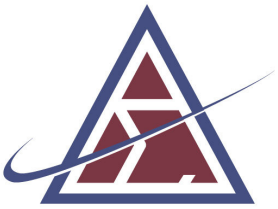


Institute of Actuaries of Australia

Modelling Claim Development Processes

Prepared by Bo Wang

Presented to the Institute of Actuaries of Australia
16th General Insurance Seminar 9-12 November 2008
Coolum, Australia

The Institute of Actuaries of Australia
Level 7 Challis House 4 Martin Place
Sydney NSW Australia 2000
Telephone: +61 2 9233 3466 Facsimile: +61 2 9233 3446
Email: actuaries@actuaries.asn.au Website: www.actuaries.asn.au

Abstract

Pricing for long-tailed products has always being problematic. Take the example of CTP, where claims can take more than ten years to settle, especially for injured children. It is not unusual to have active claims outnumber finalised claims; and, typically, the inclusion of active claims can bias the results, and equally so for the exclusion of active claims.

This paper proposes a framework that can be used to analyse active claims side by side with finalised claims for pricing purposes. This is done by modelling how a claim evolves over time so the ultimate claim size can be inferred from the various attributes of a claim. A positive side effect of this approach is that outstanding claims liability can be estimated in the model framework using a simulation approach. However, reserving is not the motivation of this paper.

The claim development process is broken down into the various component processes – a claim status process that takes on binary values; a delay process that models the time between each successive update to the incurred claims cost variable; a binary process that models whether the claims cost is revised upwards or downwards at each successive update; a positive process that models the size of the revision. These processes are modelled with appropriate distributions and explanatory variables such as injury severity, legal representation, etc.

Effectively, time is measured by the number of changes made to the estimate of the claims cost from the reporting of the claim, making this approach analogous to the operational time approach used in other reserving techniques. This approach can also be used to account for case estimation policy changes, when a certain effect can be built into the claim cost changes occurring within a specific timeframe.

When combined with an appropriate pricing framework, this approach will provide more robust inferences based on all available claims information, rather than only a subset.

1 Introduction

1.1 Background

One of the fundamental issues that need to be dealt with in a pricing exercise is to project claims to their ultimate values. The estimation of the claim costs for the underwriting period under consideration would involve analysis of the claim cost patterns in the recent past. However, the existence of open claims makes this difficult.

Open claims represent an incomplete view of the claims experience and usually cannot be analysed in conjunction with the finalised claims. Typically, case estimates on the open claims do not represent the mean of their ultimate outcomes and are typically biased. Measures of average claim size or average claims cost per policy generally show significant trends for the most recent accident periods due to the distortions caused by these open claims. This issue is most evident in long-tailed products, where a significant proportion of the reported claims remains open.

There are usually a few solutions to this issue. Firstly, the accident periods where open claims present too significant an issue can simply be excluded. Then, the historical accident periods where open claims are no longer an issue can be inflated appropriately to the underwriting period. However, this method fails when the product is undergoing change and the most recent experience is crucial in the analysis of claims experience.

Secondly, the pricing exercise can leverage off the actuarial outstanding claims valuation. The valuation would typically contain outstanding-to-case estimate ratios for each of the accident periods. The claims are then "grossed up" according to their stage of development. However, this approach is across the board and fails when various segments of the claims develop differently. For example, claims from different occupations for a professional liability portfolio may have a different development pattern.

This paper proposes a novel way of modelling claims, using characteristic information of individual claims to model the complete claim cost evolution process. Claim cost evolution refers to the history of the estimate incurred cost (payments to date plus case estimate) from when the claim is reported to when it is finalised, change by change.

By modelling claims individually, this approach falls into the category of "Individual claims models" (Taylor et al., 2006). The main difference between this approach and others in the literature is that the claim evolution is modelled rather than the final result. Too often claim modelling techniques concentrate on the end result, that is, the ultimate claim size, neglecting a vast amount of information contained in the development of the claims. This paper takes the approach of modelling claim development processes by studying how the incurred claims cost changes whenever the case estimate is updated.

One major issue that researchers noted with this category of models is the usage of "dynamic stochastic variables" (Taylor et al., 2006; Prevett and Gifford, 2007). While Taylor et al. noted that predicting these dynamic variables is problematic and neglected

these variables, they did offer a potential solution by using survival analysis to model the time until a claim is finalised. By relating claim status as being a claim characteristic (Open or Finalised), other dynamic variables can also be predicted in a similar fashion. Prevett and Gifford (2007) condensed various dynamic risk characteristics into a value of single-dimensioned "claim state", for which a transition model was built.

1.2 Claim Revisions

This paper models the claim revision process. An open claim would have a set of claim characteristics, against which the claims officer would have put a case estimate. This case estimate would be what the claims officer thinks the claim could cost. This estimate may not be a statistical mean of the claims outcome but an estimate based on another criterion. For example, it could be an estimate at the 75th percentile, so the ultimate claims cost would be below the case estimate 75% of the time.

Some time later, the claims officer would receive additional information about the claim; this new set of information would now be the basis of a revised case estimate. The information may contain updated injury details, the appointment of a plaintiff lawyer or a court decision regarding the claim. The claims officer now needs to make a few decisions, including whether or not the claim is to remain open or can now be settled; whether the new information warrants an upward or downward change and, if so, how much. Since the case estimate that the claims officer makes would be based on the new information and prior information, the use of the dynamic claims information would be insightful to the way claim estimates change.

The complete claim development process is broken down into its component processes. Four separate processes are specified and modelled:

- Delay processes (denoted by t) – a continuous or discreet variable that measures the time until the next change in the case estimate
- Claim status process (denoted by S) – a binary variable that tracks the claim status of a claim
- Movement direction process (denoted by D) – a binary variable that tracks the direction of the incurred cost change
- Size of change process (denoted by Y) – a continuous or discreet process that measures the size of the change.

The modelling of these component processes is conducted in a Generalised Linear Modelling (GLM)framework. The GLM technique is widely used in the actuarial field for both pricing and reserving and it is accepted as a robust modelling tool. In addition, using a GLM framework for these processes enables an easy interpretation of the model parameters.

It is proposed that the main time measurement will not be in conventional time units, but rather on a time scale that counts the number of changes since the occurrence of the claim. Generally, changing claim-handling speed due to staff capacity and claim frequency will affect the duration of a claim on a calendar time scale. However, if a

claim's lifecycle can be thought of as "stages" that prompt new information, the ordinal number of changes may provide greater correlation with the pattern of claim development.

1.3 Uses of the Model

By fitting the various models and obtaining model parameters, a simulation model can be developed to project open claims to their ultimate values. This approach can develop the case estimate of the open claims to an expected ultimate claims cost and hence can be used in a pricing exercise along with finalised claims. The simulation model is not dealt with in this paper. However, by fitting the models and obtaining the parameters, insights are gained into the "drivers" of claims cost evolution.

Although the initial aim of this project was to be able to model open claims together with closed claims in a unified framework, the projection of open claims to an expected ultimate claim size also makes this process a reserving tool. The model, while projecting claims development through time, would fill out the bottom half of a development triangle. However, only the Incurred But not Enough Reported (IBNER) component is projected, with Incurred But not Reported (IBNR) left for some other model.

This approach to valuation is somewhat similar to Statistical Case Estimation (Taylor and Campbell, 2002; Brookes and Prevett, 2004; and Prevett and Gifford, 2007), where claim characteristics are used to project claim payments depending on the "state" the claimant is in for Workers' Compensation schemes.

With a stochastic model as such, the use of simulation techniques can also produce a wide range of possible ultimate claims costs, both on an individual claims level and an aggregated level, potentially adding another measure of volatility to the current staple of measures.

1.4 Structure

Section 2 describes the data used and a few sample paths of how claims develop. Section 3 takes a brief look at the component processes and proposes the distributions that would be used to model them. Section 4 specifies the models used and the likelihood of the claim development process for the chosen distributions and Section 5 presents the results of fitting the data to the chosen models. Section 6 provides some of the improvements that can be made and Section 7 provides some concluding remarks. Appendix A provides the likelihood of the claim development processes and their derivaives; Appendix B shows which variables are found to be significant in the modelling of the various processes.

2 Data

The technique is applied to the NSW CTP industry data. In NSW, CTP is a motor vehicle insurance product that covers bodily injury arising out of driving, or some other limited uses, of motor vehicles. The product is considered to be long tail in nature, as the injury takes some time to stabilise, and the claim settlement process can take time following injury stabilisation, in particular, if the claim is litigated the court decision process may take considerable time.

The Motor Accidents Association (MAA) of NSW regulates the CTP scheme in NSW and keeps a Personal Injury Register to aid pricing, regulation and scheme monitoring. NSW CTP insurers submit their claims information, payment details and case estimates on a quarterly basis. The MAA has kindly granted permission for this paper to use the PIR database to develop the models and to present the results obtained from the data.

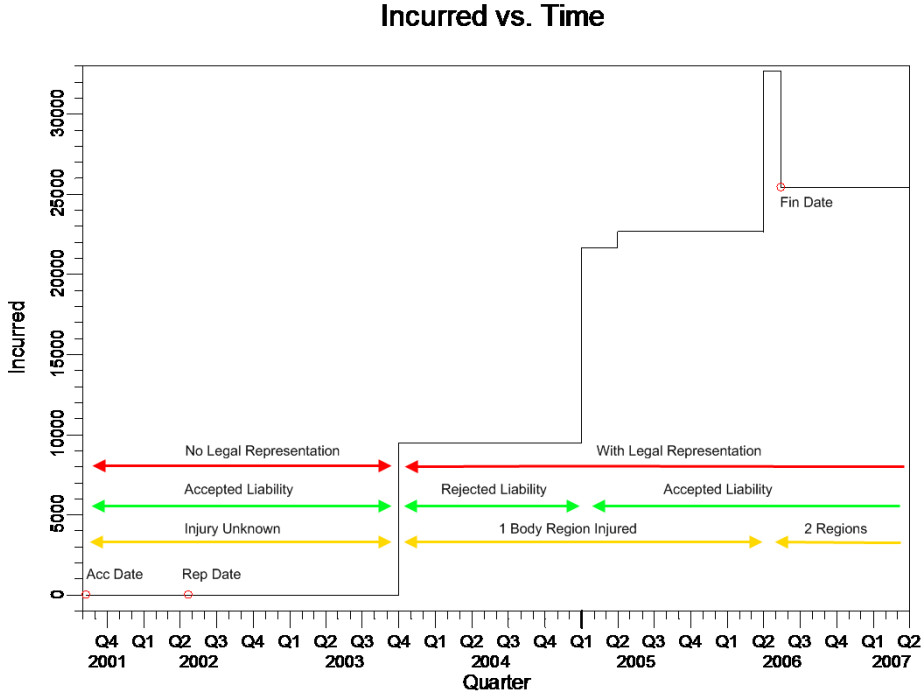
The NSW CTP scheme was deregulated with the Motor Accidents Act 1988 and, after obtaining a licence, private insurers were allowed to participate and issue CTP "Greenslips". In 1999, the Motor Accidents Compensation Act (MACA) was introduced.

The analysis in this paper is limited to 7.5 years of data, that is, accidents occurring from 1 January 2000. After the data were "cleaned" (for example Accident Notification Forms (ANF's) that did not eventuate into a claim were removed; claims finalised in the quarter they were reported were also removed) the dataset contains over 71,000 claims, of which around 60,000 claims are finalised, totalling around 200,000 revisions.

2.1 Sample Paths

The following diagrams present two "sample paths" of claim development. They provide a simple visualisation of the data.

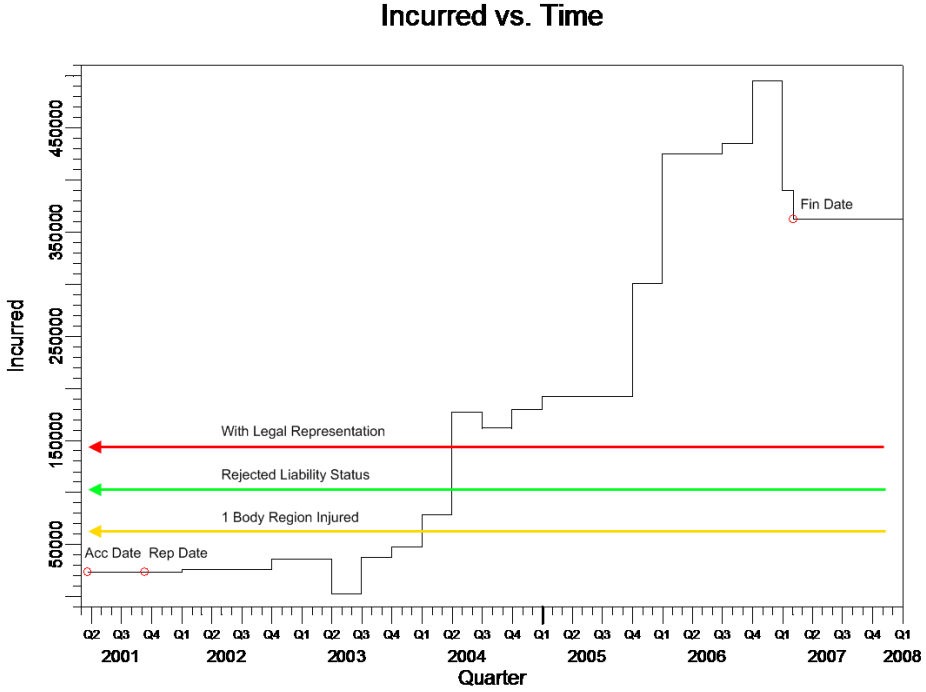
Figure 2-1 – Claim Development Sample Path 1



For the claim above, the accident occurred in Q4 2001 but was not reported until Q3 2002. It then lay dormant for another six quarters; during this period not much was known about the claim. At the end of 2003, more information about the claim was known (including that the claimant obtained legal representation and some information regarding the injury) and immediately the estimate increased to around $10,000. About a year later, the liability status was changed from "Rejected" to "Accepted" and a further sizeable increase in the estimated claims cost was made. Over another year elapsed and the injury seemed to have worsened from one body region to two body regions. The claim was settled soon after with a saving, almost five years after the claim occurred.

The next claim, however, has had very frequent revisions, almost quarterly. Yet, the main characteristics of the claim did not change throughout the life of the claim. It was finalised around six years after it occurred, with three significant upward revisions, but settled with a saving.

Figure 2-2 – Claim Development Sample Path 2



3 Exploratory Analysis

In this section, various features of the data, in particular the variable of interest, are examined and distributions are chosen for the CTP dataset. Other datasets would probably warrant selecting a different set of assumptions and distributions.

3.1 Claim Development: on Real Time vs. Number of Revisions

One of the features of the model is that the maturity of a claim is not calculated on real time, but on the number of revisions to claims cost, somewhat analogous to the concept of "operational time".

Using the number of revisions as a measurement of time, effects of varying speed of finalisation, or pace at which the claims are attended to, can be accounted for. For example, a significant reduction in claims frequency in the NSW CTP scheme occurred after the introduction of MACA. This may have resulted in, amongst other potential drivers, claims being attended to more promptly and frequently. From the data, it is observed that the delays between revisoons are shortening for the more recent years.

Such a proposition would have the effect that claims in more recent accident years are settled faster; and their development with respect to real time should be larger in the initial years of development and smaller in the later years relative to the earlier accident years, as seen from the table below. At the same time, overall average claims sizes are impacted by settlement of large, if these claims are being brought forward. This could be

misinterpreted as superimposed inflation (SI), showing excessive levels of SI in an environment where claims are settled faster.

Table 3-1 – Development Factors by Year

Acc	Development Factor (yearly)					
Year	1 : 0	2 : 1	3 : 2	4 : 3	5 : 4	6 : 5
2000	119%	113%	120%	104%	98%	101%
2001	121%	120%	113%	101%	100%	97%
2002	128%	116%	115%	101%	99%	
2003	126%	120%	114%	95%		
2004	130%	116%	112%			
2005	141%	114%				
2006	138%					
Stdev	8.2%	2.9%	3.0%	3.5%	0.8%	2.7%

*Note, the development factors are the open claims at the end of a year compared to their incurred costs at the end of the next year. That is, finalised claims are removed when calculating the development factors.

Table 3-2 - Average Size by Settlement Maturity

Acc	Development Year						
Year	0	1	2	3	4	5	6
2000	1,999	6,084	18,237	44,252	105,733	135,139	199,032
2001	1,424	6,509	25,502	77,779	124,818	165,394	237,442
2002	1,088	8,408	29,088	74,517	134,622	239,123	
2003	1,707	9,626	32,886	90,576	163,847		
2004	1,850	11,462	45,990	107,282			
2005	2,397	13,432	56,913				
2006	3,686	16,212					
Rate of Inc	12.2%	16.8%	21.7%	19.2%	13.9%	28.5%	17.6%

The Payments per Claims Finalised in Operational Time model, someimes used for the valuation of CTP claim liabilities, can be used to account for the increase in the speed of finalisation. However, the model does not work well when the profile of claims or the order of claims finalising changes. By modelling claim delays on other claims characteristics, the changing pattern in finalisation may be better understood and allowed for.

The development factors based on the number of changes is more stable. In the following table, the development factors by the number of revisions do not exhibit a clear pattern like those factors by development year. The variability of the development factors, as measured by their standard deviation, is comparable between the two time measures. However, the number of claims used to calculate each cell for the number of revisions approach is smaller compared to the development by calendar year approach, especially towards the end for each accident year. Therefore, even though the comparison understates the stability of the development factors by number of revisions, the variability of the development factors by number of revisions is small, especially for the earlier developments.

Table 3-3 Development Factors by Number of Revisions

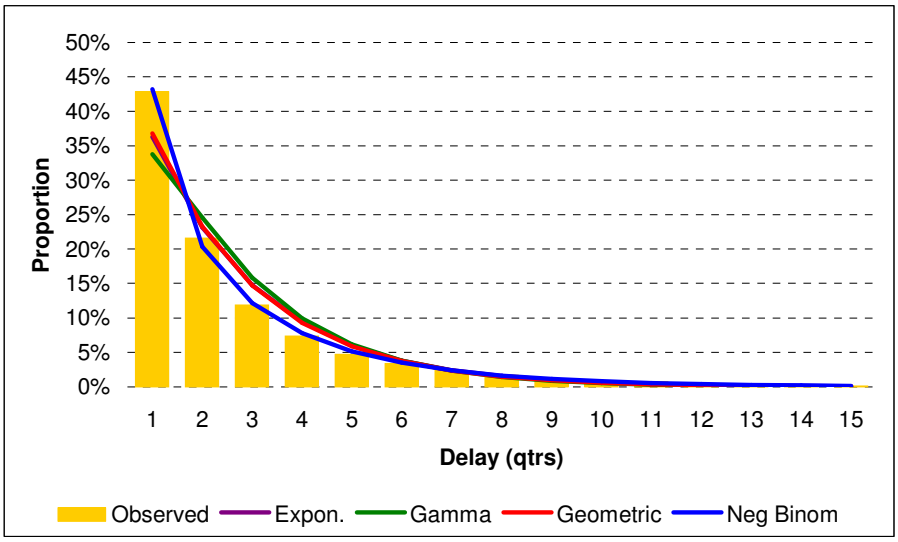
Acc Year	Development Factor (at ith revision to claims cost)									
	1	2	3	4	5	6	7	8	9	10
2000	110.1%	108.3%	107.7%	109.8%	110.7%	103.6%	105.0%	104.3%	104.5%	105.0%
2001	111.2%	113.1%	111.8%	105.7%	108.7%	108.2%	100.2%	99.5%	104.1%	100.7%
2002	110.3%	112.5%	109.7%	109.9%	107.6%	105.7%	108.7%	98.9%	102.7%	94.4%
2003	100.6%	117.1%	113.0%	108.5%	107.9%	103.5%	100.6%	99.0%	103.6%	103.3%
2004	109.8%	117.1%	119.7%	115.0%	104.1%	104.4%	98.7%	106.9%	98.7%	84.8%
2005	110.7%	115.7%	114.7%	115.7%	111.5%	107.5%				
2006	107.3%	117.6%	118.0%	119.9%						
2007	111.5%	109.0%								
Stdev	3.6%	3.7%	4.3%	4.9%	2.6%	2.0%	4.1%	3.7%	2.3%	8.2%

3.2 Delay between Revisions

When using revisions as a measure of time, the actual time between revisions needs to be analysed. The preferred format of delays between revisions would be the number of days and the exponential distribution is expected to fit the delay distribution well. The exponential would also be easy to work with.

The industry data are summarised by quarter and delays are now more discreet in that they take on values of 1, 2, 3, … (quarters). As such, Geometric and Negative Binomial distributions are candidates from the main discreet distributions. On the other hand, some continuous distributions that are expected to fit delays measured in days well are "discreetised" and compared to the discreet distributions.

Figure 3-1 – Observed vs. Expected Delays – All Data



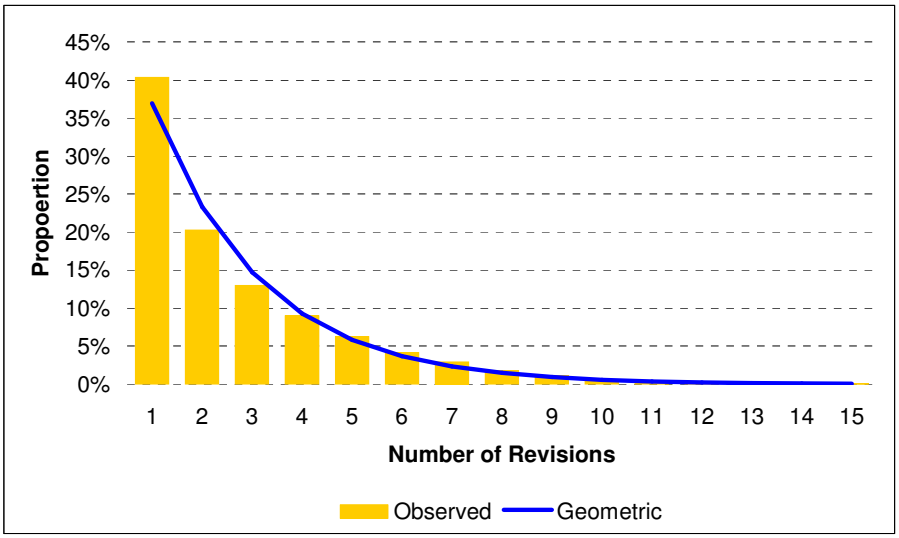
From the above graph, the Negative Binomial distribution fits best, while the others performed similarly and underestimated the level of revisions with a delay of one quarter.

3.3 Claim Status Process

A variable is used to monitor when a claim would be finalised to end the claim revision process. Initially, a variable monitoring the number of revisions during the lifetime of a claim was considered. However, this would be severely limiting as this number of revisions variable would be determined at the outset of the claim, based on claims information as at the time of the accident.

An alternative is to use an indicator variable that monitors the current status of the claim. Such a binary variable would be determined from all the current covariates and would be more realistic. When modelling a binary variable, it is equivalent to modelling the probability of finalisation at each revision; and the probability of finalisation can be quite different at different revisions. However, if the probability of settling is constant, then the number of revisions until a claim is finalised should follow a Geometric distribution. The Geometric distribution is found to fit reasonably well, as shown in the following graph.

Figure 3-2 – Observed vs. Actual Number of Revisions – All Data



Given the observed data would be driven by different means, it would have greater volatility than the standard Geometric. This is seen in the graph through the higher proportion of observations with a delay of 1. The complete data has higher variability than the standard gamma as expected and modelling the number of revisions using a claim status indicator is appropriate.

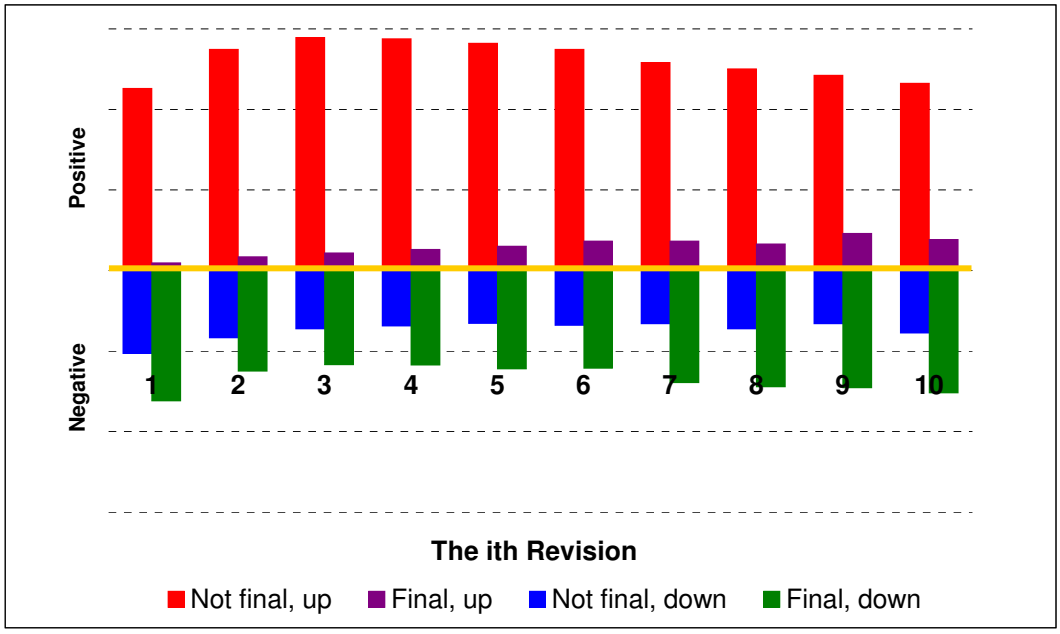
3.4 Claim Revision Direction of Change

When the timing of a revision is determined and whether the change is the final adjustment or is an intermediate revision has been modelled, the final step is to model the change. It is proposed to model the direction of the change and the magnitude of the change separately. This is because the change variable displays a highly asymmetrical distribution (positive change and negative change behave differently) and, therefore,

would be difficult to identify an appropriate distribution to model. This is observed in the industry CTP dataset; however, other datasets may be different, and modelling the change as a single process may be considered.

The direction of the change will simply be a binary variable which is modelled using covariates. From the following graph, it can be shown that the final revisions are mostly negative revisions, while other revisions are mostly positive. The mix of the directions of changes does not show a trend against the number of revisions.

Figure 3-3 – Proportion of Direction of Change, but Claim Finalisation Status



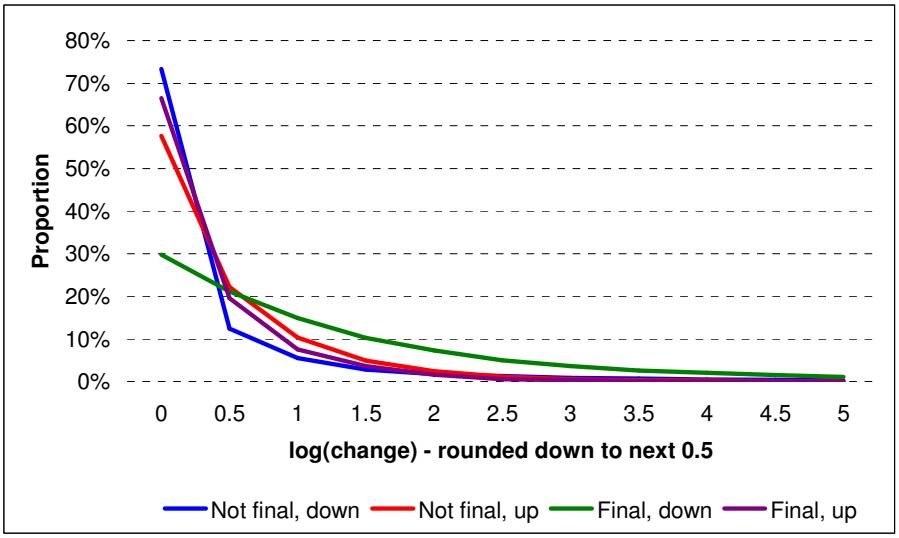
*Bar on the left is revisions prior to finalisation, bar on the right is the revision on which the claim is finalised

3.5 Size of Change

Change is defined, as in the finance literature, as the log of the new incurred cost less the log of the existing incurred cost. Such definition provides a magnitude of change that is a strictly positive continuous variable.

The following table shows the distributing of the observed size of changes, rounded down to the nearest 0.5. It can be seen that while a large proportion of changes follow the same distribution, negative changes at finalisation seem to behave quite differently. There is certainly something interesting about these "savings at finalisation" changes. These savings at finalisation occur over 80% of the time and represent a significant portion of total changes.

Figure 3-4 – Observed Changes by Claim Status and Direction of Change



The following graphs show that the Gamma, Weibull and Generalised Gamma distributions fit the size of change process quite well and almost indistinguishably. For versatility, it is hence proposed to use the Generalised Gamma distribution, of which the Gamma and Weibull distributions are special cases, for the extra flexibility when fitting the data.

Figure 3-5 – Observed vs. Expected Magnitude of Savings on Finalisation

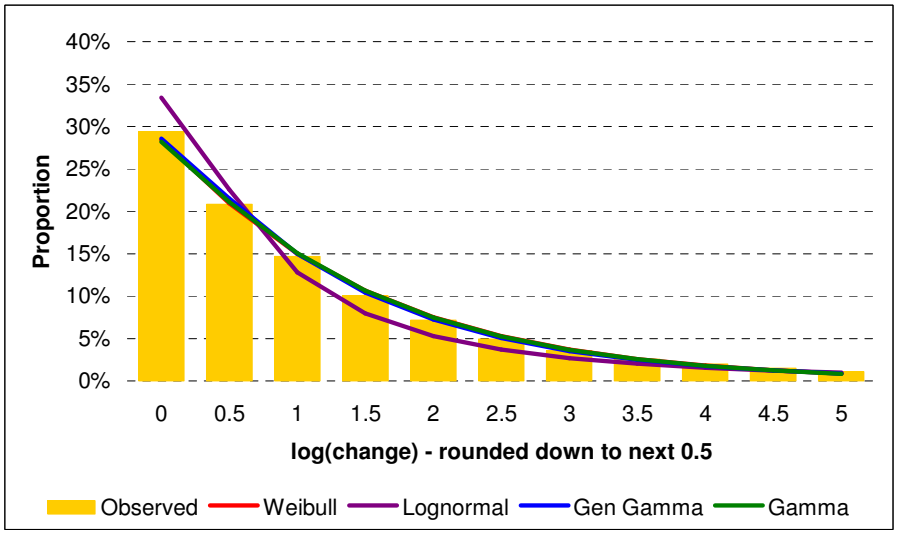
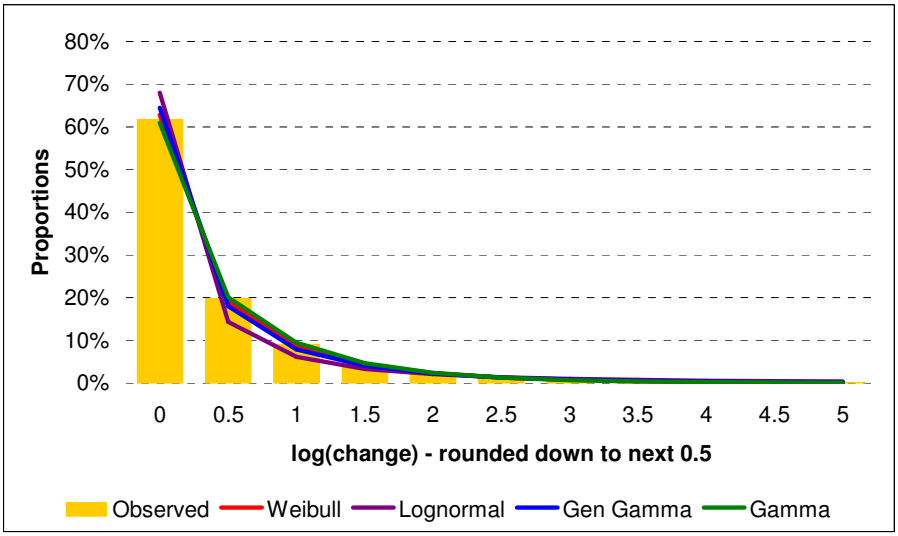


Figure 3-6 – Observed vs. Expected Magnitude of Other Revisions



4 Model Specification

4.1 Notation

Let X_{ij} be the claim cost incurred for claim i at its jth revision. The process X_{ij} is a non-negative process that maps the claims cost over its life. Let X_{i0} refer to the initial estimate of a claim when it first comes to the insurer's attention. With the industry CTP data, this value is not known, and the incurred value at the end of the reported quarter was used as a proxy.

Let n be the number of claims in the portfolio, hence, $i = 1,2,...,n$. Let T_{ij} be the time of the jth revision of the ith claim. Also, let T_i denote the date of the accident on which the claim arose. Then, the time when the claim is reported is denoted as T_{i0} and $t_{i0} = T_{i0} - T_i$ is the reporting delay — the period of time after the accident before the claim is reported to the insurer. More generally, $t_{ij} = T_{ij} - T_{i,j-1}$ is the amount of time elapsed since the last revision until the jth revision. Let m_i be the number of revisions made to the ith claim before it is finalised, hence, $j = 1,2,...,m_i$.

Also let J_{iT*} be the largest j for the ith claim such that T_{ij} is less than $T*$, in other words, the last revision made to a claim before the time $T*$. Therefore, $X_{iJ_{iT*}}$ is the incurred claims cost to date for the portfolio at date $T*$.

4.2 Conditioning on Past History

The four processes can be interpreted as one event and may be modelled together with a joint distribution. However, dealing with joint distributions would add significant complexity to the model; thus the complete change process is broken down to component processes. Conditioning is introduced between the processes and covariates, effectively creating a hierarchy of the order in which the processes happen. The conditioning is explicitly specified in the following fashion.

The hierarchical order of the claim processes is:

1. The delay process; it is assumed to model the time between the current revision and the previous revision and on a real-time scale it is the delay in the next revision taking place. This process is conditioned on only the information available at the previous revision.
2. The claim status process; it is assumed that this process will determine whether a claim is finalised at the current revision. This process is conditioned on prior information, current values of covariates and the current value of the delay process.
3. Once the claim status is determined, the next process is the direction of the change. This process is conditioned on prior information, current values of covariates, as well as the current values of the delay and claim status processes.
4. Lastly, the claim size movement distribution; the distribution for this process will be conditioned on past history, current values of covariates, as well as the current values of the preceding process variables.

For model specifications, the following filters are used. Let F_{ij} be the history of all the process variables (delay, claim status, direction of change and size of change variables) up to the j th revision. Let G_{ij} be the history of all the other covariates up to the j th revision.

4.3 The Delay Process

The Negative Binomial distribution is adopted for the delay process measured in quarters. The delay process is conditioned only on prior history, that is, information known up to the $j-1$ th revision.

$$P\left(t_{ij} = t \mid F_{i,j-1}, G_{i,j-1}\right) = \binom{t-1+r-1}{t-1} p^r \left(1-p\right)^{t-1} \qquad (1)$$

$$\text{For } k = 1, 2, \ldots$$

However, an alternative parameterisation is adopted to aid the GLM modelling where the mean parameter μ is specified directly.

$$P\left(t_{ij}=t \mid F_{i,j-1}, G_{i,j-1}\right)=\binom{t-2+e^{-d_{ij}}}{t-1}\frac{\left(e^{d_{ij}}\mu\right)^{t-1}}{\left(1+e^{d_{ij}}\mu\right)^{t-1+e^{-d_{ij}}}} \tag{2}$$

where $d=1/r$ and $\mu=r\dfrac{1-p}{p}$

4.4 The Claim Status Process

Since the claim status can either be "Open" or "Finalised" (complications of reopened claims are ignored), a Bernoulli process would be the natural selection. Let S_{ij} be the claim status process for the i th claim on its j th revision.

$$S_{ij}=\begin{cases}1 & \text{if the jth revision is the last}\\ 0 & \text{if the jth revision is not the last}\end{cases} \tag{3}$$

$$P\left(S_{ij}=1 \mid F_{i,j-1}, G_{ij}, t_{ij}\right)=q_{ij} \tag{4}$$

Based on this formulation, S_{ij} can be determined based on all the information available at time $T_{i,j-1}$, the new claims information as contained in filter G and the delay until the current revision, t_{ij}.

4.5 The Direction of Change Process

Similar to the claim status process, the revision direction process is also binary, taking values up or down. Let D_{ij} be a Bernoulli variable such that

$$D_{ij}=\begin{cases}1 & \text{if the } j\text{th revision is a positive change}\\ 0 & \text{if the } j\text{th revision is a negative change}\end{cases} \tag{5}$$

$$P\left(D_{ij}=1 \mid F_{i,j-1}, G_{ij}, t_{ij}, S_{ij}\right)=p_{ij} \tag{6}$$

4.6 The Size of Change Process

This process models the size of the change as the absolute difference between the logs of the pre and post revision claim costs; let Y_{ij} be a non-negative process that measures $\left|\log\left(X_{ij}\right)-\log\left(X_{i,j-1}\right)\right|$. Changes of less than 1% of the incurred costs are ignored; with an average claim size of around \$100,000, such a threshold would be around \$1,000.

A generalised gamma distribution has been chosen for modelling the NSW CTP due to its versatility of a three-parameter distribution. By programming the likelihood and maximisation algorithms, the Generalised Gamma can achieve a wide range of distributions. This can be done relatively easily with the statistical package S-PLUS. However, due to its data capacity issues, a dataset of around 200,000 observations and 50 covariates cannot be optimised and this approach was therefore abandoned.

As such, the standard Gamma distribution has been chosen; from the previous graphs (Figures 3-4 and 3-5), the performance of the Gamma is virtually identical to the Generalised Gamma and its density function is defined as:

$$f\left(y_{ij} \mid F_{i,j-1}, G_{ij}, t_{ij}, S_{ij}, D_{ij}\right) = \frac{1}{\Gamma\left(c_{ij}\right)\lambda_{ij}}\left(\frac{y_{ij}}{\lambda_{ij}}\right)^{c_{ij}-1} e^{-\frac{y_{ij}}{\lambda_{ij}}} \tag{7}$$

Where λ_{ij} is proportional to the mean of the variable y_{ij}

4.7 The Ultimate Claims Cost Process

Let X_i be the ultimate cost of a claim, then using the notations defined above

$$X_i = X_{i0}\prod_{j=1}^{m} e^{\left(2D_{ij}-1\right)Y_{ij}} \tag{8}$$

When constructing the likelihood, finalised claims and open claims need to be treated separately. This is because the entire history of movements of closed claims is known; however, for open claims, only the movements up to J_{iT^*} are know, plus the fact that there have been no movements between T^* and $T_{iJ_{iT^*}}$.

Hence, the likelihood for the i th claim can be expressed as

$$L(p,q,c,d,\lambda,\mu) = \prod_{j} P\left(t_{ij}\right)P\left(S_{ij}\right)P\left(D_{ij}\right)f\left(y_{ij}\right)\cdot I_{(j<m_i)}\left(P\left(t_{i,J_{iT^*}} < T^* - T_{i,J_{iT^*}}\right)\right) \tag{9}$$

Where $I_{(cond)}$ is an indicator function that takes on the value 1 if (cond) is true and 0 otherwise.

The likelihood can be expressed in this way due to the conditioning that is assumed in each of the processes as described above; notations for this conditioning are dropped from the likelihood to keep the equation manageable.

4.8 Variables and Regression

Three types of variables are used to model the above processes. These are slightly different from the three categories defined by Taylor et al, 2006.

1. Process variables themselves, that is, the current values of the four processes, are modelled upon past values of themselves as well as other processes where appropriate.
2. Static variables, or baseline variables, do not change throughout the duration of the claim. These variables may include age of claimant at date of accident, gender, location of the claimant, etc.

3. Dynamic variables, change during the evolution of the claim. The changes in these dynamic variables are the most interesting in determining the outcome of the process variables. The dynamic variables include Liability Status, Injuries and their Severities, Legal Representation, etc.

Let $\mathbf{X}$ be a vector of covariates for claim i; for ease of notation, the subscript i is left out. Then, the following relationships are assumed relating the covariates to the mean functions of the processes. These follow the standard relationships in a GLM framework.

$$\begin{aligned} \log \mu &= \alpha_{0\mu} + \alpha_{\mu} X_{\mu}^{T} = Z_{\mu} \\ \log\left(\frac{q_{ij}}{1-q_{ij}}\right) &= \alpha_{0q} + \alpha_{q} X_{q}^{T} = Z_{q} \\ \log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \alpha_{0p} + \alpha_{p} X_{p}^{T} = Z_{p} \\ \log\left(\lambda_{ij}\right) &= \alpha_{0\lambda} + \alpha_{\lambda} X_{\lambda}^{T} = Z_{\lambda} \end{aligned} \tag{10}$$

4.9 Modelling Software

SAS and S-PLUS are two of the more well-known statistical and data modelling software programs. SAS (version 9.1 and earlier) has an extensively range of modelling methods that, while flexible, are still limited to the options available. S-PLUS, on the other hand, offers a programming environment with which users can code tools that are tailored to the user's needs.

The above distribution functions and likelihoods in the previous sections are developed so they can be programmed into S-PLUS to optimise the likelihood directly to estimate parameters. However, S-PLUS has been found not to be able to handle datasets of even moderate size. The models were fitted with the standard SAS GENMOD Procedure and its various limitations are inherited.

However, SAS 9.2 should bring the welcomed addition of user programming functionality. This enables the fitting of these models through likelihood maximisation; as such, the various likelihood functions and their derivatives can be found in Appendix A. Undoubtedly, there is other software beside S-PLUS and SAS 9.2 that would be able to carry out likelihood maximisation.

5 Results

Given its ability to handle large datasets, the models are fitted using SAS. For the open claims, an adjustment has to be made in accordance with Formula (9), for the fact that they have not had a revision in the period of time since the last revision. This is done through "weighting" the observation when modelling the delay process.

The full set of variables used, with only the main effects fitted, is contained in Appendix B. Those that showed a high level of significance (with a p-value of <0.01) are marked with an "*". The fitting used a step-wise variable selection process where the full set of variables is used and insignificant ones are gradually dropped out.

The coefficients of a few of the variables that are of interest are presented below. In the tables, "Type" denotes the type of variable:

- D – Dynamic variable
- S – Static variable
- P – Process variable

The coefficients are presented for:

- t_ij – the delay process
- S_ij – the finalisation process
- D_ij – the direction of change process
- Y_ij – the size of change process, modelled separately for positive and negative changes

The delay process is modelled on information available at the previous revision only. Hence, its coefficents of the dyamic variables are based on the values at the previous revisions. The coefficents of the dynamic variables for the other processes are based on their current values.

5.1 Ith Revision vs. Development Year

Table 5-1 – Model Output – Ith Revision Variable

Type	Parameter	Level1	t_ij	S_ij	D_ij	Y_ij D_ij = 0	Y_ij D_ij = 1
D	Ith Revision	1	0.50	0.25	1.26	0.32	0.55
D	Ith Revision	2	-	-	-	-	-
D	Ith Revision	3	-0.11	0.01	0.25	-0.10	0.00
D	Ith Revision	4	-0.17	0.08	0.36	-0.16	0.00
D	Ith Revision	5	-0.23	0.17	0.44	-0.23	0.01
D	Ith Revision	6	-0.25	0.21	0.49	-0.23	0.04
D	Ith Revision	7	-0.28	0.36	0.45	-0.27	0.07
D	Ith Revision	8	-0.29	0.33	0.34	-0.23	0.09
D	Ith Revision	9	-0.26	0.41	0.51	-0.21	0.14
D	Ith Revision	10	-0.29	0.41	0.30	-0.32	0.16
D	Ith Revision	11 or more	-0.28	0.43	0.73	-0.35	0.22

*The second revision is taken as the baseline because the first revision does not use the full set of variables, as it lacks variables at the "previous revision". Because of this, the coefficients for the first revision are on a different level.

The modelling shows that the delay between revisions shortens and the probability of finalising the claim increases with the increasing number of revisions. The likelihood of having a positive change also increases; however, the change is more likely to be of smaller size, for both positive and negative changes. Putting it all together:

- Earlier revisions tend to be quite spaced out, with bigger movements
- Later revisions (as measured by the number of revisions) tend to be close together; and are more likely to be small upward adjustments

The variable "development year", which is defined as (year of revision – year of accident), was added to test whether calendar year development can add extra explanatory power to each of the models. It is of interest to note that the number of revisions variable has considerably more explanatory power compared to development year in all processes except for the finalisation process. The time elapsed since the date of accident is still a significant driver for the probability of finalising the claim at a particular revision.

5.2 Injury Severity

Table 5-2 – Model Output – Injury Variables

Type	Parameter	Level1	t_ij	S_ij	D_ij	Y_ij D_ij = 0	Y_ij D_ij = 1
D	ISS	unknown/admin	.	-	-	-	.
D	ISS	1	.	0.30	0.40	.	.
D	ISS	2	.	0.21	0.34	.	.
D	ISS	3 - 5	.	0.09	0.19	.	.
D	ISS	6 - 10	.	-0.03	0.04	.	.
D	ISS	11 - 30	.	-0.05	0.01	.	.
D	ISS	31 - 75	.	0.42	0.12	.	.
D	ISS increase	1	.	0.38	.	.	.
D	ISS increase	0	.	-	.	.	.
D	Number of Regions	0	.	-	-	-	
D	Number of Regions	1	.	0.12	0.05	0.10	
D	Number of Regions	2	.	0.18	0.12	0.13	
D	Number of Regions	3	.	0.25	0.10	0.14	
D	Number of Regions	4	.	0.44	0.18	0.18	
D	Number of Regions	5	.	0.60	0.18	0.23	
D	Num Reg increase	1	.	0.48	0.18	0.10	
D	Num Reg increase	0	.	-	-	-	

It appears that light injuries and heavy injuries (as measured by the Injury Severity Score (ISS)) are the most likely to finalise at a particular revision; this is especially so if the ISS was increased at the revision. This may be counter intuitive, as it is generally accepted that claims from severe injuries takes longer to settle. However, possible explanations for this include a small number of observations in the group with an ISS of 31 to 75. Secondly, there is a general tendency for ISS to worsen during the life of a claim and the higher ISS may represent records from the later revisions of a claim and, hence, there may be an interaction effect not accounted for in the modelling.

Similarly, a higher number of body regions injured also shows higher likelihood of finalisation, as well as for an increase in the number of regions injured. Again, possible explanations may include:

- A small number of records with 4 or 5 regions injured and also a small number of records that had an increase in the number of body regions at a particular revision, and coefficient may be volatile
- More severe injuries may be correlated with larger claims and the fitted models may not account for the interaction effect

More severe injuries are also more likely to have an upward revision compared to less severe ones. On the other hand, more regions of injury tend to have a relatively lower chance of an upward revision; however, when they do, the magnitude of these revisions tends to be bigger.

5.3 Delay between Revisions

Table 5-3 – – Model Output – Delay Between Revision Variable

Type	Parameter	Level1	t_ij	S_ij	D_ij	Y_ij	
						D_ij = 0	D_ij = 1
P	t_0	1	.	-	-	-	-
P	t_0	2	.	-0.18	0.11	0.20	-0.04
P	t_0	3	.	-0.06	0.17	0.35	0.04
P	t_0	4	.	-0.06	0.25	0.42	0.09
P	t_0	5	.	-0.03	0.25	0.48	0.13
P	t_0	6	.	0.05	0.35	0.51	0.17
P	t_0	7	.	-0.01	0.38	0.53	0.19
P	t_0	8	.	0.09	0.35	0.56	0.17
P	t_0	9	.	0.04	0.49	0.52	0.22
P	t_0	10	.	0.12	0.47	0.56	0.21
P	t_0	11 - 15	.	0.11	0.60	0.54	0.23
P	t_0	16 or more	.	-0.01	0.74	0.49	0.29
P	t_1	1	-	-	-	-	-
P	t_1	2	0.07	0.10	-0.01	0.10	0.07
P	t_1	3	0.11	0.27	0.00	0.14	0.11
P	t_1	4	0.13	0.30	0.04	0.15	0.15
P	t_1	5	0.14	0.31	0.17	0.10	0.19
P	t_1	6	0.11	0.38	0.14	0.12	0.23
P	t_1	7	0.10	0.39	0.21	0.10	0.24
P	t_1	8	0.06	0.45	0.32	0.17	0.25
P	t_1	9	0.02	0.45	0.30	0.09	0.30
P	t_1	10	-0.01	0.52	0.22	0.02	0.29
P	t_1	11 - 15	0.01	0.64	0.27	0.06	0.27
P	t_1	16 or more	0.06	0.46	0.25	-0.00	0.14

* Coefficients for t_ij only exist for t_1 as the length of the current delay is modelled on the length of the previous delay

The delay process variables t_{ij} and $t_{i,j-1}$ are extremely useful in the modelling of the process of interest. Typically, longer delayed revisions would lead to:

- being more likely to finalise,
- being more likely to be an upward revision, and
- a larger size of change, for both upward and downward changes.

These observations also apply for the delay from the previous revision.

5.4 Year Effects

Two year effects are presented below; year of accident and year of revision. The coefficient of the delay process shows that claims from the more recent accident years are subject to shorter delays and similarly for the revisions made in the more recent years. This pattern may be caused by the fact claims with revisions from the recent years tend to be smaller claims. This pattern remains even after the impact of claim size is accounted for.

However, for probability of finalisation, the coefficients between accident year and revision year are negatively correlated, suggesting a "development year effect". In fact, when a development year variable was added to the model, its explanatory power outweighed that of the number of revisions variable. For the D_{ij} process, the claims from more recent accident years are more likely to increase than decrease.

Table 5-4 – Model Output – Year Variables

Type	Parameter	Level1	t_ij	S_ij	D_ij	Y_ij	
						D_ij = 0	D_ij = 1
S	Year of Accident	2000	0.27	1.72	- 0.16	.	.
S	Year of Accident	2001	0.26	1.44	- 0.22	.	.
S	Year of Accident	2002	0.27	1.26	- 0.24	.	.
S	Year of Accident	2003	0.27	1.10	- 0.19	.	.
S	Year of Accident	2004	0.29	0.85	- 0.09	.	.
S	Year of Accident	2005	0.27	0.54	- 0.00	.	.
S	Year of Accident	2006	0.19	0.29	- 0.03	.	.
S	Year of Accident	2007	-	-	-	.	.
D	Year of Revision	2000	0.59	- 2.40	.	.	- 0.09
D	Year of Revision	2001	0.48	- 2.14	.	.	- 0.18
D	Year of Revision	2002	0.44	- 1.53	.	.	- 0.18
D	Year of Revision	2003	0.48	- 1.26	.	.	- 0.04
D	Year of Revision	2004	0.42	- 1.14	.	.	- 0.03
D	Year of Revision	2005	0.32	- 0.81	.	.	0.02
D	Year of Revision	2006	0.24	- 0.41	.	.	0.03
D	Year of Revision	2007	-	-	.	.	-

6 Further Improvements

6.1 Limitations of Using Industry Data

For this paper, the claim development processes are applied to the NSW CTP industry data. There are many reasons why this model would be more informative and powerful if it were applied to a particular insurer's dataset.

Firstly, the model accounts for how claim estimates change and this is related to the claim-handling procedures of each of the insurers. Using industry data, however, the various claim-handling procedures are inseparable in the data. Thus the modelled effects will combine all the different insurers' practices and their explanatory power will be reduced. A single insurer's data would be more homogenous and when interactions with the year variable are specified, it can reveal the changes in claim-handling procedures.

Secondly, industry data are summarised quarterly. Since it is not known whether there was a revision in the same quarter that a claim was reported, the initial information regarding the claim is not known; such information may be useful and add further explanatory power. Also, with the data summarised quarterly, the delay variable was modelled using a discreet distribution. Discreet distributions are generally more difficult to work with when accounting for the adjustment for the open claims.

Thirdly, the individual insurers' database may contain claim-management indicators. These variables may include the insurer's legal representation or claim-handling approach. The inclusion of these variables may model the effectiveness (and impact) of variables on claim-management approaches and allow decisions to be made regarding claim handling for the benefit of both claimants and insurers.

6.2 Modelling Enhancement

Due to the paper's main aim of presenting a method of modelling the claim development process by breaking it down, time has not been invested in coming up with perfect models. Generally, while interaction effects can offer great insight into how claims develop and how various types of claims are handled, this avenue has not been pursued for this paper. However, based on the industry dataset, with the possibility of extending the dataset using claims occurred before 2000, a vast amount of useful information can be gained.

For projection purposes, dynamic variables are difficult to use. To project or simulate process variables, the values of the dynamic variables will need to be projected first. Typically transition models need to be built for each dynamic variable, which is not an easy task if each one is allowed to be time-dependent or state-dependent.

One possible solution is to turn the dynamic variables into static variables by sampling their value at a common point, for example, one year after the date of the accident. Hopefully, by that stage the injuries would have stabilised and legal issues may have

progressed such that movements in these dynamic variables will be minimised. One trade-off to consider when selecting what timeframe to use is that the longer the timeframe, the better the prediction will be; yet, at the same time, more claims would not be able to be modelled because these claims are too recent.

7 Conclusions

By breaking up the claim evolution process into its component processes, this paper presents a novel way of modelling claim developments. This approach, when combined with simulation, was considered to project open claims for pricing purposes. However, in doing so, outstanding claims liabilities can be estimated. It is useful to be able to project claims outstanding on a claim-by-claim basis, as profitability by segment can then be analysed; this has always been difficult with a long-tailed insurance product.

Actuaries generally prefer not to apply statistical black boxes. Some of the stated dangers of doing so include that these models cannot account for changing trends effectively and transparently. With this model, the effects of the variables need to be tracked through time carefully in order to understand the extent of their impact through time. The benefit of this is that the drivers of change through time are better understood.

8 Bibliography

Brookes, R. and M. Prevett, 2004, *Statistical Case Estimation Modelling - An Overview of the NSW WorkCover Model*, Presented to the IAA Accident Compensation Seminar, 2004

McGuire, G., 2007, *Individual Claim Modelling of CTP Data*, Presented to the IAAust XIth Accident Compensation Seminar, 2007, Melbourne

Prevett, M. and D. Gifford, 2007, *Statistical Case Estimation for Long Term Claimants – Uncovering Drivers of Long Term Claims Cost in Accident Compensation*, Presented to the IAAust XIth Accident Compensation Seminar, 2007, Melbourne

Taylor, G. and M. Campbell, 2002, *Statistical Case Estimation*, Research Paper Series, Centre of Actuarial Studies, The University of Melbourne, No 104

Taylor, G. and G. McGuire, 2004, *Loss Reserving with GLM's: a Case Study*, Presented to the Spring 2004 Meeting of the Casualty Actuarial Society, Colorado Springs, Colorado

Taylor, G., G. McGuire and J. Sullivan, 2006, *Individual Claim Loss Reserving Conditioned by Case Estimates*, Research Paper Series, Centre of Actuarial Studies, The University of Melbourne, No 151

Part III Appendices

A Likelihoods, First Order Derivatives and Second Order Derivatives

A.1 Likelihoods and log-likelihoods of open and closed claims

The likelihood and log-likelihoods to the closed claims are

$$L_C(p,q,c,\lambda,\mu,d) = \prod_{j=1}^{m} (1-q_{ij})^{1-S_{ij}} q_{ij}^{S_{ij}} p_{ij}^{D_{ij}} (1-p_{ij})^{1-D_{ij}} \frac{1}{\Gamma(c_{ij})\lambda_{ij}} \left(\frac{y_{ij}}{\lambda_{ij}}\right)^{c_{ij}-1} e^{-\frac{y_{ij}}{\lambda_{ij}}} \cdot \prod_{j=1}^{m} \frac{(t_{ij}-2+e^{-d_{ij}})!}{(t_{ij}-1)!(e^{-d_{ij}}-1)!} \frac{(e^{d_{ij}}\mu)^{t_{ij}-1}}{(1+e^{d_{ij}}\mu)^{t_{ij}-1+e^{-d_{ij}}}} \tag{1}$$

and

$$\begin{aligned} \ell_C(p,q,c,\lambda,\mu,d) = & \sum_{j=1}^{m} (1-S_{ij}) \log(1-q_{ij}) + S_{ij} \log q_{ij} + D_{ij} \log p_{ij} + (1-D_{ij}) \log(1-p_{ij}) \\ & + \sum_{j=1}^{m} -\log(\Gamma(c_{ij})) - \log \lambda_{ij} + (c_{ij}-1)(\log y_{ij} - \log \lambda_{ij}) - \frac{y_{ij}}{\lambda_{ij}} \\ & + \sum_{j=1}^{m} \left(\sum_{h=0}^{t-2} \log(e^{-d_{ij}} + h) - \log((t-1)!) \right) \\ & + \sum_{j=1}^{m} (t_{ij}-1) \log(e^{d_{ij}}\mu_{ij}) - (t_{ij}-1+e^{-d_{ij}}) \log(1+e^{d_{ij}}\mu_{ij}) \end{aligned} \tag{2}$$

The likelihood and log-likelihood of the open claims are

$$L_O(p,q,c,\lambda,\mu,d) = \prod_{j=1}^{J_{iT^*}} (1-q_{ij})^{1-S_{ij}} q_{ij}^{S_{ij}} p_{ij}^{D_{ij}} (1-p_{ij})^{1-D_{ij}} \frac{1}{\Gamma(c_{ij})\lambda_{ij}} \left(\frac{y_{ij}}{\lambda_{ij}}\right)^{c_{ij}-1} e^{-\frac{y_{ij}}{\lambda_{ij}}} \cdot \prod_{j=1}^{J_{iT^*}} \frac{(t_{ij}-2+e^{-d_{ij}})!}{(t_{ij}-1)!(e^{-d_{ij}}-1)!} \frac{(e^{d_{ij}}\mu)^{t_{ij}-1}}{(1+e^{d_{ij}}\mu)^{t_{ij}-1+e^{-d_{ij}}}} \left(1 - P_t\left(t_{i,J_{iT^*}+1} \le T^* - T_{iJ_{iT^*}}\right)\right) \tag{3}$$

where $1 - P_t\left(t_{i,J_{iT^*}+1} \le T^* - T_{iJ_{iT^*}}\right) = \sum_{k=T^*-T_{iJ_{iT^*}}}^{\infty} \frac{(t_{ij}-2+e^{-d_{ij}})!}{(t_{ij}-1)!(e^{-d_{ij}}-1)!} \frac{(e^{d_{ij}}\mu)^{t_{ij}-1}}{(1+e^{d_{ij}}\mu)^{t_{ij}-1+e^{-d_{ij}}}}$

and

$$\begin{aligned}\ell_O(p,q,c,\lambda,\mu,d) = &\sum_{j=1}^{J_{iT^*}}(1-S_{ij})\log(1-q_{ij}) + S_{ij}\log q_{ij} + D_{ij}\log p_{ij} + (1-D_{ij})\log(1-p_{ij}) \\ &+\sum_{j=1}^{J_{iT^*}} -\log(\Gamma(c_{ij})) - \log\lambda_{ij} + (c_{ij}-1)(\log y_{ij} - \log\lambda_{ij}) - \frac{y_{ij}}{\lambda_{ij}} \\ &+\sum_{j=1}^{J_{iT^*}}\left(\sum_{h=0}^{t-2}\log(e^{-d_{ij}}+h) - \log((t-1)!)\right) \\ &+\sum_{j=1}^{J_{iT^*}}(t_{ij}-1)\log(e^{d_{ij}}\mu_{ij}) - (t_{ij}-1+e^{-d_{ij}})\log(1+e^{d_{ij}}\mu_{ij}) \\ &+\log\left(\sum_{k=T^*-T_{iJ_{iT^*}}}^{\infty}\frac{(t_{ij}-2+e^{-d_{ij}})!}{(t_{ij}-1)!(e^{-d_{ij}}-1)!}\frac{(e^{d_{ij}}\mu_{ij})^{t_{ij}-1}}{(1+e^{d_{ij}}\mu_{ij})^{t_{ij}-1+e^{-d_{ij}}}}\right)\end{aligned} \tag{4}$$

A.2 Likelihood of the claim development process expressed in linear predictor

$$\begin{aligned}\ell(p,q,c,\lambda,\mu,d) = &\sum_{i,j}\begin{pmatrix} S_{ij}Z_q - \log(1+e^{Z_q}) + D_{ij}Z_p - \log(1+e^{Z_p}) \\ +\log(\Gamma(c_{ij})) - Z_\lambda + (c_{ij}-1)(\log y_{ij} - \log Z_\lambda) - \frac{y_{ij}}{e^{Z_\lambda}} \\ +\left(\sum_{h=0}^{t-2}\log(e^{-d_{ij}}+h)\right) - \log((t-1)!) \\ +(t_{ij}-1)\log(e^{d_{ij}Z_\mu}) - (t_{ij}-1+e^{-d_{ij}})\log(1+e^{d_{ij}Z_\mu}) \end{pmatrix} \\ &+I_{(J_{iT^*}<m)}\log\left(\sum_{k=T^*-T_{iJ_{iT^*}}}^{\infty}\frac{(t_{ij}-2+e^{-d_{ij}})!}{(t_{ij}-1)!(e^{-d_{ij}}-1)!}\frac{(e^{d_{ij}Z_\mu})^{t_{ij}-1}}{(1+e^{d_{ij}Z_\mu})^{t_{ij}-1+e^{-d_{ij}}}}\right)\end{aligned} \tag{5}$$

where $I_{(J_{iT^*}<m)}$ is an indicator function that takes the value of 1 if the claim is open and 0 if claim is closed at the end of the analysis period.

With the delay process modelled with a negative binomial distribution, the likelihood would be too difficult to differentiate with respect to d_{ij} and α_μ, and then find the solutions through the Newton-Raphson method. The last part of Equation (5) is known as the Incomplete Beta function and its derivatives would need to be calculated by numerical methods. It is processed using an iterative approximation process where the last part of Equation (5) is not fit using maximum likelihood. The process would be

- Fit the parameters ignoring the Incomplete Beta function
- Using these fitted parameters and estimate the value of the incomplete beta function

- These evaluated incomplete beta functions is then used to weight the affected observations at the next fitting of the parameters, and so on.

On the other hand, if the exponential distribution was assumed for the delay process, the likelihood would be much simpler to fit using maximum likelihood methods

A.3 Derivatives of the log-likelihood

First order derivatives

$$\frac{\partial \ell}{\partial \alpha_\mu} = \sum_{i,j} \left(\left(t_{ij} - 1\right) d_{ij} - \left(t_{ij} - 1 + e^{-d_{ij}}\right) \frac{e^{d_{ij} Z_\mu}}{1 + e^{d_{ij} Z_\mu}} \right) X_\mu$$

$$\frac{\partial \ell}{\partial d} = \sum_{i,j} \left(\sum_{h=0}^{t-2} - \frac{e^{-d_{ij}}}{e^{-d_{ij}} + h} \right) + \left(t_{ij} - 1\right) Z_\mu - \left(\left(t_{ij} - 1 + e^{-d_{ij}}\right) \frac{e^{d_{ij} Z_\mu}}{1 + e^{d_{ij} Z_\mu}} - e^{-d_{ij}} \log\left(1 + e^{d_{ij} Z_\mu}\right) \right)$$

$$\frac{\partial \ell}{\partial \alpha_q} = \sum_{i,j} \frac{1}{1 + e^{Z_q}} e^{Z_q} X_q - S_{ij} X_q$$

$$= \sum_{i,j} \left(1 - S_{ij} - \left(1 + e^{Z_q}\right)^{-1} \right) X_q$$

$$\frac{\partial \ell}{\partial \alpha_p} = \sum_{i,j} \left(1 - D_{ij} - \left(1 + e^{Z_p}\right)^{-1} \right) X_p$$

$$\frac{\partial \ell}{\partial \alpha_\lambda} = \sum_{i,j} \left(-c + y_{ij} e^{-Z_\lambda} \right) X_\lambda$$

$$\frac{\partial \ell}{\partial c} = \sum_{i,j} -digamma\left(c\right) + \left(\log y_{ij} - Z_\lambda\right)$$

Second order derivatives are

$$\frac{\partial^2 \ell}{\partial \alpha_\mu^2} = \sum_{i,j} \left(t_{ij} - 1 + e^{-d_{ij}}\right)\left(1+e^{d_{ij}Z_\mu}\right)^{-2} e^{d_{ij}Z_\mu} X_\mu^2$$

$$\frac{\partial^2 \ell}{\partial \alpha_\mu \partial d} = \sum_{i,j} \left(\left(t_{ij}-1\right) - \left(t_{ij} - 1 + e^{-d_{ij}}\right)\left(e^{d_{ij}Z_\mu}\left(1+e^{d_{ij}Z_\mu}\right)^{-2}\right) + e^{-d_{ij}}\left(\frac{e^{d_{ij}Z_\mu}}{1+e^{d_{ij}Z_\mu}}\right)\right) X_\mu$$

$$\frac{\partial^2 \ell}{\partial d^2} = \sum_{i,j}\left(\sum_{h=0}^{t-2}\frac{h}{e^{-d_{ij}}+h} - 1\right) - \left(\left(t_{ij}-1+e^{-d}\right)\left(1+e^{d_{ij}Z_\mu}\right)^{-2} e^{d_{ij}Z_\mu} - e^{-d_{ij}}\left(1-\left(1+e^{d_{ij}Z_\mu}\right)^{-1}\right)\right)$$

$$+\sum_{i,j} e^{-d_{ij}}\left(\frac{e^{d_{ij}Z_\mu}}{1+e^{d_{ij}Z_\mu}}\right) + e^{-d_{ij}}\log\left(1+e^{d_{ij}Z_\mu}\right)$$

$$\frac{\partial^2 \ell}{\partial \alpha_q^2} = \frac{\partial}{\partial \alpha_q}\sum_{i,j} -\left(1+e^{Z_q}\right)^{-1} X_q$$

$$= \sum_{i,j}\left(1+e^{Z_q}\right)^{-2} e^{Z_q} X_q^2$$

$$\frac{\partial^2 \ell}{\partial \alpha_p^2} = \sum_{i,j}\left(1+e^{Z_p}\right)^{-2} e^{Z_p} X_p^2$$

$$\frac{\partial^2 \ell}{\partial \alpha_\lambda^2} = \sum_{i,j} -y_{ij} e^{-Z_\lambda} X_\lambda^2$$

$$\frac{\partial^2 \ell}{\partial c^2} = \sum_{i,j} -trigamma(c)$$

$$\frac{\partial^2 \ell}{\partial \alpha_\lambda \partial c} = \sum_{i,j} -X_\lambda$$

The other second order derivatives are zero.

A.4 Using the Generalised Gamma Distribution for Size of Change

If the generalised gamma distribution was used rather than the standard gamma, then the relevant likelihoods and derivatives are as follow. An exponential version of the parameters are used for fitting due to the strictly positive requirements of the c and k parameters.

$$f\left(y_{ij} \mid F_{i,j-1}, G_{ij}, t_{ij}, S_{ij}, D_{ij}\right) = \frac{k_{ij}}{\Gamma\left(c_{ij}\right)\lambda_{ij}}\left(\frac{y_{ij}}{\lambda_{ij}}\right)^{c_{ij}k_{ij}-1} e^{-\left(\frac{y_{ij}}{\lambda_{ij}}\right)^{k_{ij}}}$$

Let

$k = e^{\kappa}$

$c = e^{\gamma}$

$\log\left(\lambda_{ij}\right) = \alpha_{0\lambda} + \alpha_\lambda X_\lambda^T = Z_\lambda$

Then the log-likelihood is

$$\ell(\kappa,\gamma,\lambda)=\sum_{i,j}\kappa-\log\left(\Gamma\left(e^{\gamma}\right)\right)+e^{\gamma+\kappa}\left(\log y_{ij}-Z_{\lambda}\right)-\log y_{ij}-\left(\frac{y_{ij}}{e^{Z_{\lambda}}}\right)^{e^{\kappa}}$$

The relevant first order derivatives are

$$\frac{\partial\ell}{\partial\alpha_{\lambda}}=\sum_{i,j}\left(-e^{\gamma+\kappa}+e^{\kappa}y_{ij}^{e^{\kappa}}e^{-e^{\kappa}Z_{\lambda}}\right)X_{\lambda}$$

$$\frac{\partial\ell}{\partial\kappa}=\sum_{i,j}1+e^{\gamma+\kappa}\left(\log y_{ij}-Z_{\lambda}\right)-\log\left(y_{ij}/e^{Z_{\lambda}}\right)\left(y_{ij}/e^{Z_{\lambda}}\right)^{e^{\kappa}}e^{\kappa}$$

$$\frac{\partial\ell}{\partial\gamma}=\sum_{i,j}-digamma\left(e^{\gamma}\right)e^{\gamma}+e^{\gamma+\kappa}\left(\log y_{ij}-Z_{\lambda}\right)$$

The second order derivatives are

$$\frac{\partial^2\ell}{\partial\alpha_{\lambda}^2}=\sum_{i,j}-e^{2\kappa}y_{ij}^{e^{\kappa}}e^{-e^{\kappa}Z_{\lambda}}X_{\lambda}^2$$

$$\frac{\partial^2\ell}{\partial\kappa^2}=\sum_{i,j}e^{\gamma+\kappa}\left(\log y_{ij}-Z_{\lambda}\right)-\left(\log\left(y_{ij}/e^{Z_{\lambda}}\right)e^{\kappa}+1\right)\log\left(y_{ij}/e^{Z_{\lambda}}\right)\left(y_{ij}/e^{Z_{\lambda}}\right)^{e^{\kappa}}e^{\kappa}$$

$$\frac{\partial^2\ell}{\partial\gamma^2}=\sum_{i,j}-e^{\gamma}\left(trigamma\left(e^{\gamma}\right)e^{\gamma}+digamma\left(e^{\gamma}\right)\right)+e^{\gamma+\kappa}\left(\log y_{ij}-Z_{\lambda}\right)$$

$$\frac{\partial^2\ell}{\partial\alpha_{\lambda}\partial\kappa}=\sum_{i,j}\left(-e^{\gamma+\kappa}+e^{\kappa}\left(y_{ij}/e^{Z_{\lambda}}\right)e^{\kappa}+\log\left(y_{ij}/e^{Z_{\lambda}}\right)\left(y_{ij}/e^{Z_{\lambda}}\right)^{e^{\kappa}}e^{2\kappa}\right)X_{\lambda}$$

$$\frac{\partial^2\ell}{\partial\alpha_{\lambda}\partial\gamma}=\sum_{i,j}-e^{\gamma+\kappa}X_{\lambda}$$

$$\frac{\partial^2\ell}{\partial\gamma\partial\kappa}=\sum_{i,j}e^{\gamma+\kappa}\left(\log y_{ij}-Z_{\lambda}\right)$$

B Model Outputs – Significant Variables

Type	Variable	t_ij	S_ij	D_ij	Y_ij	
					D_ij = 0	D_ij = 1
Dynamic	Brain		*		*	
	Eco Loss			*		
	ISS		*	*		
	ISS increase		*			
	Ith Revision	*	*	*	*	*
	Leg Rep increase		*	*		*
	Leg Representation		*	*	*	*
	Liability	*	*	*	*	*
	Liability increase		*		*	
	Lit Level increase		*	*		*
	Litigation					*
	Litigation Level	*	*		*	
	Max Sev increase					*
	Max Severity			*		
	Number of regions		*	*	*	
	Num Reg increase		*	*	*	
	RegionG Inj		*	*	*	
	Rehab	*	*	*		
	Rehab increase					
	Spine					*
	Whiplash		*			
	Year of Revision	*	*			*
Process	D_0					*
	D_1		*	*		
	S_0			*		
	t_0		*	*	*	*
	t_1	*	*	*	*	*
	X_1	*		*	*	*
	Y_1	*	*		*	*
Static	Age at Acc.	*	*	*	*	*
	Employment Status	*				*
	Gender	*	*		*	
	Year of Accident	*	*	*		
	Region of Risk					
	Vehicle Category					