Stochastic models for claims reserving in insurance business

Tamás Faluközy\textsuperscript{1}, Ildikó Ibolya Vitéz\textsuperscript{1}, and Miklós Arató\textsuperscript{1}

\textsuperscript{1}Department of Probability Theory and Statistics
Eötvös Loránd University
Budapest, Hungary
(e-mail: fafa@math.elte.hu, vildiko@math.elte.hu)

Abstract. Insurance companies have to build a reserve for their future payments which is usually done by deterministic methods giving only a point estimate. In this paper two semi-stochastic methods are presented along with a more sophisticated hierarchical Bayesian model containing MCMC technique. These models allow us to determine quantiles and confidence intervals of the reserve which can be more reliable as just a point estimate. A sort of cross-validation technique is also used to test the models.

Keywords: insurance, reserving, quantiles, MCMC, cross-validation.

1 Introduction

For an insurance company it is crucial to determine the future payments – generally this amount is called ‘the reserve’. This is usually done by using the expected value (EV) principle which means that the reserve gives the safety appropriate to the EV. It is obvious that this is not sufficient enough, therefore insurance companies have to build a solvency capital which provides a safety puffer above the reserves for unexpected cases. In 1999 the European Union decided to reform the regulation of capital requirements for insurance companies, however, substantial work began in 2003 (Solvency II project). Although it seems settled that the new regulation will deal simultaneously with the reserves and the solvency capital. The main goal is to provide safety levels for the reserve and for the capital based on stochastic modelling. In this paper the so-called IBNR reserve is examined which refers to the claims that already incurred (occurred) but have not been reported to the insurance company yet. There are a lot of different reserves but in non-life business the main part of the ultimate reserve comes from the IBNR and from the outstanding claims, i.e. the claims that already incurred, are also reported to the company, but are not paid yet. Up to the present the determination of the IBNR reserve was usually done by using deterministic methods based upon the so-called run-off triangle. In this triangle the claims of the past are presented and they are separated according to the time period of incurrance and to the delay of report or payment. Throughout this paper the elements of this incremental run-off triangle are denoted with $C_{i,j}$.
Most of the deterministic methods uses the cumulative run-off triangle which elements are $D_{i,j} = \sum_{s=1}^{j} C_{i,s}$ (see example Figure 1).

There are a lot of different run-off triangles, for example, the time period can be different – we examined daily, monthly, quarterly and annual periods and their influence on the reserve. Our second goal was to develop stochastic reserving methods for IBNR claims according to the main idea of the Solvency II project discussed above. Two naïve stochastic models are presented first and then a more sophisticated hierarchical Bayesian model is produced which contains MCMC technique. The results of the different methods are compared and the models are also tested by using some kind of cross-validation method.

We have two different data-sets received from two different Hungarian insurance companies. The first one contains household insurance contracts and covers seven development years (from 01.10.1998. to 30.09.2005., with 686 711 contracts, 179 618 claims, 1,3 bn HUF claim amount and 10,5 mn HUF outstanding claim amount), the second one consists of motor third party liability (MTPL) contracts and covers a four-year period (from 01.01.2000. to 31.12.2003., with 500 690 contracts, 43 392 claims, 13,5 bn HUF paid claim amount and 3,5 bn HUF outstanding claim amount). These two data-sets are actually of two different kinds. The household insurance claim data-set has a short run-off which means that the claims are reported and paid within a short period of time (approximately one year), while the MTPL data-set has a substantially longer run-off. Considering that one of our aims is to analyze how the period of the run-off triangle influences the ultimate reserve, we constructed four run-off triangles to both data-sets: with daily, monthly, quarterly and yearly periods, all of them include incurred claim numbers.
Stochastic models for claims reserving

Claim amounts were also examined, moreover, in that case we dealt with both incurred and paid claims – in that context the results can be very different, and the so-called Munich chain-ladder method, which uses both triangles and tries to reduce the gap between the results given by using them separately, was also tested. However, in this paper we discuss only the claim number results, because the claim amount results are far more complicated and discussing and analyzing them would be very spacious.

2 Deterministic and semi-stochastic methods

2.1 The Chain-Ladder technique

In this section we start with reviewing one of the most widely used reserving techniques, the chain-ladder (CL) method, then we provide two naive stochastic models, which allow us to determine quantiles of the ultimate reserve. Let $D_{i,j}$ denote the number of claims incurred in period $i$ and reported until period $i+j-1$ ($i=1,\ldots,t$, $j=1,\ldots,n-i+1$). If there are claims not reported after $n$ years, we denote the approximated claim number incurred in first period and reported after $n$ periods by $D_{1,n}$.

The essential assumption of the CL technique is that the rates do not depend strongly on $i$ (the period of claim incurrence), they all are approximately $l_j$. To calculate the $l_n \approx \frac{D_{i,n}}{D_{i,n}}$ rate (the rate of very late claims) we use our experience from previous periods or the outstanding claim number. (In the latter case $l_n = 1 + \frac{T_i}{D_{i,n}}$, where $T_i$ is the outstanding claim number referring the first period.) All the other $l_j$ rates are determined by a practical weighting of the real $l_j(i)$ rates – the advantage of these formulae is that we don’t need to calculate the actual $l_j(i)$ rates, we just need the $X_{i,j}$ claim numbers:

$$l_j = \frac{D_{i,j} \cdot l_j(i) + D_{i+1,j} \cdot l_j(i+1) + \ldots + D_{t,j} \cdot l_j(t)}{D_{i,j} + D_{i+1,j} + \ldots + D_{t,j}} = \frac{D_{i,j} \cdot l_j(i) + D_{i+1,j} \cdot l_j(i+1) + \ldots + D_{n-1,j} \cdot l_j(n-1)}{D_{i,j} + D_{i+1,j} + \ldots + D_{n-1,j}}, \quad j = 1,\ldots,n-1.$$

Because of its simplicity and reliability the CL method is a very popular reserving technique, moreover, it can be shown that it is correspondent with some stochastic reserving models (see for example [Verral, 2000]).
2.2 Semi-stochastic methods

As mentioned before, our aim is to develop stochastic methods which we can use to provide quantiles of the ultimate reserve. First we introduce two pretty simple methods which poorly use stochastic techniques therefore we call them semi-stochastic methods.

First of all we recall the \( l_j(i) := \frac{D_{i,j+1}}{D_{i,j}} \) values which we used to determine the \( l_j \) ratios in the CL method. Now these \( l_j \) rates are not the quotients of the sum of the \( D_{i,j} \) values in the columns \( j \) and \( j+1 \) but they are i.i.d. random variables. Instead of the real distribution hereafter we use the empirical distribution of these variables which is a discrete uniform distribution over the known \( l_j(i) \) values. Now, we can follow two different ways. On the one hand we can consider the expected values of these variables and by filling up the lower triangle using these values we get the expected value of the ultimate claim number, e.g., the expected value of the ultimate reserve. By the same way we can also calculate the variances of the variables and that of the ultimate reserve. Considering that we are dealing with a pretty large data-set, we assume that the ultimate reserve has a normal distribution which parameters can be estimated by the way described right now. Once we have the expected value and variance we can easily calculate quantiles of the appropriate normal distribution. This method is denoted by UN (uniform-normal). The other way is to generate a suitable large amount of outcomes of the ultimate reserve using the uniform distributions and then we can look at the empirical quantiles as the quantiles of the ultimate reserve – this method is denoted by UG (uniform-generating).

Let us be more specific about the first method. Our basic assumption is that the \( L_j \) variable has a discrete uniform distribution over the set \( \left\{ l_j(i) = \frac{D_{i,j+1}}{D_{i,j}} : i = 1, \ldots, n \right\} \). Therefore its expected value is given by \( E(L_j) = \frac{1}{n-j} \sum_{i=1}^{n-j} \frac{D_{i,j+1}}{D_{i,j}} \), so we have to multiply the known claim numbers in column \( j \) by this value to get the unknown claim numbers in column \( j+1 \). One can note that this expected value is the sum of the \( l_j(i) \) rates with the same weights, i.e., the average of them, so this method exactly follows the so-called „link ratios with simple average“ method. As mentioned above we assume that the \( L_j \) variables are independent, so we
can calculate the ultimate claim number by \( \sum_{j=1}^{n} D_{n-j+1,j} \prod_{k=j}^{n-1} E(L_k) \), e.g.

the ultimate reserve by \( \sum_{j=1}^{n} \left( D_{n-j+1,j} \left( \prod_{k=j}^{n-1} E(L_k) - 1 \right) \right) \). As discussed above this value is exactly the same that one can get using the „link ratios with simple average“ method. But considering the \( l_j \) rates as \( L_j \) variables we are able to determine the variance of the ultimate claim number which is equal to the variance of the ultimate reserve (because one can get the latter as the sum of the first and the known \( \sum_{j=1}^{n} D_{n-j+1,j} \) value). As assumed earlier the \( L_j \) variables are independent, so

\[
\text{Var} \left( \prod_{k=j}^{n-1} D_k \right) = E \left( \prod_{k=j}^{n-1} L_k^2 \right) - E^2 \left( \prod_{k=j}^{n-1} L_k \right) = \prod_{k=j}^{n-1} E(L_k^2) - \prod_{k=j}^{n-1} E^2(L_k),
\]

and using \( E(L_j^2) = \frac{1}{n-j} \sum_{i=1}^{n-j} D_{i,j+1}^2 \), we get that the variance of the ultimate reserve is given by \( \sum_{j=1}^{n} \left( D_{n-j+1,j} \left( \prod_{k=j}^{n-1} E(L_k^2) - \prod_{k=j}^{n-1} E^2(L_k) \right) \right) \). So

having the expected value and variance of the ultimate reserve and assuming a normal distribution we can easily calculate any quantile of that reserve.

Let us deal with the second approach. We assume that the \( L_j \) variable has a discrete uniform distribution over the set \( \left\{ l_j(i) = \frac{D_{i,j+1}}{D_{i,j}} : i = 1, \ldots, n-j \right\} \). To get the quantiles of the ultimate reserve we shall calculate all the possible outcomes (having the same probability), i.e., we shall determine all the possible reserves for each row (each period of incurrence). Considering that the reserve for the row \( i \) is given by \( D_{i,n-i+1} \cdot (L_{n-i+1} \cdot L_{n-i+2} \cdot \ldots \cdot L_{n-1} - 1) \) and the fact that \( L_j \) can have \( (n-j)! \) different values we have \( (n-1)!(n-2)! \cdot \ldots \cdot 1! \) possible outcomes for the ultimate reserve. If we deal with daily periods (which means the run-off triangle for the MTPL data-set has \( n=1461 \) rows), it can be seen that it is impossible to calculate all the possible outcomes. Therefore we generate only 1000 outcomes: from the \( L_1, L_2, \ldots, L_{n-1} \)
variables we take a random sample, then calculate the $D_{i,n-i+1}(L_{n-i+1}L_{n-i+2}...L_{n-1}-1)$ reserve for each period $i$, finally by summarizing these period-reserves we get the ultimate reserve. From these thousand values any quantile of the real ultimate reserve can be estimated.

3 The MCMC method

The methods discussed so far are deterministic or semi-stochastic. Using Markov Chain Monte Carlo technique makes it possible to build up a more sophisticated stochastic reserving method. For the different lengths of period the model is the same, only the parameters and the input are different. We give the description of the model for daily data. We decided to create a one-level model, so the parameters of the prior distributions are fixed numbers chosen in such a way that the prior distributions of the parameters are best fitted to the data. For the calculations we need the run-off triangle of incremental claims, its elements denoted by $C_{i,j}$, and we also use the vector $e$ containing the number of contracts in the several days (or months, quarters, years). In this model we assume that the data are Poisson distributed, as we are working with positive integers. The expected value of the claims incurred on day $i$, reported $j-1$ day later is the product of $e_i$, the number of (valid) contracts on day $i$, $X_i$, the claim intensity of day $i$ per one policy, and $Y_j$, the proportion of claims reported $j-1$ days late: $E(C_{i,j}) = e_iX_iY_j, (i, j = 1,...,n)$. Assuming that the claim intensity of day $i$, $X_i$ and the proportion of claims reported $j-1$ days late, $Y_j$ are random variables, we have to determine their prior distributions. $X_1,...X_n$ are considered to be prior independent and identically Gamma distributed random variables. We also fitted the model with other prior distributions such as normal but this choice seemed to be the best working. Additionally, when examining claim numbers it is widespread to use Poisson distribution for the number of claims and Gamma for the intensity as this way the unconditional distribution of claims incurred on a fixed day is negative binomial. $Y_1,...,Y_n$ are prior Beta distributed random variables on the appropriate intervals, determined below. We have the constraint $\sum_{k=1}^{n} Y_k = 1$ as we suppose (presume) that all the claims are reported at most $n$ days late. The parameters of the prior distributions are chosen as follows. Alpha and lambda parameters of the Gamma
distribution are calculated in such a way that the expected value is the mean of the $X_i$'s starting values, and the variance is suitably high to let the data dictate. For the starting values of $X_i$ we used the naive estimate:

$$X_{start} = \sum_{i=1}^{n} C_{1,i}^t, \quad X_{start}^t = X_{start} \cdot \frac{\sum_{j=1}^{n-i+1} C_{i,j}^t}{\sum_{j=1}^{n-i+1} C_{i,j}^t}$$

where the matrix $C^t$ has row vectors $C_{iii} = c / e_i$. The prior distribution of $Y_i$, $(i = 1,...n-1)$ is Beta in the interval $[0,1 - \sum_{j=1}^{i-1} Y_j]$ with parameters $\alpha, \beta$ where the variance is fairly high and the expected value is greater than the half length of the interval. Thus, we have the following prior distributions:

$$[x] \sim \prod_{i=1}^{n} x_i^{a-1} e^{-2x_i}, \quad [y_i] \sim y_i^{a-1} (1 - y_i)^{b-1}$$

$$[y_i \mid y_1,...,y_{i-1}] = \frac{1}{1 - \sum_{j=1}^{i-1} y_j} \left(\frac{y_i}{\sum_{j=1}^{i-1} y_j}\right)^{a-1} \left(1 - \frac{y_i}{\sum_{j=1}^{i-1} y_j}\right)^{b-1}, \quad i = 2,...,n-1,$$

$$[y_i \mid y_1,...,y_{i-1}] = 1 - \sum_{j=1}^{n-i} y_j$$

where notation $[x \mid y]$ denotes the conditional density function $f_{X\mid Y}(x \mid y)$ in continuous cases, and the $P(X = x \mid Y = y)$ conditional probability in discrete cases, respectively. Providing that the elements $C_{i,j}$ have Poisson distribution with the appropriate parameter, the joint posterior distribution is:

$$[c \mid x, y] \sim \prod_{i=1}^{n} \prod_{j=1}^{n-i+1} (c_{i,j} \cdot x_i \cdot y_j)^{c_{i,j}} \cdot e^{-\sum_{j=1}^{n} c_{i,j} \cdot y_j} \cdot \prod_{i=1}^{n} x_i^{a-1} \cdot e^{-\sum_{j=1}^{n-i} y_j} \cdot \prod_{j=2}^{n-i} y_j^{a-1} \left(1 - \sum_{k=1}^{n} y_k\right)^{b-1}$$

We obtain the posterior distribution of $X_i$:

$$[x_i \mid c, y] \sim \Gamma\left(\alpha + \sum_{j=1}^{n-i+1} c_{i,j}, \lambda + c_{i,j} \cdot \sum_{j=1}^{n-i+1} y_j\right).$$
The posterior density function of $Y$ is a bit more complicated:

$$y | c, x \sim y_n^{a+b-1} \cdot \prod_{j=1}^{n-1} y_j^{a+\sum_{i=1}^{j-1} c_i x_i} \cdot e^{-\gamma_j \sum_{i=1}^{n} c_i x_i} \cdot \left(1 - \sum_{j=1}^{n-1} y_j\right)^{-a}.$$  

Clearly, the latter is not the density function of a well-known distribution, so we have to use Markov Chain Monte Carlo simulation to get the expected values and also the appropriate quantiles. We determine certain quantiles of the reserve in the following way: in each MCMC iteration we generate a subsequent sample of $X_i$ and $Y_j$, and also of $C_{i,j}$ as random sample of Poisson($\sum_{i=j}^{n} X_i Y_j$). So in each iteration we have a whole square of claims incurred in the first $n$ days and reported at the most $n-1$ days late from which it is easy to calculate the reserve as the sum of the values in the right down triangle. Therefore we performed 1 000 000 iteration of the MCMC method, leaving the first 100 000 out as the burn-in period, using the rest for determining the point-estimate and the quantiles. In each iteration we obtained a subsequent sample of the parameters and using them we generated a Poisson-distributed value with the appropriate expected value for the reserve. This way we obtained the quantiles deriving from these 900 000 values.

4 The results

In the previous sections we introduced the models used to estimate the reserve of the two data-sets, here we present the results for the monthly and the quarterly data – daily periods result a high variance when determining the quantiles and using annual data we have just four or seven rows in the run-off triangle which is insufficient for a correct MCMC algorithm.

![Fig. 2. Point estimates of the reserve for the household data-set.](image)
Figures 2 and 3 show the point estimates for the reserve given by the different methods and by using different time periods for the household and MTPL data-set, respectively. Of course we received different values working with the different periods but we also got highly different results using the various methods on the same run-off triangles which is not so obvious. It can be seen that the estimates of each algorithm grow with the length of the time period. As for the different methods it seems that the measures produced by MCMC method are consequently near the chain-ladder estimates while the two semi-stochastic methods are similar to each other, always giving higher results than the other two – as far as the household data-set is concerned. But considering the MTPL data it is the MCMC method that gives different values which are lower than the estimates of the CL technique while the two simple stochastic methods result very similar values. Therefore it looks like that choosing the best method can be done only according to the data.
For the quantiles we have three methods as from the chain-ladder technique we get only a point estimate. In Figures 4 and 5 the 50, 75, 90, 95% quantiles for the reserve can be seen for the three different methods using different time periods. As previously stated the MCMC results tend to be lower than the others and this is true for all quantiles and for both data-sets. This method differs from the other two in respect of the variation which is basically lower than that of the semi-stochastic methods. As mentioned before, the daily data give always the highest variance, but for the other periods the variance is neither decreasing nor increasing with the length of the time period.

5 Testing

The purpose of this paper is to calculate the reserve for an insurance company’s claim numbers. In the previous chapters we introduced various methods of estimation, and of course we are interested in comparing these methods, let alone the fact that our estimates are rather different. A way of testing the goodness is calculating these estimations for just the first part of our data, and the real value of the reserve using the continuation of the data. By comparing the estimates to this last value it turns out how good the different methods are for this sequence of data (a sort of cross-validation). It is unnecessary to explain how important is to know more about the reliability of these algorithms. The result of our calculations show that even in the simple case of household claims some of the basic procedures (like chain-ladder) fails. (As the MTPL data-set covers only four years, this testing method can barely applied to it.) For the test calculation we used the claims incurred and reported in the first 3 years. Here we present the results received by using monthly periods. The real
value for this part of the data is 1788 while the CL method gives only 990 which is a really bad underestimation.

Figure 6 shows the quantiles of the reserve for this first three years computed by the three stochastic methods compared to the real value. Even if the point estimate is low, the algorithm can be acceptable if the real value of the reserve lies in the [10%, 90%] confidence interval of our estimations. As discussed above the MCMC method gives the lowest results and has the lowest variance so it is not surprising that all the quantiles given by this model lie under the real value. But the 90% and 95% quantiles of the two semi-stochastic methods exceed the real value (in the case of the UN model also the 75% quantile is higher) so by using these methods it is possible to get reliable estimates for the reserve.

6 Conclusion

As a conclusion we can say that longer time periods result higher reserves, while using more detailed data (shorter periods) lowers the result but also raises the variance and therefore the uncertainty. Concerning the other goal, i.e., to develop stochastic models for claims reserving (it is obvious that the point estimates given by the deterministic methods are not reliable enough), one can see that also the stochastic models can provide very different results. The main question is how to choose between these models. We suggest that one should apply and also test all the methods (the deterministic models too), and then the most appropriate can be taken according to the type of the data set (i.e. shorter or longer run-off) and to the results of testing. In the future we want to develop our methods by using bootstrap technique and compare them with other stochastic models.
References