Modeling the Claim Duration of Income Protection Insurance Policyholders Using Parametric Mixture Models

Abstract

This paper considers the modeling 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 was 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 were 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.

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 insurance policies. Claim durations have a significant impact on both the pricing and reserving calculations routinely made by insurers. Ultimately, inaccurate modeling 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 modeling claim durations that 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 income protection insurance policies.

The first significant step in the modeling of claim termination rates was the production of a number of disability tables based on industry-wide data. In the US, we have seen the production of the Commissioner's Disability Tables, CDT (1964), which were updated with the production of the Commissioner's Individual Disability Table A, CIDA (1985). This US derived table is based on data from twenty 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 UK, the Continuous Mortality Investigation Bureau (CMIB) has produced a number of reports that describe and model both mortality and morbidity experience. Most notably the twelfth report of the CMI, CMIR12 (1991), describes the development of a multiple state model for the description of income protection insurance. CMIR12 contains a graduation by mathematical formula of claim recovery

rates. The mathematical formula employed by the CMI in this report modeled 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 IAD1989-93, was produced by a subcommittee of the Disability Committee of the Institute of Actuaries of Australia in 1995. The claim recovery rates were modeled 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 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 income protection 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 modeled explicitly in these countries. In order to predict the expected present value of claim payments under an income protection 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 modeling of claim durations with the use of survival analysis. After this introduction, Section 2 will describe the income protection insurance policy data used in the paper. Section 3 will provide the results of some initial analysis of this data and describe the modeling strategies implied by these initial analyses. Section 4 will describe the main modeling 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 modeling. Section 6 will conclude the paper and provide some avenues for further research in this area.

2. The Institute of Actuaries of Australia Claim Duration Data

The IAAust income protection insurance policy database contains information on policyholders who have purchased insurance from the main Australian providers of this form of insurance. There are about twenty different insurers that provide data to this database on an annual basis. Data is 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 that 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 the end of calendar year 1998, whichever occurred first. The data set contains information including the duration of each policyholder’s claim (if a claim was made), 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 Report of the IAAust Disability Committee, 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 8863 claims recorded, 7771 (88%) related to terminated claims, the remainder being censored. 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 that were lost at the end of each of 1995, 1996 and 1997 and that are unable to be followed further. This issue arose due to changes in claim codes adopted by companies that provided this data to the IAAust Life and Risk Committee at the end of particular calendar years. Most of these claims were able to be traced by matching claims from one 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 (less than 1%) were unable to 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 8863 claimaints included in the dataset, 2409 (27.2%) were in occupation class A, 667 (7.5%) were in occupation class B, 3165 (35.7%) were in occupation class C, and 2622 (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 Report of the IAAust Disability Committee, 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 that 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 (see Kaplan and Meier, 1958) survival curves 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 claim onset, and not time since payment of disability benefits begins.

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 that cease after two years are recorded as recoveries, when in fact they relate to the expiry of the benefit period. The effect is negligible and subsequent analysis proceeds using the data as presented in Figure 1.

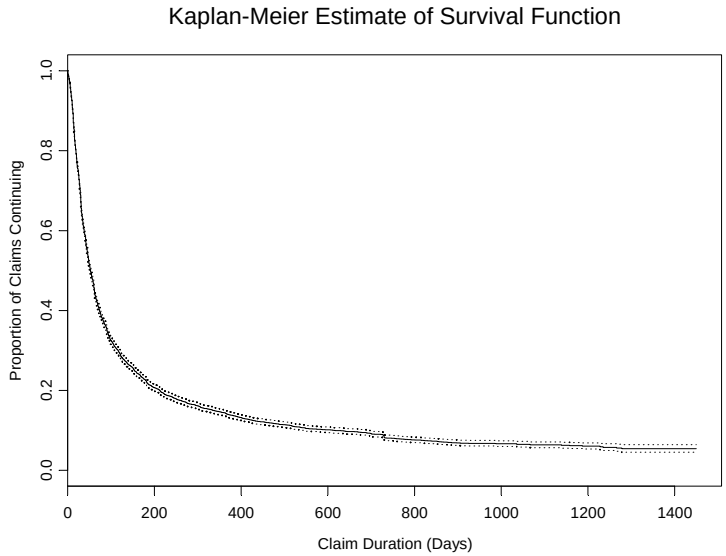


Figure 1 Kaplan-Meier Estimate of the Claim Duration Survival Function

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. This probability relates to lives who do not recover from their disability; and

- the Kaplan-Meier estimate of the survival function is very smooth. This 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, except for age, that are deemed statistically significant predictors of claim duration in the Australian industry table for disability income insurance claim rates (IAD89-93). These factors are sex, occupation class, deferment period and smoker status.

Note that the Kaplan-Meier plots shown in Figures 2 to 5 represent one-way analyses of claim duration experience observed from 1995 to 1998 inclusive.

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 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”.

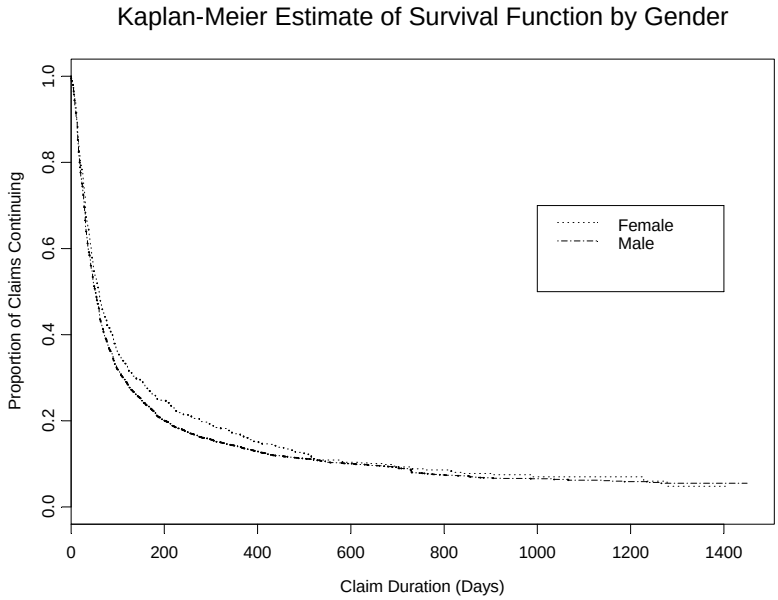


Figure 2 Kaplan-Meier Survival Function Split by Gender

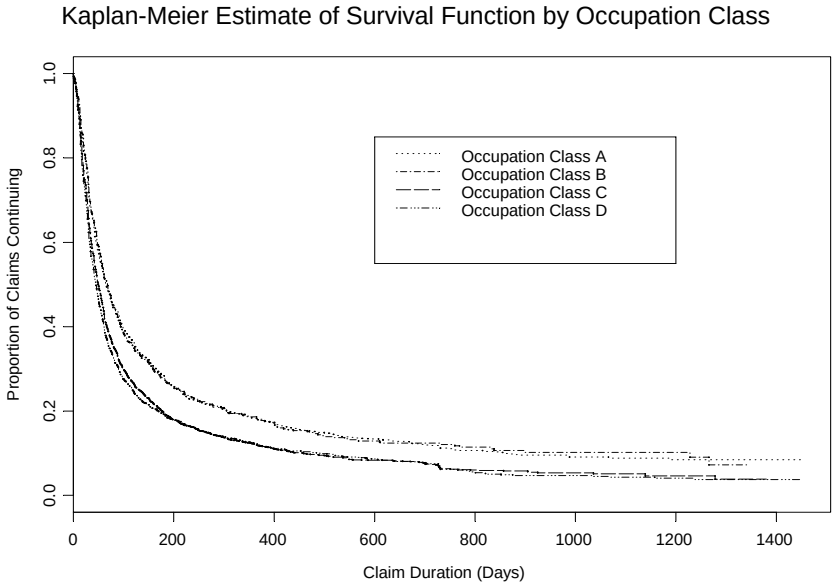


Figure 3 Kaplan-Meier Survival Function Split by Occupation Class

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 that have deferred periods of one month across all claim durations. The three month deferred period has the longest predicted claim durations. Note that these durations exclude the deferred period itself.

The initial three month continuous disability period that is required before claim payments commence under the relevant income protection insurance contract means that this group contains only more seriously disabled individuals than are present in the other deferred period groups.

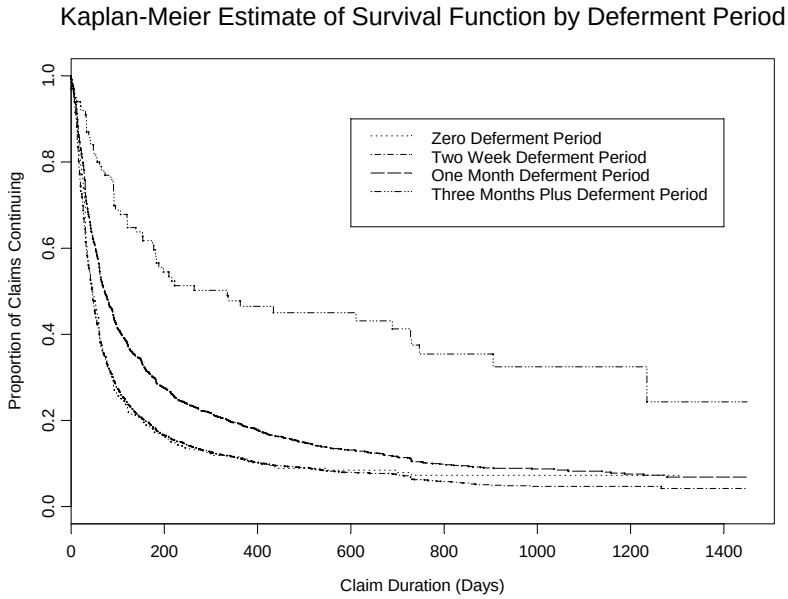


Figure 4 Kaplan-Meier Survival Function Split by Deferment Period

Figure 5, shown below, demonstrates the effect of smoker status on claim duration.

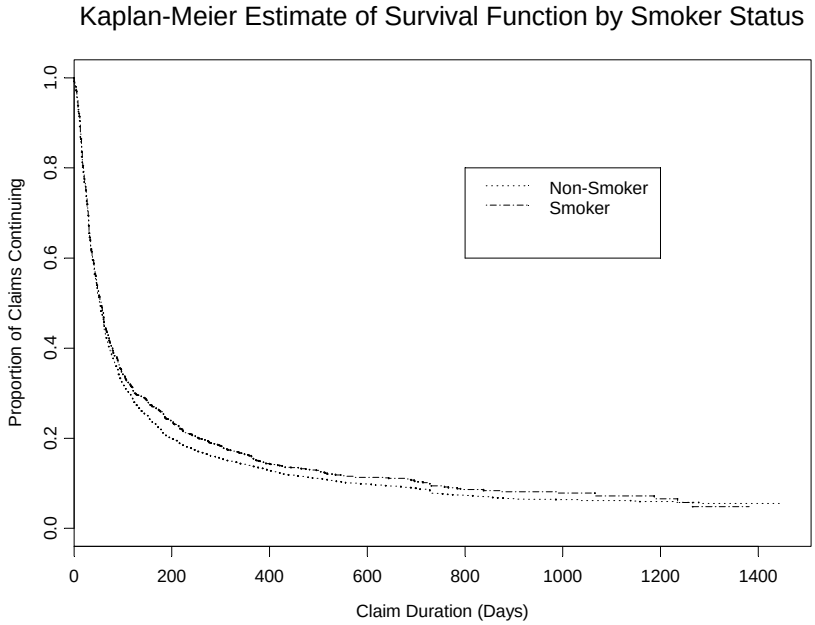


Figure 5 Kaplan Meier Survival Function Split by Smoker Status

Smoker status does not appear to have a significant impact on the longevity of claims. This conclusion is the same as reached by the IAAust Graduation SubCommittee of the Disability Committee found in the development of the IAD89-93 table. 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 modeling the duration of claims using a regression model.

The most commonly used approach to model the effect of covariates on survival probabilities is the Cox Proportional Hazards Model, (Cox, 1972). The major theoretical development that this model provides is the ability to model covariate effects in the presence of censored observations.

The data for a Cox regression model, based on a sample of size *n*, consists of $(t_j, \delta_j, z_j), j = 1,2,...,n$ where t_j is the time on study for the *j*th individual, δ_j is the event indicator ($\delta_j = 1$ if the event has occurred and $\delta_j = 0$ if the observation is right-censored) and z_j is the *p*-vector of covariates or risk factors for the *j*th individual.

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. The parameters are estimated using the maximum (partial) likelihood technique. Importantly, the Cox model assumes that the hazards are proportional; in other words, the impact of covariates on the dependent variable, in this case rate of return to work, under the Cox model do not vary with the duration of claim.

In the context of actuarial modeling of disability income 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 modeling of claim durations is to ultimately produce premium and reserve recommendations using multiple state or some other form of modeling. 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

- Grambsch and Therneau (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 (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
- Arjas (1988) suggested a plot of the estimated cumulative hazard versus 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 which follow in Figure 6 show the ratio $\frac{\Lambda(\text{Group 1})}{\Lambda(\text{Group 2})}$, 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 6, we are considering the ratio $\frac{\Lambda(\text{Occupation A})}{\Lambda(\text{Occupation B})}$ 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 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.

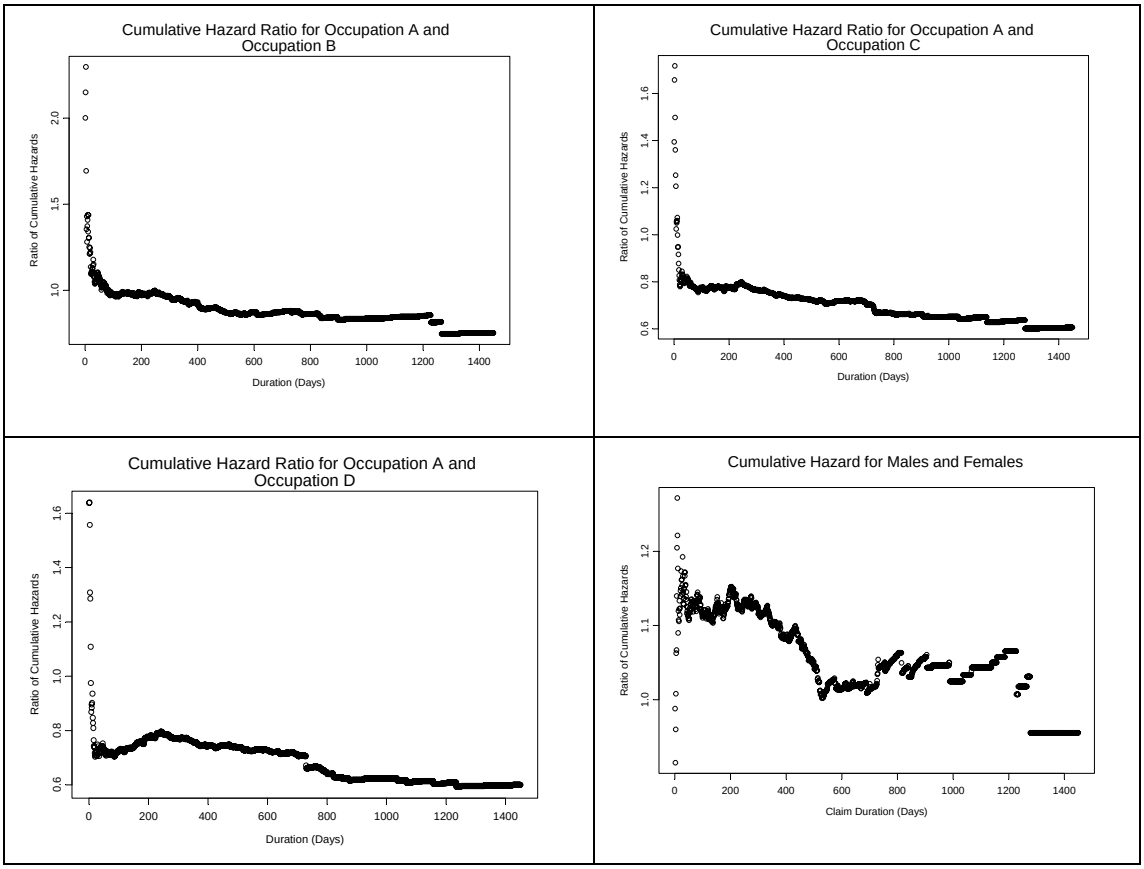


Figure 6 Cumulative Hazard Ratio Plots for Various Levels of Independent Variables

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 (Therneau, 2000). These methods however will also not solve the problem of deriving a closed form mathematical expression for the predicted hazard rates. 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, Maller and 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.

Maller and Zhou (1995) describe a statistical test for determining whether "immunes" are present in data. Immunes are long-term survivors and in the case of disability

income insurance claim termination rate analysis, they refer to those individuals who become disabled and remain disabled for the long term. The method of Maller and 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 $d_n = -2\left\{l_n\left(\tilde{\theta}_{H_0}\right) - l_n\left(\tilde{\theta}\right)\right\}$ 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 and 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 $-2(-17846.38+16357.45) = 2977.86$, 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 phenomenon and should also be expected to occur for claimants under income protection policies.

Mixture models are based on fitting a parametric distribution to the claim durations for the lives that return to work. Define *T* to be a mixed random variable for the unknown claim duration of a disabled life that 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 $F(t) = (1-\pi)\left(1-e^{-\lambda t}\right), t \geq 0,$ 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 Generalized F distribution family, Peng (1998). The density functions, cumulative distribution functions and survival functions from the Generalised F family are summarised in Table 1 below. All probability functions in the table are defined over $t \geq 0$.

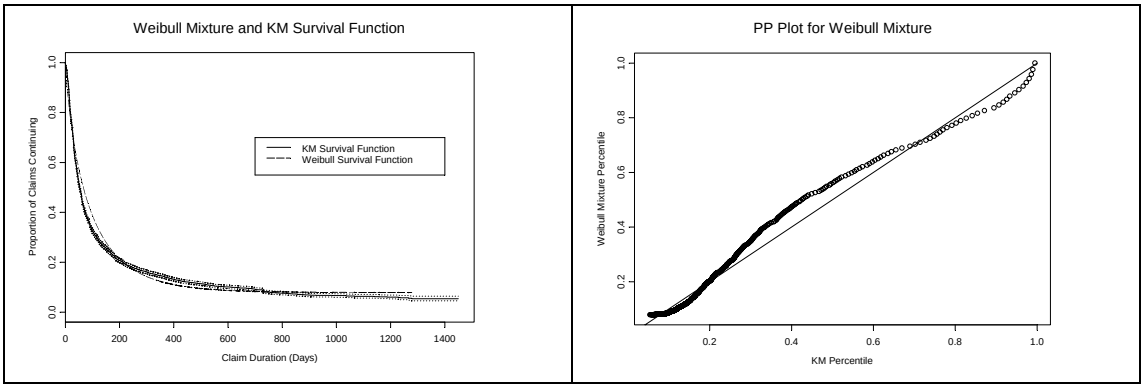
Model	**Density Function**	**Cumulative Distribution Function**	**Survival Function**
Weibull Mixture	$(1-\pi)(\lambda t)^{\alpha-1}\lambda\alpha\exp\left\{-(\lambda t)^{\alpha}\right\}$	$(1-\pi)\left[1-\exp\left\{-(\lambda t)^{\alpha}\right\}\right]$	$(1-\pi)\exp\left\{-(\lambda t)^{\alpha}\right\}$ $+\pi$
Log-Logistic Mixture	$(1-\pi)\lambda\alpha(\lambda t)^{\alpha-1}\left\{1+(\lambda t)^{\alpha}\right\}^{-2}$	$(1-\pi)\left\{1-\frac{1}{1+(\lambda t)^{\alpha}}\right\}$	$\frac{1-\pi}{1+(\lambda t)^{\alpha}}+\pi$

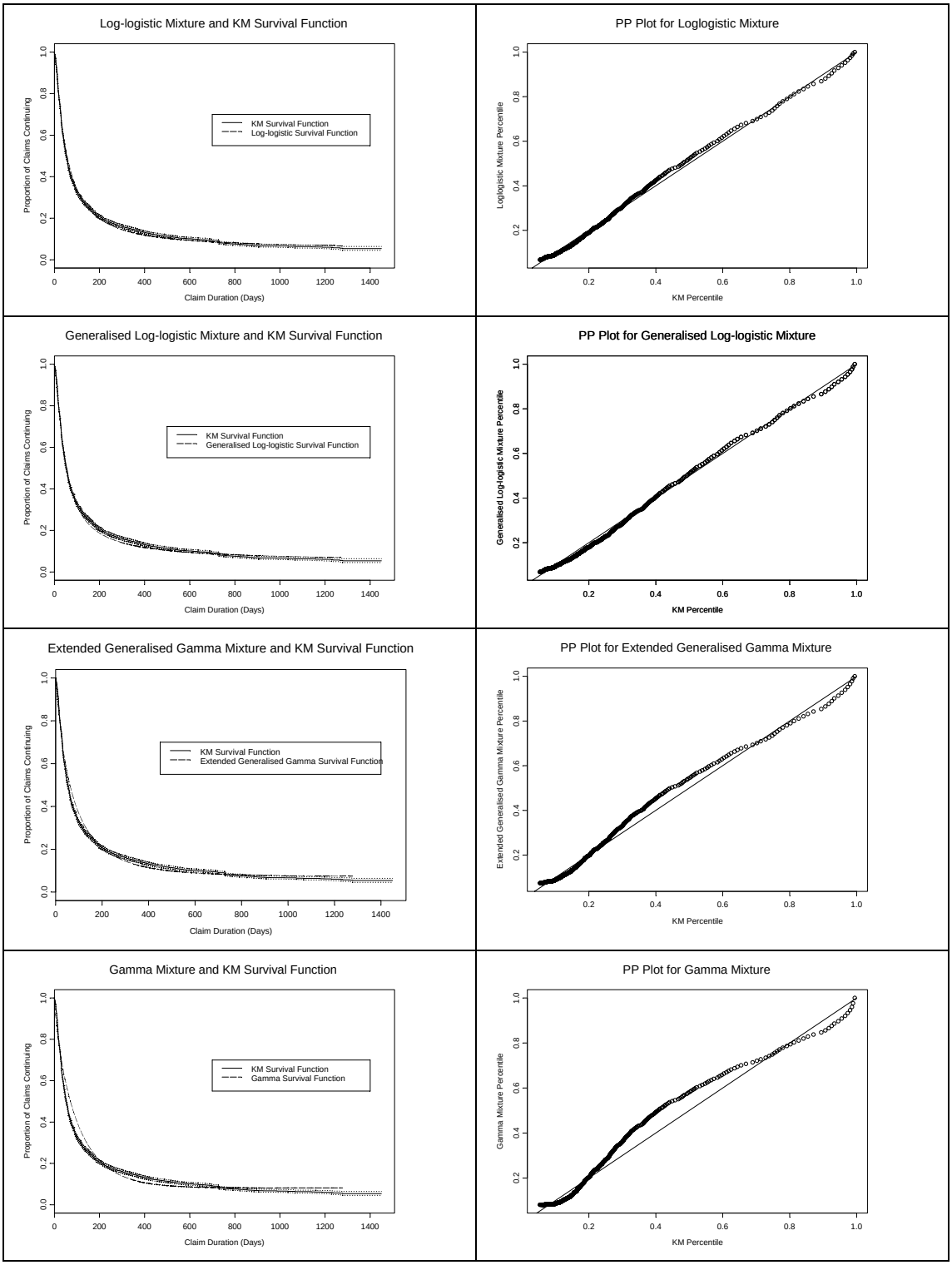
Generalised Log-logistic Mixture	$(1-\pi)\dfrac{(t\lambda)^{\alpha s-1}\alpha\lambda}{\left\{1+(t\lambda)^{\alpha}\right\}^{2s}B(s,s)}$	no simple form	no simple form
Extended Generalised Gamma Mixture	$(1-\pi)\dfrac{\alpha\lambda(\lambda t)^{\alpha s_1-1}}{\Gamma(s_1)}$ $\left[s_1\exp\left\{-(t\lambda)^{\alpha}\right\}\right]^{s_1}$	no simple form	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	no simple form
Lognormal Mixture	$(1-\pi)\dfrac{\alpha}{t\sqrt{2\pi}}\exp\left[\dfrac{-\alpha^2\left\{\log(\lambda t)\right\}^2}{2}\right]$	$(1-\pi)\Phi\left\{\alpha\log(\lambda t)\right\}$	$(1-\pi)\left[1-\Phi\left\{\alpha\log(\lambda t)\right\}\right]$
Generalised F Mixture	$(1-\pi)\dfrac{\alpha}{t}B(s_1,s_2)^{-1}\left\{\dfrac{s_1(t\lambda)^{\alpha}}{s_2}\right\}^{s_1}\cdot$ $\left\{1+\dfrac{s_1(t\lambda)^{\alpha}}{s_2}\right\}^{-(s_1+s_2)}$	no simple form	no simple form

Table 1 Summary of Potential Claim Duration Parametric Distributions

In order to determine which family of mixture densities is most appropriate, each of the models identified above 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.

The results of fitting each of the mixture models to the claim duration data are given in Figure 7 and Table 2.




Log-logistic Mixture and KM Survival Function
Proportion of Claims Continuing
Claim Duration (Days)
KM Survival Function
Log-logistic Survival Function
PP Plot for Loglogistic Mixture
Loglogistic Mixture Percentile
KM Percentile
Generalised Log-logistic Mixture and KM Survival Function
Proportion of Claims Continuing
Claim Duration (Days)
KM Survival Function
Generalised Log-logistic Survival Function
PP Plot for Generalised Log-logistic Mixture
Generalised Log-logistic Mixture Percentile
KM Percentile
Extended Generalised Gamma Mixture and KM Survival Function
Proportion of Claims Continuing
Claim Duration (Days)
KM Survival Function
Extended Generalised Gamma Survival Function
PP Plot for Extended Generalised Gamma Mixture
Extended Generalised Gamma Mixture Percentile
KM Percentile
Gamma Mixture and KM Survival Function
Proportion of Claims Continuing
Claim Duration (Days)
KM Survival Function
Gamma Survival Function
PP Plot for Gamma Mixture
Gamma Mixture Percentile
KM Percentile

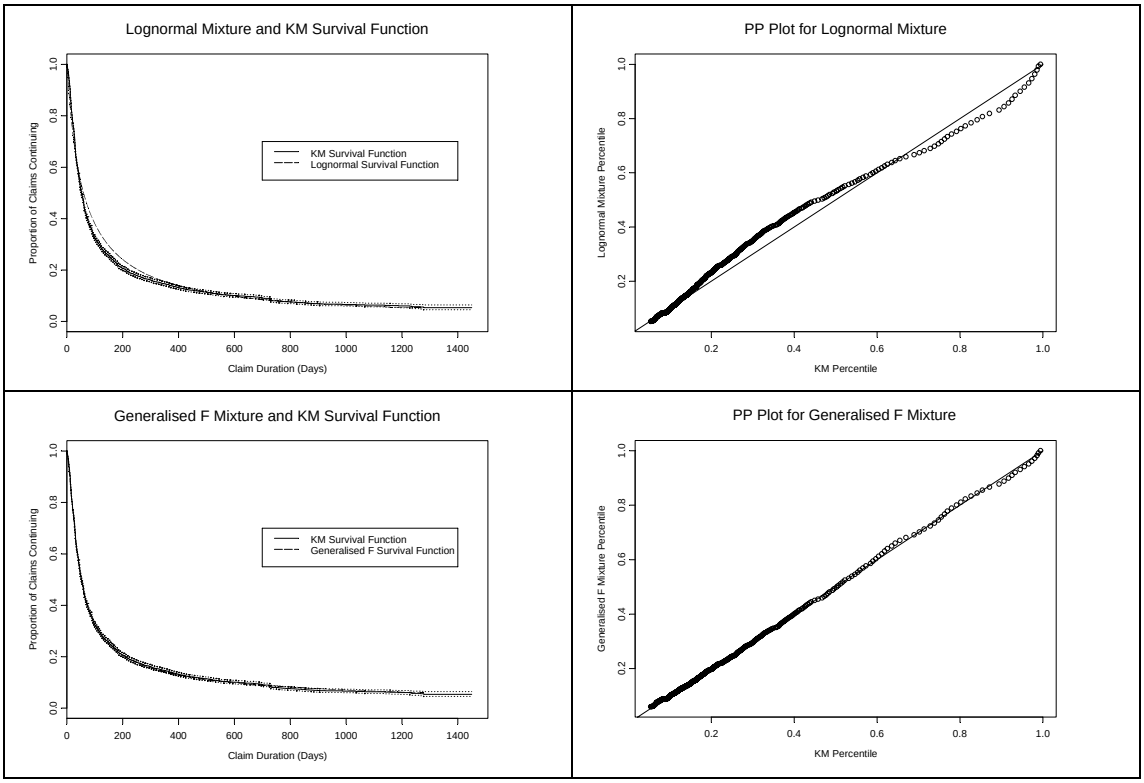


Figure 7 Assessment of Fit of Parametric Density to Claim Duration

Model	Maximised Log-Likelihood	R-squared for PP Plot	Akaike's Information Criterion (AIC)
Weibull Mixture	-16 038.77	96.455%	32 083.54
Log-Logistic Mixture	-15 598.72	99.609%	31 205.44
Generalised Log-logistic Mixture	-15 550.93	99.680%	31 109.86
Extended Generalised Gamma Mixture	-15 841.87	98.370%	31 691.74
Gamma Mixture	-16 180.51	94.877%	32 367.02
Lognormal Mixture	-16 333.63	97.210%	32 673.26
Generalised F Mixture	-15 478.77	99.915%	30 967.54

Table 2 Assessment of Fit of Parametric Density to Claim Duration

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 summary of the

data. The Generalised log-logistic distribution provides only 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, 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 that 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 disability 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, $E(\pi \mid Z) = \dfrac{1}{1+\exp(Z'\gamma)}$, 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

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

and the survival function is

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

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

$$L(s_1, s_2, \sigma, \beta, \gamma) = \sum_{i=1}^{n} \left[\delta_i \log\left\{ f(y_i; x_i, z_i, s_1, s_2, \sigma, \beta, \gamma) \right\} + (1-\delta_i) \log\left\{ S(y_i; x_i, z_i, s_1, s_2, \sigma, \beta, \gamma) \right\} \right]. \quad (6)$$

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. 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 properly model the effect of age on the return to work probability, three variables were used. The first variable was a simple linear predictor based on the age in years of the claimant at the time 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 maximising the log-likelihood of the resulting model. The definitions used in the final model are *ageind* is age for claimants below age 29. The variable *ageind2* is equal to 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 3.

	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

Table 3 Assessment of Accelerated Failure and Mixture Models for Claim Duration

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 3 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 4 and 5.

Generalized F mixture model				
The maximum loglikelihood is -15256.62				
Terms in the accelerated failure time model:				
	Coefficients	**Std.err**	**z-score**	**p-value**
Shape1	-1.27397			
Shape2	-1.59553			
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 4 Accelerated Failure 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

Table 5 Logistic Model Regression Coefficients

5. Discussion

The majority of the regressors shown in Table 5 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 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 that 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 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 8863 individuals in the study, 61 were found to possess all of the following characteristics: aged between 35 and 45, disability benefit of less than $2000 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 8, where 95% confidence bands have been included around the Kaplan-Meier fit.

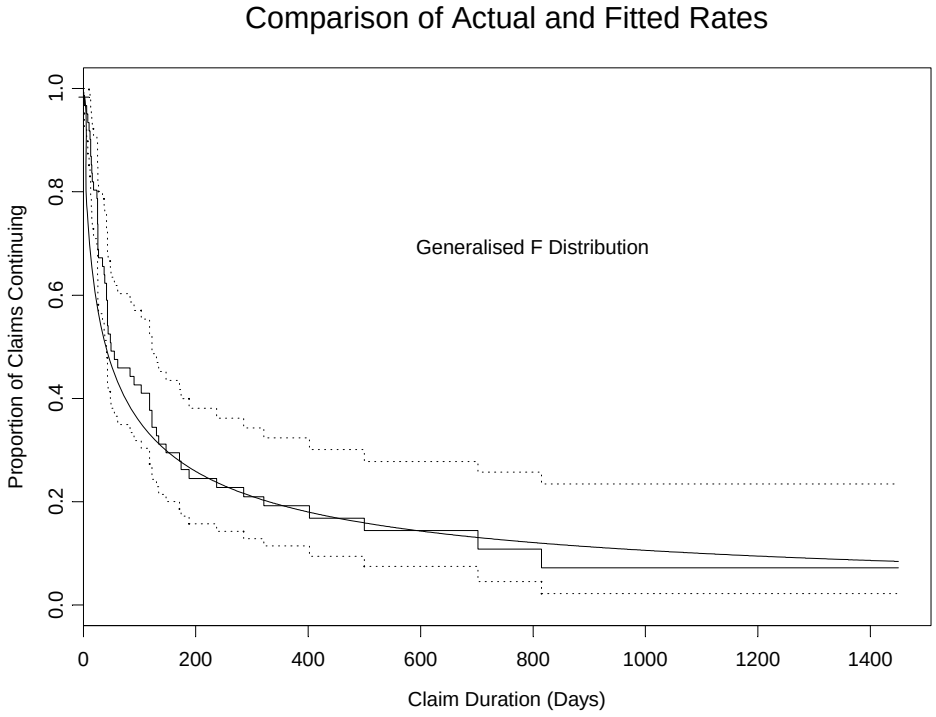


Figure 8 Comparison of Actual and Fitted Rates for the Generalised F Distribution

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 DII are impacted most by long duration claims, this imperfect fit at the shorter durations has less financial consequence for a life 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 disability claim termination data. The impact of this assumption not being satisfied on the fit of the Cox regression model is shown in the graph below. This graph compares the same data as used in Figure 9 to compare the empirical Kaplan-Meier survival function with the survival function predicted using Cox regression.

This graph shows clear evidence that the Cox regression model estimates claim termination rates that 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 the aim of the modeling is premium rating, is to compare the predicted expected present value of an annuity payable to a disability annuitant throughout their 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 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 6 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.

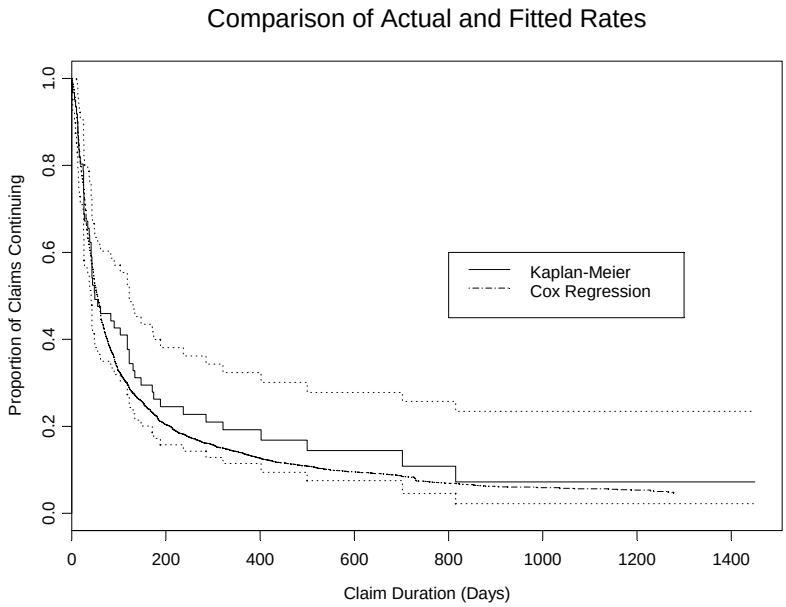


Figure 9 Comparison of Kaplan-Meier and Cox Regression Claim Duration Models

Kaplan-Meier Survival Function	Cox Regression Model	Generalised F Mixture Model
0.6163	0.4739	0.5837

Table 6 Annuity Value Comparison for three Model Fitting Procedures

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. 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. The author is also

grateful to the Institute of Actuaries of Australia for provision of the data used in this paper.

Appendix A

Field	Description	Variables (S-Plus Names)
Duration	Duration of the claim (recorded in days). This 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	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 (0 day), defpd1 (base level and deferment period between 1 and 27 days), defpd2 (28 to 89 day deferment period) and defpd3 (deferment period in excess of 90 days)	defpd0 defpd2 defpd3

References

Andersen, P.K. and 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.

Continuous Mortality Investigation Report No 12, 1991. CMIB, Institute and Faculty of Actuaries, UK.

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

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

Harrell, F.E., 1986. The PHGLM Procedure, *SUGI Supplemental Library Guide, Version 5 Edition*, Cary, NC: SAS Institute Inc.

Kaplan, E.L. and Meier, P., 1958 Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Assocation, 53, 457-481.

Maller, R. and Zhou, X., 1995. Survival Analysis with Long Term Survivors, *Wiley*, New York.

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

Robinson M.A., 1985. The 1985 CIDA Disability Table. Institute of Actuaries of Australia Quarterly Journal, December 1988.

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

Therneau, T.M., 2000. Modeling survival data: extending the Cox model. *Springer*, New York.