry and both policy duration and type is only marginally significant (Figure 5.2). Thus, on contrasting the estimates \( (\hat{\alpha}_i) \), we might reasonably conclude that lapse rates decrease with increasing age at entry without undue
interaction. Of course, if required, the nature of this marginally significant interaction can be ascertained by fitting the desired interaction terms. Details are not reproduced since they did not reveal any pronounced departures from this conclusion.

4. The main interaction lies with policy duration and type (Figure 5.2). Examination of the entries $\beta_j + \delta_1 + (\beta \delta)_{j1}$ (Table 5.2) leads us to conclude that -

(i) There is a marked reduction in lapses for all types of policies of long duration.

(ii) Lapses are markedly higher for non-profit policies than for the corresponding with-profit policies at each individual level of duration.

(iii) The two with-profit policy types show almost identical patterns of reducing lapses with increasing duration. The pattern for endowment policies is pitched at a slightly lower level than for whole-life policies.

(iv) While the non-profit whole-life policies maintain the decreasing pattern of lapses with increasing duration, this trend is partially reversed for non-profit endowment policies. Here, lapses show a small increase for policies of medium duration over those of short duration. This is clearly the main source of interaction between the two covariates. We understand that an explanation for this effect might lie with the practice at this time of using non-profit endowment policies to secure mortgages. The average length of a mortgage (i.e., before the owners move to another home) is about seven years - a duration of seven years falls into the second category (j=2) of the D variable.

Our fuller paper also reported investigations based on a binary response model and which looked at the lapse experience of a single company7.
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Categories</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>OFFICE</td>
<td>Seven</td>
<td>----</td>
</tr>
<tr>
<td>AGE AT ENTRY</td>
<td>15-19, 20-4, 25-9, 30-4, 40-4, 45-54, 55-64</td>
<td>Definition is Calendar yr of entry - office yr of birth</td>
</tr>
<tr>
<td>DURATION OF POLICY</td>
<td>0, 1, 2, 3, 4, 5, 6-8, 9-11, 12-14, 15 and over (years)</td>
<td>Definition is Calendar yr of investigation - Calendar yr of entry</td>
</tr>
<tr>
<td>SEX OF POLICYHOLDER</td>
<td>Male/Female</td>
<td>Only 6 offices able to provide this split</td>
</tr>
<tr>
<td>POLICY TYPE</td>
<td>With profit endt. non profit endt. with profit whole life, non profit whole life, temporary, open ended endt. unit linked endt.</td>
<td>Open ended and unit-linked endts are of short duration - little or no data beyond 8 yrs duration.</td>
</tr>
<tr>
<td>ORIGINAL PREMIUM-PAYING TERM</td>
<td>Under 10 yrs, 10-14, 15-19, 20-29 and over 30 yrs.</td>
<td>Only 6 offices able to provide this split. Classification only appropriate if premium still being paid.</td>
</tr>
<tr>
<td>SUM INSURED</td>
<td>£0-999, 1000-1999, 2000-4999, 5000-9999, 10,000-19999, over 20,000.</td>
<td>Bonuses excluded. For decreasing temporary insurances, original sum insured used.</td>
</tr>
<tr>
<td>PREMIUM FREQUENCY</td>
<td>Yearly, monthly, other and paid up</td>
<td>-----</td>
</tr>
<tr>
<td>AGENT TYPE</td>
<td>Broker, Chartered Accountant, Solicitor, Estate Agent, Bank, Building Society, Own Staff, Other Agent, No Agent.</td>
<td>Only 1 office able to provide this split.</td>
</tr>
</tbody>
</table>

TABLE 5.1
**Figure 5.1**

Lattice of hypothesis, testing for main effects.
Lapse data, 1976.

(*** denotes observed significance level is less than 0.1%)
Figure 5.2

Lattice of hypothesis, testing for interactions.

Lapse data, 1976.

(* denotes observed significance level is less than 5%)
\( \hat{\mu} = -2.77 \)

A: \( \hat{\alpha}_i \)  
\[
\begin{array}{cccc}
0 & -0.28 & -0.54 \\
\end{array}
\]

F: \( \hat{\gamma}_k \)  
\[
\begin{array}{ccccccc}
0 & 0.02 & -0.20 & -0.34 & -0.17 & 0.05 & -0.42 \\
\end{array}
\]

T: \( \varepsilon \)

D: \( j + \)  
\[
\begin{array}{cccccc}
0 & 0.53 & 0.11 & 0.93 & 0.40 \\
-0.27 & 0.61 & -0.01 & 0.49 & 0.33 \\
-1.04 & -0.13 & -0.93 & -0.27 & -0.12 \\
\end{array}
\]

\( \hat{\delta}_j + \hat{\delta}_k + (\hat{\delta}_j) \varepsilon \)

**TABLE 5.2**

Maximum Likelihood Estimators for Model, A+D+F+T+D*T.  
Lapse Data, 1976.
6. CONCLUSIONS

We have attempted to demonstrate the versatility of generalised linear models and the statistical package GLIM for tackling a range of modelling problems. Three different types of problem based on data from different insurance related areas have been discussed.

Section 3 is concerned with the fitting of skewed loss distributions. We have indicated how to adapt the GLIM package to fit certain types of distribution which are not immediately available within the system. Three examples are the Pareto, Burr and Weibull distributions. It is a trivial matter to fit other loss distributions such as the log normal, gamma and log gamma to uncomplicated data.

In sections 4 and 5 we deal with generalised linear models and large complex data sets. Modelling such large complex data sets may be viewed as a balancing act between model complexity and the need to encapsulate the salient underlying features present in the data. The simpler the model, the simpler the interpretation of the underlying data generating mechanism. Modelling does not necessarily have a unique solution, but a model may be deemed adequate only if it achieves this goal.

One way of assessing the adequacy is through a thorough graphical analysis of model residuals which, ideally, should be "pattern free" (section 5). Additionally, what might be termed "fine tuning" might then be attempted, and its effects formally assessed. The development of generalised linear modelling together with its associated computer soft-ware package GLIM, facilitates such modelling objectives.

The applications discussed in Section 4 are particularly novel and we believe that the GLIM-based 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 than has hitherto been attempted in this field, in which the effects of individual factors and their interactions on excess mortality may be assessed. We would highlight the meagre assumptions on which the models are based, the comparative ease with which they can be fitted and compared using GLIM and the appealing connection which these models have with the traditional actuarial standard mortality ratios. We plan further work on these lines for other impairments and are investigating further the influence of the specific base-line hazard function used in the analysis and its construction. We also seek to examine in detail the relevant residual plots.
The results discussed in Section 5 indicate how such methods may be applied to a wide range of problems where an actuarial rate or probability or one year risk premium may be represented as a function of a set of significant rating factors. Section 5 deals specifically with lapse rates but earlier authors have examined the components of motor insurance risk premiums (i.e. claim frequency and average claim cost) and we believe that there is considerable scope for extending these ideas into other areas e.g. group life risk premiums or marine insurance risk premiums.
7. ACKNOWLEDGEMENTS

We would like to acknowledge the unselfish assistance provided by Della Bloomfield in carrying out much of the analysis for Section 4 of this paper. Further, we would like to acknowledge the generosity of the Prudential insurance company in making the impaired lives data available to us (Section 4) and the Faculty of Actuaries in making the lapse data available to us (Section 5).
8. REFERENCES


