Department of Mathematical Statistics
Royal Institute of Technology, Stockholm, Sweden
Degree Project in Engineering Physics, Second Level

Claims Reserving on Macro- and Micro-Level

Author:
Annelie Johansson
annelieg@kth.se

Supervisors:
Boualem Djehiche
Tobias Janstad

September 4, 2015

Abstract

Three methods for claims reserving are compared on two data sets. The first two methods are the commonly used chain ladder method that uses aggregated payments and the relatively new method, double chain ladder, that apart from the payments data also uses the number of reported claims. The third method is more advanced, data on micro-level is needed such as the reporting delay and the number of payment periods for every single claim. The two data sets that are used consist of claims with typically shorter and longer settlement time, respectively. The questions considered are if you can gain anything from using a method that is more advanced than the chain ladder method and if the gain differs from the two data sets. The methods are compared by simulating the reserves distributions as well as comparing the point estimates of the reserve with the real out-of-sample reserve. The results show that there is no gain in using the micro-level method considered. The double chain ladder method on the other hand performs better than the chain ladder method. The difference between the two data sets is that the reserve in the data set with longer settlement times is harder to estimate, but no difference can be seen when it comes to method choice.

Keywords: Claims reserving, Chain Ladder Model (CLM), Double Chain Ladder (DCL), Micro-model

Sammanfattning

Tre reservsättningsmetoder jämförs på två dataset. De första två metoderna är den välkända chain ladder-metoden som använder sig av aggregerade utbetalningar samt den relativt nya metoden double chain ladder som förutom utbetalningarna använder sig av antalet anmälda skador. Den tredje metoden baseras på mikro-nivå och kräver information om varje enskild skada, såsom anmälningstid och antalet utbetalningsperioder. De två dataseten som används är ett som innehåller skador med typiskt kortare avvecklingstider och ett som som innehåller skador med typiskt längre avvecklingstider. Frågorna som behandlas är om man vinner något på att använda en mer avancerad metod än chain ladder och om det skiljer sig åt mellan dataseten. Metoderna jämförs genom simulering av reserven, men också genom att jämföra punktskattningar med den verkliga reserven. Resultaten visar att man i detta fall inte vinner något på att använda mikro-metoden. Double chain ladder å andra sidan presterar bättre än chain ladder. Skillnaden mellan de två dataseten är att det är svårare att estimera reserven när avvecklingstiden är längre, men ingen skillnad ses när det gäller val av metod.

Svensk titel: Reservsättning på makro- och mikro-nivå

Nyckelord: Reservsättning, Chain Ladder Method (CLM), Double Chain Ladder (DCL), Mikro-nivå

Acknowledgment

The first thank is dedicated to Totte Pikanen at Länsförsäkringar, he is the person who enabled this master thesis. The supervisor at Länsförsäkringar, Tobias Janstad, receives the author's best gratitude for his willingness to gain insight to the project during the whole process. Per Edholm, also at Länsförsäkringar, is thanked for his inspiration in the Python language. All the other colleagues at Länsförsäkringar receive the author's gratitude for their support.

The author also want to thank the supervisor professor Boualem Djehiche at KTH for his interest, fast mail correspondence and his positive spirit. Help during the project has been received by Research Associate Deniz Akdemir at Cornell University and Professor Michel Denuit at UCL, thank you!

The last but not the least thank goes to the author's family and friends that always give support and never stop believing in me.

Contents

1 Introduction 1

2 Macro-Level Methods 3
2.1 Basic Concepts 4
2.2 Chain Ladder Method 5
2.3 Double Chain Ladder Method 6
2.4 Simulation of Reserves Distribution 8

3 Micro-Level Methods 10
3.1 Basic Concepts 11
3.2 Historical Simulation Method 13
3.2.1 Likelihood 13
3.2.2 Parameter Estimation 16
3.3 Simulation of Reserves Distribution 17

4 Data Description 18
4.1 Short-Tail Data 18
4.2 Long-Tail Data 19
4.3 Data Manipulation 20

5 Model Selection 21
5.1 Short-Tail Data 22
5.2 Long-Tail Data 23

6 Results 25
6.1 Macro-Level for Short-Tail Data 25
6.2 Micro-Level for Short-Tail Data 29
6.3 Observed Reserve for Short-Tail Data 31
6.4 Macro-Level for Long-Tail Data 32

6.5 Micro-Level for Long-Tail Data 36
6.6 Observed Reserve for Long-Tail Data 38

7 Conclusions 40

Chapter 1

Introduction

The idea of insurance is that persons can share the risk that they are exposed to. Everyone pays a bonus for being covered by the insurance, and those who suffer an accident receive compensation. The problem with this idea is that the future is unknown, we do not know how many accidents will occur. Therefore, the company must find a way to estimate the risk in order to be able to pay all future claims. This is important to the company in order to be profitable, but also to be able to report the correct numbers to the authority. One part of the risk that an insurance company faces comes from the risk of a future claim to occur. Another part of the risk comes from the claims that have occurred already. It takes some time until a claim is reported and the final payment of a claim is not known until it is closed. To be prepared for this part of the risk, the insurance company estimates how much the claims that have occurred already will cost. This estimation process is called claims reserving.

Traditionally, the claims reserving has been preformed by the use of aggregated methods, also called macro-level methods. Aggregated methods use data where many payments have been summed up. The most common such method is the chain ladder method (CLM). This is a method that uses payment data that is summed up with respect to accident time and the time delay between accident and payment. This method has been proposed in many variations in the literature, and one of these extensions is double chain ladder (DCL). This method uses the same payment data as the chain ladder method, but also aggregated data of the number of reported claims. In later years, more advanced claims reserving methods have been developed, the micro-level methods. These methods do not use data that is summed up. Instead they use data from every single claim. They have not done any break through in practice and this is probably due to the fact that they need a lot more effort. It takes more time to understand a micro-level method, the implementation can be tedious and for a large data set the estimation of parameters can take a long time. Despite all these drawbacks, we still end up with the following question. If a huge data set consisting of detailed information is available to the company, why not use it?

It could be that the micro-level methods perform better and make it worth all extra effort, but the claims reserving actuary does want to know that it will end up profitable for the company before spending too much time with it. Therefore, the aim of this thesis is to compare a micro-level method with the two aggregated methods chain ladder and double chain ladder. The micro-level method that will be used is inspired by one that fits the development of payments to a multivariate skew normal distribution, but here we will use historical simulation instead of a statistical distribution. The reason is that the data does not fit the multivariate skew normal distribution. Two aggregated methods are included in the thesis since one of them is very simple and the other one a bit more complicated. This will enable us to see if we need to go all the way to the complicated micro-level method, or if a small extension from the chain ladder method is sufficient for better results.

In the articles where micro-level methods have been proposed, similar comparisons have been made comparing the micro-level method with the chain ladder method. No comparison between a micro-level method and the double chain ladder method is known to the author. Also, the insurance companies do not want to share their confidential data and the same data set occurs in more than one article. This thesis will contribute with the comparisons on two data sets that have not been seen in the literature before. One data set consists of short-tail data, which means that the time between accident and closure of the claims is typically short. The other data set consists of claims with typically longer settlement time, long-tail data. This data set should give a more insecure estimate of the reserve than the other, so it is possible that this data set makes more profit from an advanced method than the other data set.

In claims reserving, it is beyond all the expected value that is of interest, but a point estimate does not say much. The reserve could be distributed close around the expected value, but it could also have a very large variance. In the later case the reserve could be set above the expected value. Therefore, the distribution of the reserve will be simulated and compared for all three methods.

This thesis will start by two sections where the three methods that will be compared are explained, the first section includes the two macro-level methods and the second is about the micro-level method. Afterwards, the two data sets are presented and then the subject of choosing the right distributions for the micro-level method is treated. Finally the results are presented followed by the conclusion chapter.

Chapter 2

Macro-Level Methods

The chain ladder method is probably the most famous and well known claims reserving method. During the years, it has been the subject of many articles. An important such article is when a simple method of calculating the distribution-free standard error of the chain ladder method was presented (Mack, 1993). In the same time, several authors were investigating the subject of finding a stochastic model for the chain ladder method that before was interpreted as non-stochastic. In 1994, a distribution-free stochastic model that gives exactly the same results as the chain ladder was presented (Mack, 1994). Later, different stochastic chain ladder methods were compared (Hess and Schmidt, 2002). Despite all stochastic methods suggested, the work of finding the insecurity of the standard non-stochastic chain ladder keeps on moving. A few years ago this insecurity was estimated by comparing the reserves estimate at time m with the reserves estimate at time $m + 1$ (Wthrich et al., 2009). The chain ladder method has also been extended as well as combined with other methods. An example is the extension of considering both the payment triangle as well as the number of counts triangle, leading to the double chain ladder method (Miranda et al., 2012). Extensions to the double chain ladder method have been made with special focus on inflation, see (Martnez-Miranda et al., 2013) and (Miranda et al., 2015). Below, the two methods chain ladder and double chain ladder will be explained in more detail, but first some basic concepts will be illustrated. In the following, the chain ladder method will be abbreviated by CLM and the double chain ladder method will be abbreviated by DCL.

2.1 Basic Concepts

A very common concept when it comes to claims reserving on macro-level is the aggregated claims triangle. An illustrative example of such can be seen in Figure 2.1. In the figure we look at $m+1$ accident years, which are represented by the left vertical axis. As can be seen the accident years are numbered from 1 to $m+1$. This means that if we consider three accident years, 2001, 2002 and 2003, then 2001 will be denoted by 1, 2002 by 2 and 2003 by 3. Therefore, all payments related to accidents that happened in 2001 will be placed in the first row of the triangle.

Which column a payment is placed in depends on how long time it has taken between the occurrence of the accident and the payment. If less than one year has past, the payment will be placed in the first column of the row, denoted by 0. If the payment delay instead is at least one year but not two years, the payment will be placed in the second column, denoted by 1. The same logic holds for all columns. The easiest way of deciding which development period a payment should belong to is to simply say that if an accident occurred in 2000, then payments made in 2000 belongs to development period 0, payments in 2001 belongs to development period 1 etc. This is the interpretation that will be used in this thesis.

Notice that the value in the triangle at row i and column j is not the total amount paid from accident year i with j years delay. It is instead this amount summed to all payments that has been made earlier for that accident year. This is called that the triangle consists of aggregated data. Another fact worth noticing is that the period considered does not have to be a year, it could be a day, a week, a month and so on. We will use a year as time period. Also, the data does not have to be payments, even if it often is. It could also be the number of reported claims at the time as in DCL.

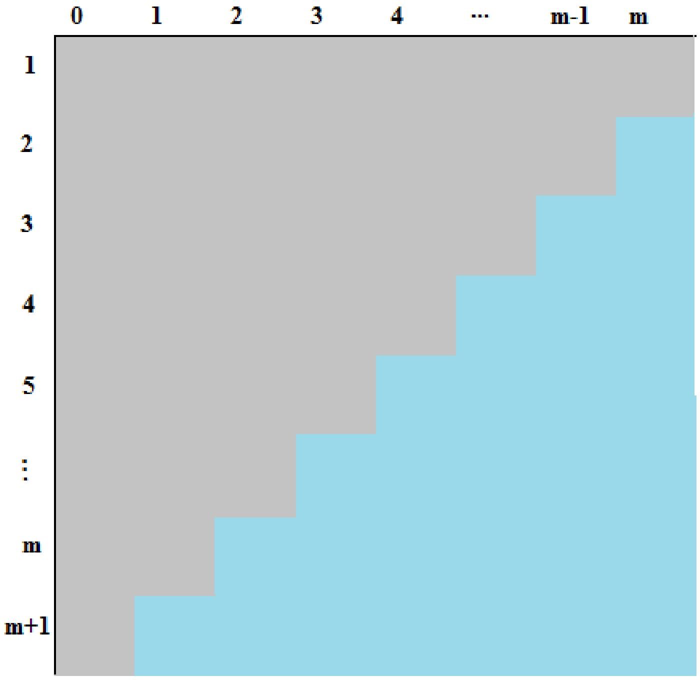

Figure 2.1: The figure shows a claims triangle with the accident years as rows and the development years as columns. The gray area is the known payments and the turquoise area is the reserve we want to estimate.

2.2 Chain Ladder Method

The chain ladder method that we are using relies on the assumption that the cumulative payment from an accident year increases with a development factor, f_j, for each development period j. The number of observed development periods must be large enough, so that we can make the assumption that all claims are closed at the end of the payment triangle. Denote the cumulative payments from the first development period of each accident year by $C_{i,0}$, then the cumulative payment C_{ij} can be calculated by

$$C_{ij} = C_{i,0} \prod_{k=0}^{j-1} f_k.$$

The develop factors are estimated from the cumulative payments triangle explained above, by

$$\hat{f}_j = \frac{\sum_{k=1}^{m-j} C_{k,j+1}}{\sum_{k=1}^{m-j} C_{k,j}}, \qquad j = 0, 1, \ldots, m-1.$$

The total amount that will be paid for each accident year can be estimated by multiplying the last known cumulative payment, $C_{i,m+1-i}$, with the development factors for the future development periods. Then the reserve for that accident year is the paid amount subtracted to the total amount and the total reserve is simply the sum of the reserves for each accident year. the formula for calculating the total reserves amount for accident year i will look like

$$\hat{C}_{i,m} = C_{i,m+1-i} \prod_{k=m+1-i}^{m-1} \hat{f}_k, \qquad i = 2, 3, \ldots, m+1.$$

Finally, we state the formula for the total reserve,

$$\text{Reserve}_{CL} = \sum_{i=2}^{m+1} C_{i,m} - C_{i,m+1-i}.$$

2.3 Double Chain Ladder Method

The method in this section is the same as the one presented in the paper (Miranda et al., 2012) and all formulas are contributed to the authors of this paper. The double chain ladder method uses the same payment triangle, Δ_m, as the chain ladder method, but it also uses a triangle of the number of claims reported (incurred counts). This triangle is denoted by ξ_m. The name double comes from the fact that two triangles are used instead of one. We start by the number of claims from accident period i that are reported in development period j and denote this by N_{ij}. This is in incremental form. The expected value of this quantity is assumed to depend on two parameters,

$$\mathrm{E}[N_{ij}] = \alpha_i \beta_j.$$

These claims will be settled at different times, so focus on the claims from all N_{ij} claims that are settled with l periods from reporting (all claims are assumed to have only one payment, so settlement is the same as payment in this case). This number is denoted by N_{ijl} and has the conditional expected value

$$\mathrm{E}[N_{ijl}|\xi_m] = N_{ij}\tilde{\pi}_l.$$

The expected value of the k:th payment from the N_{ij} claims is

$$\mathrm{E}[Y_{ij}^k] = \mu\gamma_i,$$

which means that we assume that the mean value of the payments is the same over all development periods. The parameter γ_i can be seen as an inflation parameter.

Now, when the model setup has been presented, the estimation of the included parameters will be described. We start with the development factors $\hat{f}_j$ from the payment triangle. Use these to get $\hat{\hat{\alpha}}_i$ and $\hat{\hat{\beta}}_j$ by

$$\hat{\hat{\alpha}}_i = C_{i,m+1-i} \prod_{j=m-i+1}^{m-1} \hat{f}_j,$$

$$\hat{\hat{\beta}}_0 = \frac{1}{\prod_{k=1}^{m-1} \hat{f}_k}$$

and

$$\hat{\hat{\beta}}_j = \frac{\hat{f}_j - 1}{\prod_{k=j}^{m-1} \hat{f}_k}.$$

The same calculations gives us $\hat{\alpha}_i$ and $\hat{\beta}_j$ from the incurred counts.

Once this is done, the values of $\tilde{\pi}_l$ for all l are received by solving the following equation system,

$$\begin{pmatrix} \hat{\hat{\beta}}_0 \\ \hat{\hat{\beta}}_1 \\ \vdots \\ \hat{\hat{\beta}}_{m-1} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 & 0 & \dots & 0 \\ \hat{\beta}_1 & \hat{\beta}_0 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ \hat{\beta}_{m-1} & \hat{\beta}_{m-2} & \dots & \hat{\beta}_0 \end{pmatrix} \begin{pmatrix} \pi_0 \\ \pi_1 \\ \vdots \\ \pi_{m-1} \end{pmatrix}.$$

The mean value of the payments can be estimated by

$$\hat{\mu} = \frac{\hat{\hat{\alpha}}_1}{\hat{\alpha}_1},$$

which enable us to calculate the inflation parameters as

$$\hat{\gamma}_i = \frac{\hat{\hat{\alpha}}_i}{\hat{\alpha}_i \mu}, \quad i = 1, \dots, m+1.$$

Finally, the reserve can be calculated separately for the claims that have been reported, X^1_{ij}, and the claims that have not yet been reported, X^2_{ij},

$$\hat{X}^1_{ij} = \sum_{l=i-m+j}^{j} N_{i,j-l}\hat{\pi}_l\hat{\mu}\hat{\gamma}_i$$

and

$$\hat{X}^2_{ij} = \sum_{l=0}^{i-m+j-1} \hat{N}_{i,j-l}\hat{\pi}_l\hat{\mu}\hat{\gamma}_i,$$

with $\hat{N}_{i,j-l} = \hat{\alpha}_i\hat{\beta}_{j-l}$. The total reserve is determined by summing X^1_{ij} and X^2_{ij} for all i, j in the turquoise area in Figure 2.1.

2.4 Simulation of Reserves Distribution

The process of creating a distribution for the reserve can be done by bootstrapping, either a parametric bootstrapping or a non-parametric. It is common to use errors to bootstrap a claims reserves distribution, which is a non-parametric bootstrapping method. Consider the following residual,

$$r = \frac{X_{ij} - \hat{\hat{\alpha}}_i\hat{\hat{\beta}}_j}{\sqrt{\hat{\hat{\alpha}}_i\hat{\hat{\beta}}_j}},$$

for the incremental observed payment X_{ij}. This is the unscaled Pearson residual and it has been suggested that this residual should be scaled with a scaling factor ϕ in order to take the number of observations, n_o, as well as the numbers of parameters, n_p, under consideration (England, 2001). The scaling factor ϕ is calculated by

$$\phi = \frac{\sum r^2}{n_o - n_p}.$$

In the paper by England it is suggested to start with the estimated incremental payments triangle and then generate random residuals from the set of scaled residuals. The generated residuals are added to the estimated incremental payments creating a new payment triangle. This triangle is used to produce a bootstrapped CLM reserves estimate. By repeating this process many times a reserves distribution will be determined.

The same idea can be extended to other models than the CLM. Here, we will use the same bootstrap method to find the distribution of the DCL reserve. This means that we will calculate one set of residuals for the payments triangle and one set of residuals for the incurred counts triangle. After scaling the residuals, two new triangles are generated and used in the DCL algorithm. Now, the point estimates from CLM and DCL are extended into two reserves distributions.

Chapter 3

Micro-Level Methods

The main question of this thesis is to investigate if you can gain something from considering individual claims data instead of aggregated data. Two decades ago, it was rare with methods considering individual claims, but there existed methods that used more information than can be found in the ordinary claims triangle. Already in 1978, a method was proposed (Reid, 1978) that was based on the number of claims that was open, closed with payment and closed without payments. This was a quite complex method and the author extended it later, which means that it became very flexible. An early work that described the use of marked point processes for the development of a claim was published in 1989 and it used a martingale approach (Arjas, 1989). The idea of using a marked process for the individual claim development was later extended by Norberg, who used a marked Poisson process in his two papers (Norberg, 1993) and (Norberg, 1999). A closer description of these kind of methods can be found in chapter 10 in the book (Wtrich and Merz, 2008). Another extension using a non-parametric Bayesian framework for the process was published in 1996 (Haastrup and Arjas, 1996). Closed related to the individual claims methods was the method presented in 1997 (Wright, 1997), it used the empirical distribution of all individual payments.

The past few years, the development of the so-called micro-level methods has been faster. An approach that fits the individual claims development with generalized linear models was proposed by Taylor and McGuire (Taylor and McGuire, 2004). In 2007, Larsen presented a work based on the marked Poisson processes used before (Larsen, 2007), he used some stochastic models in his work. A model that combines parametric framework with non-parametric was proposed in 2009 (Zhao et al., 2009) and the year after a method that considers copulas was presented (Zhao and Zhou, 2010). A substantial case study of the marked Poisson process method has recently been published (Antonio and Plat, 2014).

Now, a method for claims reserving that combines parts from the multivariate skew normal framework (Pigeon et al., 2013) with ordinary historical simulation will be presented. This is the method that will be used in the analysis here. An extension of the multivariate skew normal framework was presented in 2014 (Pigeon et al., 2014). First, some basic concepts are introduced and then the method is presented. Finally, the process used for simulating the reserves distribution is explained. The point estimate of this model is simply the mean value of the simulated distribution.

3.1 Basic Concepts

The information used in a micro-level method consists of several dates, the time between them and the amounts of payments. The dates that are of interest are presented as points in Figure 3.1, where a claim is represented by a line. The accident date is the date when the accident, resulting in a claim, occur. The insurance company will not know about the claim until the reporting date, when the insurance holder reports the accident to the company. The time that it takes for the insurance holder to report the claim is denoted by T_{ik}, where the index ik stands for claim ik. This means that the claim is the kth claim of all claims with occurrence period i.

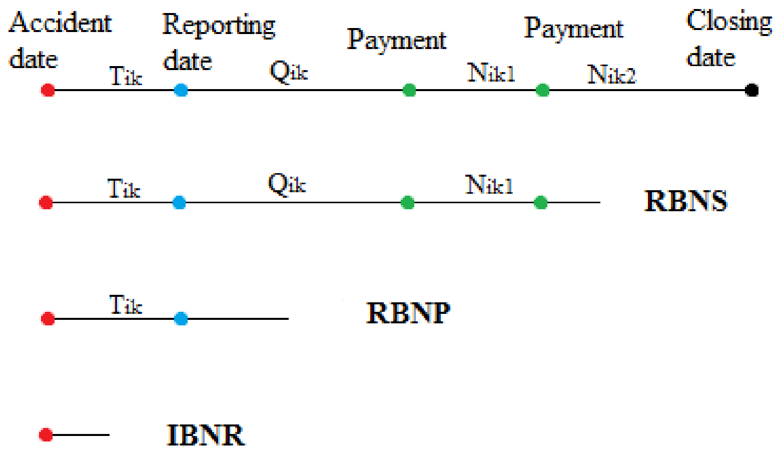

Figure 3.1: The figure shows examples of four types of claims. From above, the lines correspond to a closed claim, a Reported But Not Settled claim, a Reported But Not Paid claim and an Incurred But Not Reported claim, respectively. The accident date is presented as a red point, the reporting date as a blue one. Payments are the green points and the black point is the closing date. The letters T_{ik}, Q_{ik} and N_{ikj} represents the time between the two points they are placed between.

After the claim has been reported it may take some time until the first (if any) payment, this time will be called Q_{ik}. If there is no payment, then Q_{ik} represents the time until closure and in this case the closure will be seen as a payment of 0. For claims with one or more payments, the time from the jth payment until the next payment or closure is denoted by N_{ikj}. The number of payment periods is denoted by U_{ik}. It equals 0 if there are 0 or 1 payment, 1 if there are 2 payments, 2 if there are three payments etc.

Depending on how far in the development process a claim has come, it can be sorted into one of four types of claims, as Figure 3.1 illustrates. The claims that are considered as completed are called closed claims. The claims that have had one or more payments without being closed are called Reported But Not Settled, RBNS. The reported claims that have not yet been paid and not closed are called Reported But Not Paid, RBNP. Finally, the claims that have occurred but are not yet known to the insurance company are called Incurred But Not Reported, IBNR.

The notations that have been introduced in this section are inspired and very similar to the ones introduced in (Pigeon et al., 2013). The same kind of information is used in other micro-level methods as well.

3.2 Historical Simulation Method

With these introduced notations, the historical simulation method that will be used can be explained. But first the method by (Pigeon et al., 2013) is introduced. It relies on a discrete time line and in this thesis the time will be one month. This means that all payments made in April one year will be summed into one payment during that time step. This method is inspired by the development factors in CLM, but here development factors for the development of every single claim are considered. A vector with a claims first payment as the first component followed by development factors will be denoted by

$$\mathbf{\Lambda}_{u_{ik}+1} = [Y_{ik1} \quad \lambda_1^{ik} \quad \dots \quad \lambda_{u_{ik}}^{ik}]'. \tag{3.1}$$

This means that the second incremental payment is given by $Y_{ik1} \cdot \lambda_1^{ik}$, the third incremental payment is $Y_{ik1} \cdot \lambda_1^{ik} \cdot \lambda_2^{ik}$ etc. The logarithm of this vector is, as the name of the model indicates, fitted to a multivariate skew normal distribution. In the historical simulation method, the multivariate skew normal distribution is replaced by the observed payments. This operate better since both data sets consist of a large amount negative payments. The development vectors could of coarse be fitted to the multivariate skew normal distribution without logarithms, but this leads to a very bad fit.

The other parameters in the model, which is the time delays and number of periods introduced under *Basic Concepts*, are fitted to discrete distributions of a specific form. These discrete distributions are combinations of a truncated probability distribution and a number of degenerate components. Let us denote their probability mass functions by g_T, g_Q, g_U and g_N, and let them correspond to T_{ik}, Q_{ik}, U_{ik} and N_{ik}, respectively. Their cumulative density functions will be denoted by G with the same indexing.

In the following, the theory that is needed to analyze the model will be described. First the likelihood function for how likely the observed data is will be introduced. This is partly used to perform parameter estimates for some of the components in the model, and partly used to calculate the AIC and BIC statistics. The work of estimating all parameters is explained afterwards. Then we will go forward to the process of simulating the reserve distribution.

3.2.1 Likelihood

The expressions for the likelihood functions are taken from the work by (Pigeon et al., 2013), with the likelihood for the multivariate skew normal removed. All likelihood expressions will be on the following form,

$$\mathcal{L} = g(x; y|z). \tag{3.2}$$

The value that is observed is placed as x, this could be the reporting delay for example. The parameters that are to be estimated are placed as y, these parameters could be a vector, a matrix or even several vectors and matrices and depends on which statistical model that is used. Finally, there is often a time condition that is known, if we sample the data today, then we know that all reporting delays that are observed must have occurred before today. This initial condition is placed where z is placed in (3.2). With the notation t_i^* for the number of periods between occurrence period i and the latest observed event date, it must hold for all observed time delays from occurrence period i that they are smaller than t_i^*.

An observed reporting delay of t_{ik} has the likelihood function

$$\mathcal{L}^T = g_T(t_{ik}; \boldsymbol{\nu}|T_{ik} \leq t_i^*). \tag{3.3}$$

In the same manner, the observed payment delay q_{ik} has a likelihood function that looks like

$$\mathcal{L}^Q = g_Q(q_{ik}; \boldsymbol{\Psi}|Q_{ik} \leq t_i^* - t_{ik}) \tag{3.4}$$

and the observed number of payment periods u_{ik} has

$$\mathcal{L}^U = g_U(u_{ik}; \boldsymbol{\beta}|U_{ik} \leq t_i^* - t_{ik} - q_{ik}). \tag{3.5}$$

Finally, the likelihood function for the observed lengths, n_{ikj}, of the u_{ik} payments periods will be stated. This function includes the indicator function $I_{j>1}$, which equals 1 if $j > 1$ and 0 otherwise. Notice that when u_{ik} equals 0, there is no observed lengths n_{ikj} and hence the corresponding likelihood should not be a part of that claim's likelihood. The likelihood function looks like

$$\mathcal{L}^N = \prod_{j=1}^{u_{ik}} g_N(n_{ikj}; \boldsymbol{\phi}|0 < N_{ikj} \leq t_i^* - t_{ik} - q_{ik} - u_{ik} + j - I_{j>1} \sum_{p=1}^{j-1} n_{ikp}). \tag{3.6}$$

The condition in this likelihood looks a bit different than for the other likelihood functions, which could be confusing. Therefore, let us try to convince ourselves that the given condition is correct. First of all, N_{ikj} must be at least 1. This is due to the definition of N_{ikj} that does not accept a N_{ikj} to be 0. The upper limit in the condition consists of one part that is similar as before, but it also consists of

$-u_{ik} + j - I_{j>1} \sum_{p=1}^{j-1} n_{ikp})$. The $-u_{ik} + j$ can be thought of as a security to be sure that there will be at least one period left for the n_{ikj} that have not yet been considered. The sum is simply a subtraction of the previous n_{ikj}.

Now, the likelihood function for all different kind of claims can be stated with the help of the above introduced likelihood functions. For a closed claim with at least 1 payment period, the likelihood function consists of all above likelihoods since this claim has observation of all parameters. The likelihood function is

$$\mathcal{L}^{CL+} = \mathcal{L}^{T} \cdot \mathcal{L}^{Q} \cdot \mathcal{L}^{U} \cdot \mathcal{L}^{N}, \tag{3.7}$$

where the index $CL+$ stands for a closed claim with at least one payment period. The likelihood for a closed claim with no payment period (0 or 1 payment) is the same except that the likelihood for lengths of payment periods is omitted. If we index these types of claims by $CL0$, the likelihood is given by

$$\mathcal{L}^{CL0} = \mathcal{L}^{T} \cdot \mathcal{L}^{Q} \cdot \mathcal{L}^{U}. \tag{3.8}$$

The likelihood for a RBNS claim can be found in the same manner, depending on if there is any yet observed payment periods or not. Although, we do not know how many payment periods it finally will be for that claim, we only know that it will be at least u_{ik}^{*}, which is the number yet observed. Therefore, we cannot use the likelihood for U, but the cumulative distribution function for U can be used to express the probability

$$\Pr(u_{ik} \geq u_{ik}^{*}) = \Pr(u_{ik} > u_{ik}^{*} - 1) = 1 - \Pr(u_{ik} \leq u_{ik}^{*} - 1). \tag{3.9}$$

The second step can be done because U is discrete. This means that the likelihood for a RBNS claim with at least 1 payment period is likelihood is given by

$$\mathcal{L}^{RBNS+} = \mathcal{L}^{T} \cdot \mathcal{L}^{Q} \cdot (1 - G_U(u_{ik}^{*} - 1; \boldsymbol{\beta})) \cdot \mathcal{L}^{N}. \tag{3.10}$$

If the RBNS claim has only one observed payment, then u_{ik}^{*} will be 0 and 3.9 will simply be 1 and can be ignored. Also, there is no observed n_{ikj}, so this part of the likelihood disappears. This leads to the following expression,

$$\mathcal{L}^{RBNS0} = \mathcal{L}^{T} \cdot \mathcal{L}^{Q}. \tag{3.11}$$

The only type of claims that is left are the RBNP claims. For these claims, the only parameter that is observed is the reporting delay. But one can also be sure that the payment delay must exceed the time from the reporting time to the last observed time step in the data, which is to say that $q_{ik} > t_i^{*} - t_{ik}$. This probability can be computed with the help of a cumulative distribution function, G, as before. From this we have the final likelihood function as

$$\mathcal{L}^{RBNP} = \mathcal{L}^{T}.(1 - G_Q(t_i^* - t_{ik}; \Psi)). \tag{3.12}$$

The likelihood function for the whole data set is now determined by multiplying the likelihood for all claims with each other.

3.2.2 Parameter Estimation

The parameters in all the discrete distributions are estimated by maximum likelihood. The probability mass functions g_T, g_Q, g_U and g_N have the following form,

$$g_X(k) = \Pr(X = k) = \sum_{s=0}^{p} \nu_s I_s(k) + (1 - \sum_{s=0}^{p} \nu_s)\tilde{g}_{Y|Y>p}(k), \tag{3.13}$$

for all natural numbers k including zero. The function $\tilde{g}_{Y|Y>p}(k)$ is the truncated cumulative density function for a discrete distribution, such as Poisson or negative binomial. This is zero if $k \leq p$. The indicator function $I_s(k)$ equals one if $k = s$ and 0 otherwise. We can verify that the sum of the probability mass function $g_X(k)$ over all natural numbers k equals 1, as must be the case for a probability mass function. This is verified by

$$\textstyle\sum_{k=0}^{\infty} g_X(k) = \sum_{s=0}^{p} \nu_s + (1 - \sum_{s=0}^{p} \nu_s) \sum_{k=0}^{\infty} \tilde{g}_{Y|Y>p}(k) =$$

$$\textstyle\sum_{s=0}^{p} \nu_s + (1 - \sum_{s=0}^{p} \nu_s) \frac{\sum_{k=p+1}^{\infty} \tilde{g}_Y(k)}{\sum_{k=p+1}^{\infty} \tilde{g}_Y(k)} =$$

$$\textstyle\sum_{s=0}^{p} \nu_s + 1 - \sum_{s=0}^{p} \nu_s = 1.$$

The probability mass functions g_X in (3.13) is a weighed of the observed fractions of periods s smaller than or equal to p and a discrete probability distribution for periods larger than p. This can be motivated by the fact that many periods are small, and can work as estimates by themselves, but for larger lengths there are fewer observations and a distribution seems to be better. This also enable us to calculate the probability of observations larger than observed. The unknown parameters in this model are p, all the ν_s and the parameters in the discrete distribution. They are estimated by the following steps:

- First, a value of p is chosen.

- Then, it can easily be shown that the estimates of the ν_s that maximizes the likelihood are the fraction of observations in period s. This means that if 50 % of the claims are reported in the first period after occurrence, then $\nu_0 = 0.5$ for the reporting delay.

- Finally, the parameters in the discrete distribution are estimated. The only thing to remember in this step is to use the truncated distribution and not the original in the fitting process.

Now all steps can be repeated with another value of p and the model to choose can be evaluated with AIC or BIC as well as the mean squared error.

3.3 Simulation of Reserves Distribution

When all parameters in the distributions are estimated, the work left for receiving a reserves distribution is straight forward. For the RBNS claims we simply simulate the number of development periods, given that the number must be at least the numbers observed today. If the number of payments exceeds the observed, the remaining payments are generated from the observed claims with the same number of claims. Let us take an example. We have an observed RBNS claim with two payments. We simulate that the claim will have four payments in total, then we generate one claim from all observed closed claims with four payments. From this claim we take the two last payments. This procedure is repeated for all claims observed and then the RBNS reserve can be summed up. The same holds for RBNP claims except that for these claims, all payments from the generated claim will be used.

The simulation for the IBNR claims follows the same manner, except for the fact that the number of such claims is unknown. Therefore it has to be generated as well, which is done from a Poisson distribution with the expected value

$$\hat{\theta}_{\omega}(i)(1 - F_T(t^*; \hat{\boldsymbol{\nu}})), \tag{3.14}$$

for each occurrence period.

The three parts of the reserve can be summed into one observation from the reserves distribution. The simulation process must be done that many times that the given distribution becomes stable. If we only want to consider the future payments during a specific time line (the time line of the claims triangle for example), we must also simulate the time delays and stop simulations when the time exceeds this limit. If the number of payment periods in a simulation step becomes larger than 4, the payments after the fifth payment will equal the simulated fifth payment.

Chapter 4

Data Description

The data consists of information about individual events in the claims. An event could be the claim being reported, a payment being made to the insurance holder, or a claim being closed. The information available about the events is the development date, the date when the event occurred, and the paid amount. The accident date is known as well for every claim.

Two data sets are analyzed, one with typically shorter settlement times and one with typically longer settlement times, which we denote by short-tail and long-tail data, respectively. This is interesting since this is one of the most distinct partitions that can be made between different lines of business in insurance. The settlement time also have an important affect on the work of reserving, since a line of business with short settlement time will have less remaining amount to reserve than one with very long settlement times. Therefore, it is of interest to investigate if different methods work differently on a short-tail data set and a long-tail data set. One property that holds for both data sets is that the fraction of open claims with more than one payment is very low, and therefore it is believed that a small simplification of the likelihood function (3.10) will not disturb the comparison. The simplification mentioned is to use the likelihood function for the observed number of payment periods instead of using the cumulative distribution function. Now we move on to an individual description of each data set followed by an explanation of how the data has been manipulated.

4.1 Short-Tail Data

This is a data set with a good behavior. It consists of many observations and the characteristics of each claim, such as reporting delay and paid amounts, seem to have a normal variance. The mean reporting delay is 3 months and the mean settlement time is 8 months. Many of the observed settlements are equal to zero. This means that the settlement time follow a right skew distribution. A percentage distribution of how many payments the closed claims have can be seen in Figure 4.1. The evaluation date for this data set is March 2015.

Nr. of Payments	0	1	2	3	4	5	6+
Fraction [%]	23	54	15	5	1	1	1

Table 4.1: The table shows how large the fraction of the closed short-tail claims with a certain number of payments is.

4.2 Long-Tail Data

This is a data set with an extremely bad behavior. It consists of less observations than the data set above and the characteristics of each claim vary a lot. Some of the claims include extreme payments and it seems to be hard to predict the future for this data set. The mean reporting delay is 12 months and the mean settlement time is 15 months. The settlement time for this data set follow a right skew distribution, as the data set above. A percentage distribution of how many payments the closed claims have can be seen in Figure 4.2. The evaluation date for this data set is July 2015.

Nr. of Payments	0	1	2	3	4	5	6+
Fraction [%]	56	23	12	4	2	1	2

Table 4.2: The table shows how large the fraction of the closed long-tail claims with a certain number of payments is.

The analysis for this data set were first done as for the first one, resulting in very bad predictions of the reserve. Therefore the conclusion was made that the whole data set is not suited for being treated by the methods considered. Instead one part of the data set were picked out to be treated alone. The partition was natural since all claims belong to one of several classes in the line of business considered. The class that was chosen is the greatest class in this line of business. Still, the estimated reserve was far from the observed one. After some considerations, it could be concluded that this was due to a few claims with extremely large payments. These clams were omitted from the analysis with the argument that they must be treated individually by the help of the claims adjusters. The remaining part of the data set gave much better predictions.

4.3 Data Manipulation

Data from both data sets is available for about 30 years back in time. The judgment has been made that the earlier part of the data set is not very informative when predicting the reserve in present time. The behavior of the number of reported claims and the paid amounts differ for earlier dates and later dates. The claims that have been a part of the analysis has an accident date at January 1999 or later. To be able to perform out of sample prediction, the claims that were reported no later than December 2007 are used to estimate the reserve. It is assumed that all claims are closed at the evaluation date, which almost is the case. This means that the estimated reserve should be able to be compared to the observed reserve.

All analysis are made on the real data sets and presented for the actuaries at Länsförsäkringar. To be able to present any analysis here, the data must first be transformed so that the confidential information disappears. This is done by two steps. The first step is to transform all amounts of money by a chosen formula. It is possible that such a formula change some patterns in the data or make some claims affect the analysis more than they should have done otherwise, but the formula is chosen in order to make this impact small and in order to make the results here agree with the results on the real data. This should result in fair valuations of the methods considered. The second step in the transformation of the data is to use random chosen claims instead of using the whole data set. The number of claims analyzed is the same number of claims that is available in the data, but they are randomly chosen with replacement. All analysis are done in the Python language.

Chapter 5

Model Selection

In order to make the result chapters more concise and clear, the model selection have an own chapter. The methods considered consist of many variations, especially the micro-level method where you can choose between many statistical distributions. In this thesis, the only model choice that is made is the choice between four discrete distributions in the micro-level method. These distributions are used to model the reporting delay as well as the number of payment periods. The four distributions considered are Poisson, negative binomial, binomial and geometric, with different number of degenerate components.

The choice between several statistical models is often made by AIC and BIC statistics. These are two values that takes into account how likely the observed outcome is under the model considered, but it also takes the number of parameters into account. This means that the model with the most parameters not necessarily will be the best one, which saves us from choosing a model with too many parameters. Let the number of observations be n_o and the number of parameters in a model be n_p. With the log likelihood denoted by ℓ, the two statistics are calculated as

$$AIC = 2n_p - 2\ell \tag{5.1}$$

and

$$BIC = n_p \cdot \log(n_o) - 2\ell. \tag{5.2}$$

A model with lower AIC or BIC is better than one with larger. Besides the goal of maximizing the likelihood, the goodness of fit must be considered. The AIC and BIC can suggest a model that fits the data very badly. The goodness of fit can be illustrated by plotting the fitted distribution together with the observed values, as well as calculating the Mean Squared Error (MSE). Here, only the MSE will be presented due to secrecy. Say we have the vector u of the observed cumulative distribution function as well as the vector v of the cumulative distribution function of the fitted model. Then the MSE can be calculated as

$$MSE = \frac{1}{n_o} \sum_{i=0}^{n_m} (u_i - v_i)^2, \tag{5.3}$$

with n_m being the largest observed value in the data set. The MSE should, of coarse, be as small as possible.

5.1 Short-Tail Data

For the short-tail data it is assumed that all claims will be paid during the time limit of the payment triangles, or more specifically, as stated in the data description chapter, it is assumed that all claims are paid before evaluation. This means that we do not need to simulate the delay between payments in order to see which payments that fall inside the same time limit as in the macro-level methods. This would otherwise be necessary to be able to compare the different reserves estimates. This also enable us to compare the reserves estimates with the observed reserve. Now we only need to know the distribution of the reporting delay to be able to estimate the number of IBNR claims, as well as the distribution for numbers of payment periods.

The models are then compared with respect to the AIC, BIC and MSE. The models with the best values of the statistics are presented in Table 5.1. It seems like the Poisson distribution is a bad choice for the reporting delay, at least according to the AIC and BIC. The best distribution according to these statistics is negative binomial closely followed by the binomial. The MSE says something else, namely that the geometric distribution with 1 degenerate component is far better than the other models for the reporting delay. The reason for the difference is that AIC and BIC only says how likely the observed values are under the model considered, but not how good the fit is. In our case, we have a very skewed distribution of observed values, most of them are small but also very large values exists. The AIC and BIC prefers a model that gives a large likelihood for the smaller values, even if this gives a model that does not fit to the larger observations. The model choice will therefore fall on the distribution suggested by MSE, geometric, with 1 degenerate component.

For the number of payment periods, see Table 5.2, negative binomial seems to be a bad choice. The other three gives values of AIC and BIC that are more closely to each other, but the best value is determined from the binomial distribution. As before, the MSE says something else. The geometric distribution gives the smallest MSE for the number of payment periods. Therefore, the number of payment periods will be fitted to a geometric distribution with 1 degenerate component.

p	Distribution	AIC	BIC	MSE
25	Poisson	68,956	69,155	0.00062
29	Poisson	68,966	69,196	0.00061
2	NegBin	60,477	60,508	0.00587
3	NegBin	59,027	59,065	0.00604
0	Binomial	60,359	60,374	0.00176
1	Geometric	66,466	66,481	0.00009

Table 5.1: The AIC, BIC and MSE statistics for the short-tail data reporting delay. The number of degenerate components is p.

p	Distribution	AIC	BIC	MSE
0	Poisson	17,240	17,247	0.00061
1	Poisson	7,976	7,992	0.00091
0	NegBin	16,337	16,352	0.00236
0	Binomial	16,557	16,572	0.00104
1	Binomial	6,679	6,702	0.02099
1	Geometric	7,888	7,904	0.00017

Table 5.2: The AIC, BIC and MSE statistics for the short-tail data number of payment periods. The number of degenerate components is p.

5.2 Long-Tail Data

As for the short-tail data, we assume that all claims are paid at the end of the payment triangle as well as at the evaluation date. The negative binomial distribution has been omitted in the comparison for the long-tail reporting delay. This is due to the lack of parameter estimates, the distribution does not seem to fit the data. Among the other distributions showed in Table 5.3, the binomial gives too large values of all statistics considered. The remaining two distributions are more close to each other and as we saw in the previous section, the AIC and BIC contradict the MSE. The differences between AIC and BIC are not as large as the difference between the MSE, which leads to the decision to choose the model with the lowest MSE. Therefore, the geometric distribution with one degenerate component is chosen.

The statistics for the number of payment periods in the long-tail data follow a similar pattern as for the other model choices. Negative binomial has too large values of AIC and BIC and is not used, despite that it leads to a satisfying value of MSE. It leads to, together with the geometric distribution, the best value of MSE. The binomial distribution that with one degenerate component gives low values of AIC and BIC results in the worst value of MSE. Since the AIC and BIC for the geometric is the second lowest, and due to the fact that it leads to a good MSE value, we will again choose the model geometric distribution with one degenerate component.

p	Distribution	AIC	BIC	MSE
24	Poisson	10,127	10,262	0.00058
27	Poisson	10,131	10,281	0.00044
0	Binomial	21,113	21,123	0.00178
1	Binomial	18,465	18,481	0.00221
1	Geometric	10,746	10,757	0.00002

Table 5.3: The AIC, BIC and MSE statistics for the long-tail data reporting delay. The number of degenerate components is p.

p	Distribution	AIC	BIC	MSE
1	Poisson	876	886	0.00176
0	NegBin	1,689	1,700	0.00030
0	Binomial	1,982	1,993	0.00375
1	Binomial	776	792	0.05887
1	Geometric	869	880	0.00030

Table 5.4: The AIC, BIC and MSE statistics for the long-tail data number of payment periods. The number of degenerate components is p.

Chapter 6

Results

In this chapter, the results for the short-tail data are presented first. It begins with an investigation of how large sample size that is needed followed by the simulated reserve for both of the macro-level methods. The result for the micro-level method is presented in the same manner. Lastly, all point estimates are compared to the observed reserve. Then the results for the long-tail data are presented in the same order.

6.1 Macro-Level for Short-Tail Data

In order to make a decision on how large the sample size must be to reach a stable simulated distribution, the distribution was simulated five times for different sample sizes. The sample sizes considered are 100, 1,000 and 10,000. The resulting mean, median, 0.05 quantile and 0.95 quantile are plotted as functions of the sample size for the CLM, see Figure 6.1. The mean is probably the statistic we are most interested in when estimating a reserve, but it is interesting to see how much the distribution changes in the tails as well. The mean still varies when using a sample size of 10,000 and an even larger sample size would be beneficial. The reason for not increasing the sample size even more in this thesis is due to the fact that the simulations would take too much time and the judgment is made that this sample size is large enough to be able to compare the models considered. As the reader will see, the models leads to quite spread estimates and the spread in the statistics for the 10,000-sample distribution is small enough to not trouble the comparison.

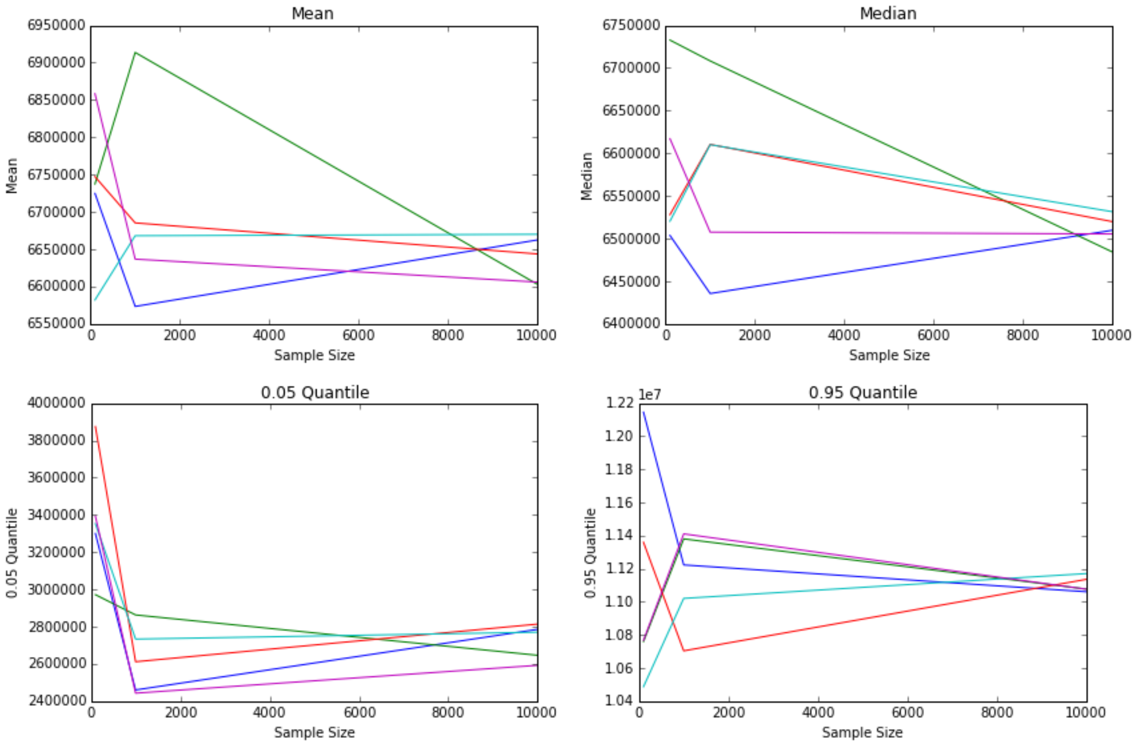

Figure 6.1: The figure shows the mean, median, 0.05 and 0.95 quantile of the reserve distribution received by the CLM. The sample sizes considered are 100, 1,000 and 10,000 and the reserve distribution is simulated 5 times for each sample size.

The sample size for the DCL distribution is evaluated in the same way, see Figure 6.2. Compare this figure with Figure 6.1, and you can see that the plots looks very similar. The reason for this is that the CLM and DCL distribution was simulated at the same time, which means that the same payment triangle was used in every simulation step for both models. This makes us suspect that the DCL is influenced more by the payment triangle than the counts triangle. This could be the explanation for DCL resulting in estimates close to the CLM, both here and in the literature. As decided for the CLM distribution, we hence also here choose to simulate the reserve with a sample size of 10,000.

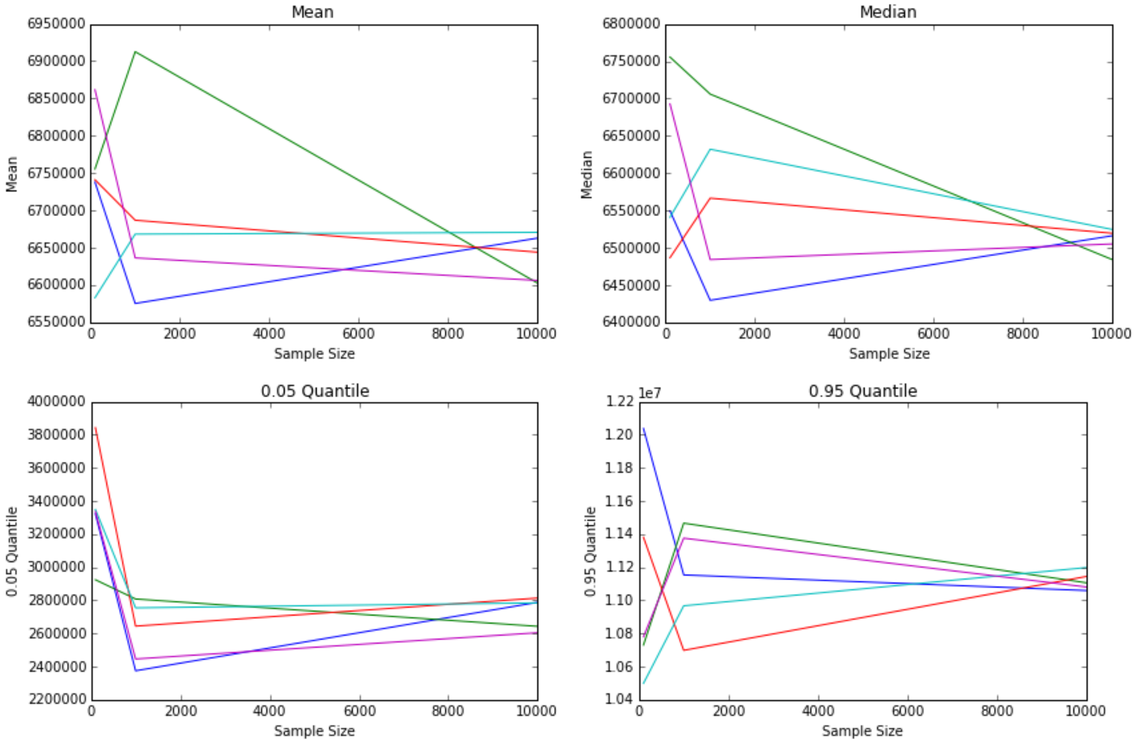

Figure 6.2: The figure shows the mean, median, 0.05 and 0.95 quantile of the reserve distribution received by the DCL. The sample sizes considered are 100, 1,000 and 10,000 and the reserve distribution is simulated 5 times for each sample size.

The reserve distribution received by CLM is shown in Figure 6.3, where the sample size used is 10,000. The distribution looks almost symmetric, but with the tendency of being right skewed. The distribution seems to have a large variance since the values of possible future reserves goes from negative values up to almost 20 millions in the figure. But it still seems that most of the simulations leads to a reserve estimate a bit under 10 millions. The DCL reserve distribution that is shown in Figure 6.4 looks the same as the CLM distribution and does not need any further comments.

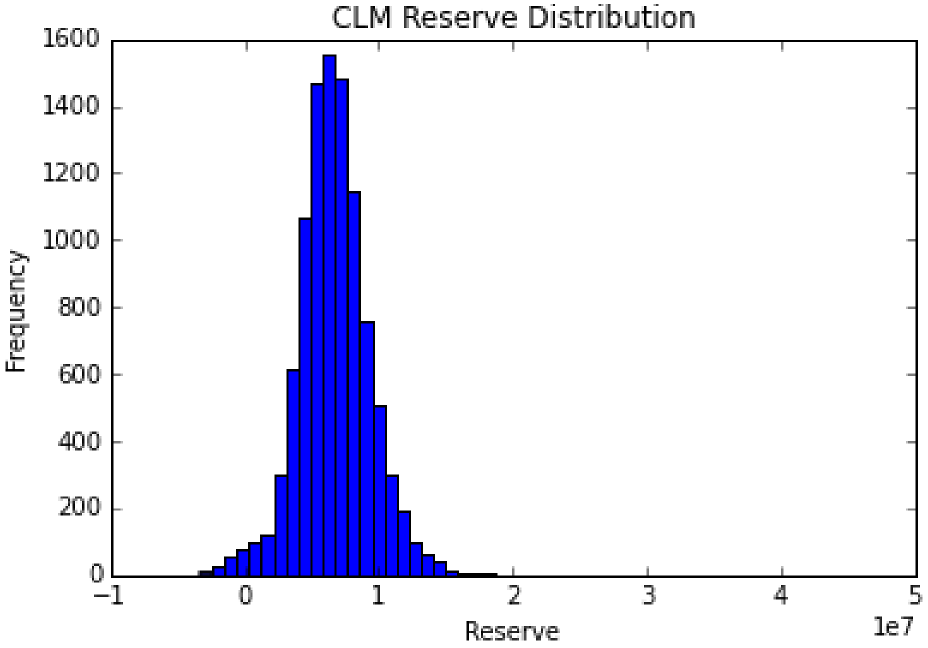

Figure 6.3: The figure shows the reserve distribution received by the CLM with a sample size of 10,000.

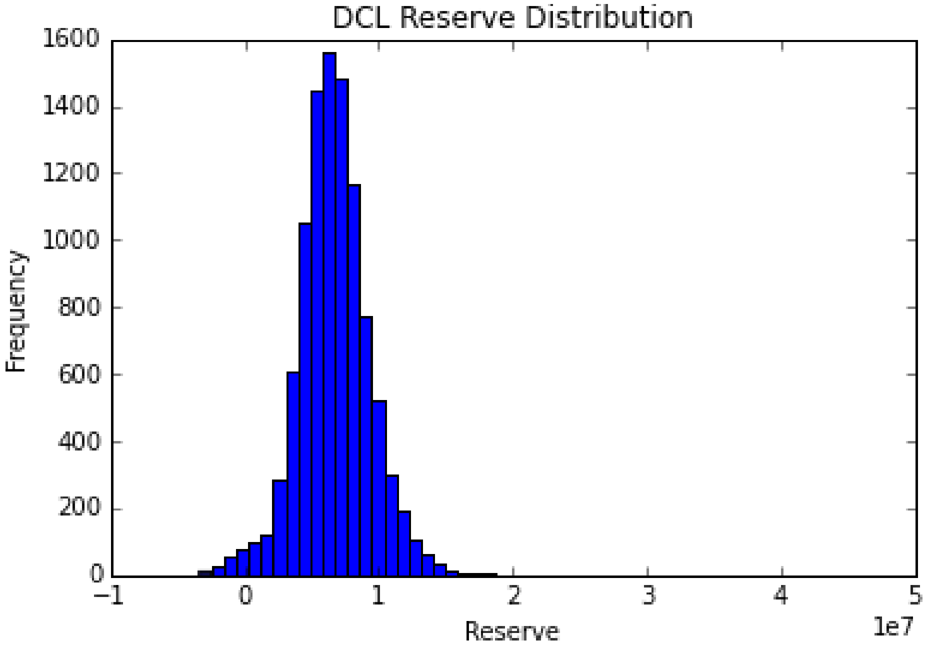

Figure 6.4: The figure shows the reserve distribution received by the DCL with a sample size of 10,000.

6.2 Micro-Level for Short-Tail Data

In contrary to the macro-models, a sample size of 10,000 seems to be satisfying for the micro-model, see Figure 6.6. The mean converges already for a sample size of 1,000. The quantiles variations are acceptable at a sample size of 10,000. The plot of the median does not look as smooth as the other three, but this is simply caused by the fact that the y-axis for the median covers a thinner interval. As for the macro-models, a sample size of 10,000 will be used in the simulations.

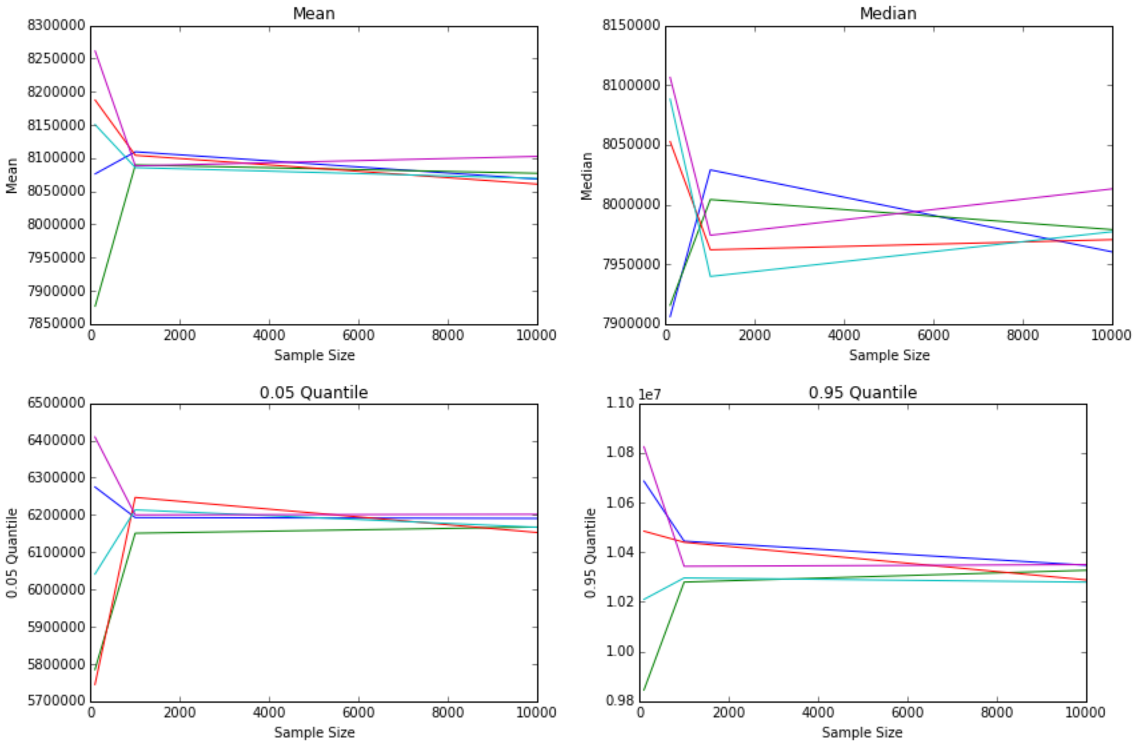

Figure 6.5: The figure shows the mean, median, 0.05 and 0.95 quantile of the reserve distribution received by the micro-model. The sample sizes considered are 100, 1,000 and 10,000 and the reserve distribution is simulated 5 times for each sample size.

The reserve distribution for the micro-model looks a bit right skewed, as for the macro-models. Another similarity is that many of the values is placed slightly under 10 millions. A difference is that this distribution has a lower variance, it does not take negative values and stays under 16 millions.

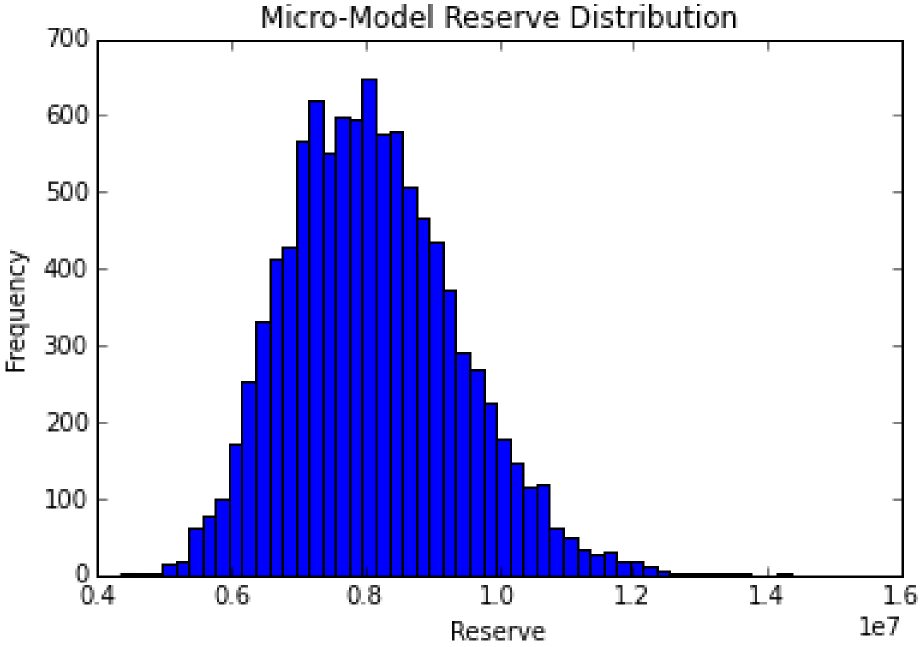

Figure 6.6: The figure shows the reserve distribution received by the Micro with a sample size of 10,000.

6.3 Observed Reserve for Short-Tail Data

The point estimates of the reserve for the three models are presented and compared to the observed reserve in Table 6.1. The micro-model does not produce a point estimate by itself, so we need do calculate it from the mean of the simulated distribution. We have simulated five distributions of size 10,000, if we calculate the mean of all these estimates together we have the mean of a distribution with sample size 50,000. The estimate received by the micro-model is far away from the real reserve. Certainly, it lies within the same tenfold as the real reserve, but the other two models perform much better. The CLM produce an estimate a little higher than the DCL, which is a pattern that also has been seen in the literature. Despite this returning pattern, this could perfectly well be caused by the data. The DCL that is the model that gives the point estimate closest to the real reserve, have an estimation error in this case of about 3.4 percents.

Observed	6,192,082
CLM	6,555,806
DCL	6,401,330
Micro	8,075,538

Table 6.1: The table shows the observed reserve and the estimated reserve from the three models. The point estimate for the micro-model is the mean from the five simulations with sample size 10,000.

6.4 Macro-Level for Long-Tail Data

With respect to how varied the long-tail data set is, the statistics in Figure 6.9 converges very well. It is the 0.95 quantile that converges poor which one can see by checking the scale on the y-axis. This is not surprisingly, since it is here the most extreme values of the distribution lies. Since we, as stated in the previous chapter, believe that the mean is the most important statistics to consider, we make the decision that 10,000 as a sample size is sufficient.

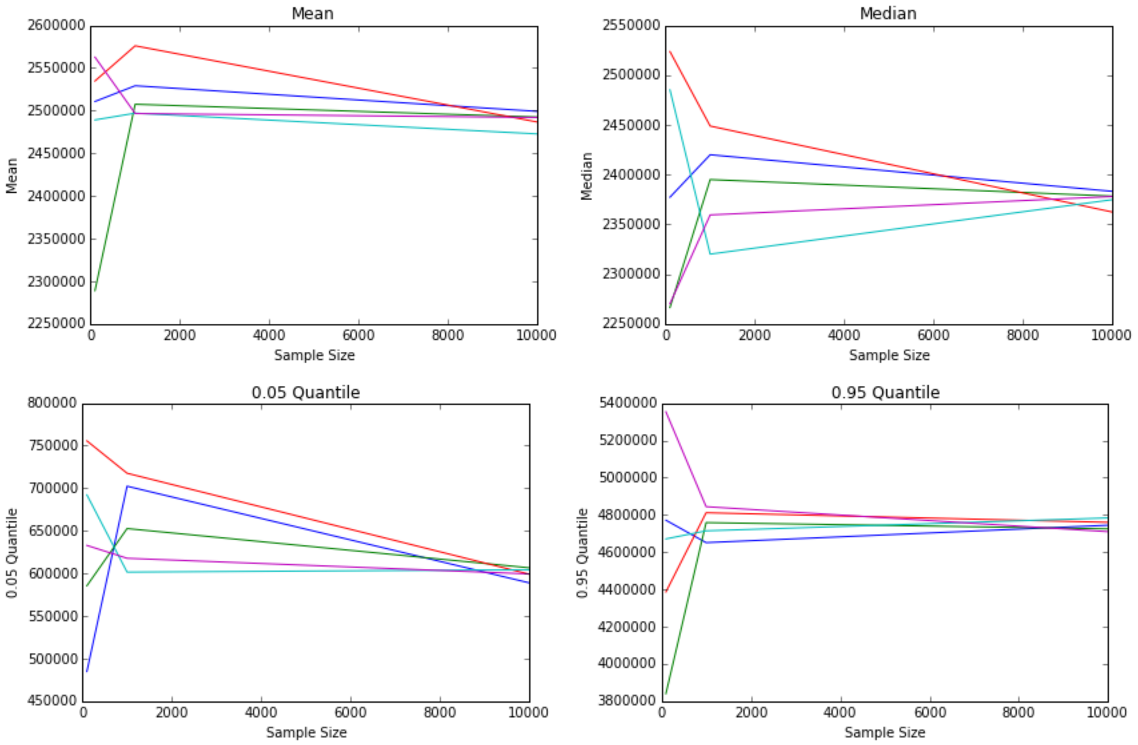

Figure 6.7: The figure shows the mean, median, 0.05 and 0.95 quantile of the reserve distribution received by the CLM. The sample sizes considered are 100, 1,000 and 10,000 and the reserve distribution is simulated 5 times for each sample size.

The same statistics for the DCL are shown in Figure 6.10, and as in the previous chapter it looks similar to the CLM plots. This is due to the fact that the same payment triangle is used for both methods in the simulation process, as stated before. The mean here converges a little better, but is very similar to the CLM mean. The opposite hold for the median, the convergence for the median was better for the CLM. The 0.05 quantile converges to a lower value this time and the 0.95 quantile seem to converge to a higher value this time. As for the CLM, a sample size of 10,000 will be used.

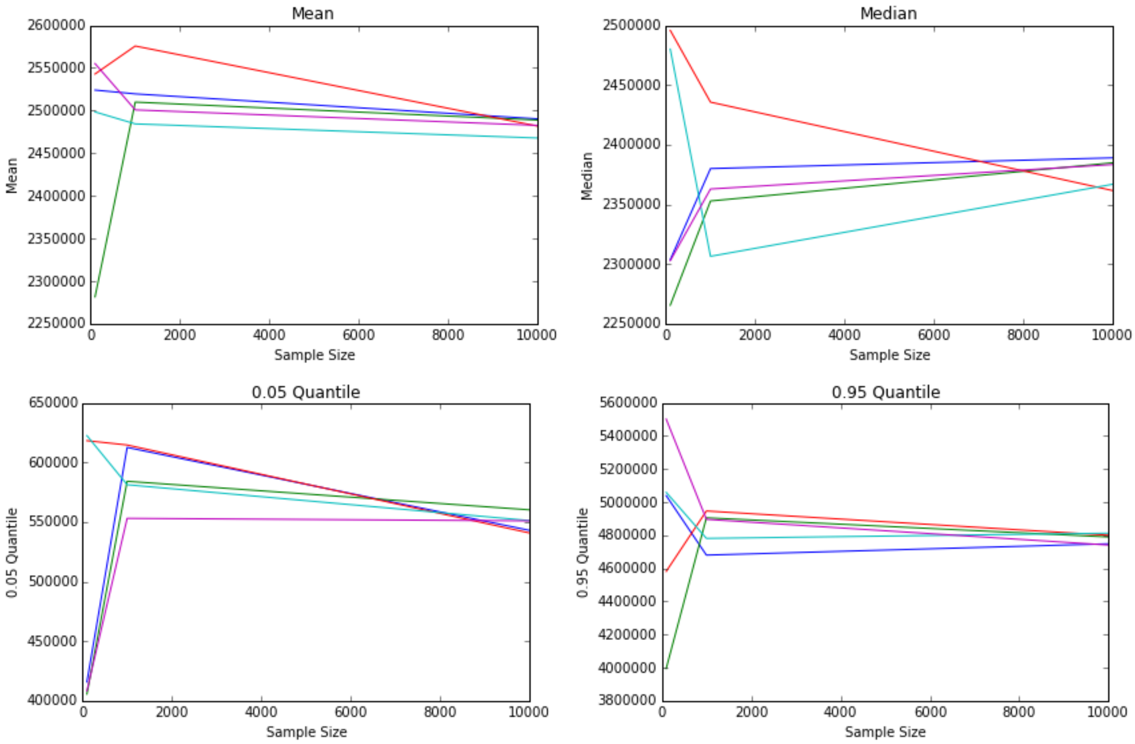

Figure 6.8: The figure shows the mean, median, 0.05 and 0.95 quantile of the reserve distribution received by the DCL. The sample sizes considered are 100, 1,000 and 10,000 and the reserve distribution is simulated 5 times for each sample size.

Despite some small differences between the statistics for CLM and DCL, their distributions looks the same, see Figure 6.9 and 6.10. It has a symmetric look with a tendency of right skewed behavior. many values are centered at minus 5 millions up to 15 millions, which is a large spread. The distribution has in reality an even larger variance since the most extreme values had to be removed to get a clear graph.

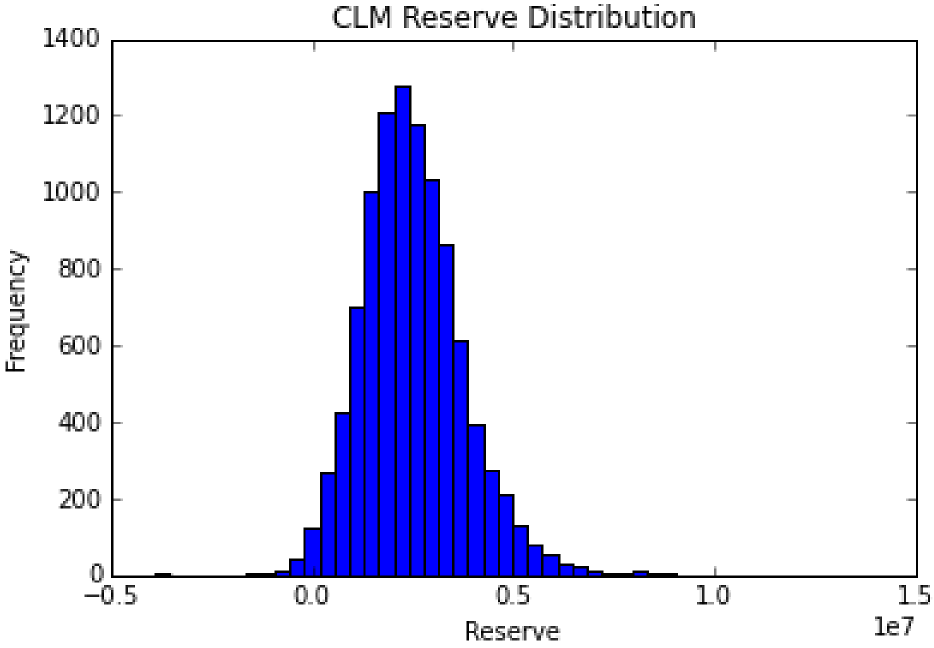

Figure 6.9: The figure shows the reserve distribution received by the CLM with a sample size of 10,000.

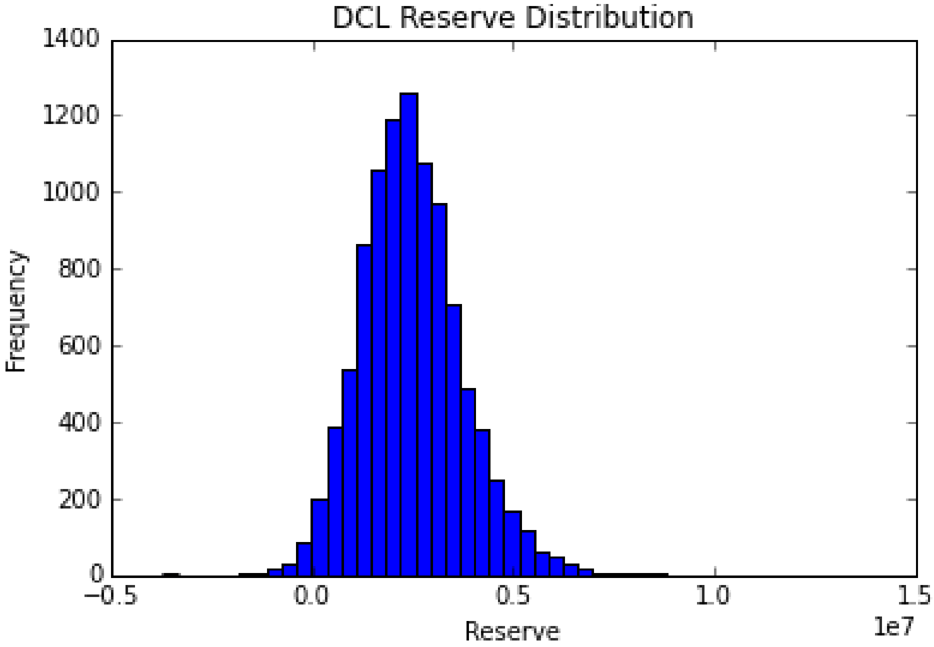

Figure 6.10: The figure shows the reserve distribution received by the DCL with a sample size of 10,000.

6.5 Micro-Level for Long-Tail Data

The mean, median, 0.05 and 0.95 quantile for the micro-model are shown in Figure 6.11. The convergence of the mean is acceptable already for a sample size of 1,000. The median has a similar behavior, but looks even better for a sample size of 10,000. The same holds for the quantiles, you can see an improvement for their convergence from a sample size of 1,000 to 10,000. As before, we choose to simulate with a sample size of 10,000 since it seems like this sample size will enable a comparison while the simulations does not take to much time.

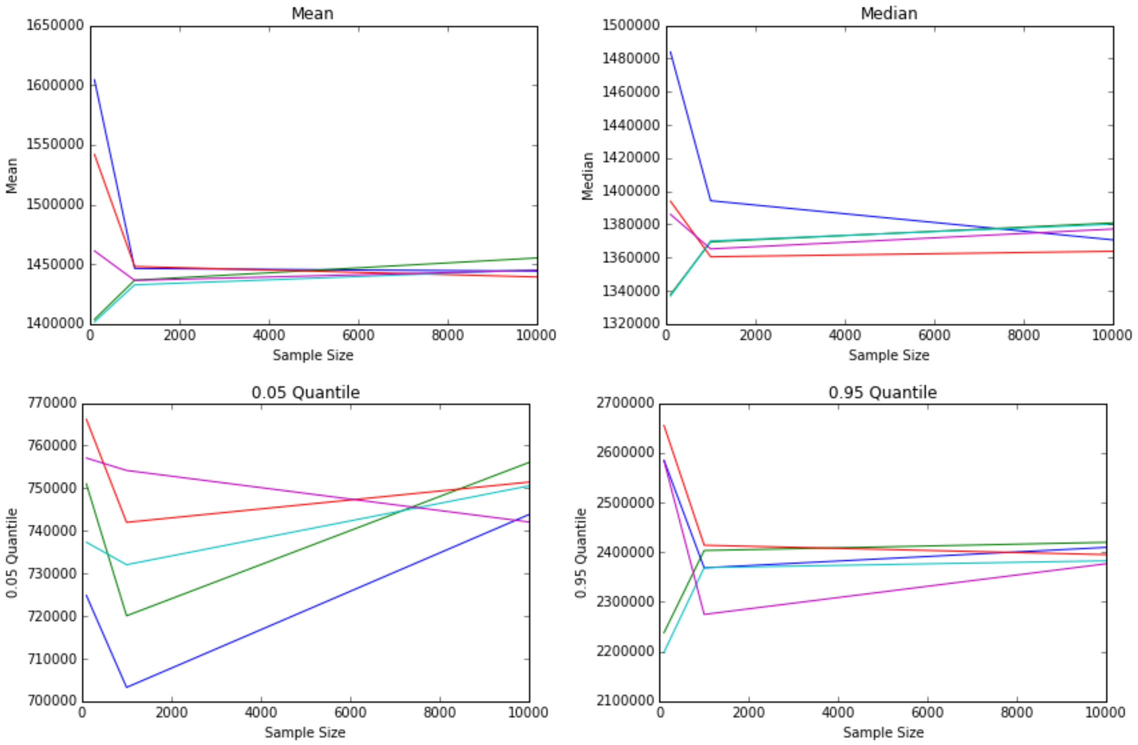

Figure 6.11: The figure shows the mean, median, 0.05 and 0.95 quantile of the reserve distribution received by the micro-model. The sample sizes considered are 100, 1,000 and 10,000 and the reserve distribution is simulated 5 times for each sample size.

The reserve distribution in Figure 6.12 shows a clear right skewed behavior. The values that are covered are all positive and stays under 5 millions. most of the values are centered between 1 million and 2 millions.

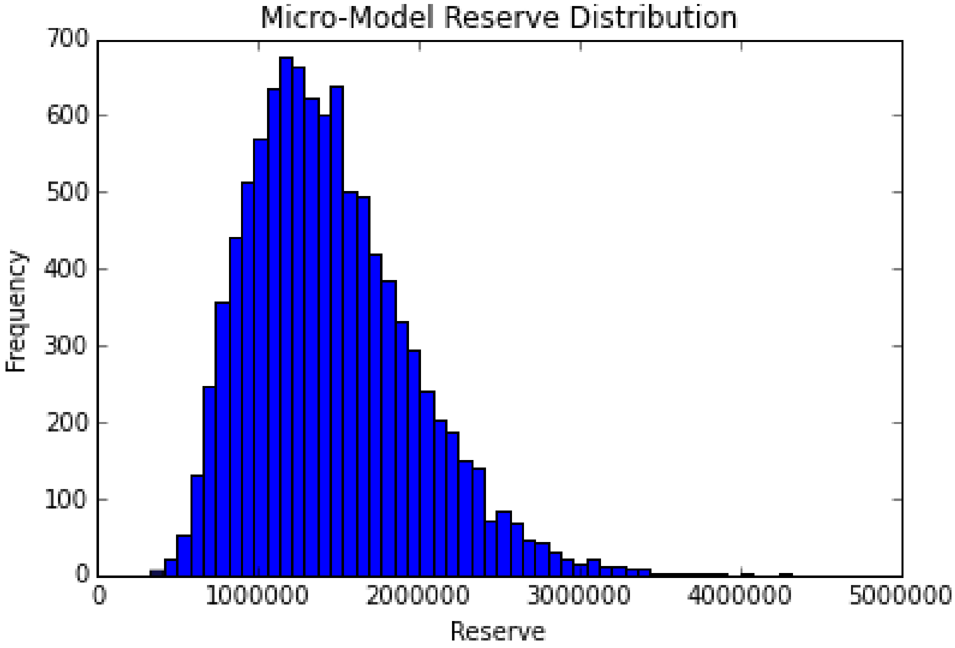

Figure 6.12: The figure shows the reserve distribution received by the micro-model with a sample size of 10,000.

6.6 Observed Reserve for Long-Tail Data

The point estimates for all models, as well as the observed reserve, are shown in Table 6.2. The point estimate of the micro-model is calculated as in the previous chapter. This time the micro-model gives an estimate that is lower than the real reserve. It is still not the best model. The pattern of CLM producing higher estimates remains, and also here the DCL is the model which gives the point estimates that is closest to the real reserve. The estimation error of DCL is about 8.9 percents, which is higher than the error for the best model of the short-tail data. This feature is expected since the long-tail data include a more varying behavior.

Observed	1,796,845
CLM	2,305,610
DCL	1,957,216
Micro	1,445,736

Table 6.2: The table shows the observed reserve and the estimated reserve from the three models. The point estimate for the micro-model is the mean from the five simulations with sample size 10,000.

Chapter 7

Conclusions

Before drawing any conclusions, be aware of the fact that we have assumed that all claims are closed at evaluation date and at the end of the payment triangle. Even if it seems to be a reasonable assumption, it could cause disturbance in the results. With this said and with the results presented at hand, we can say a few things about how the methods preformed on the two data sets.

- The micro-level method does not estimate a better reserve than the two other methods for any of the two data sets.
- The micro-level method has a lower variance than the two other methods.
- The DCL point estimates of the reserve are lower than the estimates from CLM for both data sets.
- The DCL estimates were closest to the observed reserve for both data sets.
- The variance of the CLM reserve and DCL reserve are similar to each other and larger than for the micro-level method.
- The observed reserve lies inside of the limits of the simulated distribution for all methods and both data sets.

One of the questions that this thesis was meant to investigate was if you gain something from using a micro-level method. From the results received, the answer to this question is no. But remember these results only hold for this particular micro-level method and these two data sets. We cannot say anything about other micro-level methods. The problem with the micro-level method used is that we assume that likely future payments have been observed in the past. This means that we confine ourselves that all possible future payments must be one that has already been seen. This is not very intuitive, the payments can vary a lot and the variance of the reserves distribution can be much larger than what has been observed in the past. An interesting future research topic is to find a distribution that can replace the historical simulation used here.

Let us now discuss the idea of using a micro-level method in more general sense, not focusing on any particular method. The micro-level methods use very much information in the data, such as number of reported claims and singular payments. For a micro-level method to be useful, we must be able to say that all this information observed in the past follow a pattern that will remain in the future. If suddenly the reporting delays increase, the micro-level method will fail. And what happens if there has been a large fire, a sudden depression or if there has been large changes in the line of business considered? The conclusion is that one must know the data well and consider the model choice carefully before expecting to gain anything from a micro-level method.

The DCL however, seems to be beneficial compared to the ordinary CLM. Still remember that this only holds for the data set at hand and not necessarily hold in general. Another benefit from using the DCL is that one can calculate the reserve that lies outside the ordinary payment triangle and that one can separate the calculated reserve in one part for the already reported claims and one part for the not yet reported claims. The conclusion when it comes to the DCL is that since it seems to have many advantages and is a very simple method to implement, it could be a good idea to use it.

The other question that this thesis is studying is if there is any differences in which model to use depending on the settlement time of the data set. This question has no clear answer. It seems like the variability in the data set makes it harder to forecast the reserve in general and we cannot say that any method is better on this. We can remind ourselves how problematic the long-tail data set was in the beginning before dividing it into parts and removing the most extreme claims. The answer to the question must therefore be that it is probably more important to choose how to divide the data and which claims to treat separately, than which method you are using. This is an important topic for future research.

This thesis has contributed with a comparison of three claims reserving methods on two very different data sets. This has illustrated how different the features of data sets can be and how different estimation results the methods can come up with depending on the features of the data set. We can create a lot of advanced methods, but nothing compares to having a good knowledge and understanding for the underlying data set. Without any knowledge of the data set, it is impossible to draw any conclusions about which method to use.

Bibliography

Antonio, K. and Plat, R. (2014). Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal*, 2014(7):649–669.

Arjas, E. (1989). The Claims Reserving Problem in Non-Life Insurance. *ASTIN Bulletin*, 2:139–152.

England, P. (2001). Addendum to analytic and bootstrap estimates of prediction errors in claims reserving. *Insurance: Mathematics and Economics*.

Haastrup, S. and Arjas, E. (1996). Claims reserving in continuous time: a nonparametric Bayesian approach. *ASTIN Bulletin*, 2:139–164.

Hess, K. T. and Schmidt, K. D. (2002). A comparison of models for the chain-ladder method. *Insurance: Mathematics and Economics*, 31(3):351 – 364.

Larsen, R. C. (2007). An individual claims reserving model. *ASTIN Bulletin*, 1:113132.

Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin*, 23(2):213–225.

Mack, T. (1994). Which stochastic model is underlying the chain ladder method? *Insurance: Mathematics and Economics*, 15:133 – 138.

Martnez-Miranda, M. D., Nielsen, J. P., and Verrall, R. (2013). Double chain ladder and bornhuetter-ferguson. *North American Actuarial Journal*, 17(2):101–113.

Miranda, M. D. M., Nielsen, J. P., and Verall, R. (2012). Double Chain Ladder. *ASTIN Bulletin*.

Miranda, M. D. M., Nielsen, J. P., Verrall, R., and Wthrich, M. V. (2015). Double chain ladder, claims development inflation and zero-claims. *Scandinavian Actuarial Journal*, 2015(5):383–405.

Norberg, R. (1993). Prediction of outstanding liabilities in non-life insurance. *ASTIN Bulletin*, 1:95115.

Norberg, R. (1999). Prediction of outstanding liabilities II. Model extensions variations and extensions. *ASTIN Bulletin*, 1:525.

Pigeon, M., Antonio, K., and Denuit, M. (2013). Individual Loss Reserving with the Multivariate Skew Normal Framework FRAMEWORK. *ASTIN Bulletin*, 3:399–

428.

Pigeon, M., Antonio, K., and Denuit, M. (2014). Individual loss reserving using paid incurred data. *Insurance: Mathematics and Economics*, 58(0):121 – 131.

Reid, D. (1978). Claim reserves in general insurance. *Journal of the Institute of Actuaries*, 105:211–296.

Taylor, G. and McGuire, G. (2004). Loss reserving with glms : a case study. *Meeting of the Casualty Actuarial Society, Colorado Springs.*

Wright, T. S. (1997). Probability distribution of outstanding liability from individual payments data. *Claims Reserving Manual*, 2. Institute of Actuaries, London.

Wthrich, M. V., Merz, M., and Lysenko, N. (2009). Uncertainty of the claims development result in the chain ladder method. *Scandinavian Actuarial Journal*, 2009(1):63–84.

Wtrich, M. V. and Merz, M. (2008). *Stochastic claims reserving methods in insurance.* Wiley Finance.

Zhao, X. and Zhou, X. (2010). Applying copula models to individual claim loss reserving methods. *Insurance: Mathematics and Economics*, 46(2):290 – 299.

Zhao, X. B., Zhou, X., and Wang, J. L. (2009). Semiparametric model for prediction of individual claim loss reserving. *Insurance: Mathematics and Economics*, 45(1):1 – 8.