STATISTICAL ANALYSIS OF LIFE ASSURANCE LAPSES

BY A. E. RENSHAW B.SC., PH.D. AND
PROFESSOR S. HABERMAN M.A., PH.D., F.I.A.

(of The City University, London)

1. INTRODUCTION

DATA have been generously supplied by the Faculty of Actuaries Withdrawals Research Group. These data cover the lapse or withdrawal experience for the calendar year 1976 of seven Scottish life offices. An extensive analysis has been published by the Research Group[1] although, for reasons we shall discuss later, we believe that the approach outlined here is better able to describe the structure of the data than the detailed (and somewhat pedestrian) 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 bonus.

The characteristics are summarized in Table 1 together with the categories into which each has been divided. Further information is provided in the Research Group's report. The total exposed to risk is in excess of 750,000. As noted in Table 1 there are some missing data.

The calendar year method of defining duration, rather than the more natural policy year method, was adopted because some of the offices could not have provided the data in the required form. With a calendar year rate interval, half a year's exposure was counted for each case at duration 0. In interpreting the results against policy duration, a possible source of bias should be noted which arises from the offices' widely varying practice regarding the retention of business on the live file after non-payment of premiums. Thus some offices included no-payment cases on the live file and counted them as withdrawals while other offices ignored such cases. Considerable variations in the rates for early durations (particularly durations 0 and 1) exist which would probably not have occurred if it had been possible to use complete years' premiums paid. This practical problem attracted some comment in the Faculty discussion referred to above.

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

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

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 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 so as 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 analagous to parametric graduation in that estimates from the data are smoothed so as to satisfy an assumed relationship.

Models such as those described in the subsequent sections have been used to describe claim rates in, for example, general insurance but not in life assurance. References from the United Kingdom actuarial literature include Johnson and Hey,(2) Grimes,(3) Bennett(4) and Coutts.(5)

We take advantage of the GLIM statistical package in carrying out the modelling required. Details are provided in §§3, 4 and 5.

The analyses published in 1978 indicate 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 categorization of these four factors is shown in Table 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 assurance class includes family income benefits, reducing and level temporary assurances as well as those with conversion options—this is, therefore, a heterogeneous group of policies. Where a single policy combined a basic type of assurance and some type of temporary assurance, the Research Group considered the policy as one of the appropriate basic type and ignored the temporary assurance portion. Altered policies were grouped by their current policy class in the investigation.

Two further points should be noted about these data.

Firstly, if we were considering attempting to identify, at the inception of a group of policies, which types of policy were more or less likely to be withdrawn, we would be concerned with a wider set of factors many of which were not recorded in this particular investigation. Thus, from this viewpoint of 'underwriting for withdrawal' we might be interested in policyholder's income, area of residence, occupation, tenure (including home ownership), number of years at

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

Table 1. *Withdrawals Data: Policy Characteristics*

Characteristic	Categories	Comments
Office	Seven	—
Age at entry	15–19, 20–24, 25–29, 30–34, 40–44 45–54, 55–64	Definition is Calendar year of entry—office year of birth
Duration of policy	0, 1, 2, 3, 4, 5, 6–8, 9–11, 12–14, 15 and over (years)	Definition is Calendar year of investigation—Calendar year of entry.
Sex of policyholder	Male/Female	Only 6 offices able to provide this split.
Policy type	With-profit endt, non-profit endt, with-profit whole-life, non-profit whole-life, temporary, open ended endt, unit linked endt.	Open ended and unit-linked endts are of short duration—little or no data beyond 8 years' duration.
Original premium-paying term	Under 10 years, 10–14, 15–19, 20–29 and over 29 years.	Only 6 offices able to provide this split. Classification only appropriate if premium still being paid.
Sum assured	£0–£999, £1,000–£1,999, £2,000–£4,999, £5,000–£9,999, £10,000–£19,999, over £20,000.	Bonuses excluded. For decreasing temporary assurances, original sum assured used.
Premium frequency	Yearly, monthly, other and paid up	—
Agent type	Broker, Chartered Accountant, Solicitor, Estate Agent, Bank, Building Society, Own Staff, Other Agent, No Agent.	Only 1 office able to provide this split.

that address, reason for effecting the policy. In this sense the study is rather restricted. Further, it would also be of value to know the reason for withdrawal, although this is likely to be difficult to ascertain.

Secondly, the investigation is a cross-sectional one in the classical, actuarial sense. Such investigations have been widely discussed in the actuarial literature (for example, Benjamin and Pollard).[6] The withdrawal of a policy is unlike a death claim in that it is voluntary. Hence groups of policies with given characteristics may vary both in their overall propensity to withdraw and in the timing of these withdrawals (i.e., the distribution over time, or by policy duration). This element of volition means that cross-sectional investigations are deficient in attempting to describe such phenomena. A cohort approach is more natural and more satisfactory. The situation is similar to that in demography where, although a cross-sectional approach may be adequate for mortality investigations, it is unsatisfactory in attempting to describe phenomena like fertility, first marriage, remarriage and divorce. Indeed in these cases such an approach can often lead to fallacious conclusions (Cox).[7] The same is true for withdrawals. This deficiency here should be borne in mind throughout this

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

report—strangely, no comments were made along these lines in the Faculty paper nor in the subsequent discussion.[1] We shall return to this point in §6.

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 $$\begin{cases} i=1\text{: early (15 to 29 yrs)} \\ i=2\text{: medium (30 to 39 yrs)} \\ i=3\text{: late (40 to 64 yrs)} \end{cases}$$

D—duration of policy; 3 categories $$\begin{cases} j=1\text{: short (1 to 3 yrs)} \\ j=2\text{: medium (4 to 8 yrs)} \\ j=3\text{: long (9 or more yrs)} \end{cases}$$

F—office, these are 7 denoted by $k=1, 2, \ldots\ldots, 7$*

T—type of policy; 5 categories $$\begin{cases} \left.\begin{array}{l} l=1\text{: with-profit} \\ l=2\text{: non-profit} \end{array}\right\}\text{endowment} \\ \left.\begin{array}{l} l=3\text{: with-profit} \\ l=4\text{: non-profit} \end{array}\right\}\text{whole-life} \\ l=6\text{: temporary} \end{cases}$$

The cross-classification of covariates gives rise to a set of cells or units $\{u\colon u\,(i, j, k, l)\}$. 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 7*, giving rise to a total of $N=3\times3\times7\times5-3=312$ 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.

3. LINEAR MODELS WITH NORMAL ERROR STRUCTURE

A full theoretical discussion of linear models with normal error structure is included in Technical Appendix A which serves as an adjunct to this section.

As a first step in the setting up of a linear model, attention is focused on the decomposition

$$\text{(response)}=\text{(systematic component)}+\text{(error component)}$$

which may be written either as

$$y_u=m_u+\varepsilon_u$$

for each unit u, or as the vector identity

$$\mathbf{y}=\mathbf{m}+\boldsymbol{\varepsilon}.$$

* Offices 6 and 7 were renumbered 7 and 6.

The response Y_u and error component ε_u are treated as random variables and the systematic component m_u is treated deterministically.

Let $E(\varepsilon_u)=0$ so that $m_u=E(Y_u)$ for all u.

In order to employ the decomposition it is necessary to:

(i) select a suitable response variable Y_u, whose realization y_u is a function of the lapse data (w_u, n_u);
(ii) cater for the covariates by incorporating them into the nominated structure, M, of the systematic components m_u;
(iii) select an error structure, which (hopefully) has independent, homoscedastic, normally distributed components $\varepsilon_u \sim N(0, \sigma^2)$ so that $Y_u \sim N(m_u, \sigma^2)$ for all u.

The viability of any such proposed overall model can be assessed by fitting and observing residual plots. If satisfactory, other (simpler) models, obtained by specifying different structures, H, for the m_u's can be investigated by traditional hypothesis-testing theory (using F-tests for example).

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

(i) the annual lapse rate w_u/n_u;
(ii) the lapse frequency n_u/w_u; and
(iii) the log odds of lapsing* $\log(w_u/(n_u-w_u))$.

The first two choices of response variable failed to produce satisfactory residual plots when fitted for a variety of model structures. By way of illustration, the plot of residuals against fitted values for the lapse frequency response (choice (ii)) and an additive, main effects model structure (which will be discussed in detail in a subsequent paragraph) is reproduced as Figure 1. This clearly cannot be described as 'pattern free', indicating that the residuals are not independent of the fitted values as required (see Technical Appendix A). The plot also casts serious doubts on the homoscedastic error assumption. Here, since $E(\varepsilon_u)=\sigma^2$ for all u and the residuals r_u more or less behave like ε_u (see Technical Appendix A), we would expect the r_u's to lie roughly in a horizontal band, on either side of the origin, when plotted against fitted values.

Residual plots for the log odds response variable (choice (iii)) with an additive, main effects model structure, and normal homoscedastic error structure are also reproduced (see Figures 2a to 2e). While these are not completely satisfactory it can be argued that, subject to a few outliers, they do offer very reasonable supporting evidence for the model. Summary details of the initially adopted overall model therefore are:

* 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.

$$\text{lapse odds}=\frac{w_u}{n_u}\bigg/\left(1-\frac{w_u}{n_u}\right)\doteqdot\frac{w_u}{n_u}.$$

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

Data:	(w_u, n_u)
Covariates:	A, D, F, T with $u=(i, j, k, l)$
Decomposition:	$y_u = m_u + \varepsilon_u$
Response:	$y_u = \log\,(w_u/(n_u - w_u))$
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.

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

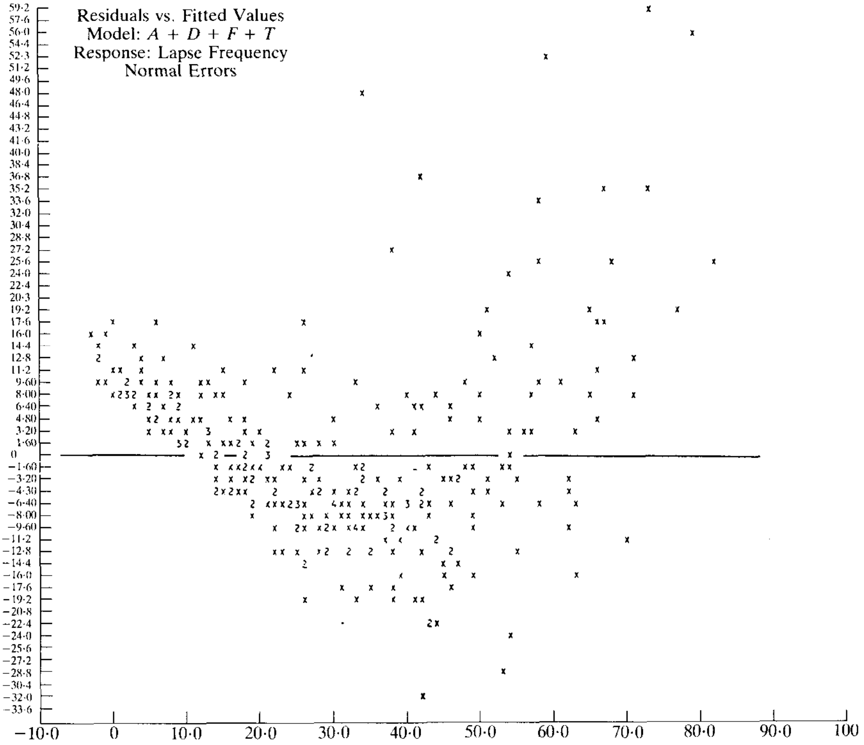


Figure 1. Plot of Residuals vs. Fitted Values for $A+D+F+T$ model of Lapse Frequency (Normal Errors).

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

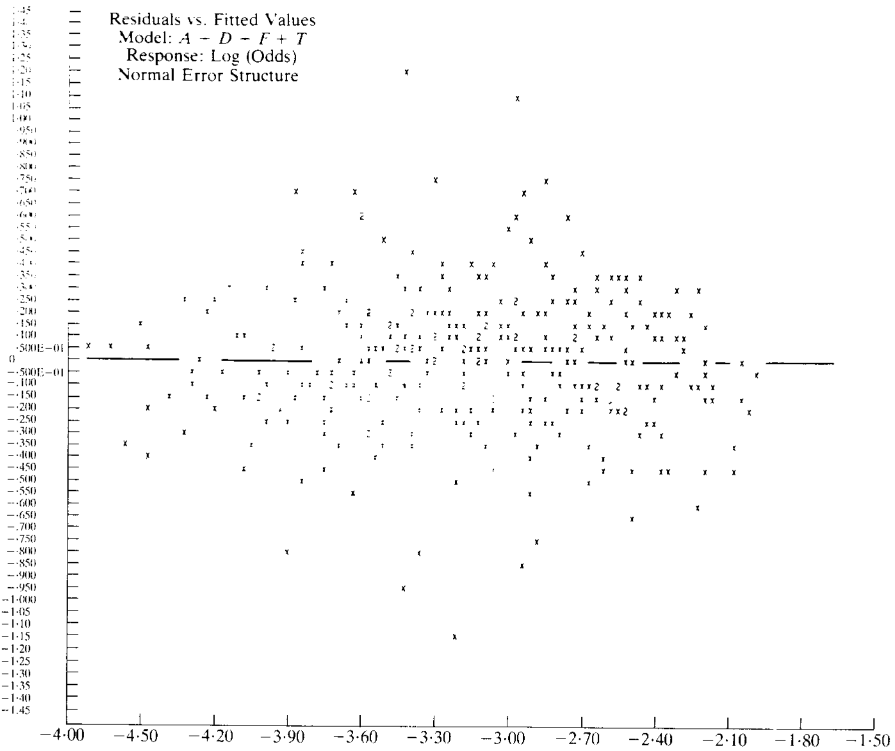


Figure 2a. Plot of Residuals vs. Fitted Values for $A+D+F+T$ model of Log Odds (Normal Errors).

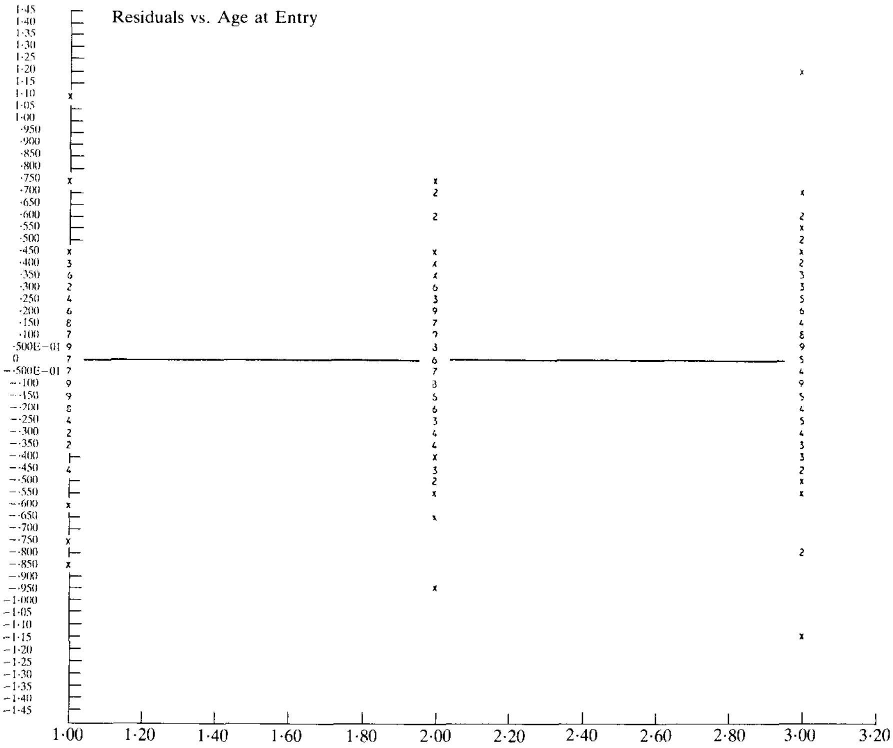


Figure 2b. Plot of Residuals vs. Age at Entry for $A+D+F+T$ model of Log Odds (Normal Errors).

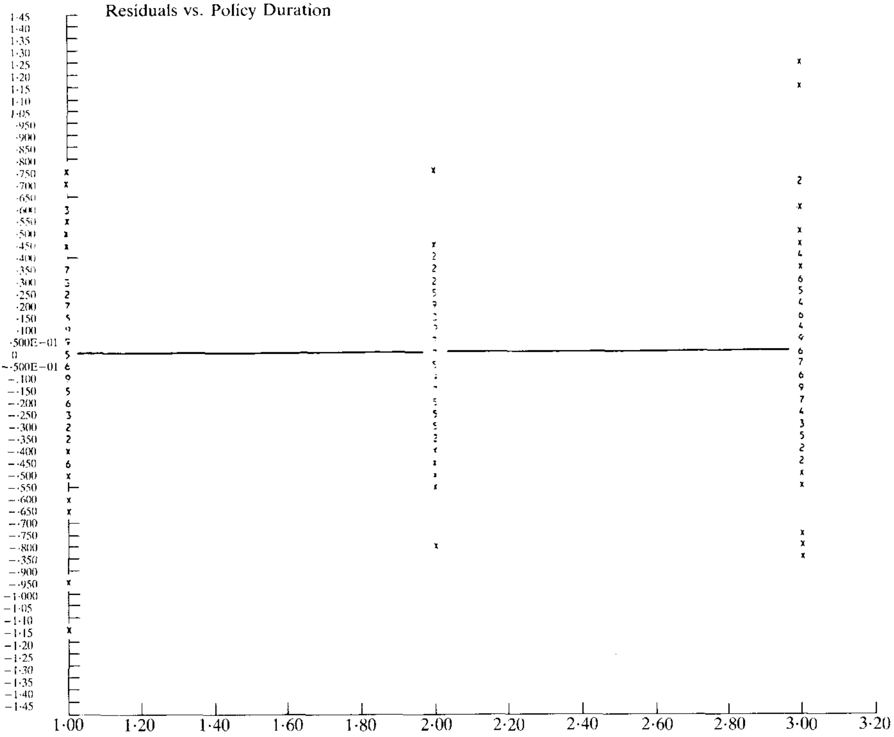


Figure 2c. Plot of Residuals vs. Policy Duration for $A+D+F+T$ model of Log Odds (Normal Errors).

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

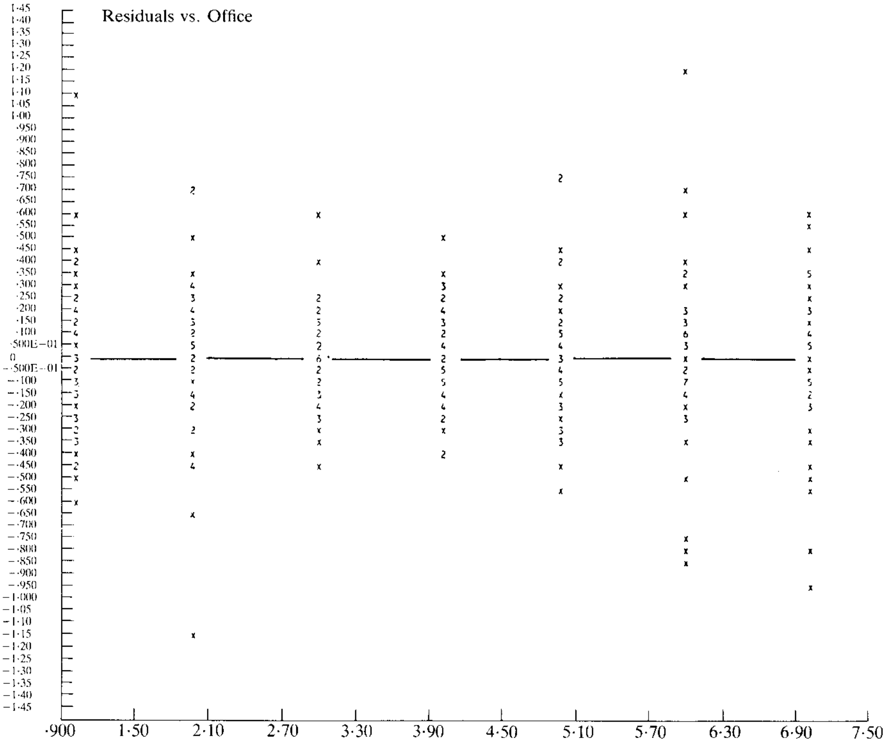


Figure 2d. Plot of Residuals vs. Office for $A+D+F+T$ model of Log Odds (Normal Errors).

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

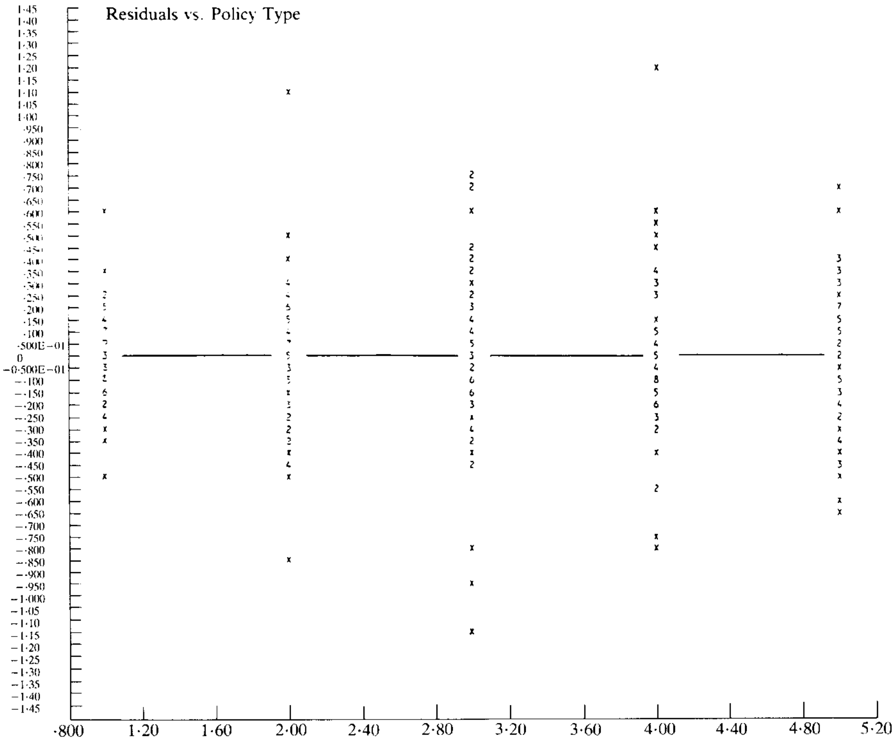
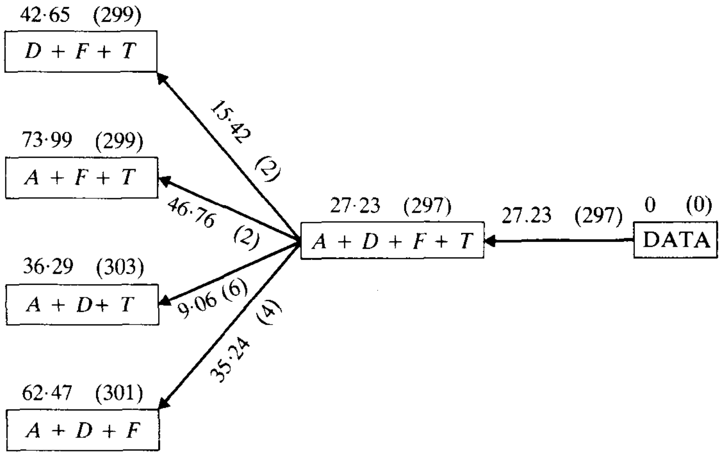


Figure 2e. Plot of Residuals vs. Policy Type for $A+D+F+T$ model of Log Odds (Normal Errors).

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 plot. Test details are displayed on a so-called lattice of hypotheses (Figure 3) 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 alongside 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’.

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press



Source of Main Effect	M.S.	F-Ratio	D.F.
A	7·71	84·0***	2,∞
D	23·38	255·0***	2,∞
F	1·51	16·5***	6,∞
T	8·81	96·0***	4,∞
Residuals	.0917		

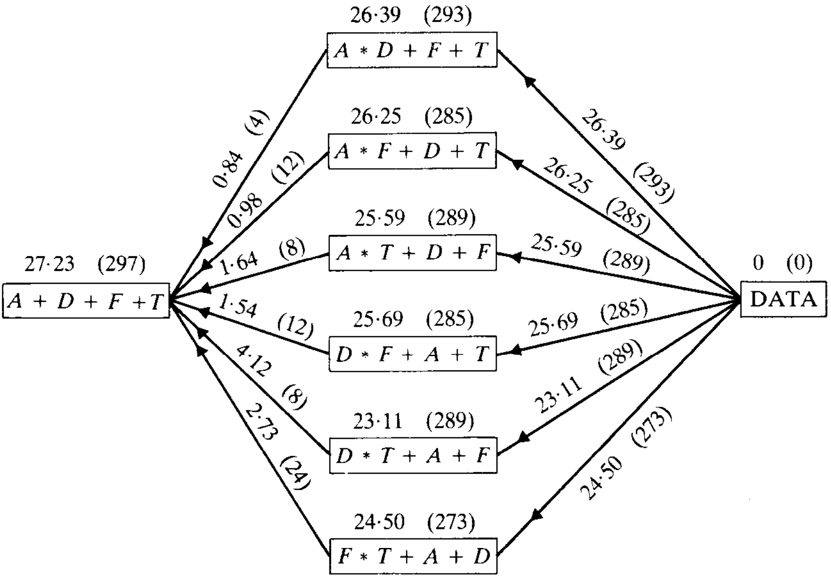
Figure 3. Tests of Significance of Main Effects in $A+D+F+T$ model of Log Odds (Normal Errors).

To answer the second of the above questions, we consider formal significance F-tests for the interaction between the various covariates. These are summarized in the same way (Figure 4). Here for example, the model $A*D+F+T$, which has the parametric representation

$$m_u = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij}$$

is composed of the main effects terms plus an interaction term between covariates A and D. (Note: In GLIM notation $A*D = A + D + A \cdot D$ so that the interaction term is $A \cdot D$.) The F-ratios for all three interaction terms involving the covariate F have been calculated and are shown in Figure 4—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 $D \cdot T$ is clearly the most significant, overwhelmingly so. Thus we might reasonably

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press



Source of Interaction	F-Ratio	D.F.	5%	(1%)
A.D	2·33	4, ∞	2·37	
A.F	0·89	12, ∞	1·75	
A.T	2·32*	8, ∞	1·94	(2·51)
D.F	1·42	12, ∞	1·75	
D.T	6·44***	8, ∞	1·94	
F.T	1·27	24, ∞	1·52	

Figure 4. Tests of Significance of Interaction Terms in Linear models of Log Odds (Normal Errors).

conclude that the essential features of the data set are encapsulated within the model structure

$$A+D+F+T+D\cdot T$$

with its parametric representation

$$m_u=\mu+\alpha_i+\beta_j+\gamma_k+\delta_l+(\beta\delta)_{jl}.$$

Fitted values (using GLIM)

$$\hat{m}_u=\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j+\hat{\gamma}_k+\hat{\delta}_l+(\widehat{\beta\delta})_{jl}$$

are based on the maximum likelihood (or least squares) estimators shown in Table 2.

Table 2. *Maximum likelihood estimators for the parameters in the normal linear response model of log odds*: D∗T+A+F

				$\hat{\mu}=-2\cdot77$				
A:	$\hat{\alpha}_i$	0	−·28	−·54				
F:	$\hat{\gamma}_k$	0	·02	−·20	−·34	−·17	·05	−·42
				T: *l*→				
D:	*j*↓	0	·53	·11	·93	·40		
		−·27	·61	−·01	·49	·33		
		−1·04	−·13	−·93	−·27	−·12		
				$\hat{\beta}_j+\hat{\delta}_l+(\widehat{\beta\delta})_{jl}$				

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 these lies with the appropriate *F*-tests of Figure 4. 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 3 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 (1)), 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 the inter-office differences are emphasized in order to indicate the extent of variation that might be anticipated between offices. This is pursued further in § 5 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 4). 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 4). Examination of the entries $\hat{\beta}_j+\hat{\delta}_l+(\widehat{\beta\delta})_{jl}$ (Table 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, that for endowment policies being 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 the probable explanation for this effect lies 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.

4. BINARY RESPONSE MODEL

Here the lapse data (w_u, n_u) are treated as binary responses with the observed number of lapses in each category, or unit u, being modelled as a binomial response variable

$$w_u \sim \text{Bin}\ (n_u, p_u).$$

A full theoretical discussion of this model is provided in Technical Appendix B.

This time the possible effects of the covariates are entered into the model through the lapse probabilities p_u by means of a link-function. The GLIM computer package was used first to fit, and then to conduct a graphical analysis of residuals, for a variety of model structures under the logit link function. As anticipated, the previously used structure $D*T+A+F$ with logit link

$$\log\left(\frac{p_u}{1-p_u}\right)=\mu+\alpha_i+\beta_j+\gamma_k+\delta_l+(\beta\delta)_{jl}$$

provided an adequate fit. The plot of standardized residuals

$$r_u = \frac{w_u - \hat{m}_u}{\sqrt{\hat{m}_u\,(1-\hat{m}_u/n_u)}}$$

against fitted values

$$\hat{m}_u = \widehat{E(w_u)} = n_u\,\hat{p}_u$$

only is reproduced here (Figure 5). This is satisfactory, subject to a few outliers. Estimates for the parameters are reproduced (Table 3) alongside those (in brackets) for the same structured normal model (Table 2) in order to facilitate comparison. It is reassuring to note that these estimates lead to the same conclusions as before.

Tests of significance of the dependence of lapse rates (binary response) on all four covariates are summarized on lattice diagrams (Figure 6). In each instance, the value of the likelihood ratio test statistic (the deviance) is written on the lattice branches, together with the corresponding number of degrees of freedom. The asymptotic reference distribution is chi-square. The results offer overwhelming evidence that all four factors are significant. Theoretical details of these tests are contained in Technical Appendix B.

Tests to ascertain the relative importance of the various possible interaction terms are also summarized on a lattice diagram (Figure 7). Here the deviance and degrees of freedom are recorded against each model structure (as before) and their differences displayed on the branches. Since the asymptotic distribution is known to be unreliable and is at the centre of current research, these entries may be used only as a 'screening device' indicating that the interaction between covariates D and T is the most dominant.[8]

Table 3. Maximum likelihood estimators for the parameters in the Binary Response Model (Logit Link): D*T+A+F.

$\hat{\mu} = -2{\cdot}81$ $(-2{\cdot}77)$

A:	$\hat{\alpha}_i$	0 (0)	−·26 (−·28)	−·45 (−·54)				
F:	$\hat{\gamma}_k$	0 (0)	·03 (·02)	−·22 (−·20)	−·33 (−·34)	−·18 (−·17)	·00 (·05)	−·35 (−·42)
		T: l→						
D: j	↓	0 (0)	·56 (·53)	·27 (·11)	·82 (·93)	·40 (·40)		
		−·21 (−·27)	·68 (·61)	−·02 (−·01)	·44 (·49)	·35 (·33)		
		−1·04 (−1·04)	−·09 (−·13)	−·97 (−·93)	−·30 (−·27)	−·11 (−·12)		

$\hat{\beta}_j + \hat{\delta}_l + \widehat{(\beta\delta)}_{jl}$

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

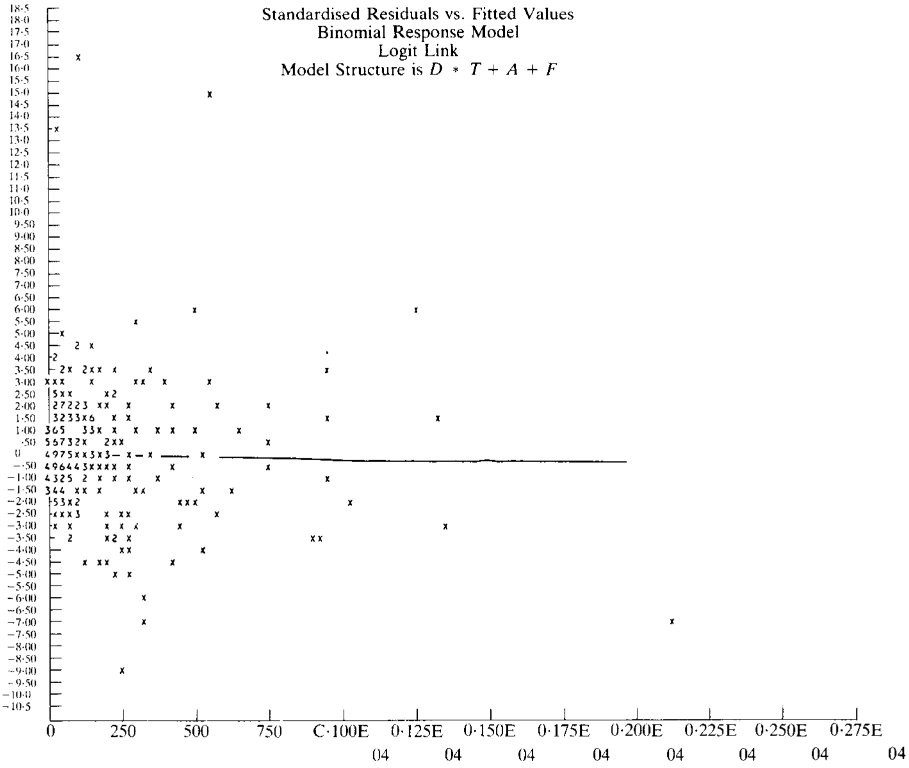


Figure 5. Plot of Standardized Residuals vs. Fitted Values for Binary Response model (Logit Link): $D*T+A+F$.

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

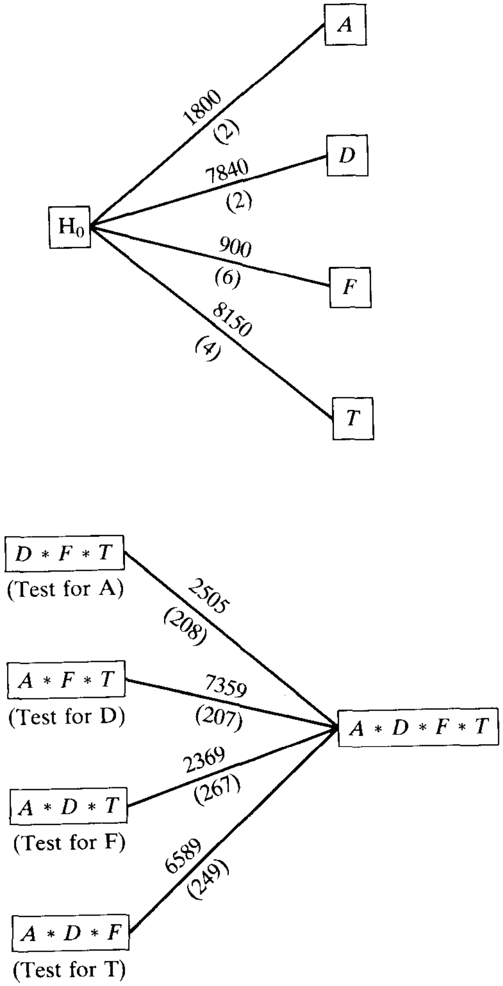


Figure 6. Tests of Significance of Dependence of Lapse Rates on all four Covariates for Binary Response model (Logit Link).

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

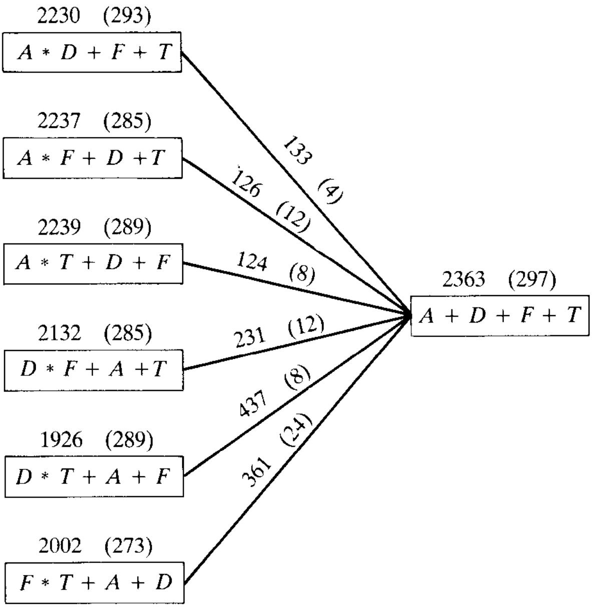


Figure 7. Tests of Significance of Interaction Terms in Binary Response model (Logit Link).

5. ANALYSIS OF DETAILED LAPSE DATA FOR A SPECIFIC OFFICE

More detailed lapse rates than hitherto used were available for analysis. It was decided to concentrate on those for a specific office (office 4). Office 4 was chosen for this more detailed analysis because lapse rates for an additional policy type (unit-linked policies), albeit incomplete, were available for this chosen office. According to the Faculty Research Group's analyses, office 4 represents the second lowest set of lapse rates (this is confirmed by the estimates of $\hat{\gamma}_k$ in Tables 2 and 3). The covariates investigated, together with their crossed categories were as follows:

A—age at entry, 6 levels $i=1$ to 6 corresponding to the age groupings (years):
20–24, 25–29, 30–34, 35–39, 40–44, 45–54

D—policy duration, 9 levels $j=1$ to 9 corresponding to durations (years):
1, 2, 3, 4, 5, 6–8, 9–11, 12–14, 15 and over

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

T—policy type; 6 categories
$k=1$ with-profit } endowment
$k=2$ non-profit }
$k=3$ with-profit } whole-life
$k=4$ non-profit }
$k=5$ temporary
$k=6$ unit-linked

Lapse data were not available for unit-linked policies of duration in excess of 12 years, giving rise to a total of $6^2.9-6.2=312$ units $\{u\colon u=(i,j,k)\}$.

Again the additive, non-interactive model $A+D+T$ with normal error structure and log (lapse odds) response was found to provide a reasonable initial working model on the basis of the residual plots (Figure 8). These plots, only one of which is reproduced here, exposed six clear outliers which were traced to the only six cross-classified cells where lapses had so far failed to register due to

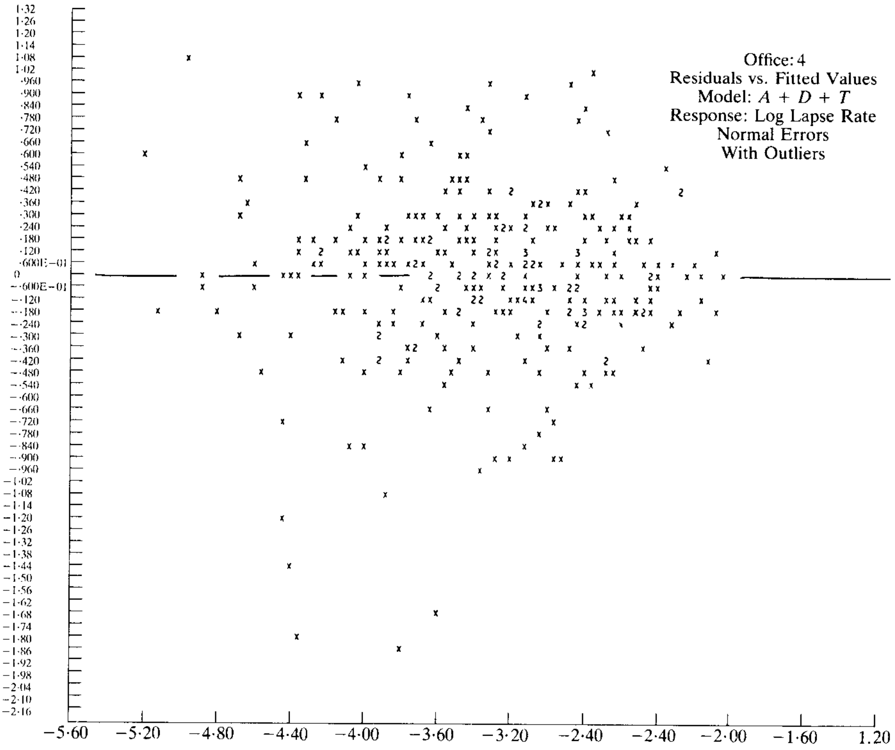


Figure 8. Plot of Residuals vs. Fitted Values for $A+D+T$ model for Office 4 of Log Odds (Normal Errors) with Outliers.

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

insufficient exposure. Residual plots (Figures 9(a)–9(d)) with the outliers omitted from the fit may be deemed to be satisfactory leading to the adoption of the model in the first instance with some assurance.

Attempts to simplify the model by omitting main effects, term by term, were all vigorously rejected. Details of the formal *F*-tests are summarized in Figure 10(a). Interaction terms were also investigated. Summary details of the formal tests applied are given in Figure 10(b); the result is that only one interaction term is statistically significant, overwhelmingly so, leading to the acceptance of the model

$$M\colon\ m_u = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\gamma)_{ik}$$

written as

$$A{*}T + D \text{ or } A + D + T + A \cdot T$$

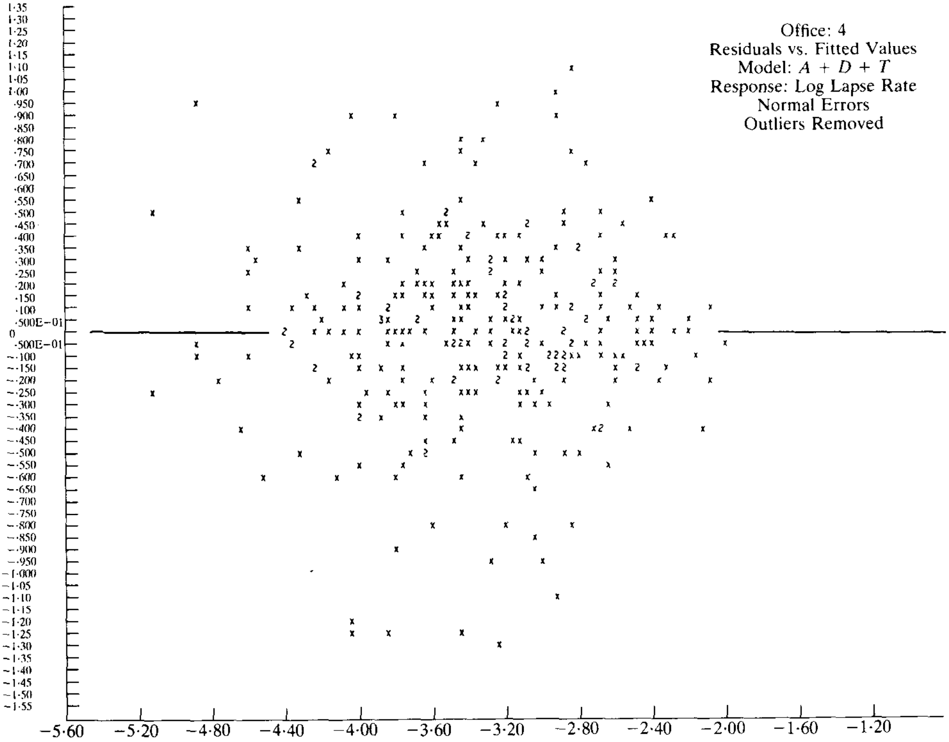


Figure 9a. Plot of Residuals vs. Fitted Values for $A+D+T$ model for Office 4 of Log Odds (Normal Errors)—Outliers Removed.

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

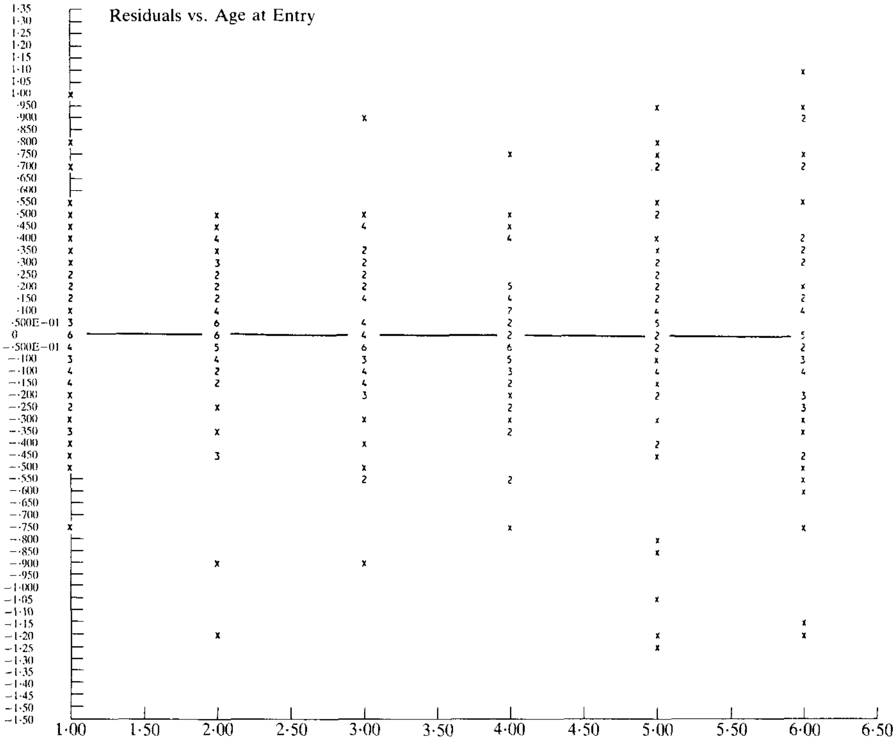


Figure 9b. Plot of Residuals vs. Age at Entry for $A+D+T$ model for Office 4 of Log Odds (Normal Errors)—Outliers Removed.

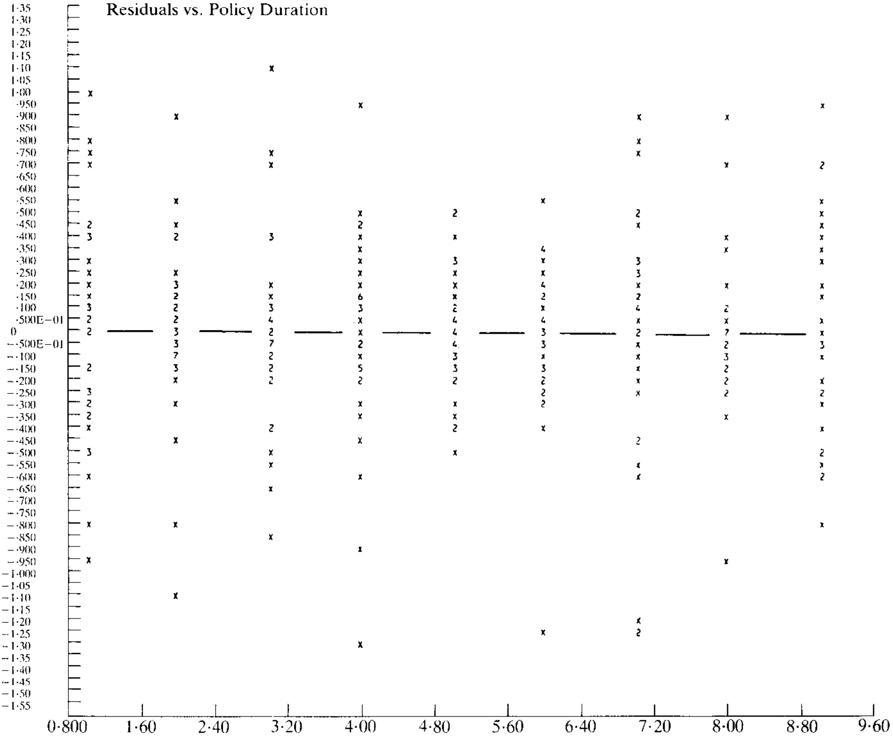


Figure 9c. Plot of Residuals vs. Policy Duration for $A+D+T$ model for Office 4 of Log Odds (Normal Errors)—Outliers Removed.

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

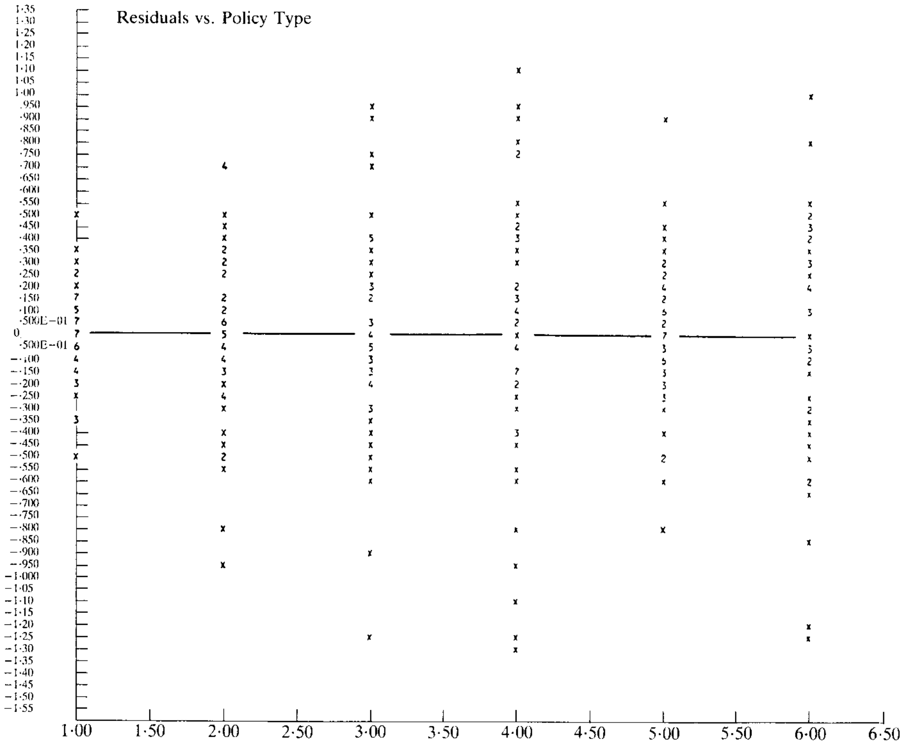
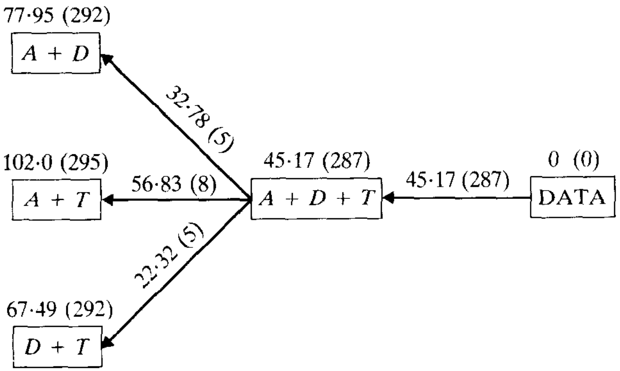


Figure 9d. Plot of Residuals vs. Policy Type for $A+D+T$ model for Office 4 of Log Odds (Normal Errors)—Outliers Removed.

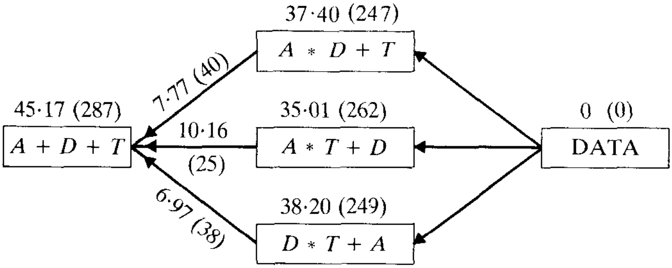
https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press



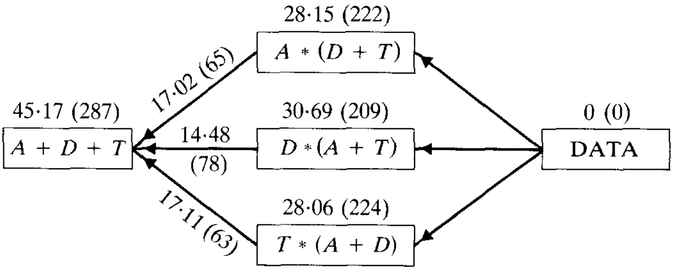
Source of Main Effect	M.S.	F-Ratio	D.F.
T	6·56	41·7***	5, ∞
D	7·10	45·1***	8, ∞
A	4·46	28·3***	5, ∞
Residuals	·1574		

Figure 10a. Tests of Significance of Main Effects in $A+D+T$ model for Office 4 of Log Odds (Normal Errors).

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press



Source of Interaction	F-Ratio	D.F.	5%	(1%)
$A.D$	1·28	40, ∞	1·39	
$A.T$	3·04***	25, ∞	1·52	(1·79)
$D.T$	1·20	38, ∞	1·40	



Source of Interaction	F-Ratio	D.F.	5%	(1%)
$A.D + A.T$	2·07***	65, ∞	1·31	(1·47)
$D.A + D.T$	1·26	78, ∞	1·29	
$T.A + T.D$	2·17***	63, ∞	1·31	(1·47)

Figure 10b. Tests of Significance of Interaction Terms in $A+D+T$ model for Office 4 of Log Odds (Normal Errors).

in GLIM notation. Parameter estimates (in the non-unique, GLIM version) are tabulated for interpretation in Table 4, from which we conclude that:

1. Interaction between policy type and age at entry, which was only marginally significant for the condensed data set (§ 3) takes precedence. This may be partially an artefact of the different ways in which the data were edited. Further we remark that the two data sets, for the specific office, are not strictly comparable with respect to both age at entry (at the extremes) and policy type (because of the inclusion of unit-linked).

2. Policies of short duration (2–3 years) are most prone to lapse, with a general reduction in the propensity to lapse with increasing duration thereafter. This feature is essentially non-interactive with respect to both policy type and age at entry.

3. There is a tendency for the lapse rate to decrease slightly with increasing age for both non-profit whole-life and temporary policies but with a marked additional reduction, in excess of these trends, at age 40–44 years. This feature is confirmed (see Figure 3 on p. 270 of reference (1)) for all offices for non-profit whole-life policies. This is a matter for further investigation for the specific office.

4. There is a steady reduction in lapse rates with increasing age at entry for unit-linked policies.

5. For the non-profit policies, lapse rates are greater for endowment policies with a younger age at entry but greater for whole-life policies with an older age at entry.

6. For the non-profit policies, the lapse rates are higher than for the corresponding with-profit policies. The temporary policies are similar to the non-profit group. The unit-linked policies have higher lapse rates that the other policy types at the youngest ages; at ages 40–54 their experience is similar to that of with-profit policies. In attempting to understand these differences, the reasons why the policies were originally effected would be helpful, and similarly the nature of any guaranteed benefits on early surrender. For the older ages at entry the similarity between the experience of the unit-linked and with-profit policies

Table 4. Maximum likelihood estimators for the parameters in the Linear Response Model of Log Odds: A∗T+D *for office 4*

$\hat{\mu}$:				−3·58				
$D:\hat{\beta}_j$	0	·53	·60	·22	·00	−·39	−·38	−·89

	$T:k\rightarrow$					
$A:i\downarrow$	0	·90	·27	·66	·89	1·23
	−·03	·78	−·02	·54	·71	·71
	−·28	·53	−·08	·49	·36	·51
	−·41	·45	−·09	·44	·31	·09
	−·44	·36	−·35	·12	·09	−·44
	−·82	·15	−·28	·44	·45	−·64

$\hat{\alpha}_i+\hat{\gamma}_k+(\widehat{\alpha\gamma})_{ik}$

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

may again reflect the purpose for which the policies were originally effected and the savings nature of these policy types.

6. HETEROGENEITY AND POLICY DURATION

As indicated in § 1, the data under discussion came from a cross-sectional investigation rather than a cohort investigation. This impairs interpretation of the results. However, even if the data were derived from a cohort study there would be problems in applying the results derived for a subgroup of policyholders to an individual policyholder. This bias arises because the members of a well-defined subgroup are inevitably mixed with respect to their propensity to experience the decrement under study. The short discussion that follows relies heavily on the ideas of Vaupel and Yashin.[9]

Consider a group of policies of a given type, issued by a certain office to policyholders in the same year with the same age at entry. We assume mortality is negligible and we are to consider the variation of withdrawal rates for this group of policies with policy duration. This situation corresponds to the cohort version of the data available from the Faculty of Actuaries' investigation. We are controlling for three of the four significant factors that have been identified and considering the variation in withdrawal rates with the fourth factor—viz. policy duration.

Without loss of generality, we assume that this group of policies is made up of two homogeneous subgroups (subcohorts).

Let $\mu_1(t)$ and $\mu_2(t)$ be the forces of withdrawal at duration t for the two subcohorts, and let $\bar{\mu}(t)$ be the observed force of withdrawal for the entire cohort. The important question is: how does the duration profile of $\bar{\mu}(t)$ compare with those of $\mu_1(t)$ and $\mu_2(t)$?

Let $p_i(t)$ be the survival probabilities for the two subcohorts ($i=1, 2$):

$$p_i(t) = \exp\left[-\int_0^t \mu_i(s)\,\mathrm{ds}\right] \qquad i = 1, 2.$$

Let $\pi(t)$ be the proportion of the surviving cohort at duration t that is in subcohort no. 1, so

$$\pi(t) = \frac{\pi(0)\, p_i(t)}{\pi(0)\, p_i(t) + (1-\pi(0))\, p_2(t)}$$

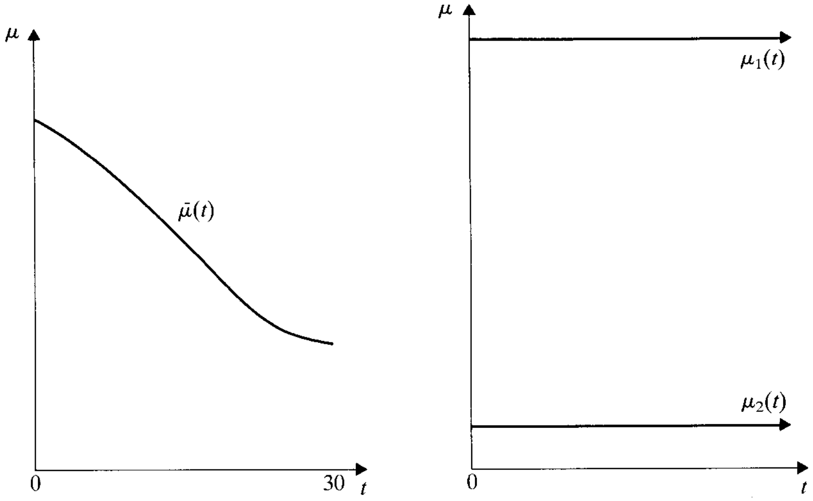
Clearly

$$\bar{\mu}(t) = \pi(t)\,\mu_1(t) + (1-\pi(t))\,\mu_2(t).$$

The dependency of the force of withdrawal for the whole cohort on the forces of withdrawal for the individual subcohorts is affected by the variation in $\pi(t)$ and $1-\pi(t)$, i.e. the changing proportion of the population that is in each of the subcohorts.

An example is given in Figure 11 (which is not unlike the appearance of Figure 1 in reference (1)). In Figure 11, the cohort aggregate reveals a force of

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press



$$\mu_1(t) = \cdot 10$$
$$\mu_2(t) = \cdot 01$$
$$\pi(0) = \cdot 8$$

Figure 11. Apparent Decline in the Overall Force of Withdrawal in the Presence of Heterogeneity.

withdrawal that declines from ·082 at duration 'zero' to ·030 at duration 30 (years). Does this imply that the hazard of withdrawal for individual policyholders decreases with increasing policy duration? Not necessarily. As in Figure 11, there might be two homogeneous types of policyholder:

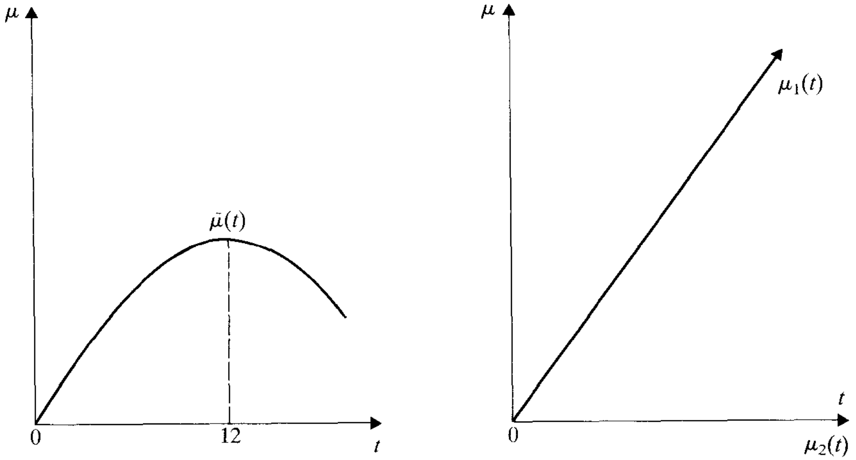
$$\mu_1(t) = \cdot 10$$
$$\mu_2(t) = \cdot 01$$

one with the force of withdrawal ten times that of the other. For individuals in each group the force of withdrawal is a constant. The observed decline is an artefact caused by the original population being heterogeneous.

A similar example is shown in Figure 12 where the cohort picture indicates a force of withdrawal that peaks at a duration of about 12 years. This does not necessarily imply that an individual policyholder will experience a force of withdrawal that peaks in this way. There might be two homogeneous types of policyholder:

$$\mu_1(t) = \cdot 02t$$
$$\mu_2(t) = 0$$

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press



$$\mu_1(t) = \cdot 02t$$
$$\mu_2(t) = 0$$
$$\pi(0) = \cdot 9$$

Figure 12. Apparent Maximum for the Overall Force of Withdrawal in the Presence of Heterogeneity.

one with a linearly increasing hazard rate, the other immune to withdrawal! Again the observed peak is an artefact (like the ‘seven year itch’ in divorce rates).

Further examples of the potential effects of heterogeneity are given by Vaupel and Yashin.[(9)]

The following more general points may be made given the above examples:

1. Regardless of how many different factors or attributes are considered, individuals (in this case policyholders) who are grouped together will differ according to other unobserved or neglected characteristics. Some of these differences will affect, in this case, the propensity to withdraw. This heterogeneity leads to *selection*, in that the surviving population will differ from the initial population. This means that (a) observations of the surviving population cannot be translated directly into conclusions about the behaviour or characteristics of the individuals making up the original population and (b) that the overall variation in, here, withdrawal rates with time cannot be used to make direct inferences about the variation of withdrawal rates with time for the individuals making up the population.

2. The aggregate withdrawal patterns may, if accepted without question, lead to erroneous policy decisions if for example, in Figures 11 and 12 a change in marketing leads to an increase in $\pi(0)$.

3. It is not clear how important these effects are for an understanding of

withdrawal rates. They are certainly important elsewhere, e.g. mortality, where pioneering work has been published by Beard[(10)] and Redington.[(11)]

7. CONCLUSIONS

Modelling 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 this is through a thorough graphical analysis of model residuals which, ideally, should be 'pattern free'. Additionally, what might be termed 'fine tuning' might then be attempted, and its effects formally assessed. The development of generalized linear modelling, together with its associated computer soft-ware package GLIM, facilitates such modelling objectives.

These techniques have been applied to the 'lapse' data of the Faculty of Actuaries' Withdrawal Research Group and conclusions have been drawn about the relative importance of the four factors identified by the Withdrawals Research Group (age at entry, duration of policy, office, type of policy) their independence and possible interactions. Possible sources of bias in interpreting the results have also been discussed.

REFERENCES

(1) Crombie J. G. R. *et al.* (1979) Faculty of Actuaries Withdrawal Research Group. 'An Investigation into the Withdrawal Experience of Ordinary Life Business'. *T.F.A.* **36,** 262–295 (with discussion).

(2) Johnson P. D. & Hey G. B. (1971) Statistical Studies in Motor Insurance. *J.I.A.* **97,** 199–232 (with discussion).

(3) Grimes T. (1971) Claim Frequency Analysis in Motor Insurance. *J.S.S.* **19,** 147–154.

(4) Bennett M. (1978) Models in Motor Insurance. *J.S.S.* **22,** 134–159.

(5) Coutts S. M. (1984) Motor Insurance Rating: An actuarial approach. *J.I.A.* **111,** 87–148.

(6) Benjamin B. & Pollard J. H. (1980) *Analysis of Mortality and other Actuarial Statistics.* Chapters 2–4. Heinemann.

(7) Cox P. R. (1976) *Demography.* Fifth Edition, Chapter 6. Cambridge University Press.

(8) McCullagh P. & Nelder J. R. (1983) *Generalised Linear Models,* pp. 26–28. Chapman and Hall.

(9) Vaupel J. W. & Yashin A. I. (1985) Heterogeneity's Ruses; Some Surprising Effects of Selection on Population Dynamics. *The American Statistician,* **36,** 176–185.

(10) Beard R. E. (1961) A Theory of Mortality Based on Actuarial, Biological and Medical Characteristics. *International Population Conference,* Vol. 1, pp. 611–625.

(11) Redington F. M. (1969) An Exploration into the Patterns of Mortality. *J.I.A.* **95,** 243–298 (with discussion).

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

TECHNICAL APPENDIX A

Linear models with normal error structure

Consider the response decomposition

$$\mathbf{y}=\mathbf{m}+\boldsymbol{\varepsilon}$$

with declared model structure M. Here $\mathbf{y}\ \varepsilon\ R^N$ where N is the number of units; while the structure M is said to define a *linear model* if the systematic component $\mathbf{m}$ is chosen so that

$$m\ \varepsilon\ L_M \subseteq R^N.$$

Further let d_M ($\leqslant N$) denote the *dimensionality* of the linear vector space L_M and $v_{\mathrm{M}}=\mathrm{N}-\delta_{\mathrm{M}}$ the associated degrees of freedom.

The decomposition $\mathbf{y}=\mathbf{m}+\boldsymbol{\varepsilon}$ is succinctly represented geometrically in Figure A1 in which the vector $\mathbf{m}$ is constrained to lie somewhere in the subspace or hyper-plane $L_M \subseteq R^N$.

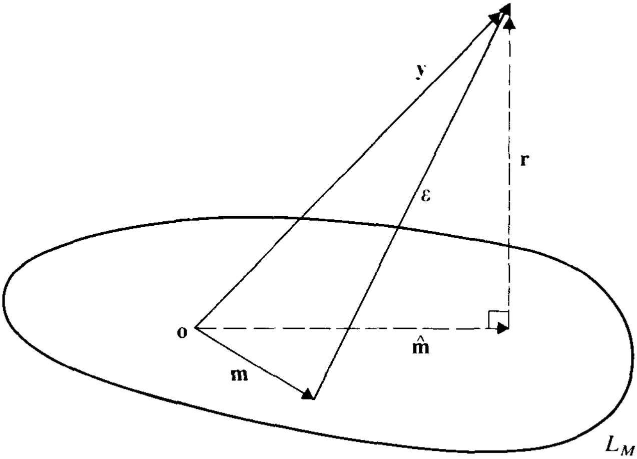


Figure A1. Vector Space Representation of the Response Decomposition: $\mathbf{y}=\mathbf{m}+\boldsymbol{\varepsilon}$.

Fitting and assessing linear models

Fitting a linear model M to an observed response vector $\mathbf{y}$ is equivalent to selecting a suitable value $\hat{\mathbf{m}}$, for $\mathbf{m}\ \varepsilon\ L_M$. Let $\mathbf{r}=\mathbf{y}-\hat{\mathbf{m}}$ denote the corresponding error vector. Call $\hat{\mathbf{m}}=(\hat{m}_u)$ the *fitted values* and $\mathbf{r}=(r_u)$ the *residuals*. How is $\hat{\mathbf{m}}$ (and hence $\mathbf{r}$) to be selected?

Extensive use is made of the method of least squares in which $\hat{\mathbf{m}}$ is chosen to lie

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

at the foot of the perpendicular from $\mathbf{y}$ on to the linear space L_M (see Figure A1). Hence, by this method, fitted values are constructed to satisfy the two criteria:

(C1) $\hat{\mathbf{m}} \,\varepsilon\, L_M$
(C2) $\mathbf{r}=(\mathbf{y}-\hat{\mathbf{m}})$ is perpendicular to any $\mathbf{m} \,\varepsilon\, L_M$.

These in turn imply that:

(i) the fitted values have the imposed model structure M;
(ii) the residuals (r_u) estimate the underlying error structure (ε_u) but subject to d_M constraints; and
(iii) the residuals (r_u) are statistically independent of the fitted values $(\hat{m}_u)$.

Consequently these properties are used to assess the adequacy of any proposed overall model M through a thorough examination of various residual plots.

An example of model fitting

We shall consider fitting the model $M=A+D+F+T$ with parametric form

$$M:\ m_{ijkl}=\alpha_i+\beta_j+\gamma_k+\delta_l.$$

First the model is expressed in the non-parametric form

$$M:\ m_{ijkl}=\bar{m}_{i\ldots}+\bar{m}_{.j..}+\bar{m}_{..k.}+\bar{m}_{\ldots l}-3\,\bar{m}_{\ldots.}$$

where dot denotes averaging over the relevant subscript.

Then criteria C1 and C2 are used as follows:

Criterion C1 implies that the fitted values $\hat{m}_u$ satisfy

$$\hat{m}_{ijkl}=\hat{\bar{m}}_{i\ldots}+\hat{\bar{m}}_{.j..}+\hat{\bar{m}}_{..k.}+\hat{\bar{m}}_{\ldots l}-3\,\hat{\bar{m}}_{\ldots.}\,.$$

Criterion C2 can be rewritten as

$$\sum_u y_u m_u=\sum_u \hat{m}_u m_u \qquad \text{for any } \mathbf{m}=(m_u)\,\varepsilon\, L_M$$

where

$$\sum_u y_u m_u=\sum_{ijkl} y_{ijkl}(\alpha_i+\beta_j+\gamma_k+\delta_l)$$
$$=\sum_i y_{i+++}\,\alpha_i+\sum_j y_{+j++}\,\beta_j+\sum_k y_{++k+}\,\gamma_k+\sum_l y_{+++l}\,\delta_l$$

and

$$\sum_u \hat{m}_u m_u=\sum_i \hat{m}_{i+++}\,\alpha_i+\sum_j \hat{m}_{+j++}\,\beta_j+\sum_k \hat{m}_{++k+}\,\gamma_k+\sum_l \hat{m}_{+++l}\,\delta_l.$$

Here plus denotes summation over the relevant subscript. Further since

$$\sum_u y_u m_u=\sum_u \hat{m}_u m_u$$

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

is *identically* equal (for any $\mathbf{m} \varepsilon L_M$), coefficients of like parameters are equated to give

$$\hat{m}_{i+++} = y_{i+++},\ \hat{m}_{+j++} = y_{+j++},\ \hat{m}_{++k+} = y_{++k+},\ \hat{m}_{+++l} = y_{+++l}.$$

Dividing both sides of each identity by the relevant number of terms delivers the identities

$$\bar{\hat{m}}_{i\ldots} = \bar{y}_{i\ldots},\ \bar{\hat{m}}_{.j..} = \bar{y}_{.j..},\ \bar{\hat{m}}_{..k.} = \bar{y}_{..k.},\ \bar{\hat{m}}_{\ldots l} = \bar{y}_{\ldots l},$$

while averaging each identity over the one remaining suffix gives

$$\bar{\hat{m}}\ldots. = \bar{y}\ldots.$$

Hence the fitted values are

$$\hat{m}_{ijkl} = \bar{y}_{i\ldots} + \bar{y}_{.j..} + \bar{y}_{..k.} + \bar{y}_{\ldots l} - 3\,\bar{y}_{\ldots.}\,.$$

We have assumed no missing values, with each suffix ranging over a predetermined number of (crossed) levels. Any missing values are treated by setting them equal to what would be their fitted values and solving the resulting equations.

The method develops a formula for the fitted values avoiding the use of matrices and the theory of generalized inverses.

Hypothesis testing and analysis of variance

Suppose that on the basis of an examination of residual plots the overall linear model M is no longer in doubt. The question then arises whether a simpler linear model H will suffice. How is this hypothesis H to be tested?

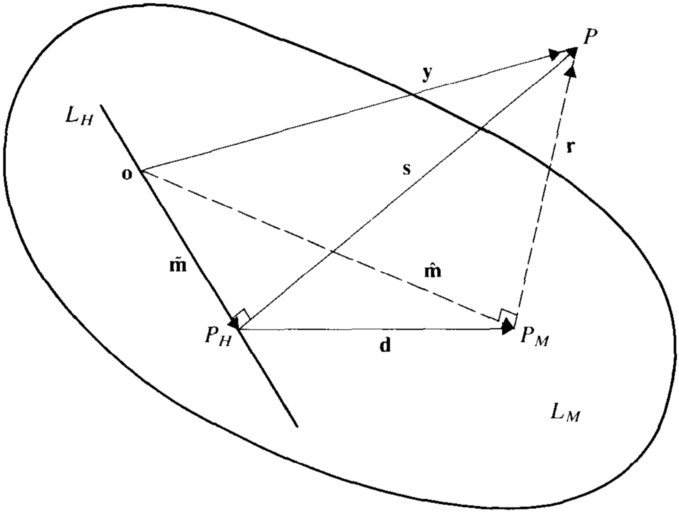


Figure A2. Vector Space Representation of Testing of Significance in a Linear Model.

Denoting the (least squares) fitted values and residuals under the two linear models M and H by $(\hat{\mathbf{m}}, \mathbf{r})$ and $(\tilde{\mathbf{m}}, \mathbf{s})$ respectively it is proposed that we examine the succinct geometrical representation (Figure A2) of these vectors together with their associated linear spaces $L_H \subseteq L_M \subseteq R^N$ where $d_H < d_M < N$ and $v_H > v_M > 0$.

Here L_H is represented as a hyper-line contained within the hyper-plane L_M. The points P_H and P_M are the feet of the perpendiculars from the observed response point P in R^N on to the linear spaces L_H and L_M respectively. Clearly if the point P_H lies 'close' to the point P_M then the simpler model H will suffice. But how close is 'close'? A reference distribution based on the length of the vector $\mathbf{d} = \overrightarrow{P_H P_M}$ is required.

Call **d** the *departure* vector *from* **H** *in* **M**. The vector identities

$$\mathbf{d} = \mathbf{s} - \mathbf{r}$$

$$\mathbf{d} = \hat{\mathbf{m}} - \tilde{\mathbf{m}}$$

are immediately obvious from the relevant hyper-triangles in the figure, either of which may be used to compute **d**. It follows from the rearranged form

$$\mathbf{s} = \mathbf{r} + \mathbf{d}$$

of the first of these identities, where **r** is perpendicular to **d**, that the square lengths of each vector satisfy the identity

$$\| \mathbf{s} \|^2 = \| \mathbf{r} \|^2 + \| \mathbf{d} \|^2.$$

This is written as

$$R_H = R_M + D_H^M$$

the mere expression of Pythagoras' Theorem for a right angled triangle in higher dimensions. Here $R_H = \| \mathbf{s} \|^2$ and $R_M = \| \mathbf{r} \|^2$ are the *residual sums of squares* or *deviances* under the respective models and D_H^M the *departure sum of squares*. The orthogonality of vectors **r** and **d** (**r** is perpendicular to **d**) implies that the statistics R_M and D_H^M are statistically independent. Further, under an independent identically distributed (IID) normal homoscedastic error structure ($\varepsilon_u \sim N(o, \sigma^2)$ and IID for all u) it follows that the ratio

$$\frac{D_H^M}{d_M - d_H} \Big/ \frac{R_M}{N - d_M} = \frac{D_H^M}{v_H - v_M} \Big/ \frac{R_M}{v_M}$$

has the F-distribution on $(v_H - v_M, v_M)$ degrees of freedom *under* hypothesis H. This forms the basis of the conventional F-test for the hypothesis H generally presented as an analysis of variance table displayed below:

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

Lattice	Sources	Sum of Squares	Degrees of Freedom	Mean Square	F-ratio
H ↑	due to M in H	D_H^M	$\nu_H-\nu_M$	$D_H^M/$ $\nu_H-\nu_M$	*
M ↑ DATA	residuals	R_M	ν_M	R_M/ν_M	
	Total	R_H	ν_H		

Such an ANOVA table and its associated lattice of hypotheses

$$\begin{tikzcd}
R_H & & R_M & & O \\
\boxed{H} & & \boxed{M} \arrow[ll, "R_H-R_M"', "\nu_H-\nu_M"] & & \boxed{\text{DATA}} \arrow[ll, "R_M"', "\nu_M"] \\
\nu_H & & \nu_M & & O
\end{tikzcd}$$

can be adapted to cater for more than one hypothesis. The GLIM computer package outputs the residual sum of squares (deviance) and number of degrees of freedom each time a particular model structure is fitted, so that the relevant lattice of hypotheses and equivalent ANOVA table(s) are readily constructed.

TECHNICAL APPENDIX B

Binomial models

The saturated model S

Attention is focused in this Appendix on the dichotomy with w_u lapses (withdrawals) observed in the n_u independent replications (exposures) for each unit u (policy characterization). What follows is largely a question of notation.

Let α, with $\alpha=1$ for a lapse and $\alpha=2$ for a non-lapse, denote the possible states of the dichotomous response variable R and now write $n_{u\alpha}$ $(\alpha=1,\ 2)$ with $n_{u+}=n_{u1}+n_{u2}$ for the observed number of lapses and non-lapses (respectively) in the n_{u+} replications. Constant response probabilities $p_{u\alpha}$ are assumed for each fixed u, subject only to the necessary constraints $p_{u+}=p_{u1}+p_{u2}=1$ for all u. Then the following family of binomial distributions are defined:

$$S\colon n_{u\alpha}\sim \text{Bin}\ (n_{u+},\ p_{u\alpha}).$$

Denoting expected values by

$$m_{u\alpha}=E(n_{u\alpha})=n_{u+}p_{u\alpha}$$

and the model parameters collectively by $\mathbf{p}=(p_{u\alpha})$, it follows that the likelihood function, under S

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

$$L(\mathbf{p}) = \prod_{u,\alpha} \frac{n_{u+}!}{n_{u\alpha}!} p_{u\alpha}{}^{n_{u\alpha}}, \text{ with } p_{u+} = 1$$

can be reparameterized in terms of expected values and written as

$$L(\mathbf{m}) = K \prod_{u,\alpha} m_{u\alpha}{}^{n_{u\alpha}}, \qquad m_{u+} = n_{u+} \tag{B1}$$

retaining only those parts of the expression involving the new parameters $\mathbf{m} = (m_{u\alpha})$.

Concentratiing for the time being on the untransformed version of the likelihood function which can be rewritten, for instance as

$$L(\mathbf{p}) \propto \prod_{u} p_{u1}{}^{n_{u1}} (1 - p_{u1})^{n_{u2}},$$

it follows from the following trivial result:

Lemma: The function $f(\ )$ defined as

$$f(x;\, a,\, b) = x^a\, (1-x)^b\ a,\, b > 1,\, 0 \leqslant x \leqslant 1$$

has a maximum at $x = a/(a+b)$,

that the maximum likelihood estimators of the $p_{u\alpha}$'s under S are

$$p^*_{u\alpha} = n_{u\alpha}/n_{u+}\,.$$

Using these estimators for the expected values, the fitted values under S are

$$m^*_{u\alpha} = n_{u+} p^*_{u\alpha} = n_{u\alpha},$$

the data themselves. For this reason, the model S is said to be saturated. Further, the model has dimension $d_S = N$, the number of units u, since there are N independent unknown parameters p_{u1} say (with $p_{u2} = 1 - p_{u1}$), and consequently $N - d_S = 0$ degrees of freedom. The model S is denoted by $A{*}D{*}F{*}T$ in GLIM notation for the application in question.

Hypothesis testing

The detailed composition of a unit, u, is determined by the explanatory factors. Thus for the condensed data set with factors A, D, F and T discussed previously, $u = (i, j, k, l)$.

Next, consideration is given to the class of hypotheses (submodels), L_S, defined by ignoring one or more of the explanatory factors. To denote this, partition u and write $u \equiv (r, s)$ in which r identifies the units associated with the *retained* factors and s identifies the levels associated with the factors *omitted*. Thus, for example, if attention is focused just on factors A and F and the response R is assumed to be independent of factors D and T, $r = (i, k)$ and $s = (j, l)$. One

https://doi.org/10.1017/S0020268100042566 Published online by Cambridge University Press

possible abbreviation for such a submodel or hypothesis H, within S, would be

$$H\colon R \perp (D, T) \mid (A, F)$$

to be interpreted as—the hypotheses H is such that the response R is independent ($\perp$) of factors D and T but is dependent (or conditional) on factors A and F. An alternative parametric representation of H, in terms of probabilities $p_{u\alpha}$, which have now to be expanded to $p_{rs\alpha}$, is

$$H\colon p_{u\alpha} \equiv p_{rs\alpha} = p_{(r)\alpha}.$$

Here $p_{(r)\alpha}$ denotes the probability that the response is α, *conditional* (hence the brackets) on the factors of immediate interest and whose levels are characterized by units r, but *independent* of s. The expanded notation applied to the data, means we can write $n_{u\alpha} \equiv n_{rs\alpha}$ and

$$H\colon n_{rs\alpha} \sim \text{Bin}\,(n_{rs+}, p_{(r)\alpha})$$

with expected values (under H)

$$m_{rs\alpha} = E\,(n_{rs\alpha}) = n_{rs+}p_{(r)\alpha}$$

The likelihood under H, *conditional* on a given r, is

$$L(\mathbf{p}_{(r)}) = \prod_{s,\,\alpha} \frac{n_{rs+}!}{n_{rs\alpha}!}\, p_{(r)\alpha}{}^{n_{rs\alpha}} \propto \prod_{\alpha} p_{(r)\alpha}{}^{n_{r+\alpha}}$$

where $n_{r+\alpha}$ denotes summation over s; while the full likelihood becomes

$$L(\mathbf{p}) = \prod_{r} L(\mathbf{p}_{(r)}) \propto \prod_{r,\,\alpha} p_{(r)\alpha}{}^{n_{r+\alpha}} \propto \prod_{r} p_{(r)1}{}^{n_{r+1}}\,(1 - p_{(r)1})^{n_{r+2}}$$

for instance. Applying the lemma from before, it follows that the maximum likelihood estimators under H are

$$\hat{p}_{(r)\alpha} = n_{r+\alpha}/n_{r++},$$

so that the expected or fitted values under H are

$$\hat{m}_{rs\alpha} = n_{rs+}\hat{p}_{(r)\alpha} = n_{rs+}n_{r+\alpha}/n_{r++}.$$

Clearly the submodel H is not saturated and has dimension d_H equal to the number of (sub) units r with $\nu_H = N - d_H$ degrees of freedom.

The likelihood ratio test statistic for H within S,

$$\lambda = \frac{\displaystyle\sup_{\mathbf{m}\,\varepsilon\,S} L(\mathbf{m})}{\displaystyle\sup_{\mathbf{m}\,\varepsilon\,H} L(\mathbf{m})} = \frac{L\,(\mathbf{m}^*)}{L\,(\hat{\mathbf{m}})}$$

for which ${}^{S}_{H}Y^2 = 2\log\lambda$ asymptotically has the chi-square distribution with ν_H

degrees of freedom under H so the hypothesis test is readily constructed. Introducing the expanded suffix notation $u = (r, s)$ into equation (B1), gives the following:

$$\mathrm{L}(\mathbf{m}) = K \prod_{r,\,s,\,\alpha} m_{rs\alpha}{}^{n_{rs\alpha}}.$$

The values derived for $m^*_{rs\alpha}$ and $\hat{m}_{rs\alpha}$ are substituted into this expression in turn and it follows that

$$\lambda = \prod_{r,\,s,\,\alpha} \left(\frac{n_{rs\alpha}\, n_{r++}}{n_{rs+}\, n_{r+\alpha}} \right)^{n_{rs\alpha}}$$

from which the test statistic can be simply obtained

$${}^{S}_{H}Y^2 = 2 \sum_r \left[\sum_{s,\,\alpha} n_{rs\alpha} \log n_{rs\alpha} - \sum_s n_{rs+} \log n_{rs+} - \sum_\alpha n_{r+\alpha} \log n_{r+\alpha} + n_{r++} \log n_{r++} \right].$$

Its value, the model deviance, together with the degrees of freedom ν_H are outputted by the GLIM package on fitting the model $A{*}F$ for the case study. (The binomial error structure must be declared and the logit link selected either consciously or by default.)

It should be noted that we have only expressed an interest in the class of hypotheses L_S determined by the deletion of one or more of the explanatory factors of a saturated model S, thus enabling us to test for the possible independence of the response R of certain of the factors. It is as well to remember, in so doing, that the ${}^{S}_{H}Y^2$ statistic is as much a reflection of the saturated model S, either preselected by design or imposed externally, as it is a reflection of the status of the submodel H of interest. To neutralize any possible arbitrary effect that the choice of S may be thought to have on an analysis of submodels, it is possible to compare two submodels H_1, $H_2 \varepsilon L_S$ (same S) with $\nu_1 > \nu_2$ say, using the difference ${}^{S}_{H_1}Y^2 - {}^{S}_{H_2}Y^2$. Asymptotically, this has the chi-square distribution with $\nu_1 - \nu_2$ degrees of freedom under H_1.

More specifically, it is possible to compare submodels H with the so-called *minimal model* H_0, which, in the present context, conjectures that the response R is independent of all explanatory factors. Under H_0, $s = u$ and all reference to r is deleted in the formula above for ${}^{S}_{H_0}Y^2$, giving

$${}^{S}_{H_0}Y^2 - {}^{S}_{H}Y^2 = 2 \left[\sum_{r,\,\alpha} n_{r+\alpha} \log n_{r+\alpha} - \sum_r n_{r++} \log n_{r++} - \sum_\alpha n_{++\alpha} \log n_{++\alpha} + n_{+++} \log n_{+++} \right]$$

with $\nu_{H_0} - \nu_H$ or $d_H - 1$ degrees of freedom. The resulting test statistic is clearly invariant of the number of components in s, and hence the saturated model (S), since *all* the data frequency counts in the formula involve summation over s.