A.A.S. **2**, I, 1-24 (2007)

MODELLING THE CLAIM DURATION OF INCOME PROTECTION INSURANCE POLICYHOLDERS USING PARAMETRIC MIXTURE MODELS

By D. G. W. Pitt

ABSTRACT

This paper considers the modelling of claim durations for existing claimants under income protection insurance policies. A claim is considered to be terminated when the claimant returns to work. Data used in the analysis were provided by the Life and Risk Committee of the Institute of Actuaries of Australia. Initial analysis of the data suggests the presence of a long-run probability, of the order of 7%, that a claimant will never return to work. This phenomenon suggests the use of mixed parametric regression models as a description of claim duration which include the prediction of a long-run probability of not returning to work. A series of such parametric mixture models was investigated, and it was found that the generalised *F* mixture distribution provided a good fit to the data and also highlighted the impact of a number of statistically significant predictors of claim duration.

KEYWORDS

Income Protection Insurance; Mixture Models; Claim Termination Rates; Maximum Likelihood

CONTACT ADDRESS

David Pitt, Centre for Actuarial Studies, Department of Economics, The University of Melbourne, Parkville 3502, Victoria, Australia. Fax: +61-3-8344-6899; E-mail: dgpitt@unimelb.edu.au

1. Introduction

Insurers will benefit considerably from having a good understanding of the durations of claims likely to be experienced by claimants under their income protection (IP) insurance policies. Claim durations have a significant impact on both the pricing and the reserving calculations routinely made by insurers. Ultimately, inaccurate modelling of claim durations could also contribute to insurer insolvency and a lack of consumer confidence. A mathematical model of claim durations also enables the profit testing of a set of premium rates to be readily automated. This paper provides a strategy for modelling claim durations which is demonstrated to provide a good summary of observed claim duration patterns, and hence will be of value to insurers in their quest for suitable pricing and reserving methods in respect of their IP insurance policies.

The first significant step in the modelling of claim termination rates was

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

the production of a number of disability tables based on industry-wide data. In the United States of America, we have seen the production of the Commissioner's Disability Tables in 1964, which were updated with the production of the Commissioner's Individual Disability Table A, CIDA, in 1985 (see Robinson, 1988). This U.S. derived table is based on data from 20 companies over the period 1973 to 1979. The termination rate for a particular claim duration since disablement is derived as a product of factors corresponding to the profile of each claim. These factors include duration since onset of disability, age of claimant, deferred period, occupation of claimant, gender, and an indicator of whether the claim is related to accident or sickness.

In the United Kingdom, the Continuous Mortality Investigation (CMI) Bureau has produced a number of reports which describe and model both mortality and morbidity experience. Most notably, the twelfth report of the CMI, CMIR12 (CMIB, 1991), describes the development of a multiple state model for the description of IP insurance. CMIR12 contains a graduation by mathematical formula of claim recovery rates. The mathematical formula employed by the CMI in this report modelled the impact of age at the date of falling sick, duration of disability and the deferred period written into the insurance policy of the claimant in the description of the claim termination rate.

The Institute of Actuaries of Australia (IAAust) has also developed its own industry morbidity table. This table, known as IAD 1989-93, was produced by a subcommittee of the Disability Committee of the IAAust in 1995. The claim recovery rates were modelled using a series of linear functions relating the recovery rate to the age of the claimant. The coefficients in the estimated linear models were allowed to vary according to the age and the gender of the claimant and the deferred period selected by the claimant at the time of policy inception.

Besides industry developed tables, a number of other investigations into recovery rates have been conducted. Gregorius (1993) describes a multiple state model used for the analysis of IP policies in the Netherlands. The recovery rates are described using a piecewise constant force of recovery. Segerer (1993) describes the methodology used in Germany, Austria and Switzerland. Recovery rates are not modelled explicitly in these countries. In order to predict the expected present value of claim payments under an IP policy, ordinary life table annuity values are used as the starting point. These annuity values are then reduced by a factor to allow for the fact that payment is made contingent on both survival and continuing disability.

This paper will consider the modelling of claim durations with the use of survival analysis. After this introduction, Section 2 will describe the IP insurance policy data used in the paper. Section 3 will provide the results of some initial analysis of these data and will describe the modelling strategies implied by these initial analyses. Section 4 will describe the main modelling

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

technique employed in this paper, namely mixed parametric regression models of claim continuation. The results of fitting these models will be presented. Section 5 will provide discussion of the results from the mixture modelling. Section 6 will conclude the paper, and will provide some avenues for further research in this area.

2. The Institute of Actuaries of Australia Claim Duration Data

The IAAust IP insurance policy database contains information on policyholders who have purchased insurance from the main Australian providers of this form of insurance. There are about 20 different insurers which provide data to this database on an annual basis. Data are recorded for each policyholder, based on the information provided in the insurance proposal form. In addition, and most importantly for this study, dates of claim commencement and claim cessation are recorded for each policyholder who commenced claim.

For this research, all claims which began in calendar year 1995 were extracted from the IAAust database. There were 8,863 new claims recorded in respect of calendar year 1995. These claims were followed until termination or until the end of calendar year 1998, whichever occurred first. The data set contains information including: the duration of each policyholder's claim; the age of the claimant on the date of disability onset; the definition of disability used in assessing whether the policyholder is eligible for a benefit under the policy; the gender of the policyholder; the occupation class of the policyholder (classified into four levels — see The Institute of Actuaries of Australia, 1997); the frequency of benefit payment; the rate of benefit payable monthly; the type of benefits payable (increasing in line with inflation or level); the smoker status of the insured life; and the deferment period specified in the insurance contract. Appendix A provides a table of the characteristics (potential rating factors) recorded for each of these claimants, along with the coded variable name and a brief description of the variable.

Of the 8,863 claims recorded, 7,771 (88%) related to terminated claims, the remainder being censored. Terminated claims include those claims where claimants recovered and returned to work. All other forms of claim cessation, including death of the claimant, result in a censored claim for the purpose of this study. The most common cause of censoring was that the claim reached the end of 1998 and was continuing at that time. There were a small number of claims which were lost at the end of each of 1995, 1996 and 1997, and which are unable to be followed further. This issue arose due to changes in claim codes adopted by companies which provided these data to the IAAust Life and Risk Committee at the end of particular calendar years. Most of these claims could be traced by matching claims from one

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

calendar year to the next on the basis of date of birth, date of entry to the policy, sex, occupation class and smoker status. However, a small proportion (comprising less than 1% of the total number of claims which began in 1995) could not be successfully matched. The age profile of claimants ranged from 17 to 70, with an average age of 40. The distribution of ages for new claimants was approximately bell shaped. Of the 8,863 claimants included in the dataset, 2,409 (27.2%) were in occupation class A, 667 (7.5%) were in occupation class B, 3,165 (35.7%) were in occupation class C, and 2,622 (29.6%) were in occupation class D. Occupation Class A relates to professional white collar and sedentary occupations. Class B relates to other sedentary occupations, including supervision of manual workers. Class C relates to light manual workers and class D relates to moderate manual workers. See The Institute of Actuaries of Australia (1997) for further discussion on occupation class descriptions. Just over 50% of the claims related to disability definitions where the ‘inability to perform any occupation’ test is applied in determining whether the claim can continue after an initial period.

Males account for 87% of the data, while monthly benefit payments are clearly the most common, also accounting for 87% of the data. We note also that 54.7% of the claimants had chosen benefits which increase in line with inflation. Only 5% of the claimants would have required thorough medical examinations before claim payments commenced. Level premiums accounted for 13.6% of the data, the remainder relating to stepped premiums. The smoker prevalence rate amongst claimants was 19.5%. Sickness caused 58.9% of the claims, the remainder being due to accident.

3. Initial Analysis of the Data

In order to understand the duration profile of disability claims, Kaplan-Meier survival curves (see Kaplan & Meier, 1958) have been created for the duration of disability claims. Kaplan-Meier curves can be used to provide a non-parametric estimate of the survival function for claims. The event of interest in this survival analysis is clearly claim termination. The duration variable is used to measure time since payment of disability benefits begins.

In order to produce Figure 1, all claims data have been aggregated. This Figure shows the Kaplan-Meier estimated survival function for claims where both sexes, all deferred periods and all other possible classifications of the data have been grouped together. Immediately apparent from Figure 1 is the drop in claims in force after 730 days; that is after two years. This issue was investigated, and claims which cease due to the expiry of a two-year benefit period were not included in the data used to create Figure 1. It is suspected, therefore, that a small proportion of claims which cease after two years are recorded as recoveries, when, in fact, they relate to the expiry of the benefit

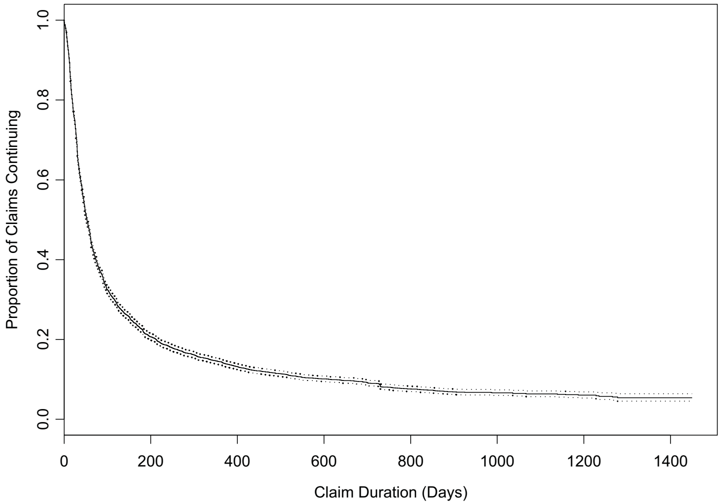

Figure 1. Kaplan-Meier estimate of the claim duration survival function

period. The effect is negligible, and subsequent analysis proceeds using the data as presented in Figure 1.

Figure 1 includes lines showing the 95% confidence intervals for the estimated claim duration survival function. From the Kaplan-Meier analysis we note that:

— there appears to be a non-zero long-term survival probability of about 0.07, which relates to lives who do not recover from their disability; and
— the Kaplan-Meier estimate of the survival function is very smooth, which suggests that parametric survival function models may work well in this context.

The results of an initial investigation of the impact of the various rating factors given in the table in Appendix A on claim termination rates are now presented. Again Kaplan-Meier estimation is used. Kaplan-Meier estimates of the survival function are created for the claims relating to levels of each of the rating factors which are deemed statistically significant predictors of claim duration in the Australian industry table for IP insurance claim rates (IAD 89-93). These factors are age, sex, occupation class, deferment period and smoker status.

Note that the Kaplan-Meier plots shown in Figures 2 to 6 represent one-way analyses of claim duration experience observed from 1995 to 1998 inclusive.

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

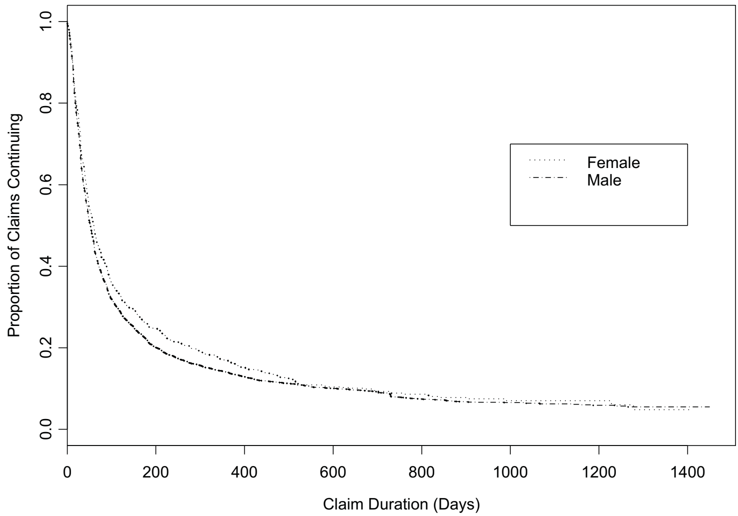

Figure 2. Kaplan-Meier survival function split by gender

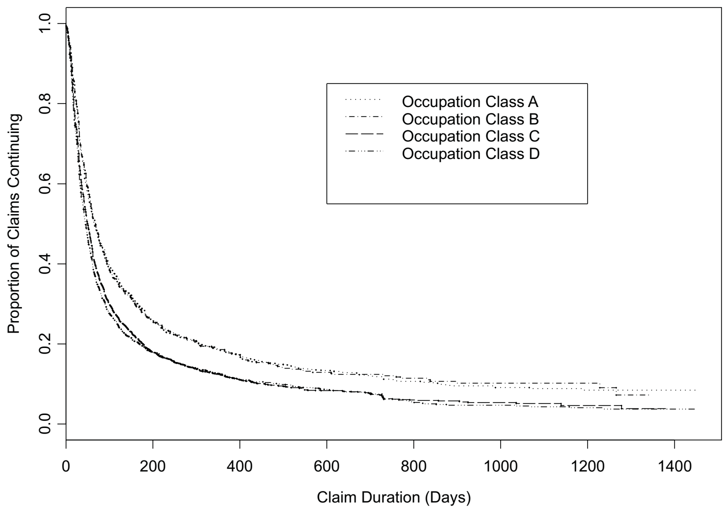

Figure 3. Kaplan-Meier survival function split by occupation class

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

The estimated survival functions for males and females are very close with mild evidence that males have higher recovery rates than females between six months and one and a half years after onset of disability, but that in the long term there is very little difference.

The Kaplan-Meier estimates by occupation class indicate that occupations can be grouped into two groups, 'A and B' compared with 'C and D'.

The most noticeable feature of the Kaplan-Meier estimates by deferred period is the significantly larger long-term claim probability associated with the longest (greater than three months) deferred period group. There is also evidence of longer claim durations amongst those claimants with policies which have deferred periods of one month. Note that these durations exclude the deferred period itself. The initial three-month continuous disability period which is required before claim payments commence under the relevant IP insurance contract means that this group contains only more seriously disabled individuals than are present in the other deferred period groups.

Examination of Figure 4 raises an interesting issue regarding IP insurance claims data with different deferred periods. It is clear from Figure 4 that the duration of claims arising from three-month deferred period policies is

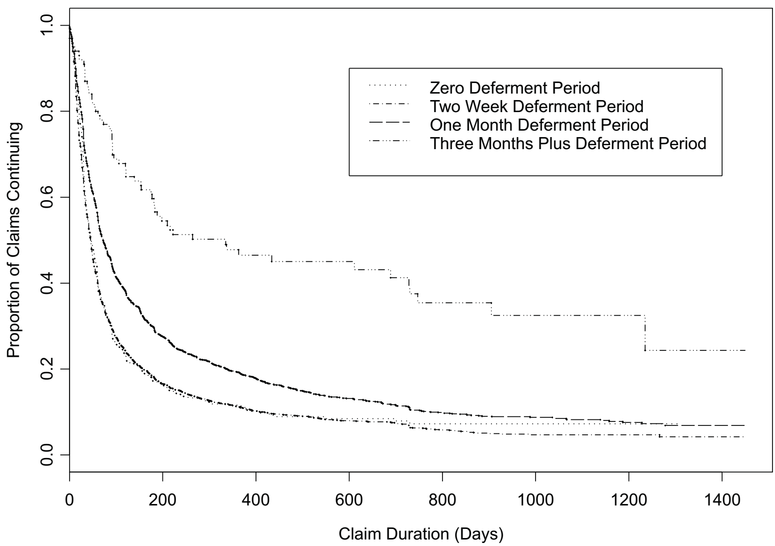

Figure 4. Kaplan-Meier survival function split by deferment period

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

Table 1. This table gives the probability that a period of disability extends for various durations, given that it has continued for a period of three months for policies with varying deferred periods

	Duration of disability				
	4 months	5 months	6 months	9 months	12 months
Deferred period 2 weeks	0.779	0.653	0.579	0.415	0.341
Deferred period 1 month	0.801	0.681	0.607	0.435	0.362
Deferred period 3 months	0.929	0.788	0.752	0.560	0.460

considerably longer than the duration of claims arising from either zero, two-week or one-month deferred period policies. A natural question to ask is whether, for example, the probability that a claim, where the policy has a three-month deferred period, will continue for two months from when payments commence is approximately equal to the probability that a claim, where the policy has a deferred period of one month and payments have been made for two months already, will continue for a further two months. This analysis attempts to see whether duration of disability is related to the deferred period of the policy. Table 1 reports the estimated survival probabilities for claims, conditional on the life being disabled for a period of three months for each of the two-week, one-month and three-month deferred periods. The estimated conditional survival probabilities are calculated using the estimated Kaplan-Meier survival functions shown in Figure 4. It is clear from Table 1 that longer deferred period claims lead to claims with longer duration. This indicates that insured lives who have been disabled for a period of three months are more likely to continue with their claim for long periods into the future if the deferred period on their policy is of longer duration.

Figure 5 demonstrates the effect of smoker status on claim duration. Smoker status does not appear to have a significant impact on the longevity of claims. This conclusion is the same as that reached by the IAAust Graduation SubCommittee of the Disability Committee in the development of the IAD 89-93 table.

Figure 6 shows the effect of age on claim duration. We note that there is a steady and consistent increase in the height of the survival function across the range of possible claim durations with increasing age. Of course, marginal analyses such as those presented above do not give a complete picture of how the covariates (jointly) relate to the claim termination rates. We now turn our attention to modelling the duration of claims using a regression model.

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

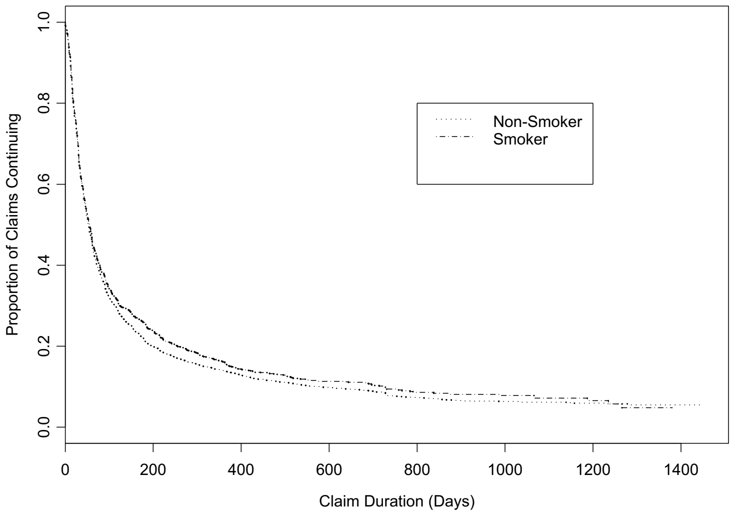

Figure 5. Kaplan-Meier survival function split by smoker status

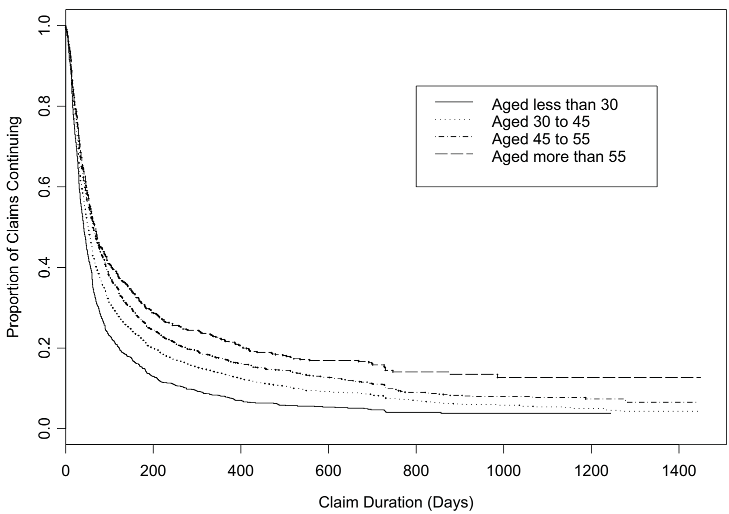

Figure 6. Kaplan-Meier survival function split by age groups

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

The most commonly used approach to model the effect of covariates on survival probabilities is the Cox Proportional Hazards Model, see Cox (1972).

The relation between the distribution of event time and the covariates, or risk factors z, can be described in terms of a model, in which the hazard rate at time t for an individual is:

$$\lambda(t; z) = \lambda_0(t)\exp(z\beta) \tag{1}$$

where $\lambda_0(t)$ is the baseline hazard rate, a function for which the mathematical form is not specified, which outputs the hazard function for the standard set of conditions $z = 0$ and β is a p-vector of unknown coefficients. For further description of the Cox model see, for example, Pitt (2006).

In the context of actuarial modelling of IP insurance, this model has two major shortcomings. First, the Cox model does not produce a closed form mathematical formula for either the predicted hazard rate or the survival function. A significant motivation for the modelling of claim durations is ultimately to produce premium and reserve recommendations using multiple state or some other form of modelling. In order for such work to be performed, it is preferable to have a mathematical model linking the various transitions between the states of the model. The second possible limitation of the Cox model is the potential invalidity of the proportional hazards assumption.

A number of methods for testing the validity of the proportional hazards assumption have been proposed. Methods proposed based on statistical tests have included:

— Cox (1972) suggested testing the statistical significance of an interaction between time (or log{time}) and the various covariates specified in the model. If such an interaction term is statistically significantly different from zero, then there is evidence that the impact of the covariate on survival duration varies with time; and
— Therneau & Grambsch (1994) and also Harrell (1986) have developed statistical tests based on the Schoenfeld partial residuals. These residuals are a measure of the difference between observed and expected values of the covariate at each time point. The idea of the tests is to detect a correlation between the Schoenfeld partial residuals (or some transformation thereof) and the rank order of the failure times.

Graphical procedures have also been proposed for testing the proportional hazards assumption. These have included:

— Andersen & Gill (1982) suggested a plot of cumulative baseline hazards in different groups;
— a plot of the difference of the log cumulative baseline hazard versus time; and

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

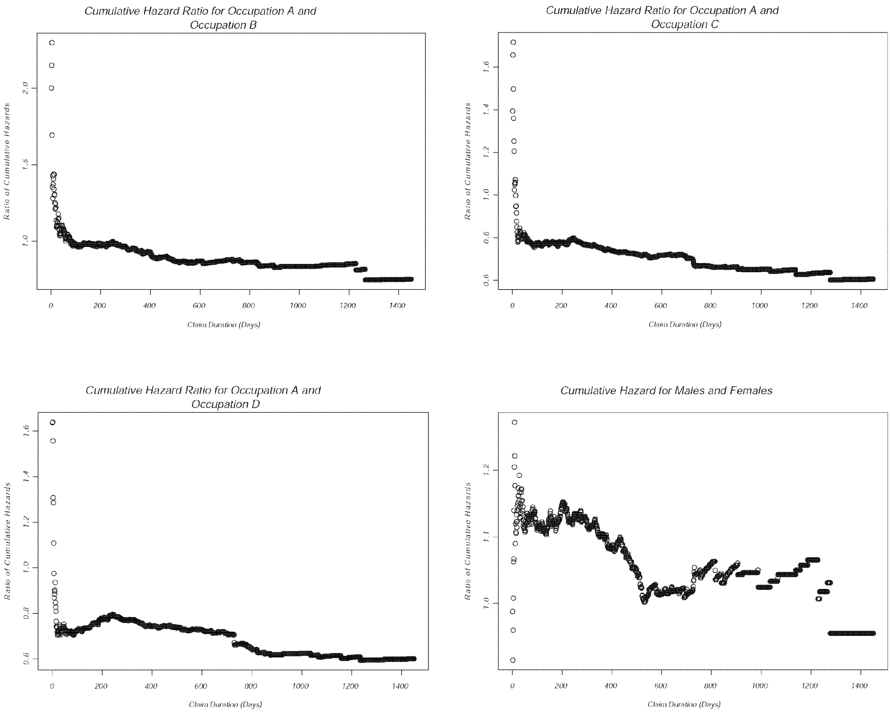

Figure 7. Cumulative hazard ratio plots for various levels of independent variables

— Arjas (1988) suggested a plot of the estimated cumulative hazard versus the number of failures.

For categorical covariates with only a small number of levels, graphical checks are more suitable than tests based on the correlation of residuals.

Integration of both sides of (1) leads to cumulative hazard rates, which are also proportional. Hence, if the proportional hazards assumption is valid, we would expect graphs depicting the ratios of cumulative hazards to be horizontal.

Note that the graphs in Figure 7 show the ratio $\dfrac{\Lambda(\text{Group } 1)}{\Lambda(\text{Group } 2)}$, where Group 1 represents the first named classification in the graph title and Group 2 refers to the second named covariate classification in the graph title, and $\Lambda(x)$ is an empirical estimate of the cumulative hazard for disabled lives with characteristic set x. So, for example, in the first graph in Figure 7, we

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

are considering the ratio $\dfrac{\Lambda\,(\text{Occupation A})}{\Lambda\,(\text{Occupation B})}$ as a function of claim duration. Again note that these cumulative hazard comparisons are one-way analyses.

The cumulative hazard ratio graphs for occupation class show immediately that the cumulative hazard ratio seems to decrease with time. The occupation class graphs all show cumulative hazard ratios less than one. This indicates that the cumulative hazards are greater for occupation classes B, C and D than for class A. These graphs also indicate that the higher rate of return to work for claimants in Occupation Classes B, C and D compared to Occupation Class A becomes more significant as the duration of claim increases.

The cumulative hazard ratio graphs for benefit rate also indicate that middle and higher income earners have a lower rate of return to work. The effect of middle income compared to low level income is close to proportional across time. It is difficult to discern a pattern in the cumulative hazard ratio of high income earners compared to low income earners. There is certainly evidence of non-proportionality in the cumulative hazard ratio. The effect of smoking on the hazard rate is close to proportional. The cumulative hazard ratio graph for gender indicates that males have a higher rate of return to work than females, but that the effect reduces significantly with the duration of claim. Hence, there is also evidence of non-proportionality in the effect of gender on the rate of return to work.

This graphical analysis shows clear violations of the assumption of proportional hazards for some of the key rating factors used in the proposed proportional hazards model. Extensions to the Cox regression model allowing for time varying regression coefficients have also been proposed, see Therneau & Grambsch (2000). These methods, however, will also not solve the problem of deriving a closed form mathematical expression for the predicted hazard rates. While a significant advance made by the Cox regression model is the ability to model covariates without having to specify the form for the baseline hazard, we decide not to continue with this approach, given that the underlying modelling assumptions are not satisfied by this particular dataset. We therefore proceed with a parametric analysis of claim termination rates.

4. Model and Results

In Section 2 we noted that the Kaplan-Meier estimates of the survival function were relatively smooth and also plateaued at long durations at a probability greater than zero, approximately 0.07. This non-zero long-run probability of survival is referred to in the literature, see Maller & Zhou (1995), as an 'immune probability'. This section describes survival analysis models, which take this feature of the data into account and therefore are suitable for describing claim termination rate data.

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

Maller & Zhou (1995) describe a statistical test for determining whether 'immunes' are present in data. Immunes are long-term survivors, and, in the case of IP insurance claim termination rate analysis, they refer to those individuals who become disabled and remain disabled for the long term. The method of Maller & Zhou is described for the case of the exponential distribution of claim duration, and involves comparing the likelihood for a model where the immune probability is zero with the maximum likelihood achievable when the immune probability is allowed to vary on the range from zero to one. The test statistic is based on the usual likelihood ratio test, and is written $d_n = -2\{l_n(\tilde{\theta}_{H_0}) - l_n(\tilde{\theta})\}$, where $\tilde{\theta}$ are the maximum likelihood estimates (MLEs) obtained from fitting an exponential mixture model, $\tilde{\theta}_{H_0}$ is the corresponding MLE under the null hypothesis of no immunes, and $l_n(\theta)$ is the log-likelihood function evaluated at θ. Maller & Zhou show that the asymptotic distribution of d_n, under the null hypothesis of no immunes, is a 50-50 mixture of a chi-square random variable with one degree of freedom and a point mass at zero. Applying this test to the claim termination rate data, we get a test statistic of $-2(-17{,}846.38 + 16{,}357.45) = 2{,}977.86$, highly significant under the chi-square point mass mixture distribution. This conclusion is not surprising after considering the Kaplan-Meier survival functions in Section 2. 'Total and permanent disability' is also a commonly insured event, and therefore long-duration claims are well known phenomena, and should also be expected to occur for claimants under IP policies.

Mixture models are based on fitting a parametric distribution to the claim durations for the lives who return to work. Define T to be a mixed random variable for the unknown claim duration of a disabled life who has just reached the end of the deferred period and is about to receive claim payments for the first time under this current period of disability. This distribution is then mixed with a point mass probability that the life will never return to work. For the case of the exponential mixture distribution, the density function is $f(t) = (1 - \pi)\lambda e^{-\lambda t}$, $t \geq 0$, and the associated distribution function is $F(t) = (1 - \pi)(1 - e^{-\lambda t})$, $t \geq 0$, where π is the immune probability and λ is the usual exponential rate parameter. The survival function for the exponential mixture distribution is $\pi + (1 - \pi)e^{-\lambda t}$, $t \geq 0$.

In order to achieve a good fit to the data, we will also consider a number of other potential mixture models from the generalised F distribution family (Peng *et al.*, 1998). The density functions and survival functions from the generalised F family are summarised in Table 2. All probability functions in the table are defined over $t \geq 0$.

In order to determine which family of mixture densities is most appropriate, each of the models identified was fitted to the claim duration data. At this stage, covariate information was ignored in the analysis. The fitted claim survival function was then compared with the Kaplan-Meier estimate of the survival function from Section 2.

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

Table 2. Summary of potential claim duration parametric distributions

Model	Density function	Survival function
Weibull mixture	$(1-\pi)(\lambda t)^{\alpha-1}\lambda\alpha\exp\{-(\lambda t)^{\alpha}\}$	$(1-\pi)\exp\{-(\lambda t)^{\alpha}\}+\pi$
Log-logistic mixture	$(1-\pi)\lambda\alpha(\lambda t)^{\alpha-1}\{1+(\lambda t)^{\alpha}\}^{-2}$	$\dfrac{1-\pi}{1+(\lambda t)^{\alpha}}+\pi$
Generalised log-logistic mixture	$(1-\pi)\dfrac{(t\lambda)^{\alpha s-1}\alpha\lambda}{\{1+(t\lambda)^{\alpha}\}^{2s}B(s,s)}$	no simple form
Extended generalised gamma mixture	$(1-\pi)\dfrac{\alpha\lambda(\lambda t)^{\alpha s_1-1}}{\Gamma(s_1)}$ $[s_1\exp\{-(t\lambda)^{\alpha}\}]^{s_1}$	no simple form
Gamma mixture	$(1-\pi)\dfrac{(s_1\lambda)^{s_1}t^{s_1-1}}{\Gamma(s_1)}\exp(-s_1\lambda t)$	no simple form
Lognormal mixture	$(1-\pi)\dfrac{\alpha}{t\sqrt{2\pi}}\exp\left[\dfrac{-\alpha^2\{\log(\lambda t)\}^2}{2}\right]$	$(1-\pi)[1-\Phi\{\alpha\log(\lambda t)\}]$
Generalised F mixture	$(1-\pi)\dfrac{\alpha}{t}B(s_1,s_2)^{-1}\left\{\dfrac{s_1(t\lambda)^{\alpha}}{s_2}\right\}^{s_1}\left\{1+\dfrac{s_1(t\lambda)^{\alpha}}{s_2}\right\}^{-(s_1+s_2)}$	no simple form

The results of fitting the log-logistic, generalised log-logistic and generalised F mixture models to the claim duration data are given in Table 3 and Figure 8. The results of fitting the remaining models from Table 2 are shown in Pitt (2006).

From the results shown here and in Pitt (2006), it is clear that the three-parameter distributions, excluding the log-logistic distribution, all significantly overestimate the survival function for claims of duration less than six months. The PP plots highlight this deficiency very clearly. This phenomenon occurs because the first six months after claim inception accounts for approximately 80% of claim terminations. The extended generalised gamma fit exhibits similar properties to the Weibull, gamma and lognormal models. The log-logistic distribution provides the best three-parameter distribution

Table 3. Assessment of fit of parametric density to claim duration

Model	Maximised log-likelihood	R-squared for PP plot	Akaike's information criterion (AIC)
Log-logistic mixture	−15,598.72	99.609%	31,205.44
Generalised log-logistic mixture	−15,550.93	99.680%	31,109.86
Generalised F mixture	−15,478.77	99.915%	30,967.54

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

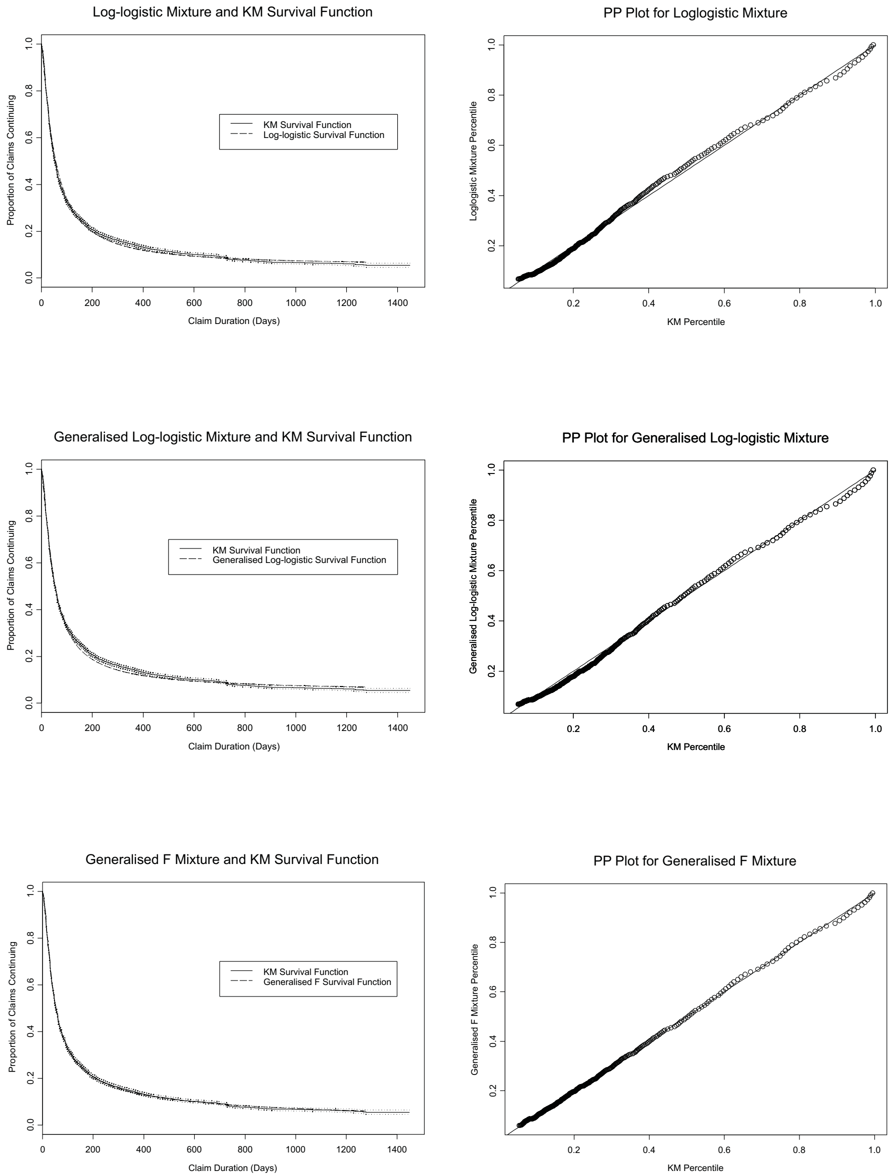

Figure 8. Assessment of fit of parametric density to claim duration

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

summary of the data. The generalised log-logistic distribution provides only a marginal improvement over the log-logistic distribution. The generalised F is clearly the best of the distributions considered in terms of fit. Note that the generalised F distribution leads to a very small estimated immune probability. However, the tail of the standard (non-mixed) generalised F distribution is sufficiently long that the resulting model still predicts that a small percentage of claims will continue for a long period. The fitted model predicts a 5.1% probability of claim continuation after ten years.

Based on the above findings and those in Pitt (2006), the analysis of the impact of covariates on claim duration will be performed using the log-logistic, generalised log-logistic and generalised F mixture distributions. We now describe the mixture models which are fitted and tested in this section. Assume that T is a random variable for the time (measured in days) it takes for a new IP insurance claimant to return to work. We consider the transformation $Y = \log T$. The survival function for Y is modelled using:

$$S(y) = (1 - \pi)S_u(y) + \pi \tag{2}$$

where $S_u(y)$ is the survival function of Y, given that the person returns to work. The density function for Y is:

$$f(y) = (1 - \pi)f_u(y) \tag{3}$$

where $f_u(y)$ is the density function for the time until return to work, conditional on the individual returning to work at some stage. The long-term disability probability π is modelled using a logistic regression:

$$\mathrm{E}(\pi \mid Z) = \frac{1}{1 + \exp(Z'\gamma)}$$

where Z is a covariate vector and γ is a vector of regression coefficients. The part of the model relating to return to work is often called the accelerated failure part of the survival model in the literature. The random variable T is said to have a generalised F distribution, with μ and σ as location and scale parameters and s_1, s_2 as shape parameters, if $W = \dfrac{\log T - \mu}{\sigma}$ is the logarithm of a random variable having an F distribution, with $2s_1$ and $2s_2$ degrees of freedom. The density of W is then:

$$f(w; s_1, s_2) = \left(\frac{s_1 e^w}{s_2}\right)^{s_1} \left(1 + \frac{s_1 e^w}{s_2}\right)^{-(s_1+s_2)} B(s_1, s_2)^{-1} \tag{4}$$

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

and the survival function is:

$$S(w; s_1, s_2) = \int_0^{s_2(s_2+s_1e^w)^{-1}} x^{s_2-1}(1-x)^{s_1-1}B(s_2, s_1)^{-1}dx \tag{5}$$

where $-\infty < \mu < \infty, \sigma > 0,\ s_1 > 0,\ s_2 > 0$ and $B(s_1, s_2)$ is the beta function evaluated at s_1 and s_2. For claimants who may return to work, we assume that the failure time T follows a generalised F distribution where the covariate vector X impacts the failure time through the relationship $\mu = X'\beta$, where β is a vector of regression coefficients. The model is fitted using maximum likelihood estimation. The log-likelihood function for the model is:

$$\begin{aligned} L(s_1, s_2, \sigma, \beta, \gamma) = \sum_{i=1}^{n} [&\delta_i \log\{f(y_i; x_i, z_i, s_1, s_2, \sigma, \beta, \gamma)\} \\ &+ (1-\delta_i)\log\{S(y_i; x_i, z_i, s_1, s_2, \sigma, \beta, \gamma)\}] \end{aligned} \tag{6}$$

where δ_i is an indicator variable equal to one if the claimant is observed to return to work and zero otherwise. Note that if $s_1 = s_2 = s$, then the generalised F distribution reduces to the generalised log-logistic distribution. If, in the generalised log-logistic, we have $s = 1$, then the model further reduces to the log-logistic distribution.

The covariates described in Appendix A, along with all possible two-way interaction variables, were tested in each of the three model families described above. Model selection was performed on the basis of the marginal significance of regression variables. This is equivalent to testing the statistical significance of estimated regression coefficients using the likelihood ratio test. Two-way interaction variables were also considered as possible regression variables. However, due, most likely, to the high correlation between the interaction variables and the underlying main effects, these interaction variables did not continue to have a significant effect throughout the model selection process, and hence were not included in the final model.

The only continuous predictor used in the model was age. In order to model the effect of age on the return to work probability properly, three variables were used. The first variable was a simple linear predictor, based on the age in years of the claimant at the time when the disability commenced. The remaining two variables used were break-point predictor terms. These terms enable a different sensitivity of the return to work probability to increases in age at different levels of age. The terms were labelled *ageind* and *ageind2* in S-Plus. The variable *ageind* is equal to the age of the claimant if the claimant is 'young' and *ageind2* is equal to the age of the claimant if the claimant is 'old'. The definitions of 'young' and 'old' were formed by

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

Table 4. Assessment of accelerated failure and mixture models for claim duration

	Maximised log-likelihood	Likelihood ratio test statistic relative to generalised *F* model	AIC
Accelerated failure: no covariates; no logistic model			
Generalised *F*	−15,478.44	—	30,964.9
Generalised log-logistic	−15,631.24	305.6	31,268.5
Log-logistic	−15,701.86	446.8	31,407.8
Accelerated failure: no covariates; logistic: no covariates			
Generalised *F*	−15,478.77	—	30,967.5
Generalised log-logistic	−15,550.93	144.3	31,109.9
Log-logistic	−15,598.72	239.9	31,203.4
Accelerated failure: covariates included; logistic: no covariates			
Generalised *F*	−15,297.81	—	30,627.6
Generalised log-logistic	−15,343.79	92.0	30,717.6
Log-logistic	−15,403.06	210.5	30,834.12
Accelerated failure: covariates included; logistic: covariates included			
Generalised *F*	−15,260.81	—	30,565.6
Generalised log-logistic	−15,291.47	61.3	30,624.9
Log-logistic	−15,351.74	181.9	30,743.5

maximising the log-likelihood of the resulting model. The definitions used in the final model are *ageind* is the age for claimants below age 29. The variable *ageind2* is equal to the age for claimants above age 44.

The likelihood ratio test and the Akaike's Information Criterion (AIC) were used to assess the models fitted. The results are summarised in Table 4.

Note also that these likelihood ratio test statistic values can be compared to critical values derived from the chi-squared distribution. This statistical test will be conservative, because the true distribution of the likelihood ratio test statistic has greater density at zero and the shortest durations than does a chi-square variable.

It is clear from Table 4 that the generalised *F* mixture model, with covariates for both the accelerated failure time part of the model and the logistic part of the model, is optimal. A summary of this fitted model is given in Tables 5 and 6.

5. Discussion

The majority of the regressors shown in Tables 5 and 6 have a statistically significant effect on the rate of return to work at the 5% significance level. For variables which are highly subdivided, for example occupation, which has four classes, the statistical significance of the variable

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

Table 5. Accelerated failure model regression coefficients

Generalised *F* mixture model

The maximum loglikelihood is −15,256.62

Terms in the accelerated failure time model

	Coefficients	Std. err.	*z*-score	*p*-value
Log(scale)	0.00278	0.003667	0.7579	0.4485249
(Intercept)	−0.00354	0.147557	23.4308	0.0000000
age	0.00278	0.003667	0.7579	0.4485249
ageind	−0.00354	0.002286	−1.5489	0.1213947
ageind2	0.00202	0.001226	1.6476	0.0994444
occupB	0.12864	0.063015	2.0414	0.0412076
occupC	−0.04742	0.042424	−1.1178	0.2636703
occupD	−0.12454	0.044126	−2.8224	0.0047672
benrate2	0.06753	0.040174	1.6809	0.0927729
benrate3	0.11961	0.041763	2.8639	0.0041843
benrate4	0.27448	0.056277	4.8774	0.0000011
benratetop2	0.12041	0.050253	2.3961	0.0165717
sick	0.04555	0.031021	1.4685	0.1419671
defpd2	0.35779	0.033275	10.7526	0.0000000
defpd3	1.02768	0.183896	5.5884	0.0000000

Table 6. Logistic model regression coefficients

Terms in the logistic model

	Coefficients	Std. err.	*z*-score	*p*-value
(Intercept)	10.00406	1.586363	6.3063	0.0000000
age	−0.07391	0.015437	−4.7875	0.0000017
smokernew	−0.92219	0.298505	−3.0894	0.0020058
conttypenew1	−0.58790	0.271987	−2.1615	0.0306560
sick	−2.51927	1.482109	−1.6998	0.0891704
defpd0	−2.52043	1.643859	−1.5332	0.1252161
defpd2	−0.83440	0.301255	−2.7697	0.0056100
defpd3	−2.54072	0.424056	−5.9915	0.0000000

is strongly affected by the amount of data for that particular class. For that reason, we note that occupation class C does not appear to have a significantly different rate of return to work than occupation class A, despite the contrasting results from the Kaplan-Meier analysis shown in Figure 3.

It is also of interest that there are independent variables which are statistically significant predictors of the rate of return to work in the accelerated failure time part of the model which are not significant in the logistic part of the model. In particular, the model shows that smoker status, which until now in Australian studies has not been considered a significant

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

determinant of claim termination rates, leads to a statistically significant increase in the probability of long-term disability.

Apart from the likelihood ratio test, it is also possible to assess the quality of the fit of the model by dividing the data into groups according to the values of the covariates included in the final model. Out of the 8,863 individuals in the study, 61 were found to possess all of the following characteristics: aged between 35 and 45; disability benefit of less than $2,000 per month; disability caused by sickness; deferred period of two weeks; occupation class A; and non-smoker. For these 61 lives, the Kaplan-Meier fit to the survival function is compared to the survival function predicted by the generalised *F* model. The result of this comparison is shown in Figure 9, where 95% confidence bands have been included around the Kaplan-Meier estimate.

The fit of the generalised *F* distribution is clearly very good except at the shortest durations, where the model predicts higher rates of return to work than does the empirical Kaplan-Meier survival function. Since pricing and reserving for IP insurance are impacted most by long-duration claims, this imperfect fit at the shorter durations has less financial consequences for a life

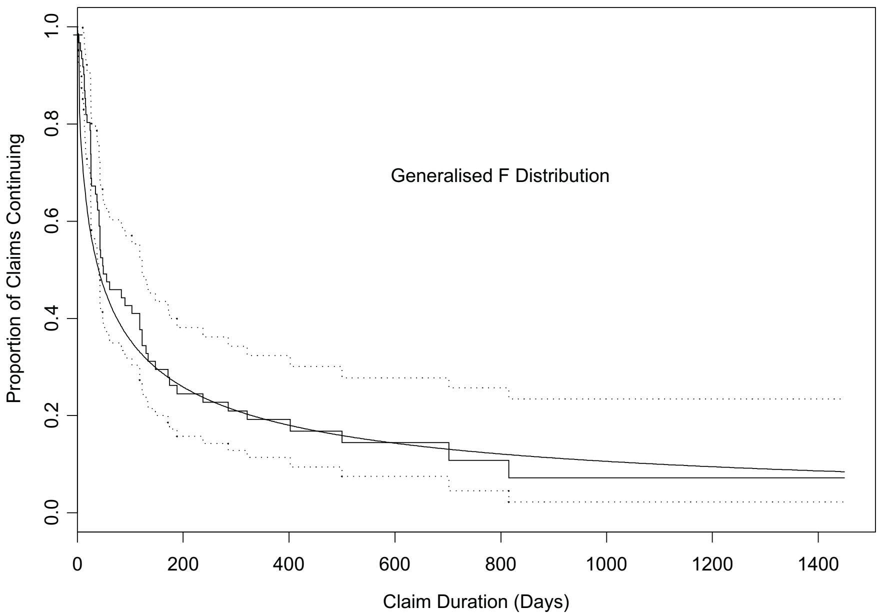

Figure 9. Comparison of actual and fitted rates for the generalised *F* distribution

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

office than would imprecise model fitting in the tail of the claim duration probability distribution, and so may not be of practical significance.

In Section 2 we demonstrated that the proportional hazards assumption of the Cox regression model was not satisfied by the covariates in the IP insurance claim termination data. The impact of this assumption not being satisfied on the fit of the Cox regression model is shown in Figure 10. This graph compares the same data as were used in Figure 9 to compare the empirical Kaplan-Meier survival function with the survival function predicted using the Cox regression.

This graph shows clear evidence that the Cox regression model estimates claim termination rates which are significantly higher than the Kaplan-Meier estimate between durations six months and two years.

A useful way to compare the fits of various models, given that the aim of the modelling is premium rating, is to compare the predicted expected present value of an annuity payable to a disability annuitant throughout his/her period of disability. We consider a disability income insurance policy with a four-year benefit period. The annuity is assumed to be payable continuously, with the valuation performed at a force of interest of 5% per annum. Mortality is ignored, which is a reasonable assumption at this stage,

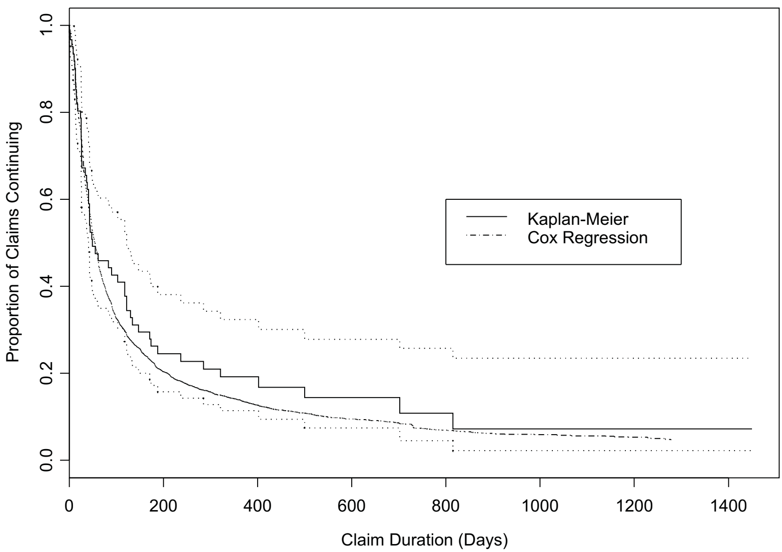

Figure 10. Comparison of Kaplan-Meier and Cox regression claim duration models

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

Table 7. Annuity value comparison for three model fitting procedures

Deferred period	Kaplan-Meier survival function	Cox regression model	Generalised F mixture model
4 weeks	0.6163	0.4739	0.5837
2 weeks	0.5176	0.3851	0.4977
1 week	0.3674	0.3092	0.3123

given that we are considering lives aged between 35 and 40, and also that our aim is to assess the relative merits of the Cox regression model and the generalised F mixture model in describing claim durations. Table 7 gives the expected present value of an annual annuity of one dollar payable throughout the period of disability under each model. It is clear that the generalised F mixture model is preferable in this case to the Cox regression model, as evidenced by a much closer estimate of the annuity value to the underlying annuity value. The results for the same policy with a two-year and one-year benefit period provide a similar message, and are also shown in Table 7.

6. Conclusion

One of the most noteworthy features of this analysis is the difference in statistically significant regressor variables between the accelerated failure time part of the model and the logistic regression for the immune probability component of the model. Comparison with the results from the CMI for claim longevity indicates a close alignment at short claim durations, although some deviation at longer claim durations. The Kaplan-Meier curves shown in Section 3 broadly match claim survival curves generated based on the CMI findings in CMIR12 (CMIB, 1991). Further research conducted by the author extends this investigation to quantile regression, where the significance of rating variables is assessed at various quantiles of the distribution of claim durations, rather than just at the conditional mean.

Acknowledgement

This work was prepared as part of a PhD thesis at the Australian National University. The author would like to thank Dr M.A. Martin for his supervision and the anonymous referees for their comments, which certainly improved this paper significantly. The author is also grateful to the Institute of Actuaries of Australia for provision of the data used in this paper.

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

References

Andersen, P.K. & Gill, R.D. (1982). Cox's regression model counting process: a large sample study. *Annals of Statistics*, **10**, 1100-1120.

Arjas, E. (1988). A graphical method for assessing goodness of fit in Cox's proportional hazards model. *Journal of the American Statistical Association*, **83**, 204-212.

CMIB (1991). *Continuous Mortality Investigation Report No 12*. Institute and Faculty of Actuaries, U.K.

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

Grambsch, P.M. & Therneau, T.M. (1994). Proportional hazard tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515-526.

Gregorius, F.K. (1993). Disability insurance in the Netherlands. *Insurance: Mathematics and Economics*, **13**, 101-116.

Harrell, F.E. (1986). The PHGLM procedure. *SUGI Supplemental Library Guide, Version 5 edition*, Cary, NC: SAS Institute Inc.

Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.

Maller, R. & Zhou, X. (1995). *Survival analysis with long term survivors.* Wiley, New York.

Peng, Y., Dear, K.B. & Denham, J.W. (1998). A generalized F mixture model for cure rate estimation. *Statistics in Medicine*, **17**(8), 813-830.

Pitt, D. (2006). Actuarial models for the analysis of disability income insurance. University of Melbourne working paper series: http://www.economics.unimelb.edu.au/actwww/wps2006.html

Robinson, M.A. (1988). The 1985 CIDA disability table. *Institute of Actuaries of Australia Quarterly Journal*, December 1988.

Segerer, G. (1993). The actuarial treatment of disability risk in Germany, Austria and Switzerland. *Insurance: Mathematics and Economics*, **13**, 131-140.

The Institute of Actuaries of Australia (1997). Report of the Disability Committee. Transactions of the Institute of Actuaries of Australia, 489-576.

Therneau, T.M. & Grambsch, P.M. (2000). *Modeling survival data: extending the Cox model.* Springer, New York.

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press

APPENDIX A

Field	Description	Variables (S-plus names)
Duration	Duration of the claim (recorded in days), which is the number of days from when the sickness began until recovery (or censoring), less the deferment period	durn2
Age	Age at the date of claim commencement	age
Terminate	An indicator of whether the claim was observed to terminate or was censored	terminate
Disability definition	Own occupation for which the insured person is reasonably suited by education, training or experience, or any occupation after an initial period (indicator variable for any occupation after initial period)	poldesnew3
Sex	Indicator variable for gender; Male = 1	sex1
Occupation class	Occupation is grouped into four levels: A, B, C or D as described in IAAust Disability Reports	occupA, occupB, occupC, occupD
Frequency of benefit payment	Classified as (1) weekly, (2) monthly or (3) annually	benhp1, benhp2
Benefit rate	Monthly benefit rate in dollars	benrate
Benefit type	Level or Increasing Benefits (indicator variable for increasing benefits)	bentypnew2
Medical evidence	Medical exam required or automatic acceptance (indicator for medical exam required)	medevid1
Contract type	Level premiums or stepped premiums (indicator variable for level premiums)	conttypenew1
Smoker status	Smoker or non-smoker (indicator variable for smoker)	smokernew
Sickness or accident	Sickness claim or accident related claim (indicator is for sickness)	sick
Deferred period	Classified according to defpd0 (zero day), defpd1 (base level and deferment period between one and 27 days), defpd2 (28 to 89 day deferment period) and defpd3 (deferment period in excess of 90 days)	defpd0 defpd1 defpd2 defpd3

https://doi.org/10.1017/S1748499500000233 Published online by Cambridge University Press