

AN INDIVIDUAL CLAIMS RESERVING MODEL

BY

CHRISTIAN ROHOLTE LARSEN

ABSTRACT

Traditional Chain Ladder models are based on a few cells in an upper triangle and often give inaccurate projections of the reserve. Traditional stochastic models are based on the same few summaries and in addition are based on the often unrealistic assumption of independence between the aggregate incremental values. In this paper a set of stochastic models with weaker assumptions based on the individual claims development are described. These models can include information about settlement and can handle seasonal effects, changes in mix of business and claim types as well as changes in mix of claim size. It is demonstrated how the distribution of the process can be specified and especially how the distribution of the reserve can be determined. The method is illustrated with an example.

KEYWORDS

Stochastic Claims Reserving, Chain Ladder, Marked Poisson Process (MPP), Decomposition, Markov Chain, Logistic Regression, Generalised Linear Models (GLMs), Generalised Pareto Distribution (GPD), Business Mix, Bootstrap.

1. INTRODUCTION

In this paper a particular approach to stochastic claims reserving is taken where relatively complete information of individual claims is used. The model is described in theoretical terms, however as can be seen, it has many practical applications.

TRADITIONAL RESERVING MODELS:

Traditional stochastic models (e.g. England and Verrall 2002), including Mack's model (e.g. Mack 1993), are based on the crucial and yet often ignored assumption that the incremental amounts in the 'upper triangle' are stochastically independent.

Let us consider a simple example of incremental claims costs arranged traditionally:

Accident period	No. of periods to development	
	1	2
1	0	2
2	2	0
3	1	1
4	0	?

If no further information is available, most people would suggest a reserve of 2. However, the traditional Chain Ladder and stochastic models would lead to a reserve of 0!

The example illustrates the importance of the assumption of independent incremental values and of using the underlying data, if it is available.

If the figures were individual claim figures a reserve of 2 would also seem to be a sensible answer. The model presented in this paper would also give this result.

The situation where the incrementals are correlated, perhaps not as obviously as illustrated above, is common and in these situations the Chain Ladder model and traditional stochastic models are not appropriate. This is the main motivation for developing more precise models.

INDIVIDUAL CLAIMS MODELS:

While individual claims development has been the subject for reserving models recently, e.g. Mahon (2005) where the claims distribution is modelled, the starting point in this paper is the stochastic Poisson process. The claims not yet reported are therefore integrated in the model. This is similar to the nonparametric Bayesian approach by Haastrup and Arjas (1996) where the number of partial payments are considered. The approach is also similar to that of Norberg (1999) where the partial claims are modelled using the Dirichlet distribution.

The presented model below is a parametric model utilising General Linear Models (GLMs). It projects the effect of changes in portfolio size, changes of mix of business, changes of mix of claim types, seasonal effects and changes of empirical claims distribution. The pure period inflation can be estimated and isolated from inflation caused by changes in portfolio mix.

Applying GLMs on individual claims data has been used before e.g. Taylor and McGuire (2004) where a stochastic model of the total amount paid per finalised claim is fitted using GLMs. The model presented below does not require information concerning settlement.

The model includes an estimation of the multi dimensional distribution of the future incurred amounts per development period given the incurred

accumulated amount at the beginning of the period and given other information concerning the claim, such as claim type and policy information. Based on this distribution the mean of the future changes, i.e. the required reserve in excess of the individual reserve, and even the distribution of this or any function of it or of the total projection can be calculated. Where a theoretical calculation in best case would be extremely complicated a projection via simulation can be obtained. The model addresses the common situation where the incremental amounts concerning a specific claim are not stochastically independent and is dynamic in the sense that the future incremental amounts are stochastically dependent on the past incremental amounts. The model can take the claims settlement into consideration and this will be discussed briefly.

The model deals specifically with the fact that the development of large claims is often very different from the development of other claims.

While the model here is based on Incurred Amounts it could equally be based on Payments.

Although the model is complex compared to traditional reserving models, each model component is manageable and the parameters can be estimated using traditional distributions such as Generalised Pareto Distributions and smoothing methods such as GLMs.

The paper is structured as follows: In section 2 the basic model is formulated as a Marked Poisson Process (MPP) and the stochastic reserve is defined. In section 3 a discrete version is created by making several assumptions and the process likelihood is described. In section 4 it is demonstrated how the distributions can be modelled using GLMs and other traditional smoothing methods. In section 5 examples are outlined based on policy and claims data from a Marine portfolio. In section 6 the resulting distributions of the reserve and of the IBNR reserve are outlined. In section 7 the Bootstrap method as a tool to create the estimation uncertainty is briefly outlined.

2. THE MARKED POISSON PROCESS

The claims process is described in the framework of MPPs. The advantage is that general statements and results are available from this theory which facilitates the construction of the likelihood. Decomposing the process as described below turns out to be particularly useful.

2.1. Definition of the MPP

Norberg (1999) defines a MPP as follows:

A *claim* is a pair $C = (T, Z)$, where T is the time of occurrence of the claim and Z is the so-called mark describing its development from the time of occurrence until the time of final settlement.

The *claims process* is a random collection of claims $\{(T_i, Z_i)\}_{i=1, \dots, N}$, the index i indicating chronological order so that $0 < T_1 < T_2 < \dots$

It is assumed that the times are generated by an inhomogeneous Poisson process with intensity $w(t)$ at time $t > 0$.

It is assumed that the distribution of the mark Z_t only depends on t through T_t i.e. is of the form $Z_t = Z_{T_t}$, where $\{Z_t\}_{t>0}$ is a family of random elements that are mutually independent and also independent of the Poisson process, and $Z_t \sim P_{Z:t}$.

The claim process is then called a Marked Poisson Process (MPP) with intensity $w(t)$ and position-dependent marking $P_{Z:t}$ and we write

$$\{(T_i, Z_i)\}_{i=1, \dots, N} \sim Po(w(t), P_{Z:t}; t > 0). \quad (2.1)$$

We shall exclusively consider marks of the form

$$Z_i = (J_i, Y_{J_i}, Y_{J_i+1}, \dots, Y_{D,i}, G_i) \quad (2.2)$$

with domain $J_i \in \{1, \dots, D\}$, $Y_{J_i} + \dots + Y_{J_i+n,i} \geq 0$, $n \in \{0, \dots, D - J_i\}$, $G_i \in C$, where i is the claim identification index, J_i is the stochastic reporting delay in years i.e. $J_i = 1$ if the claim i is reported within the calendar year of occurrence, $J_i = 2$ if reported the year after etc. and where $Y_{k,i}$ is the stochastic incremental incurred amount in the development period $k \in \{J_i, \dots, D\}$ (we implicitly assume that the claims are settled after D development years) and $G_i \in C$ is a discrete stochastic characteristic of the claim, for example claim-type and information from the policy the claim is covered under.

We shall denote $t(i)$ the outcome of T_i and $I_i = 1 + [T_i]$ i.e. the stochastic year of occurrence concerning claim i and $i(i)$ the outcome of I_i . Similarly $j(i)$ is the outcome of J_i i.e. the reporting delay concerning claim i and $g(i)$ the outcome of G_i i.e. the characteristic of the claim i .

The claim identification index i will frequently be omitted i.e. $Y_k = Y_{k,i}$, $i = i(i)$, $j = j(i)$ etc.

CLAIMS SETTLEMENT:

We will briefly discuss the situation where the settlement of the claim is included in the mark: We consider the indicator variables $U_{k,i}$ where $U_{k,i} = 1$ if the claim i is closed by the end of the development period k and $U_{k,i} = 0$ otherwise and consider marks of the form

$$Z = (J, Y_j, Y_{j+1}, \dots, Y_D, U_j, U_{j+1}, \dots, U_D, G) \text{ where index } i \text{ has been omitted.} \quad (2.3)$$

2.2. The stochastic reserve

The stochastic outstanding claims reserve R_D (in excess of the individual case reserve) at the end of year D , is defined as the sum of all incremental amounts $Y_{k,i}$ incurred after time D concerning claims that have occurred by the end of

year D i.e. where the sum goes over all i and k where $i(i) \leq D$ and $D + 1 - i(i) < k \leq D$. For simplicity the index i has been omitted below:

$$R_D = \sum_{i \leq D, D+1-i < k \leq D} Y_k \tag{2.4}$$

The stochastic IBNR reserve $IBNR_D$ at time D is defined as the sum of all incremental amounts in the future concerning claims that have occurred at time D but are reported after time D i.e. where $j(i) > D + 1 - i(i)$:

$$IBNR_D = \sum_{j > D+1-i, i \leq D, D+1-i < k \leq D} Y_k \tag{2.5}$$

We are interested in the conditional distribution of both the R_D -reserve and of the $IBNR_D$ -reserve given the information at time $t = D$ or more generally in the conditional distribution of the MPP $Po(w(t), P_{Z,t}; D \geq t \geq 0)$ given the process' value at the end of period D , i.e. at time $t = D$.

2.3. Decomposing the process

The MPP can be decomposed into independent sections. The advantage of this is that the likelihood can be split into corresponding products which can be maximised in isolation.

We will consider a partitioning of the calendar year into q intervals or seasonal periods, for example quarters, months or even days. The motivation for this partitioning is that while the distributions can change over time $(0, D)$ it would be reasonable to assume that the changes through the shorter intervals are negligible.

Let $0 = s_0 < \dots < s_q = 1$ be fixed values. The intervals $[i - 1 + s_{m-1}, i - 1 + s_m], i = 1, \dots, D, m = 1, \dots, q$ will be denoted by i_m .

The interval that the point of time t_i belongs to is denoted $i_m(t)$ or just i_m . Similarly we shall denote $m(t)$ (or just m) the interval number concerning t_i .

We now decompose the process by the values of (i, m, j, g) . In other words, for each combination of i, m, j and g we consider the process where $T \in i_m$ and $Z = (j, Y_j, Y_{j+1}, \dots, Y_D, g)$. The process is here called the (i, m, j, g) -component process and the claims are called the (i, m, j, g) -claims. The number of (i, m, j, g) -claims will also be denoted $N_{i,m,j,g}$ or $N(i, m, j, g)$.

We will also identify the claims by (i, m, j, g, n) where $n = 1, \dots, N_{i,m,j,g}$ so that, after rearranging the indices, the incremental amounts are $Y_{i,m,j,k,g,n}$.

We then have the following expressions for the outstanding claims reserve R_D and for the IBNR reserve $IBNR_D$:

$$R_D = \sum_{\substack{n=1, \dots, N(i, m, j, g) \\ i=1, \dots, D, j=1, \dots, D \\ m=1, \dots, q, g \in G \\ k > D-i+1}} Y_{i, m, j, k, g, n} \tag{2.6}$$

and

$$\text{IBNR}_D = \sum_{\substack{n=1, \dots, N(i, m, j, g) \\ i=1, \dots, D \\ m=1, \dots, q, g \in G \\ k \geq j > D-i+1}} Y_{i, m, j, k, g, n} \quad (2.7)$$

Below we will show that under certain assumptions the conditional distributions of the sums given the process' value at time D can be regarded as sums of stochastically (conditional) independent variables.

It follows (Norberg (1999)) that the (i, m, j, g) -component processes are MPPs, that they are independent and that the (i, m, j, g) -claims occur with an intensity which is the claim intensity multiplied by the probability that the claim is a (j, g) -claim i.e.

$$\begin{aligned} w_{i, m, j, g}(t) &= w(t)P_{Z:t}\{J=j, G=g\} \\ &= w(t)P_{Z:t}\{J=j|G=g\}P_{Z:t}\{G=g\}, t \in i_m. \end{aligned} \quad (2.8)$$

By decomposing the process by (i, m, j, g) the distribution of the mark in the component process $(j, Y_{J,t}, Y_{J+1,t}, \dots, Y_{D,t}, g)$ is equal to the distribution of $(Y_{J,t}, Y_{J+1,t}, \dots, Y_{D,t})$ given (i, m, j, g) . It also follows that $N_{i, m, j, g}$ are independent and Poisson distributed and independent of the marks.

3. A DISCRETE MODEL

From the above it is seen that we must specify the following:

- 1: $w(t)$
- 2: $P_{Z:t}\{G=g\}$
- 3: $P_{Z:t}\{J=j|G=g\}$ and
- 4: $P_{Z:t}\{Y_{J,t}, Y_{J+1,t}, \dots, Y_{D,t}\}$ given (i, m, j, g) .

This is done in the following four sections.

The sections 3.1 and 3.2 concern the distribution of the time of occurrence and of the type of claim. In section 3.3 a 'Chain-Ladder' assumption is made which is suitable for reserving purposes where information concerning the 'future' is missing and where extrapolation of information concerning the past into the future is required.

Section 3.4 deals with the multi-dimensional distribution of the incurred amounts in the years after occurrence for a single claim. Assumptions are made to discretise the distribution and to reduce the D -dimensional problem into a more practical two-dimensional problem.

3.1. $w(t)$

We assume that the intensity $w(t)$ is constant in year i , except for the same seasonal variation within the year, i.e. we assume that there exist positive figures $w_i, i = 1, \dots, D$ and $\sigma_m, m = 1, \dots, q$ so that:

$$w(t) = w_i \sigma_m, \text{ for } t \in i_m. \tag{3.1}$$

3.2. $P_{Z:t}\{G = g\}$

While allowing for changes in business mix and/or claim type mix over the period $[0, D]$ we assume that $P_{Z:t}\{G = g\}$ is constant for $t \in i_m$ i.e. that the change in business mix through these shorter periods i_m is negligible.

Let $e_{i,m}$ be the exposure in the interval i_m and $e_{i,m,g}$ the exposure concerning the (i, m, j, g) -claims.

We will then use the parameterisation

$$P_{Z:t}\{G = g\} = c(e_{i,m,g}/e_{i,m})f(i, m, g), \text{ } t \in i_m \tag{3.2}$$

where $c > 0, f(i, m, g) > 0$ and $f(1, 1, g_1) = 1$ for a *reference level* g_1 of G .

3.3. $P_{Z:t}\{J = j | G = g\}$

In order to estimate the distribution of J given G , i.e. of the reporting delay for each group of business g , we assume that this distribution is independent of year of occurrence.

We therefore assume that the distribution of the reporting delay for each group g of business is ‘the same’ for each year. Formally, we assume that $P_{Z:t}\{J = j | G = g\}, t \in i_m$, only depends on t through $t - [t]$ i.e. that there exists a function r' of $(j, g, t - [t])$, for which

$$P_{Z:t}\{J = j | G = g\} = r'(j, g, t - [t]), \text{ } t \in i_m \tag{3.3}$$

This implies that the conditional distribution of the development delay J given G is independent of the year of occurrence i . The assumption made therefore corresponds to the assumption that is often made when e.g. the aggregated number of claims is modelled using simple Chain-Ladder.

RESULT 1:

As a consequence of the assumptions made in sections 3.1, 3.2 and 3.3 the number, $N_{i,m,j,g}$, of (i, m, j, g) -claims is Poisson distributed with mean

$$\begin{aligned} E(N_{i,m,j,g}) &= \int_{t \in (i,m)} e_{i,m} w_{i,m,j,g}(t) dt \\ &= e_{i,m,g} c w_i \sigma_m f(i, m, g) \int_{t \in (i,m)} r'(j, g, t - [t]) dt \\ &= e_{i,m,g} c w_i \sigma_m f(i, m, g) r(m, j, g) \end{aligned} \tag{3.4}$$

where

$$r(m, j, g) = \int_{t \in (i, m)} r'(j, g, t - [t]) dt. \quad (3.5)$$

It should be emphasised that the development pattern $r(m, j, g)$ concerning g -claims can be dependent on g and m but not on i .

3.4. $P_{Z:t}\{Y_{j,t}, Y_{j+1,t}, \dots, Y_{D,t}\}$ given (i, m, j, g) .

The remaining part of the distribution to be specified is of the mark $(Y_{j,t}, Y_{j+1,t}, \dots, Y_{D,t})$, in the (i, m, j, g) -components processes.

In order to handle the multi-dimensional time dependent distribution of the mark $(Y_{j,t}, Y_{j+1,t}, \dots, Y_{D,t})$ we make two assumptions:

1: $P_{Z:t}(Y_j, \dots, Y_D)$ only depends on t through (i, m) i.e.

$$P_{Z:t}(Y_j, \dots, Y_D) = P(Y_j, \dots, Y_D), t \in i_m \quad (3.6)$$

It then follows that the $(Y_{j,t}, Y_{j+1,t}, \dots, Y_{D,t})$ are identically distributed given (i, m, j, g) . It is seen that the assumption is fulfilled if, for example, all increases of the incurred amounts occur at the beginning of the period i_m .

2: The conditional distribution of $(Y_k | Y_{k-1}, \dots, Y_j)$, $k = j+1, \dots, D$, only depends on (Y_{k-1}, \dots, Y_j) through a function of (Y_{k-1}, \dots, Y_j) i.e. a function h exists so that

$$P_{Z:t}(Y_k | Y_{k-1}, \dots, Y_j) \sim P_{Z:t}(Y_k | h(Y_{k-1}, \dots, Y_j)). \quad (3.7)$$

Let $S_k = Y_k + \dots + Y_j$ i.e. the accumulated incurred amount. The assumption is for example fulfilled for $h(Y_{k-1}, \dots, Y_j) = S_k$ if the process (S_k) , $k = j, \dots, D$ is a Markov Chain.

We have omitted the index t from the Y_k and S_k .

RESULT 2:

Under the assumptions 1) and 2) we have

$$\begin{aligned} P_t(Y_D, Y_{D-1}, \dots, Y_j) &= P_t(Y_D | Y_{D-1}, \dots, Y_j) P_t(Y_{D-1}, \dots, Y_j) = \dots \\ &= P(Y_D | h(Y_{D-1}, \dots, Y_j)) * P(Y_{D-1} | h(Y_{D-2}, \dots, Y_j)) * \dots * P(Y_j), t \in i_m \end{aligned} \quad (3.8)$$

The main advantage is that the conditional distributions are independent while still maintaining a possibility that the incremental amounts are dependent on the past developments.

Below we shall only deal with the situation where

$$h(Y_{k-1}, \dots, Y_j) = S_k. \quad (3.9)$$

However, h could be extended in different ways, e.g. h could include the information whether or not the total incurred amount has been positive at some stage in the past.

CLAIMS SETTLEMENT:

If the indicators U_k ($= 1$ if the claim is closed, otherwise 0) were included in the mark we could for example consider functions of the form $h(Y_{k-1}, \dots, Y_j, U_{k-1}, \dots, U_j) = (h_1(Y_{k-1}, \dots, Y_j), h_2(U_{k-1}, \dots, U_j))$.

RESULT 3: THE LIKELIHOOD

Combining Result 1 and 2 it is seen that the (unconditional) joint distribution of the process can be specified by for each component (i, m, j, g) specifying the independent Poisson distributions of $N_{i,m,j,g}$, and by specifying the distribution of Y_j and of Y_k given $h(Y_{k-1}, \dots, Y_j)$, $k = j + 1, \dots, D$ which are all independent and also independent of the $N_{i,m,j,g}$.

It follows that the likelihood, $L_{i,m,j,g}$, for the observations for the (i, m, j, g) -component is

$$L_{i,m,j,g} = Po(n_{i,m,j,g}, e_{i,m,g} c w_i \sigma_m f(i, m, g) r(m, j, g)) \tag{3.10}$$

$$\prod_{D-i+1 > k \geq j, l=1, \dots, N(i, m, j, g)} P(y_{i,m,j,k,g,l} | h(y_{i,m,j,k-1,g,l}, \dots, y_{i,m,j,j,g,l}), i, m, j, g)$$

Since the components are independent the total likelihood is the product of all the component likelihoods.

RESULT 4: THE CONDITIONAL DISTRIBUTION OF THE PROCESS

We are interested in the conditional distribution of the process given the process at time $t = D$. Obviously, it is only the conditional distribution of the part of the process where $t > D$ we are concerned about.

Since the N 's and Y 's are independent so are the conditional distributions of the N 's and Y 's given the information at time $t = D$.

The conditional distribution of the 'future' $N_{i,m,j,g}$, $j > D - i + 1$ is the same as the unconditional distribution since the N 's are independent.

Let $Y_{s+1} = y_{s+1}, \dots, Y_j = y_j$ be the observed values at time D , i.e. $s + 1 = D$, $j \leq D$. It follows (by successive conditioning) that the conditional distribution of the 'future' Y 's are determined by the conditional distributions $P(Y_k | h(Y_{k-1}, \dots, Y_s, Y_{s+1} = y_{s+1}, \dots, Y_j = y_j))$, $k > D - i + 1$, which are independent and also independent of the N 's.

When the underlying conditional distribution is fully specified (an example is outlined below) the future N 's and Y 's can be simulated and the corresponding reserves calculated by summarising relevant Y 's. Approximations to the distribution of the reserves can then be obtained by repeating the simulations.

3.5. Observation plan:

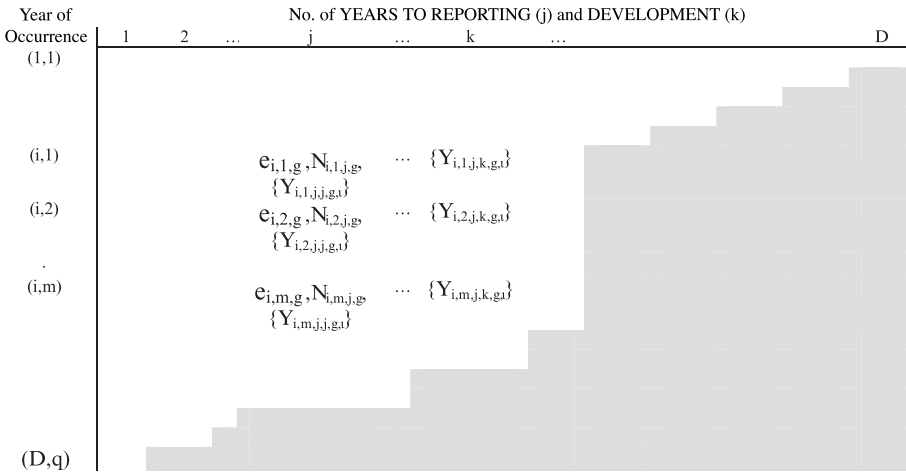
We recall that

- i is the year of occurrence, $1, \dots, D$
- m the season, $1, \dots, q$
- j the number of years to reporting the claim, $1, \dots, D$
- k the number of years to development, $1, \dots, D$
- g a characteristic of the claim, say claim type

- $e_{i,m,g}$ the exposure for g -claims in year i and season m
- $N_{i,m,j,g}$ the number of (i, m, j, g) claims
- $Y_{i,m,j,k,g,t}$ the incremental incurred amount for (i, m, j, g) claims with k years to development, $t = 1, \dots, N_{i,m,j,g}$

We only observe $N_{i,m,j,g}$ for the cells in the past, i.e. where $i + j - 1 \leq D$ and $Y_{i,m,j,k,g,t}$ where $i + k - 1 \leq D, j \leq k$.

The observations can be arranged in a set of upper triangles, one for each j :



The index (i, m, j, g, t) from $Y_{i,m,j,k,g,t}$ will frequently be excluded below.

4. MODELLING THE PROCESS USING GLM

We will for simplicity exclude the seasonal effect i.e. $q = 1$ and omit the index m .

The discrete stochastic characteristic G of the claim is of the form $G = G_1 \times G_2 \times \dots \times G_n$ corresponding to n covariates, for example, $n = 2, G_1 =$ Class of business and $G_2 =$ Claim Type.

4.1. The distribution of the number of claims $N_{i,j,g}$

The number of claims $N_{i,j,g}$ is fully specified by the mean structure since it is Poisson distributed. It is assumed that there are no interactions i.e. that the mean has the form

$$E(N_{i,j,g_1,g_2,\dots,g_n}) = e_{i,j,g_1,g_2,\dots,g_n} c f_I(i) f_J(j) f_1(g_1) \dots f_n(g_n) \tag{4.1}$$

where $(g_1, \dots, g_n) \in G_1 \times \dots \times G_n$, $f_I, f_J, f_1, \dots, f_n$ are positive functions of the covariate levels, $c > 0$ and where $e_{i,j,g_1,g_2,\dots,g_n}$ is the exposure in number of insurance years. Please note that the exposure is independent of the reporting delay j and of any covariate which does not originate from the policy covering the claim such as claim type.

This is a GLM with exposure as offset, log as link function and Poisson as distribution.

REMARKS:

In order to acquire a reasonable fit, interactions can be included in the model. However interactions between I and J could have implications in forecasting which would need careful consideration. For simplicity interactions are ignored in this paper.

For each covariate there is a *reference-level* for which the factor is 1. Therefore the mean concerning the reference cell (i.e. the cell consisting of the combination of all the reference-levels) is proportional to the exposure i.e. $E(N) = ec$ for the reference cell.

Please note that in the situation where there are no covariates g the model gives the same estimates as the Chain Ladder model based on volume weighted averages. This follows from the fact that the sums N_i' and N_j' of the *estimated* values $N_{i,j}'$ in the upper triangle in both models are equal to the *observed* sums N_i and N_j .

By offsetting by the exposure $e_{i,j,g_1,g_2,\dots,g_n}$ the pure period effect $f_I(i)$ can be quantified as well as the effect implied by changes in the mix of claims.

Changes in the intensity of occurrence can be smoothed by treating the I -factor f_I as a continuous variable.

4.2. The distribution of the incremental amounts $Y_{i,j,j,g}$

We are considering the distribution of the amount incurred in the reporting period, $k = j$. The probability for the event $\{Y_{i,j,j,g} = 0\}$ is positive and the distribution of $Y_{i,j,j,g}$ can be specified by the probability $P(Y_{i,j,j,g} = 0)$ and by the conditional distribution $P(Y_{i,j,j,g} | Y_{i,j,j,g} > 0)$.

The conditional distribution of $Y_{i,j,j,g}$ (given $Y_{i,j,j,g} > 0$) (or a transformation of it) could be assumed to be a member of the exponential family and then the parameters estimated using maximum likelihood estimation. This might be a reasonable assumption for some portfolios. Alternatively, if this assumption is

not suitable, the quasi-likelihood approach could be taken. However, since the reserve distributions will be determined by simulation, difficulties when simulating the incremental amounts would need to be overcome.

In this paper we will take another approach by modelling the small and large amounts separately. It turns out that the Generalised Pareto Distribution provides a accurate description of the large claims and simulation from this is straightforward.

A large value L is chosen. First we specify the probabilities for the three disjoint events $\{Y_{i,j,j,g} = 0\}$, $\{0 < Y_{i,j,j,g} < L\}$ and $\{L \leq Y_{i,j,j,g}\}$. They are uniquely determined by the conditional probabilities $p_{>0} = P(Y_{i,j,j,g} > 0)$ and $p_{>L} = P(Y_{i,j,j,g} > L \mid Y_{i,j,j,g} > 0)$ via the expressions

$$\begin{aligned} P(Y_{i,j,j,g} = 0) &= 1 - p_{>0}, \\ P(0 < Y_{i,j,j,g} < L) &= (1 - p_{>L})p_{>0} \text{ and} \\ P(L \leq Y_{i,j,j,g}) &= p_{>L}p_{>0}. \end{aligned} \quad (4.2)$$

The probabilities $p_{>0}$ and $p_{>L}$ are both assumed to be of the form $1/(1+p)$ where

$$p = p_{i,j,j,g_1,g_2,\dots,g_n} = cf_I(i)f_J(j)f_1(g_1)\dots f_n(g_n). \quad (4.3)$$

This is a Logistic Regression Model with the logit function as link function.

Secondly we define the conditional distributions of $Y_{i,j,j,g}$ given the above events:

$\{0 < Y_{i,j,j,g} < L\}$:

The conditional distribution of $Y_{i,j,j,g}$ given $\{0 < Y_{i,j,j,g} < L\}$ is assumed to be Gamma distributed with mean and variance of the form

$$E_c(Y_{i,j,j,g_1,g_2,\dots,g_n}) = cf_I(i)f_J(j)f_1(g_1)\dots f_n(g_n) \text{ and} \quad (4.4)$$

$$V_c(Y_{i,j,j,g_1,g_2,\dots,g_n}) = E_c(Y_{i,j,j,g_1,g_2,\dots,g_n})^2 \varphi \quad (4.5)$$

where E_c and V_c denotes the conditional mean and variance given $\{0 < Y_{i,j,j,g} < L\}$.

The support for the Gamma distribution is $\{0 < y\}$ and therefore the choice of distribution is not entirely consistent. However, if L is large this is not necessarily a significant problem in practice.

$\{L \leq Y_{i,j,j,g}\}$:

The conditional distribution of $Y_{i,j,j,g}$ given $Y_{i,j,j,g} \geq L$ is assumed to be a Generalised Pareto Distribution i.e. the distribution function is of the form

$$F(y) = 1 - [1 + (y-L)/(\alpha\beta)]^{-\alpha}, \quad \alpha > 0, \beta > 0. \quad (4.6)$$

The distribution does not depend on $S_{i,j,k-1,g}$ as defined in (3.9), however, this could be implemented if required.

4.3. Distributions of $Y_{i,j,k,g}$ given $S_{i,j,k-1,g} = 0, k = j + 1, \dots, D$

The distribution of the incremental amounts $Y_{i,j,k,g}$ incurred after the reporting period are defined similarly to the distribution of $Y_{i,j,j,g}$ above, however a covariate concerning the development delay k is also incorporated, for example it is assumed that the conditional distribution of $Y_{i,j,k,g}$ given $\{0 < Y_{i,j,k,g} < L\}$ is Gamma with mean and variance of the form

$$E_c(Y_{i,j,k,g_1, g_2, \dots, g_n}) = cf_i(i)f_j(j)f_K(k)f_1(g_1) \dots f_n(g_n) \text{ and} \tag{4.7}$$

$$V_c(Y_{i,j,k,g_1, g_2, \dots, g_n}) = E_c(Y_{i,j,k,g_1, g_2, \dots, g_n})^2 \varphi \tag{4.8}$$

where E_c and V_c denote the conditional mean and variance given $\{0 < Y_{i,j,k,g} < L\}$.

It could further be tested/assumed that the factors concerning I, J and the covariates are the same for the distributions of $Y_{i,j,j,g}$ and for the conditional distribution of $Y_{i,j,k,g}$ given $S_{k-1} = 0$ and that the distributions only differ via the factors concerning the development k .

4.4. Distribution of $Y_{i,j,k,g}$ given $S_{i,j,k-1,g}, S_{i,j,k-1,g} > 0, k = j + 1, \dots, D$

This situation is different from the above since the incremental amounts can be negative.

First we specify the probability for the following five disjoint sets with joint probability 1: $\{Y_{i,j,k,g} = 0\}, \{0 < Y_{i,j,k,g} < L\}, \{L \leq Y_{i,j,k,g}\}, \{0 > Y_{i,j,k,g} > -S_{i,j,k-1,g}\}$ and $\{Y_{i,j,k,g} = -S_{i,j,k-1,g}\}$.

It is seen that the probabilities are uniquely determined by the conditional probabilities $p_{>0} = P(Y_{i,j,k,g} > 0 | S_{i,j,k-1,g}), p_{>L} = P(Y_{i,j,k,g} > L | Y_{i,j,k,g} > 0, S_{i,j,k-1,g}), p_{=0} = P(Y_{i,j,k,g} = 0 | Y_{i,j,k,g} \leq 0, S_{i,j,k-1,g})$ and $p_{S>0} = P(Y_{i,j,k,g} > -S_{i,j,k-1,g} | Y_{i,j,k,g} < 0, S_{i,j,k-1,g})$ since we have the expressions:

$$\begin{aligned} P(Y_{i,j,k,g} = 0 | S_{i,j,k-1,g}) &= p_{=0}(1 - p_{>0}), \\ P(0 < Y_{i,j,k,g} < L | S_{i,j,k-1,g}) &= (1 - p_{>L})p_{>0}, \\ P(L \leq Y_{i,j,k,g} | S_{i,j,k-1,g}) &= p_{>L}p_{>0}, \\ P(0 > Y_{i,j,k,g} > -S_{i,j,k-1,g} | S_{i,j,k-1,g}) &= p_{S>0}(1 - p_{>0})(1 - p_{=0}) \text{ and} \\ P(Y_{i,j,k,g} = -S_{i,j,k-1,g} | S_{i,j,k-1,g}) &= (1 - p_{S>0})(1 - p_{>0})(1 - p_{=0}) \end{aligned} \tag{4.9}$$

Let $0 = x_0 < x_1 < \dots < x_h$ be fixed values (where x_h is sufficiently large) and let $SG_{i,j,k,g}$ be the right interval point that $S_{i,j,k-1,g}$ belongs to, i.e.

$$SG_{i,j,k,g} = \min\{x_s : (Y_{i,j,k-1,g} + \dots + Y_{i,j,j,g}) \leq x_s\} \tag{4.10}$$

We assume that $p_{=0}, p_{>0}, p_{>L}$ and $p_{S>0}$ only depends on $S_{i,j,k-1,g}$ via the grouping $SG_{i,j,k,g}$ and that they all have the form $1/(1 + p)$ where

$$p = p_{i,j,k,x_s, g_1, g_2, \dots, g_n} = cf_i(i)f_j(j)f_K(k)f_{SG}(x_s)f_1(g_1) \dots f_n(g_n) \tag{4.11}$$

Secondly we define the conditional distributions of $Y_{i,j,k,g}$ given the above events (for $S_{i,j,k-1,g} > 0$):

$$\{0 < Y_{i,j,k,g} < L, S_{i,j,k-1,g}\}$$

The conditional distribution of $Y_{i,j,k,g}$ given $(0 < Y_{i,j,k,g} < L, S_{i,j,k-1,g})$ is assumed to be Gamma distributed with mean and variance of the form

$$E_c(Y_{i,j,k,x_s, g_1, g_2, \dots, g_n}) = cf_I(i)f_J(j)f_K(k)f_{SG}(x_s)f_1(g_1) \dots f_n(g_n) \text{ and} \quad (4.12)$$

$$V_c(Y_{i,j,k,x_s, g_1, g_2, \dots, g_n}) = E_c(Y_{i,j,k,x_s, g_1, g_2, \dots, g_n})^2 \varphi, \quad (4.13)$$

where E_c and V_c denote the conditional mean and variance and where the covariate SG , capturing the accumulated incurred amount at the beginning of the period, is incorporated. The distribution only depends on $S_{i,j,k-1,g}$ through x_s .

$$\{Y_{i,j,k,g} \geq L, S_{i,j,k-1,g}\}$$

The conditional distribution of $Y_{i,j,k,g}$ given $(Y_{i,j,k,g} \geq L, S_{i,j,k-1,g})$ is assumed to be a Generalised Pareto Distribution which is not dependent on $S_{i,j,k-1,g}$ i.e. the density function is of the form

$$F(y) = 1 - [1 + (y - L)/(\alpha\beta)]^{-\alpha}, y > L, \alpha > 0, \beta > 0. \quad (4.14)$$

The distribution does not depend on $S_{i,j,k-1,g}$, however, this could be implemented if required.

$$\{0 > Y_{i,j,k,g} > -S_{i,j,k-1,g}, S_{i,j,k-1,g}\}$$

The range for $Y_{i,j,k,g}$ given $(0 > Y_{i,j,k,g} > -S_{i,j,k-1,g}, S_{i,j,k-1,g})$ is obviously $]-S_{i,j,k-1,g}, 0[$ [and therefore the range for $-\log((Y_{i,j,k,g} + S_{i,j,k-1,g})/S_{i,j,k-1,g})$ is $]0, \infty[$.

The conditional distribution of $Y_{i,j,k,g}$ given $(0 > Y_{i,j,k,g} > -S_{i,j,k-1,g}, S_{i,j,k-1,g})$ is specified by assuming that the distribution of $-\log((Y_{i,j,k,g} + S_{i,j,k-1,g})/S_{i,j,k-1,g})$ is Gamma distributed with mean and variance of the form as above.

REMARKS:

In order to acquire a reasonable fit, interactions can be included in the model. However interactions between I and J could have implications in forecasting which would need careful consideration. For simplicity interactions are ignored in this paper.

The pure period effect $f_I(i)$ can be quantified as well as the combined effects of $f_I(g_1) \dots f_n(g_n)$ which is the effect of changes in the mix of claims. The pure period effect (i.e. the pure claims inflation) in the incurred amounts can be smoothed by treating the I -factor f_I as a continuous variable.

CLAIMS SETTLEMENT:

Let us briefly consider the situation where the indicators for closed claim U_k are included in the mark and consider the functions

$$h(Y_{k-1}, \dots, Y_j, U_{k-1}, \dots, U_j) = ((Y_{k-1} + \dots + Y_j), U_{k-1}). \tag{4.15}$$

The distributions can be specified by, for example, assuming that Y_k and U_k are conditionally independent given $((Y_{k-1} + \dots + Y_j), U_{k-1})$ and then specifying the marginal distributions. The marginal distribution concerning Y_k and the ‘event’-probabilities can be specified in the same way as above where an extra covariate f_U concerning U is included in the GLMs. The marginal distributions of U_k given $((Y_{k-1} + \dots + Y_j), U_{k-1} = 1)$ and of U_k given $((Y_{k-1} + \dots + Y_j), U_{k-1} = 0)$ can be modelled using Logistic Regression.

5. AN EXAMPLE

We will illustrate the model based on a Marine portfolio with policy and claims information available from the period 1992-2004.

5.1. Data

Data consists of the following:

<i>Policy records</i>	<i>Claims records</i>
Policy Id	Policy Id
Start date	Claim Id
End date	Claim date
	Reporting date
Vessel type	Claim type
Vessel tonnage	Incurred amount
Class of business	Transaction date

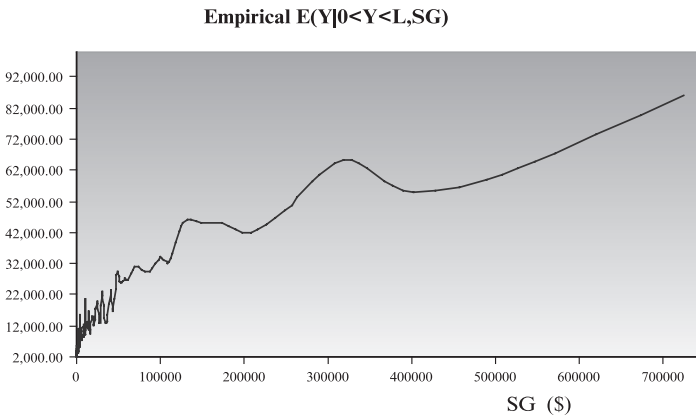
Based on this data the sufficient statistics are created:

<i>N</i>	<i>Y</i>
	l : claim id
i : year of occurrence	i : year of occurrence
j : reporting period	j : reporting period
e : exposure in years	k : development period
g_1 : grouped claim type	g_1 : grouped claim type
g_2 : grouped vessel type	g_2 : grouped vessel type
g_3 : grouped vessel tonnage	g_3 : grouped vessel tonnage
g_4 : class of business	g_4 : class of business
N : Number of claims	Y_k : incurred amount in the development period k
	S_{k-1} : accumulated incurred amount at the end of development period $k-1$
	SG_k : the discretised value of S_{k-1} .

The N -data has been created as follows: The exposure is first summarised by all combinations of (i, g_2, g_3, g_4) . Then the exposure for the combination of $(i, j, g_1, g_2, g_3, g_4)$ is defined as the exposure for the projection (i, g_2, g_3, g_4) i.e. the exposure is independent of reporting delay j and claim-type g_1 .

The Y -data has been created as follows: For all observable combinations of i and k (i.e. where $i + k \leq D + 1$) where there are no records of incurred amount a record is generated with $Y_k = 0$ and thereafter the S_{k-1} and SG_k values are calculated.

As an example the empirical mean as a function of SG is outlined below. The greater the accumulated incurred amount at the beginning of the period the greater the future average increase given that it is positive and less than \$500,000. However, this trend does not continue for $SG > 700,000$ where the mean is approximately \$85,000 and independent of SG .



5.2. Estimation method

The estimated parameters concerning the Poisson, Gamma and Logit models are the maximum likelihood estimates. The parameters concerning the Generalised Pareto Distributions are fitted using non linear regression analysis where the ‘distance between the empirical d.f. and model d.f.’ is minimised. The process of fitting the parameters in the model-components will be illustrated below by a few examples.

5.3. Example 1

In order to specify a reasonable model thorough empirical analyses are required. As a first example we will illustrate the impact of the SG -criteria on the likelihood that the incremental amount is greater than $L = \$500,000$ given that it is greater than 0 i.e.

$$p_{>L} = P(Y_{i,j,k,g} > L | Y_{i,j,k,g} > 0, S_{i,j,k-1,g} > 0). \tag{5.1}$$

The probability $p_{>L}$ is of the form $1/(1 + p)$ where

$$p = p_{i,j,k,s_g,g_1,g_2,\dots,g_n} = cf_i(i)f_j(j)f_k(k)f_{SG}(x_s)f_1(g_1)\dots f_4(g_4). \tag{5.2}$$

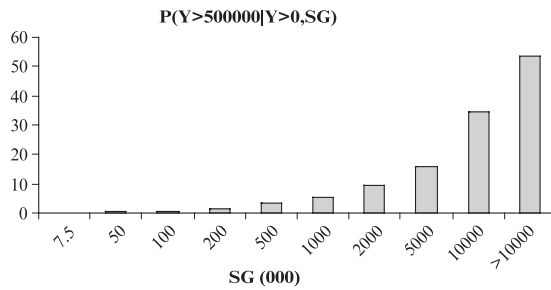
The sufficient statistic is the number of ‘trials’ T and number of ‘hits’ H :

$$T_{i,j,k,s_g,g_1,g_2,g_3,g_4} = \sum I(Y_{i,j,k,s_g,g_1,g_2,g_3,g_4} > 0, S_{i,j,k-1,g_1,g_2,g_3,g_4} > 0) \tag{5.3}$$

$$H_{i,j,k,s_g,g_1,g_2,g_3,g_4} = \sum I(Y_{i,j,k,s_g,g_1,g_2,g_3,g_4} > L, S_{i,j,k-1,g_1,g_2,g_3,g_4} > 0) \tag{5.4}$$

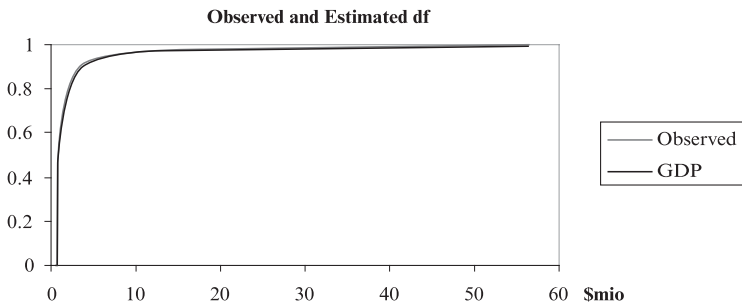
where the summary is over the claims identification i .

The observed (equal to estimated) hit-rate, H/T , for each level of the SG -criteria are outlined below. The hit-rate increases dramatically by the SG -value i.e. for claims where the incurred amount at the beginning of the period is large there is a much higher likelihood that the incremental value is greater than \$500,000 given that the incremental value is positive.



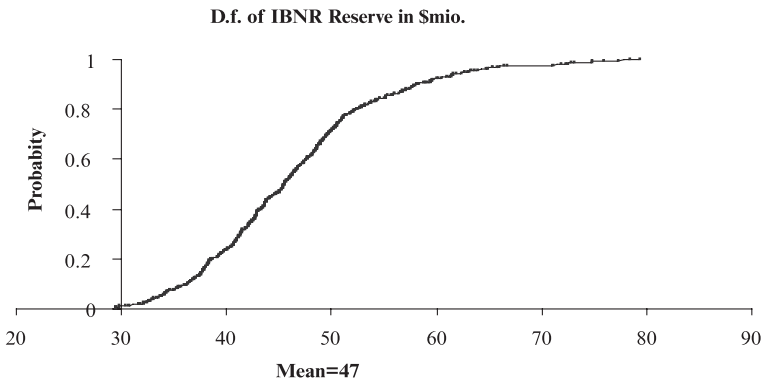
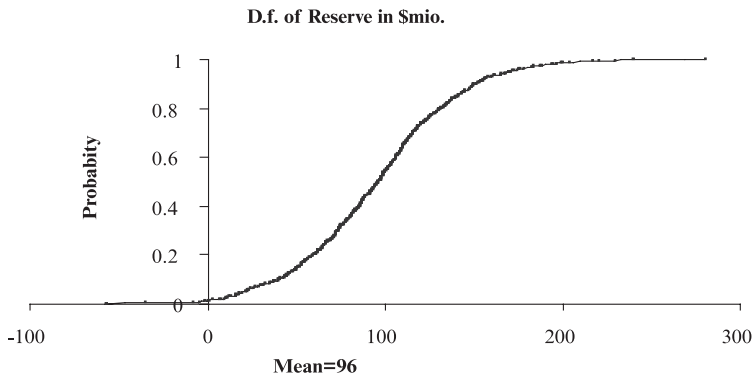
5.4. Example 2

We will now focus on the conditional distribution of $Y_{i,j,k,g}$ given that it is greater than $L = \$500,000$ and given that $S_{i,j,k-1,g} = 0$. 371 yearly incremental amounts of this kind have been observed. A section of the empirical d.f. and the fitted Generalised Pareto d.f. are outlined below.



6. ESTIMATING THE DISTRIBUTION OF THE RESERVES R_D AND $IBNR_D$

Despite the fact that the joint distribution is fully specified by the one-dimensional conditional distributions, an exact calculation of the conditional distribution of the R_D -reserve and of the $IBNR_D$ -reserve given all information in the past, is not simple. The distribution of the total reserve R_D and of the $IBNR_D$ reserve, based on 500 simulated 'ultimate' projections, is outlined below. 500 repetitions seem sufficient for estimating the fractions up to 95% but reliable estimates for higher fractions would need more simulations. The average time per simulation is approximately 0.5 min. on a Pentium 4, CPU 3.00 GHz, 1.00 GB of RAM.



7. ESTIMATION OF UNCERTAINTY

Since the observation is random, the estimated parameters, as functions of the observation, are also random. The implication is that the projected distribution of the reserve as described above is uncertain. To quantify this uncertainty the Bootstrap method could be applied since the claims are assumed to be

stochastically independent, however this has not yet been implemented. This solution would only be practical if all the programs corresponding to the model components described in section 4 could be run automatically in a batch and it would be a huge number of GLM- and GPD-estimations.

The steps would be as follows:

- 1) From a Poisson distribution with mean equal to the total number N of observed claims a number M is sampled.
- 2) M claims $\{(T_m, Z_m)\}_{m=1, \dots, M}$ from the set of observed claims $\{(T_i, Z_i)\}_{i=1, \dots, N}$ are sampled with replacement.
- 3) The model parameters $\{p\}$ are estimated based on the sampled claims $\{(T_m, Z_m)\}_{m=1, \dots, M}$.
- 4) One reserve outcome is simulated according to the model and fitted parameters $\{p\}$ (as described above) given the original observation.
- 5) Step 1-4 are repeated e.g. 500 times.

The resulting distribution would be a reasonable approximation to the total uncertainty i.e. the process variation as well as the estimation uncertainty if the number of repetitions in step 5) is sufficiently large and if the empirical distribution is 'close' to the distribution of the underlying process, i.e. if there are 'many' claims.

8. CONCLUSION

It is concluded that the distribution of the outstanding claims liabilities, when detailed individual claims information is available, can be assessed by describing the claims process as a MPP with relatively weak assumptions and using GLM and GPD to specify the components of the model.

The models can include information about settlement and can handle seasonal effects, changes in mix of business and claim types as well as changes in mix of claim size.

The models are more suitable than Chain Ladder models and traditional stochastic models based on aggregated data when the incremental amounts are not independent.

While the distribution of the process has not been specified in an exact closed form, the distribution of any function of the process, including the reserve Gross and Net of Reinsurance, can be approximated via simulation.

Bootstrap could be a practical way forward to assess the combined process variation and estimation uncertainty, however more work would be required.

ACKNOWLEDGEMENT

The author would like to thank Professor Ragnar Norberg for an interesting discussion about MPPs.

9. REFERENCES

- Claims Reserving Manual, Volume I and II, *The (British) Faculty and Institute of Actuaries*.
- DOBSON, A. (1990) An introduction To Generalized Linear Models, London: *Chapman and Hall*.
- EFRON, B. and TIBSHIRANI R.J. (0000) An introduction to the Bootstrap: *Chapman and Hall*.
- ENGLAND, P.D. and VERRALL R.J. (2002) Stochastic claims reserving in general insurance, *British Actuarial Journal* **8(3)**, 443-518.
- HAASTRUP, S. and ARJAS, E. (1996) Claims Reserving in Continuous Time: A Nonparametric Bayesian Approach, *ASTIN Bulletin* **26(2)**, 139-164.
- HESSELAGER, O. (1995) Modelling of discretized claim numbers in loss reserving, *ASTIN Bulletin* **25(2)**, 119-135.
- HOEDEMAKERS, T., BEIRLANT, J., GOOVAERTS, M.J. and DHAENE, J. (2005) On the distribution of discounted loss reserves using generalized linear models, *Scandinavian Actuarial Journal* **1**, 000-000.
- MACK, T. (1993) Distribution-free calculation of the standard error of chain ladder reserve estimates, *ASTIN Bulletin* **23(2)**, 213-225.
- MAHON, J.B. (2004) Transition Matrix Theory And Individual Claim Loss Development, *Casualty Actuarial Society Forum, Spring 2005*, (<http://www.casact.org.pubs/dpp/dpp04/>).
- NORBERG, R. (1986) A Contribution to Modelling of IBNR Claims, *Scan. Actuarial J. 1986*: 155-203.
- NORBERG, R. (1993) Prediction of outstanding liabilities in non-life insurance, *ASTIN Bulletin* **23(1)**, 95-115.
- NORBERG, R. (1999) Prediction of outstanding claims II: Model variations and extensions, *ASTIN Bulletin* **29(1)**, 5-25.
- TAYLOR, G. and MCGUIRE, G. (2004) Loss Reserving with GLMs: A Case Study, Spring 2004 *Meeting of the Casualty Actuarial Society, Colorado Springs, Colorado 16-19 May 2004* (<http://www.casact.org.pubs/dpp/dpp04/>).

LARSEN & PARTNERS LIMITED
Actuarial Consultants
30 Manor Road
Salisbury
Wilts SP1 1JS
UK
Email: clarsen@fastmail.fm