Survival Analysis of Left Truncated Income Protection Insurance Data

[March 29, 2012]

[1]Qing Liu

[2]David Pitt

[3]Yan Wang

[4]Xueyuan Wu

Abstract

One of the main characteristics of Income Protection Insurance (IPI) claim duration data, which has not been considered in the actuarial literature on the topic, is left-truncation. Claimants that are observed are those whose sickness durations are longer than the deferred periods specified in the policies, and hence left-truncation exists in these data. This paper investigates a series of conditional mixture models when applying survival analysis to model sickness durations of IPI claimants, and examines the consequence of treating the IPI data with lengthy deferred periods as complete data and therefore ignoring the left-truncation by fitting the corresponding unconditional distributions. It also quantifies the extent of the bias in the resulting parameter estimates when ignoring the left-truncation in the data. Using the UK Continuous Mortality Investigation (CMI) sickness duration data, some well-fitting survival model results are estimated. It is demonstrated that ignoring the left-truncation in certain IPI data can lead to

[1]Qing Liu
Centre for Actuarial Studies, Faculty of Business and Economics, The University of Melbourne.
Email address: liuq@student.unimelb.edu.au
[2]David Pitt
Department of Applied Finance and Actuarial Studies, Faculty of Business and Economics, Macquarie University.
Email address: david.pitt@mq.edu.au
[3]Yan Wang
School of Mathematical and Geospatial Sciences, RMIT
Email address: yan.wang@rmit.edu.au
[4] Xueyuan Wu
Centre for Actuarial Studies, Faculty of Business and Economics, The University of Melbourne.
Email address: xueyuanw@unimelb.edu.au

substantially different statistical estimates. We therefore suggest taking left-truncation into account by fitting conditional mixture distributions to IPI data. Furthermore, the best fitting model is extended by introducing a number of covariates into the conditional part to do regression analysis.

Key words: Income protection insurance (IPI); left-truncated; mixture distribution.

1 Introduction

Income Protection Insurance (IPI) plays a significant role in maintaining the quality of life of individuals, of working age, who become unable to work due to a non-work related injury or an illness. It achieves this by providing such insured lives with a proportion of their usual salary during the time that they are unable to work. Actuaries are often required to assess the future expected cash outflows associated with a portfolio of IPI contracts, and to set up a reasonable premium in order to make sure that the insurance company can meet its future obligations while remaining solvent. The main uncertainty of the IPI business comes from the fact that the likelihood of claim and the resultant claim duration can vary considerably for different IPI policies. Therefore having a good understanding of the durations of IPI claims has a significant impact on the actuaries' pricing and reserving calculations. This paper will consider the modelling of sickness durations with the use of survival analysis.

We already have some understanding of the level of claim termination rates from existing industry tables and other industry level studies. In the UK, the Continuous Mortality Investigation (CMI) is responsible for conducting research on mortality and morbidity experience for the UK life insurers. In particular, the CMI Report 12 (1991) established a new methodology for the analysis of IPI data in the form of a three state model. The recovery intensity estimated was also presented in this report. This model assumed that there was only one sick state to represent all causes of sickness, which means all the claims in the same portfolio will be subject to the same termination assumption regardless of their different causes of sickness. Ling *et al*. (2010) extended the model from CMI 12 and estimated recovery intensities by cause of sickness using IPI data provided by the CMI. The recovery intensities for each cause of sickness were estimated using the Cox proportional hazards regression model (Cox, 1972) and generalised linear models. The graduation formulae for the recovery intensity presented in Ling *et al*. (2010) modelled the impact of age at the date of falling sick, duration of disability and the deferred period on the recovery intensity.

Besides investigations into recovery intensities, there has also been related research into the modelling of claim durations. Pitt (2007) analysed a set of IPI data provided by the Life and Risk Committee of the Institute of Actuaries of Australia and suggested modelling claim durations using survival analysis. Pitt (2007) reported that there are approximately 7% of claimants who will never return to work, and therefore

suggested the use of mixed parametric regression models, which included the prediction of a long-run probability of not returning to work, as a description of claim duration. The objective of Pitt (2007) was to model claim durations, which were defined as the time since claim onset. This is different from the concept of sickness duration that we consider in this paper, which is defined as the time since the claimants first become sick. This difference in the objectives arises due to the different lengths of the deferred periods specified in the UK and the Australian IPI data. Most Australian IPI data have short deferred periods such as 2 weeks and 4 weeks whereas the UK IPI data commonly have lengthy deferred periods such as 13 weeks or 26 weeks. To examine the impact of the long deferred periods on the sickness durations, we first assume the data were complete, that is, were without left-truncation. We then propose a new approach making use of conditional mixture distributions. Thus we are able to quantify the extent of bias introduced in the parameter estimates by ignoring the left-truncation in the data. Using the UK sickness duration data, some useful survival models are estimated. It is shown that ignoring the left-truncation in certain IPI data can lead to substantially over-estimated median and upper quantiles of the distribution of sickness duration. When taking left-truncation into account through the conditional distributions, the conditional version of the relatively new mixture model, suggested by Shao and Zhou (2004), called the Burr XII mixture model fits the data best. Furthermore, the model is extended by introducing a number of covariates into the conditional Burr XII part, and the method and results of this regression analysis are also presented.

The rest of the paper is organised as follows. Section 2 provides a brief summary of the data. Section 3 presents a series of conditional mixture models that are used to analyse the left truncated data and the performance of these models are examined in this section. The results are also compared with those obtained by ignoring the left-truncation. Section 4 demonstrates how to incorporate different covariates into the mixture models, and Section 5 concludes the paper.

2 Data

This set of IPI data is obtained from Continuous Mortality Investigation (CMI), which was set up to carry out research into mortality and morbidity experience for the UK life insurers. It contains claim records for which payments have been made from record years 1975 to 2002 inclusive. Data are recorded for each policyholder based on the information provided in the insurance proposal form. Most of these claims were able to be traced by matching claims from one calendar year to the next on the basis of the birthday of the policyholder, date of falling sick, year of entry, sex and occupation category. Therefore we are able to calculate the sickness duration as the difference between claim cessation date and date of falling sick. Our income protection insurance data are left truncated because if a policyholder fell sick but recovered before the end of a specified deferred period, that individual would not be able to claim from the insurance company. That is to say, our data does not contain such records of early recovery. The most common deferred periods in the UK are 1, 4, 13, 26 and 52 weeks.

This IPI data comprise 70 different causes of sickness. Ling *et al*. (2010) have previously conducted some analysis on the recovery intensities using this same data set and pointed out that cause of sickness is an important source of heterogeneity among IPI claimants, and so suggested estimating the recovery intensities by cause of sickness. We have chosen cause of claim 21 (benign neoplasms and neoplasms

of unspecified nature) because of its high claim volume relative to other causes as an illustration to demonstrate the impact of ignoring the left truncation in the data when modelling, and to report our conditional mixture model based survival analyses. This method could equally be applied to other causes of sickness.

There are a total of 1523 claim records for the chosen cause of sickness, including 689 new claims. The remaining 834 claims continued from previous years prior to 1975. Out of the 689 new claim records, 418 of the claimants recovered during the investigation period.

- The age profile of claimants ranges from 21 to 65 with an average age of 46. The distribution of the ages is approximately bell-shaped.
- There are only two occupation classes: 758 (90%) are in occupation class 0, and 128 (10%) are in occupation class 1.
- Males account for 82% of the data.
- The year of entry to the IPI policy for the claimants ranges from 1936 to 2001 with an average year of entry being 1976. The distribution of year of entry is also approximately bell-shaped.

3 Model and Results

Pitt (2007) reported that there were approximately 7% of claimants who will never recover based on the Australian IPI data. Maller and Zhou (1995) described this non-zero long-run probability of survival as an "immune probability". We will therefore consider a number of mixture models, which take this feature of the data into account. For the survival analyses conducted here the event of interest is claim termination.

Define T to be a mixed random variable for the unknown sickness duration of a disabled life. To build a mixture model, we need to fit a parametric distribution to the sickness durations from the recorded lives that returned to work, and mix it with a point mass probability that one life will never return to work. For the case of the Weibull mixture distribution, the density function has the form

$$f(t) = p(\lambda t)^{\alpha-1}\lambda\alpha\exp\{-(\lambda t)^{\alpha}\}, t \geq 0 \ ,$$

and the associated distribution function is $F(t) = p(1 - \exp\{-(\lambda t)^{\alpha}\})$, $t \geq 0$, where p is the proportion of claimants who will eventually recover, α and λ are the usual Weibull parameters. The survival function for the Weibull mixture distribution is $S(t) = p\exp\{-(\lambda t)^{\alpha}\} + 1 - p$, $t \geq 0$.

In order to decide whether there is a significant percentage of claimants who will never recover for our UK IPI data, we employed the likelihood ratio test described in Maller and Zhou (1995). The method of Maller and Zhou was introduced for the case of the Weibull distribution of claim duration and involved comparing the maximum restricted likelihood able to be obtained for a model where the immune probability is set to be zero with the maximum likelihood achievable when this restriction is removed. The test statistic is based on the usual likelihood ratio test and is written as $d_n = -2\left(l(\tilde{\theta}_{H_0}) - l(\breve{\theta})\right)$, where $\breve{\theta}$ are the maximum likelihood estimates (MLEs) obtained from fitting a Weibull mixture model, θ_{H_0} are the

corresponding MLEs under the null hypothesis of no immunes, and $l(\theta)$ is the loglikelihood function evaluated at θ. Maller and Zhou showed that the asymptotic distribution of d_{n}, under the null hypothesis of no immunes, is a 50-50 mixture of a chi-square distribution with one degree of freedom and a point mass of 0.5 at zero. Applying this test to our IPI data, we get a test statistic of $d_n = -2(-3242.651 + 3224.83) = 35.64$ for the Weibull mixture distribution, which is significant under the chi-square point mass mixture distribution.

In addition to the Weibull mixture model, we also considered some other mixture models, including the log-logistic mixture model, the log-normal mixture model and the Burr XII mixture model. The long term survivor probability was found to be significant under all of these mixture models. The density functions and survival functions for these mixture distributions are summarised in Table 1 below. All probability functions in the table are defined over $t \geq 0$. In particular, the Burr XII mixture model introduced by Shao and Zhou (2004) is a relatively new model in survival analysis, and was not included in the analysis in Pitt (2007).

Table 1: Density functions and survival functions

Model	Density Function	Survival Function
Weibull Mixture	$p(\lambda t)^{\alpha-1}\lambda\alpha\exp\{-(\lambda t)^{\alpha}\}$	$p\exp\{-(\lambda t)^{\alpha}\} + 1 - p$
Log-logistic Mxture	$p\lambda\alpha(\lambda t)^{\alpha-1}\{1 + (\lambda t)^{\alpha}\}^{-2}$	$\frac{p}{1 + (\lambda t)^{\alpha}} + 1 - p$
Log-normal Mixture	$p\frac{\alpha}{t\sqrt{2\pi}}\exp\left\{\frac{-\alpha^2(\log(\lambda t))^2}{2}\right\}$	$1 - p\Phi\{\alpha\log(\lambda t)\}$
Burr XII Mixture	$p\alpha\lambda^{\alpha}t^{\alpha-1}\{1 + \beta(\lambda t)^{\alpha}\}^{-\left(1+\frac{1}{\beta}\right)}$	$p\{1 + \beta(\lambda t)^{\alpha}\}^{-\frac{1}{\beta}} + 1 - p$

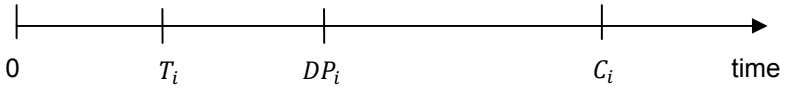
If the IPI data that the insurers observe were complete, we could fit the above distributions to the sickness durations of claimants directly. We call such a model an unconditional fit to the data. However, in reality, the IPI data that insurers have are left-truncated at different deferred periods specified in the policies, and hence are incomplete. Cox and Oakes (1984) explained that left truncation arises when individuals come under observation only some known time after the natural time origin of the phenomenon under study. In the context of IPI, for each of the *n* claimants we observe a vector (DP_i, Y_i, δ_i), where

$$Y_i = \min(T_i, C_i) \text{ and } \delta_i = \begin{cases} 1 \text{ if } DP_i < T_i \leq C_i, \\ 0 \text{ if } DP_i < C_i \leq T_i, \end{cases} \qquad i = 1,2,\dots,n.$$

Here Y_i is the recorded sickness duration for individual *i*. Its truncation time is the deferred period DP_i specified in the policy. Had the insured recovered before the deferred period, that individual would not be

able to claim from the insurance company and thus this period of sickness would not be recorded in the dataset. Figure 1 shows this situation.

Figure 1



Therefore, any contribution to the likelihood must be conditional on the deferred period having been exceeded. Our IPI data are also potentially right censored, that is, the individual either recovers at time T_i as shown in Figure 2 or are right-censored at the end of record time C_i as described in the example shown in Figure 3.

Figure 2

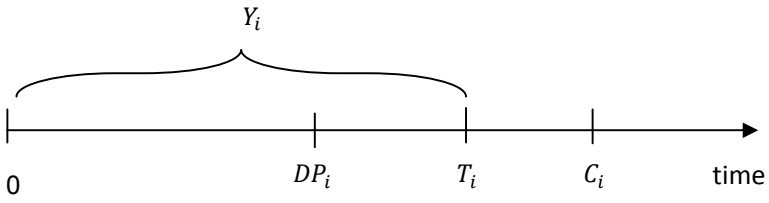
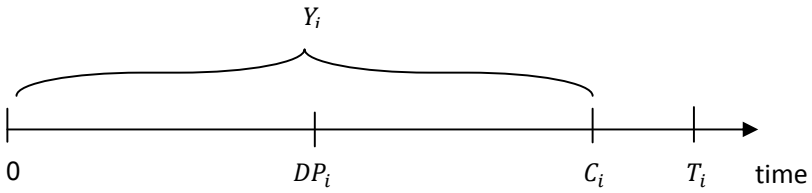


Figure 3



The contribution to the overall likelihood for a portfolio of policyholders is either

$$\frac{pf(t_i)}{pS(dp_i)+1-p}$$

for claimants who recovered at time T_i before reaching the censoring time C_i, or

$$\frac{pS(c_i)+1-p}{pS(dp_i)+1-p}$$

for right-censored claimants at time C_i, where $f(.)$ and $S(.)$ are the density function and the survival function for the fitted model respectively. The proportion of claimants who will eventually recover is denoted p. We denote δ_i as an indicator variable which takes the value 1 when the policyholder is observed to recover and takes the value 0 when the policyholder's sickness duration random variable is right censored. We call this a conditional distribution because each contribution to the overall likelihood for a portfolio of policyholders is conditional on the policyholder's sickness duration having exceeded the deferred period. The probability of this condition happening for policyholder *i* is given in the denominator of the above two equations as $pS(dp_i)+1-p$. Then, for a sample of n independent and identically distributed (DP_i, Y_i, δ_i), where $i = 1,2,\dots,n$, the joint likelihood function is given by

$$L = \prod_{i=1}^{n} \left(\frac{pf(t_i)}{pS(dp_i) + 1 - p} \right)^{\delta_i} \left(\frac{pS(c_i) + 1 - p}{pS(dp_i) + 1 - p} \right)^{1-\delta_i}. \quad (1)$$

The log-likelihood function is then

$$l = \sum_{i=1}^{n} \delta_i \log\{pf(t_i)\} + \sum_{i=1}^{n} (1 - \delta_i) \log\{pS(c_i) + 1 - p\} - \sum_{i=1}^{n} \log\{pS(dp_i) + 1 - p\}.$$

The maximum likelihood estimates of the parameters under different parametric models can be found by numerical methods such as the Newton-Raphson procedure.

We demonstrate the significant difference between using conditional distributions and unconditional distributions through the survival function plots of the sickness durations. In order to assess which method provides a better fit to the data, K-M (see Kaplan and Meier, 1958) survival curves are used to provide a non-parametric estimate of the survival function for claim durations. Notice here, the K-M estimator needs to be adjusted to reflect the presence of left truncation. This estimator along with a modified estimator of its variance was proposed by Tsai, Jewell, and Wang (1987). For each of the n individuals we observe the triple (DP_i, Y_i, δ_i), where

$$Y_i = \min(T_i, C_i) \text{ and } \delta_i = \begin{cases} 1 \text{ if } DP_i < T_i \leq C_i \\ 0 \text{ if } DP_i < C_i \leq T_i \end{cases}.$$

Let $t_{(1)}, \cdots, t_{(r)}$ denote the $r \leq n$ distinct ordered and uncensored recovery times, so that $t_{(j)}$ is the jth ordered recovery time. We now define the modified risk set $\mathrm{R}(t_{(i)})$ at $t_{(\mathrm{i})}$ by

$$\mathrm{R}(t_{(i)}) = \{j | d_j \leq t_{(i)} \leq y_j\}, j = 1, \cdots, n, i = 1,2, \cdots, r,$$

where d_i denotes the number of left-truncated claimants before t_i but recovered at t_i.

The modified K-M estimator of the survivor function has the form

$$\hat{\mathrm{S}}(t) = \prod_{\mathrm{i}=1}^{k} \left(\frac{n_{\mathrm{i}} - d_{\mathrm{i}}}{n_{\mathrm{i}}} \right),$$

where $t_{(k)} \leq t < t_{(k+1)}$. Let n_i denote the number of claimants in $R(t_{(i)})$, which is the number of left-truncated claimants that are alive, but not censored, just before time $t_{(i)}$. The fitted conditional and unconditional survival functions identified above were compared with the modified Kaplan-Meier estimate of the survival function.

The results of fitting each of the conditional and unconditional versions of the mixture models summarised in Table 1 are given from Figure 4 to Figure 11**Error! Reference source not found.**Figure 4. The left column of Figure 4 to Figure 11 show the graphs of the survival functions by fitting the mixture distributions compared with the modified K-M estimates. The right column of these figures show the P-P plots for these fits. Comparing Figure 4 with Figure 5, Figure 6 with Figure 7, Figure 8 with Figure 9, and

Figure 10 with Figure 11, it is very clear that treating the data as complete by fitting the unconditional distributions will lead to results that are far off track from the true data whereas treating the data as left-truncated by fitting the conditional distributions give us much better results in terms of a closer fit. It is demonstrated that ignoring the truncation data will result in survival functions that are significantly overestimated for claims of all different durations. The PP plots highlight this deficiency very clearly. Using the conditional distributions improves the fitness substantially, especially for claims with short durations. This improvement in fit is also clearly evident when we look at the maximum log-likelihood results provided in

Table 2. The maximum log-likelihood values calculated from using the conditional distributions are always higher compared to using the unconditional ones. We have also calculated AIC values in Table 2 for all the models considered, and the one that achieved the lowest value is the conditional Burr XII mixture model. This result is consistent with the graphs, where we can see that the right bottom corner is the P-P plot for the conditional Burr XII mixture model, and it is the closest to the 45 degree straight line compared to the other traditional models in survival analysis. We are going to use the conditional Burr XII mixture model in next section to do our regression analysis.

Figure 4: Conditional model fitting a)

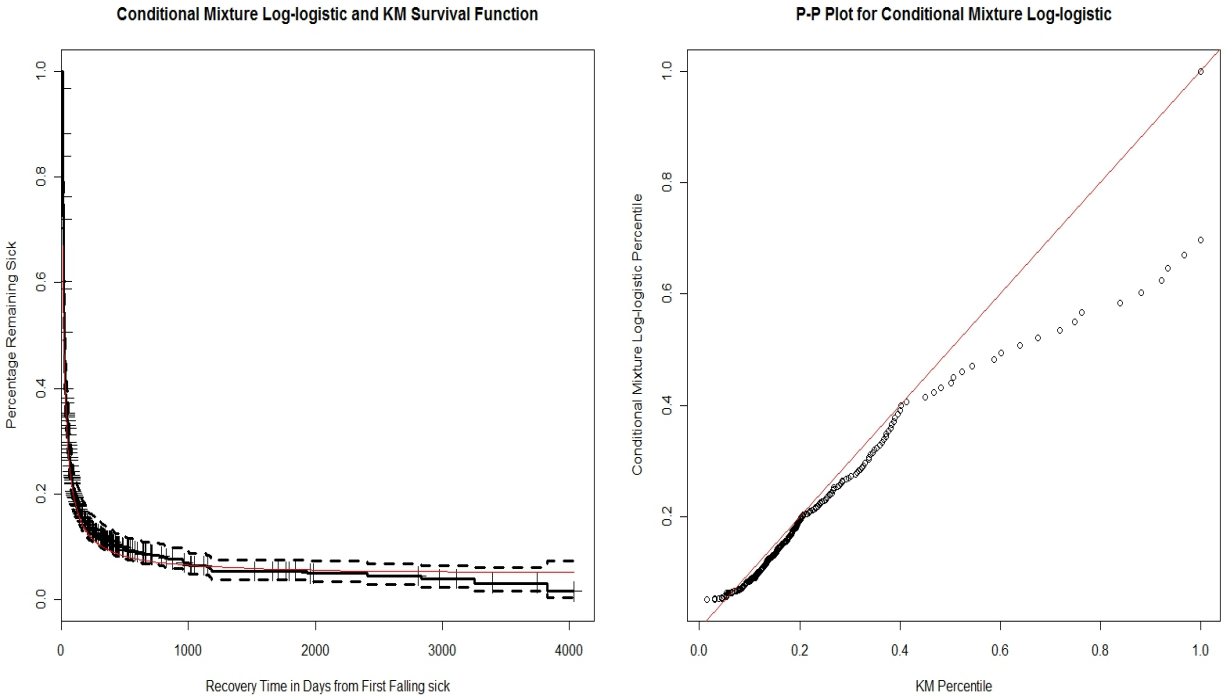


Figure 5: Unconditional model fitting a)

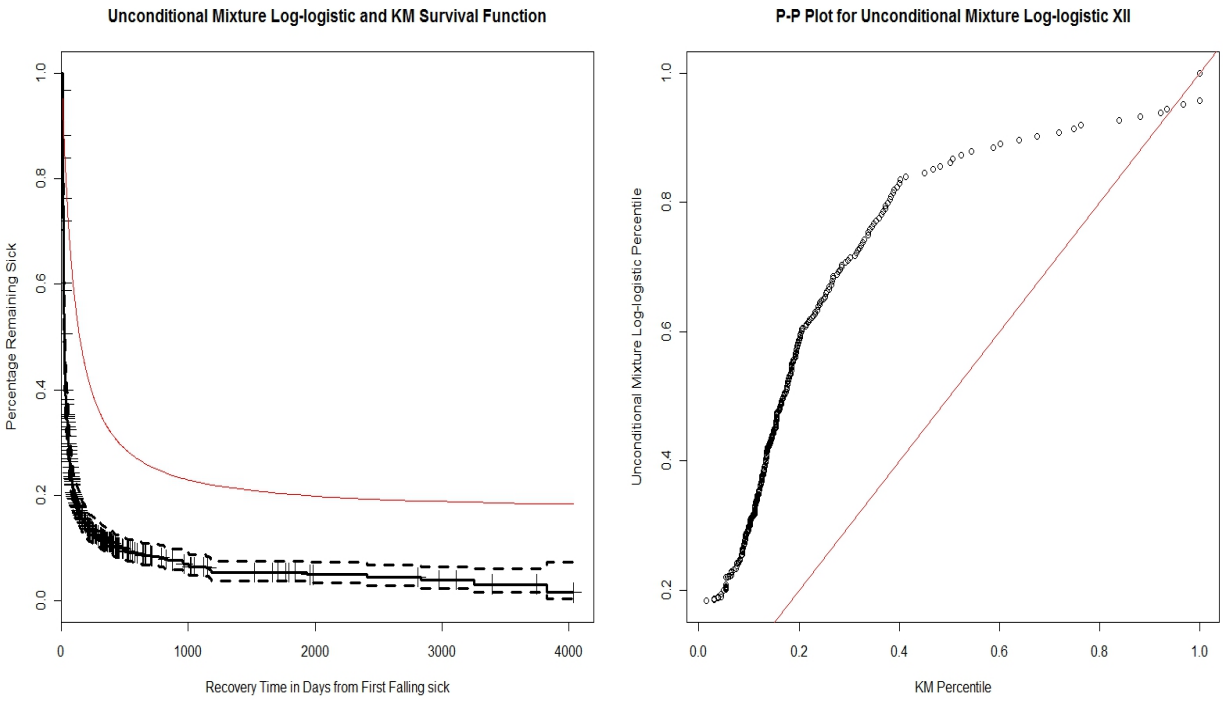


Figure 6: Conditional model fitting b)

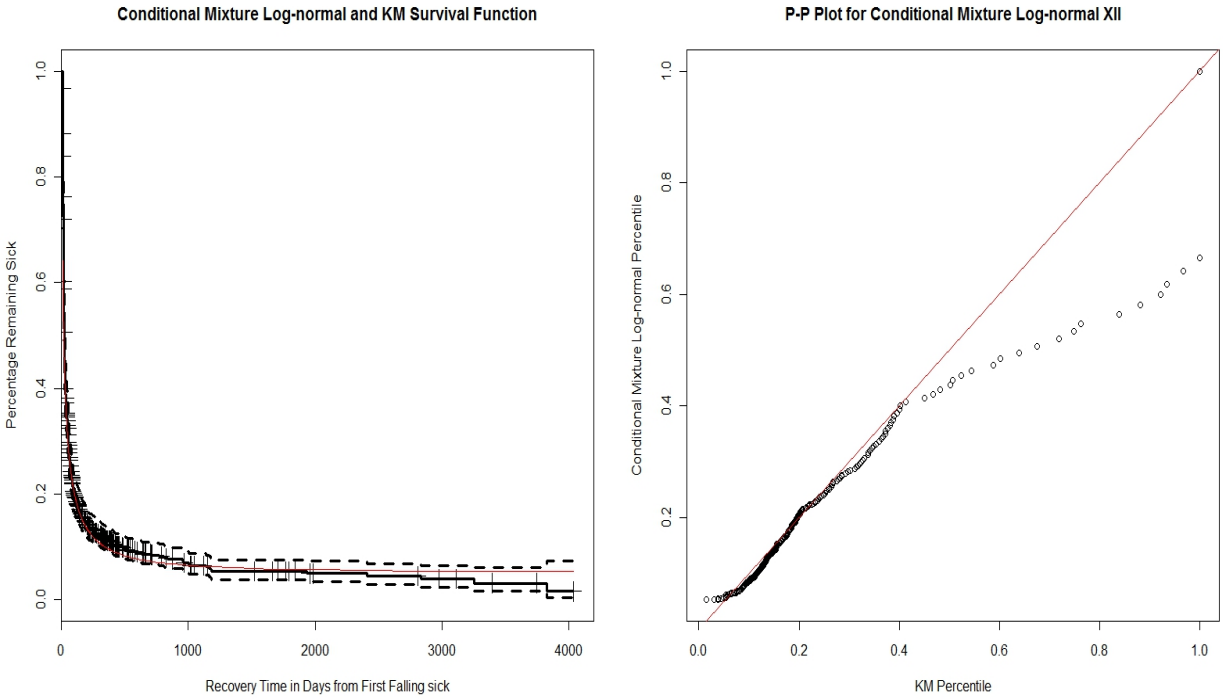


Figure 7: Unconditional model fitting b)

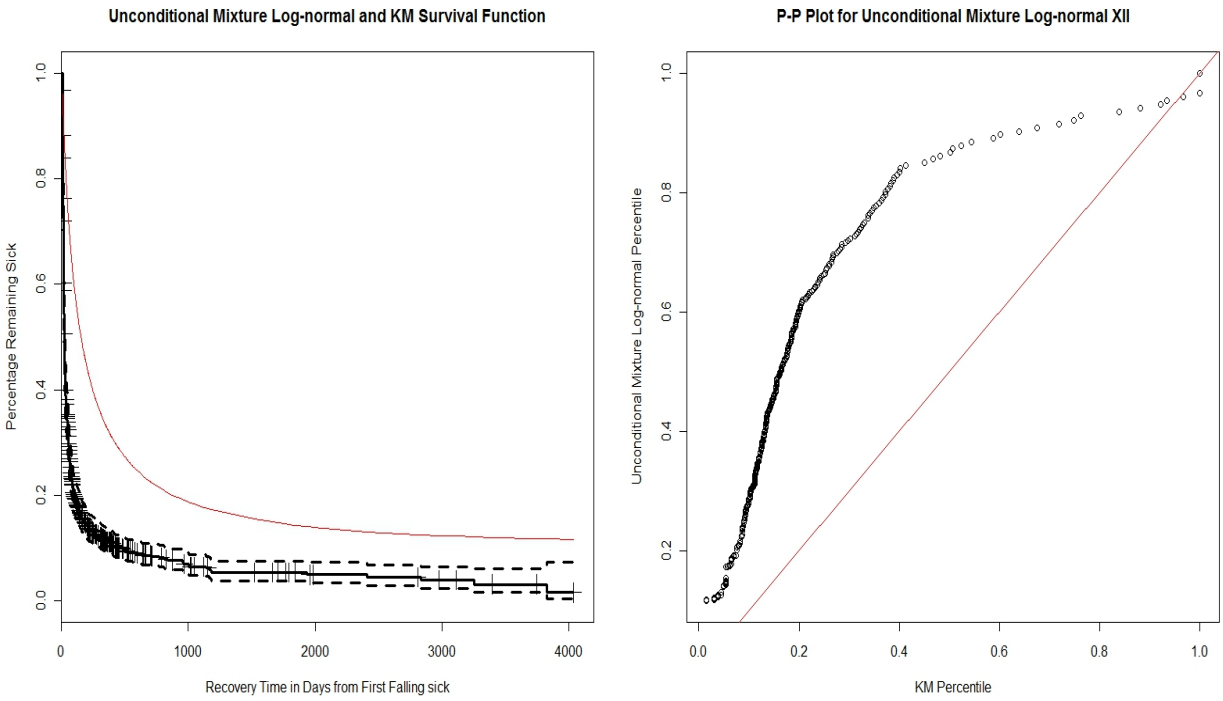


Figure 8: Conditional model fitting c)

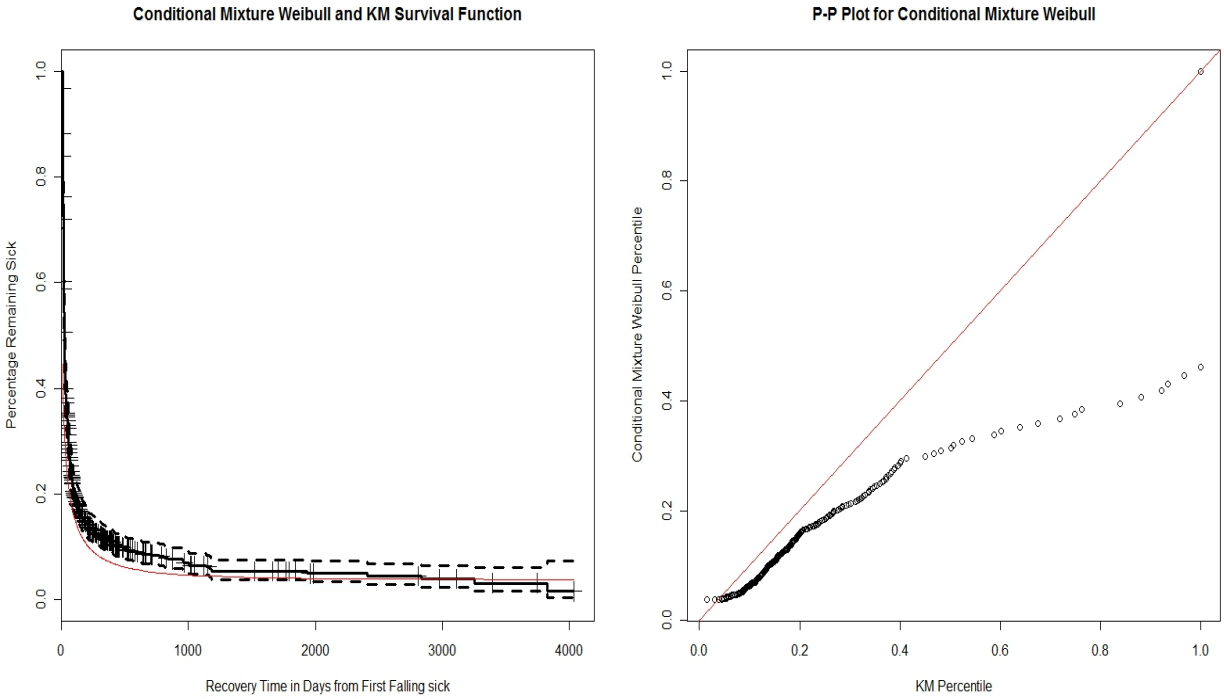


Figure 9: Unconditional model fitting c)

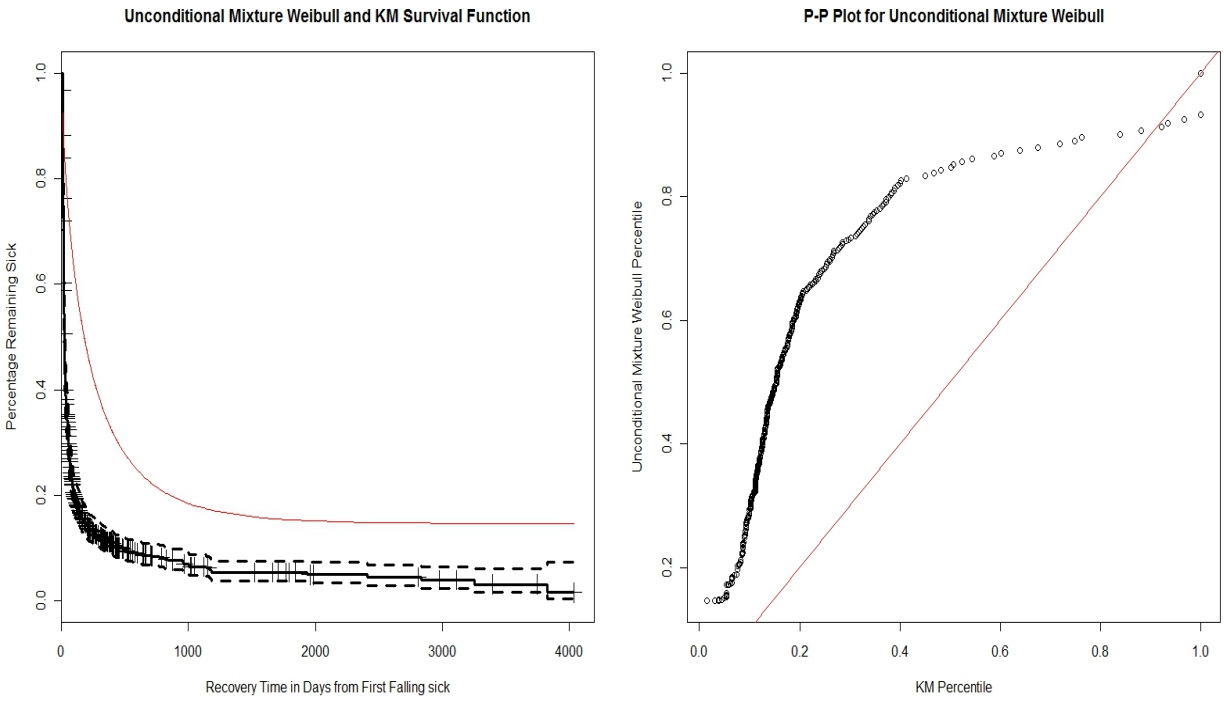


Figure 10: Conditional model fitting d)

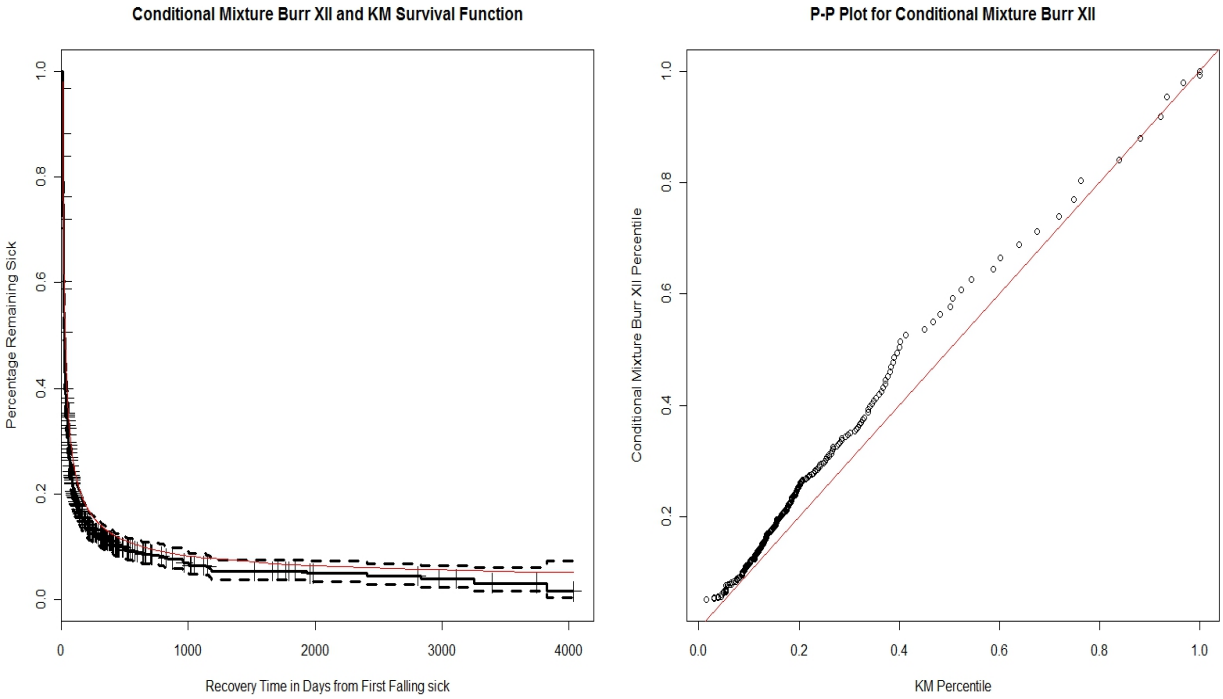


Figure 11: Unconditional model fitting d)

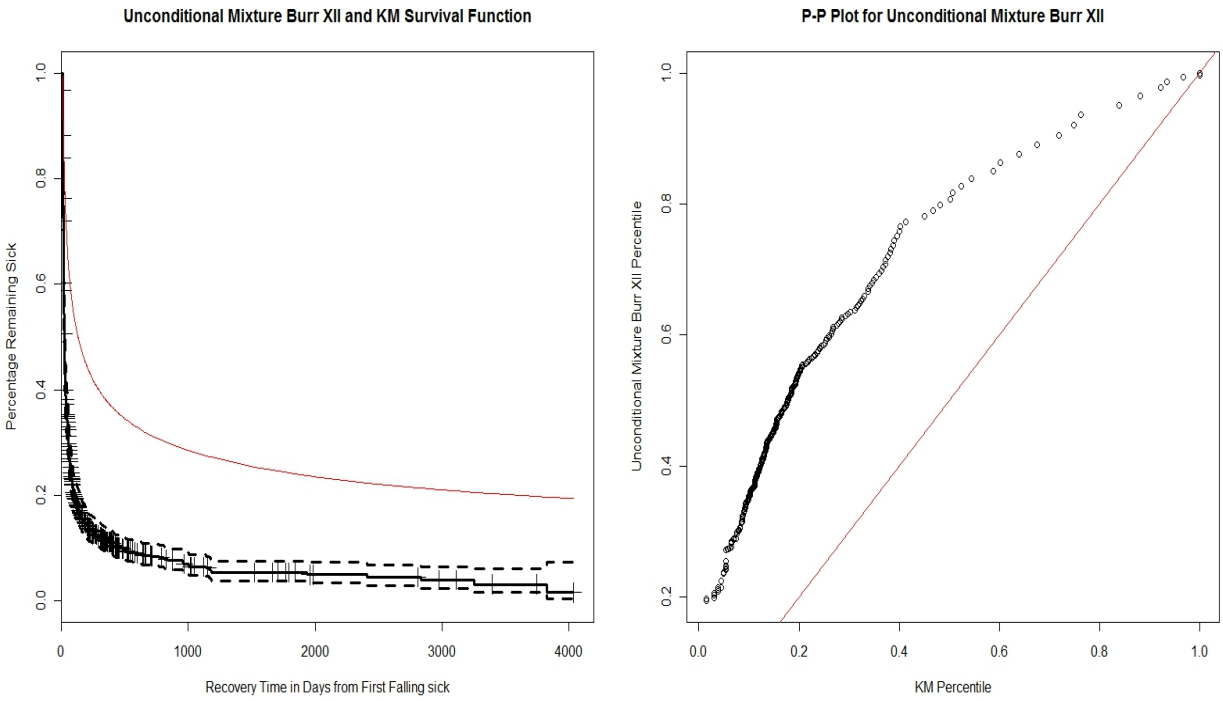


Table 2: Model Comparison

	Unconditional		Conditional	
	MLL	AIC	MLL	AIC
Mixture Log-logistic	-3610.96	7227.92	-3190.67	6387.35
Mixture Log-normal	-3596.15	7198.30	-3200.29	6406.57
Mixture Weibull	-3652.01	7310.02	-3224.83	6455.67
Mixture Burr XII	-3562.54	7131.08	-3155.02	6318.04

In Table 3, we show the change in the parameter values when the conditional mixture models are employed instead of the unconditional ones.

For the log-logistic distribution, log-normal distribution and Weibull distribution, they can all be expressed in the form

$$\log T = \mu + \sigma Z,$$

where Z is a standard member of a location and scale family of distributions ($\mu = 0$ and $\sigma = 1$). For $\log T$, the location parameter is $\mu = log\, \lambda$ and the scale parameter is $\sigma = \frac{1}{\alpha}$. From Table 3, we can see that under the correct data specification, that is, considering left-truncation, the location parameters are lower and the scale parameters are higher. The estimated fraction of claimants who recovered before the deferred period, denoted by $F(dp)$ is larger under the conditional fit for any value of deferred periods. This is what we expected since the conditional estimation accounts for the true information loss while the unconditional fit assumes all instances of sickness persisted for a duration at least equal to the deferred period in the policy. The results are consistent with findings in Chernobai *et al*. (2006) with the catastrophe claims data. It is also clearly demonstrated in the table that ignoring the left-truncation would result in over-estimated median and upper quantiles of sickness durations. This may have serious consequences. Suppose the insurance company sets up premium and reserve levels based on the results from fitting the unconditional distribution to the incomplete data, they would over-estimate the future claim payments as a result of over-estimating the claimants' sickness durations, especially so for those claimants with long sickness durations. Therefore it is essential to set up likelihood functions that incorporate left-truncation and to estimate parameters on this basis.

Table 3: Model Summary Statistics

		Conditional	Unconditional		Conditional	Unconditional	Conditional	Unconditional	Conditional	Unconditional
Mixture Log-logistic	α	0.9337	1.1047	F(1W)	31.83%	5.11%	median	median	90-percentile	90-percentile
	λ	0.0632	0.0101	F(4W)	63.02%	19.94%	15.05	83.3	158.27	608.74
	p	0.9542	0.8304	F(13W)	83.67%	47.80%				
				F(26W)	90.73%	66.32%				
				F(52W)	94.92%	80.90%				
Mixture Log-normal	α	0.5174	0.6262	F(1W)	35.22%	3.67%	median	median	90-percentile	90-percentile
	λ	0.0686	0.0082	F(4W)	63.23%	17.83%	2.81	6.39	35.95	56.38
	p	0.9488	0.8966	F(13W)	82.84%	42.71%				
				F(26W)	90.43%	59.89%				
				F(52W)	95.20%	75.32%				
Mixture Weibull	α	0.3539	0.7298	F(1W)	55.89%	7.96%	median	median	90-percentile	90-percentile
	λ	0.0811	0.0047	F(4W)	73.73%	20.40%	0.64	4.30	2.92	9.82
	p	0.9622	0.8537	F(13W)	86.85%	41.68%				
				F(26W)	92.52%	59.11%				
				F(52W)	96.36%	77.31%				
Mixture Burr XII	α	8.7320	6.3709	F(1W)	0.72%	0.29%	median	median	90-percentile	90-percentile
	λ	0.0818	0.0575	F(4W)	49.92%	23.50%	26.79	130.56	374.35	44200.00
	β	14.3093	23.0594	F(13W)	75.60%	44.76%				
	p	0.9719	1.0000	F(26W)	84.02%	54.38%				
				F(52W)	89.53%	62.33%				

4 Conditional Burr XII Regression for Left-Truncated and Right Censored Data

Shao and Zhou (2004) first introduced the Burr XII mixture model for survival analysis, but they did not include any covariates in the model. Since there are a number of covariates associated with our data, such as age, gender, and year of entry, we are going to extend the conditional Burr XII mixture model in this section to incorporate such covariates into this parametric model. Beirlant *et al*. (1998) proposed two regression models for the Burr distribution. We begin by following the same approach as Beirlant *et al*. (1998) in their so-called parametrisation I. However, we need to modify their approach to cater for our jointly left-truncated and right-censored data. In order to be consistent with our previous notations used for the parameters, our notation used for the Burr XII distributions are different from that used in Beirlant *et al*. (1998). The shape parameter α is allowed to vary with covariates x, that is, $T|x$ follows a Burr $(\lambda, \beta, \alpha(x))$. An exponential link-function is used for positive parameter α:

$$\alpha(x) = \exp\{\gamma^{\mathbb{T}} x\},$$

where x is a covariate vector and γ is a vector of regression coefficients. The model is fitted using maximum likelihood estimation. The log-likelihood function for the model is given below, and it is similar to Equation (1). However, we now relax the assumption of (DP_i, Y_i, δ_i) being identically distributed in order to model the potential impact of different covariates on sickness duration.

$$l = \sum_{i=1}^{n} \delta_i \log\{pf(t_i; x_i, \gamma_i, \lambda, \beta)\} + \sum_{i=1}^{n} (1-\delta_i) \log\{pS(c_i; x_i, \gamma_i, \lambda, \beta) + 1 - p\}$$
$$- \sum_{i=1}^{n} \log\{pS(dp_i; x_i, \gamma_i, \lambda, \beta) + 1 - p\}.$$

where $f(.)$ and $S(.)$ are the density function and the survival function for the fitted Burr XII mixture model respectively. The covariates age, sex, year of entry, dp0 (indicator variable for deferred period 0), dp1 (indicator variable for deferred period 1), dp13 (indicator variable for deferred period 13) and dp26 (indicator variable for deferred period 26) were tested. Covariates age and year of entry are scaled as follows

$$x_{\text{age}} = \frac{(\text{age} - 43)}{26}, \; x_{\text{entry}} = \frac{(\text{entry} - 1988)}{13},$$

so that x_{age} ranges from -1 to 1 when age ranges from 17 to 69 and x_{entry} ranges from -1 to approximately 1 when entry year ranges from 1975 to 2002. Two-way interaction variables were also considered as possible regression covariates. The regression model that produced the smallest AIC value was chosen to be the final model. A summary of the final fitted model is given in

Table **4**. It is not surprising that after taking these covariates into account, we have achieved a higher maximum log-likelihood value compared with the results we obtained in Section 3 for all the models without taking any covariate into account. The AIC value for the conditional Burr XII mixture regression model is 6236.026. This is lower than the AIC value achieved by the general model, which was 6318.04. Notice that, it is also possible to link the long term survivor proportion p to the covariates by a logistic function if we are interested in the covariates effects on proportion p. However, Table 4 shows that $p = 1$ after taking the covariates into account for the Burr XII distribution, hence there is little point to complicate the model further by introducing more covariates into the description of the ultimate recovery probability, p.

All of the covariates shown in

Table **4** have statistically significant effects on sickness durations at the 5% significance level except interaction term *age*dp13*. There are 5 different levels of deferred period, the statistical significance of the variables *dp0, dp1, dp4, dp13, dp26* and *dp52* are strongly affected by the amount of data for that particular deferred period category. For that reason, we only observe that *dp0, dp1* and *dp13* appear to have significant impacts on sickness durations. To interpret these covariate coefficients, we take covariate *age* as an example. The covariate coefficient for age is -0.4121, this means as age increases by 1 unit, the shape parameter α decreases by $\exp(-0.4121) = 0.66$. In order to visualise this effect, we have plotted the survival functions for two different age groups in

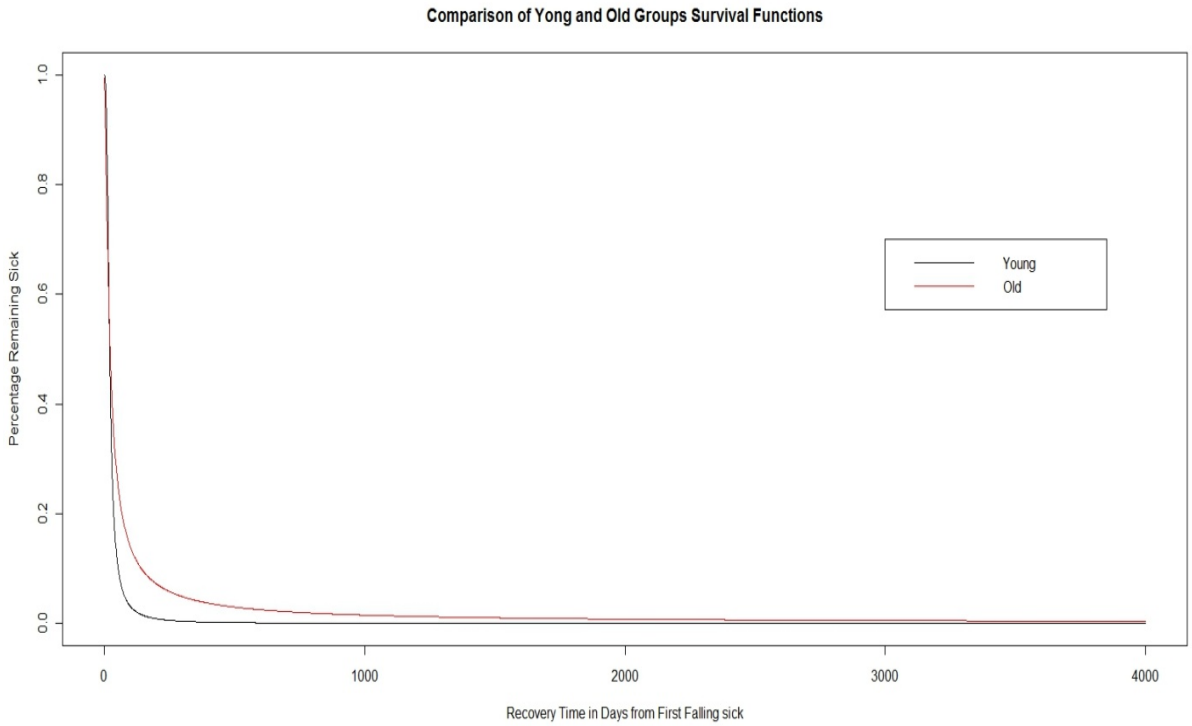
Figure **12**. The shape parameter α is calculated for claimants who are aged 22 and 64 respectively, just as an illustration for young and old groups, all the other covariates are holding constant at a set of following values: deferred period is 1 week, gender is male, and entry year is 1988. As illustrated in

Figure **12**, the age covariate has an impact on the shape of the fitted survival functions. The young group has a higher recovery rate compared with the old group. The difference in the recovery rates between the two groups are negligible at very short sickness durations and becomes more apparent as the duration increases. When the difference between the two age groups reaches its maximum at around 200 days, it starts to decrease as time goes on. Similar or opposite conclusions can be drawn for all the other covariates according to the signs of their coefficients. These graphs are useful to visualise the impact of each covariate on the shape parameter of Burr XII mixture distribution and give us a better understanding of the survival functions for different groups of claimants.

Table 4: Regression Model Results

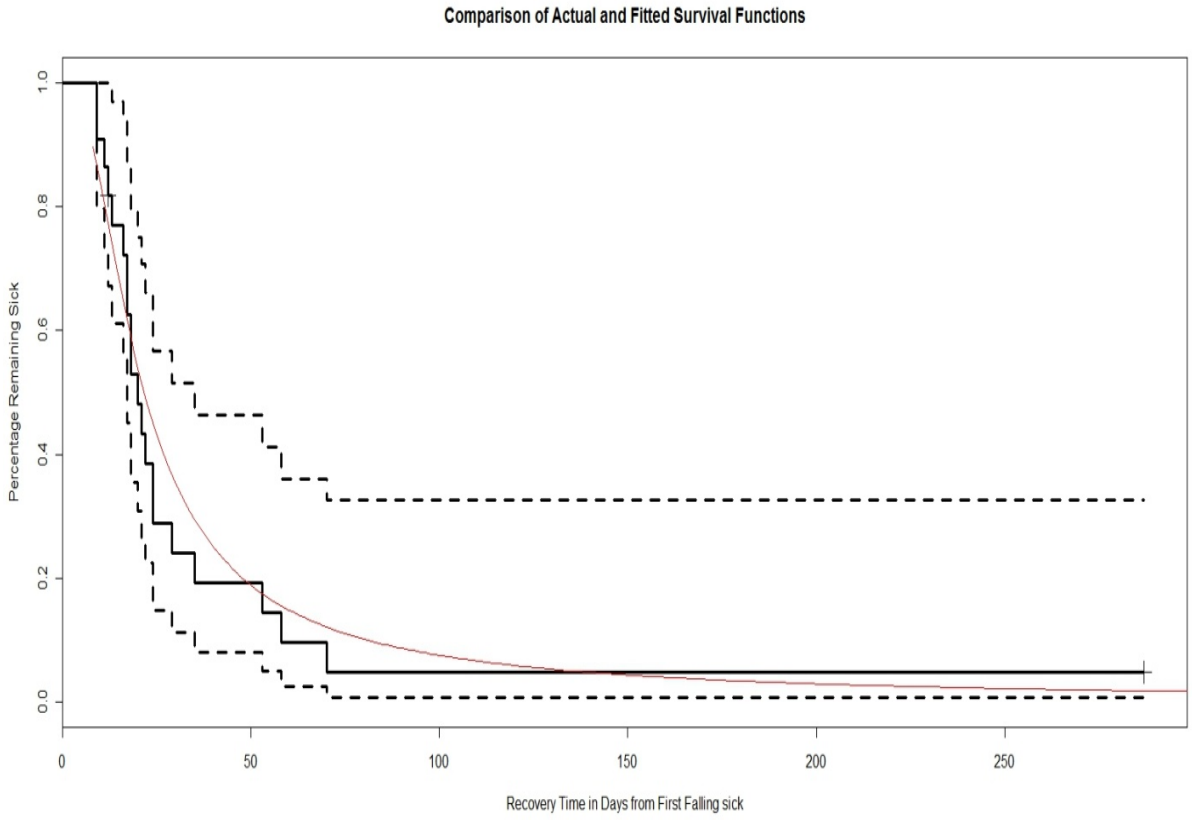
	MLL	AIC		
	-3105.013	6236.026		
Final model	Coefficients	Std.err	Z-score	p-value
λ	0.0527	0.0029	18.0735	0.0000
β	1.8043	0.0954	18.9216	0.0000
p	1.0000	0.0061	162.7580	0.0000
age	-0.4121	0.0995	-4.1409	0.0000
sex	-0.0784	0.0460	-1.7064	0.0440
entry	0.3615	0.0928	3.8971	0.0000
dp0	1.1715	0.1494	7.8401	0.0000
dp1	0.9729	0.0682	14.2709	0.0000
dp13	-0.3935	0.0987	-3.9971	0.0000
age*sex	-0.2607	0.1096	-2.3778	0.0087
sex*entry	-0.3074	0.1092	-2.8158	0.0024

Figure 12: Comparison of Yong and Old Groups Survival Functions



In order to assess the quality of the fit of the model, we can divide the data into groups according to the values of the covariates included in the final model. There were 22 claimants found to possess all of the following characteristics: aged between 38 and 48, deferred period of 13 weeks, entry year is between 1985 to 1990 and male. For these 22 claimants, the modified Kaplan-Meier fit to the survival function is compared to the conditional survival function predicted by the mixture Burr XII model. The result of this comparison is shown in Figure 13, where 95% confidence bands have been included around the modified Kaplan-Meier fit. It is very clear from Figure 13 that the fit of the Conditional Burr XII mixture regression model is very good even for a specific group of data. Therefore the regression model could be useful for prediction of sickness durations of future claims.

Figure 13: Comparison of Actual and Fitted Survival Functions



5 Conclusion

In this paper, we brought our attention to one important characteristic of IPI data, which has not been modelled in this context in the actuarial literature. We focused on the fact that the IPI data available to insurance companies are left-truncated at different deferred period levels, and demonstrated that using conditional distributions instead of unconditional distributions provides more accurate parameter estimates especially for the IPI data that contain some very lengthy deferred periods. It was shown that treating the available IPI data as complete can lead to substantially over-estimated median and upper quantiles of sickness durations, which could lead to serious consequences for providers of IPI. We identified some useful survival model results discovered using the UK sickness duration data. We demonstrated that if the insurance company sets up premium and reserve levels based on the results from fitting the unconditional distribution to the incomplete data, they would over-estimate the future claim payments as a result of over-estimating the claimants' sickness durations, especially so for those claimants with long sickness durations. Therefore it is essential to realise that the IPI data are left truncated and conditional distributions should be used. This point becomes crucial for the UK IPI data where long deferred period of 26 and 52 weeks are quite common. We have demonstrated that after

taking left-truncation into account, the conditional version of the relatively new mixture model called Burr XII mixture provided most flexibility, and was capable of improving data fitting substantially over the other well-known traditional models. We have therefore extended this model by introducing a number of covariates into the conditional Burr XII mixture model, and presented the method and results of this regression analysis. This paper will be of value to insurers considering both pricing and the valuation of their IPI policies.

6 References

Beirlant, J., Y. Goegebeur, R. Verlaak, P. Vynckier (1998). "Burr regression and portfolio segmentation," *Insurance: Mathematics and Economics* 23: 231-250.

Chernobai A., K. Burnecki, S. Rachev, S. Trueck, R. Weron (2006). "Modelling Catastrophe Claims with Left-truncated Severity Distributions," *Computational Statistics*, 21: 537--555.

Cox, D. R. (1972). "Regression Models and Life Tables (with Discussion)," *Journal of the Royal Statistical Society*. Series B (Methodological), 34(2): 187-220.

Cox, D. R. and D. Oakes (1984) *Analysis of Survival Data*. London: *Chapman & Hall.* 177-178.

Kaplan, E.L. and P. Meier (1958). "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, 53: 457-481.

Ling, S.Y., H.R. Waters and A.D. Wilkie (2010). "Modelling Income Protection Insurance claim termination rates by cause of sickness I: Recoveries," *The Annals of Actuarial Science*, 4 (2): 199 - 240.

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

Pitt, D. (2007). "Modeling the Claim Duration of Income Protection Insurance Policyholders Using Parametric Mixture Models," *Annals of Actuarial Science*, 2(1): 1-24.

Shao, Q. and X. Zhou (2004). "A new parametric model for survival data with long-term survivors," *Statistics in Medicine*, 23: 3525-3543.

Tableman, M. and J.S. Kim (2004). *Survival analysis using S*, London: Chapman & Hall. 98-103.

Tsai, W.Y., N.P. Jewell, and M.C. Wang (1987). "A note on the product-limit estimator under right censoring and left truncation," *Biometrika*, 74: 883-886.

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