UiO : Department of Mathematics
University of Oslo

Contributions to RBNS Modelling

Vaamanan Murugendran

Master's Thesis, Spring 2016

Cover design by Martin Helsø

The front page depicts a section of the root system of the exceptional Lie group E_8, projected into the plane. Lie groups were invented by the Norwegian mathematician Sophus Lie (1842–1899) to express symmetries in differential equations and today they play a central role in various parts of mathematics.

Abstract

Reserving against future claims and settlements is vital for insurance companies, in the sense that it influences how they may price their products and the solvency of the insurance company. This thesis will present and compare two models that can estimate the outstanding liabilities: The well known and well used Chain Ladder model which uses aggregated data and the Kaminsky approach that divides the problem into modelling counts and losses separately and regard the delay in IBNR and RBNS as multinomial phenomenon governed by delay probabilities. Mean square error will be used to compare these methods. The thesis will provide a theoretical basis for each method and an analysis when implemented on fire and car insurance data provided by a Norwegian non-life insurance company. A large portfolio approximation will be done analytically, which will confirm with the observation done in the numerical study that for large portfolios it will be more accurate to model claim counts and sizes separately than using aggregates to estimate the outstanding liabilities. The more heavy-tailed the claim size distribution is, the more superior will the Kaminsky approach be.

Key words and phrases: Chain Ladder, Kaminsky approach, estimation error, large portfolio approximation, multinomial distribution, IBNR, RBNS, Bootstrapping, Monte Carlo

Acknowledgements

This thesis is written for the Master's degree in Modelling and Data Analysis (Modellering og Dataanalyse) with the programme option: Finance, Insurance and Risk (Finans, Forsikring og Risiko). The thesis corresponds to 60 credits and was written in the period between January of 2015 and May 2016. The topic of the thesis is the accuracy in the Chain Ladder method and the Kaminsky approach when estimating the outstanding liabilites.

I want to thank my thesis supervisor Erik Bølviken. We have had some very interesting and educative conversations about the actuary community concerning this topic and others. His style of thesis guidance has given me the motivation needed to solve the problem of this thesis. I also want to thank Nils Haavardson for informative conversations and information about the current practices within insurance companies regarding reserving.

I also want to thank Emil Aas Stoltenberg for reading through my thesis and pointing out possible improvements. This was much appreciated. Last but not least, I want to thank Lise-Marie Lia for the support throughout the writing of this thesis and for proof reading several times. I will gladly extend the same support and help to her when it is her turn.

Vaamanan Murugendran
May 2016, University of Oslo

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Non-life insurance

In this thesis we are going to consider the claims reserving problem for a branch of insurance products known as non-life insurance. In the UK, non-life insurance branch is known as General Insurance and in the USA as Property and Causality Insurance. In Europe it is mainly known as non-life insurance, which is the term that will be used in this thesis. Non-life insurance contain all kinds of insurance products except life insurance. The reason for separating them is that life insurance products are somewhat different from non-life insurance contracts. The differences can be seen in the type of claims, risk drivers, terms of contracts, etc. As a consequence, life and non-life insurance products are modelled quite differently.

The non-life insurance branch operates in the following line of business (Wütherich and Merz, 2008):

- Motor/car insurance, for example: third party liability.
- Property insurance, for example: against fire, water, flooding and etc.
- Liability insurance, for example: private and commercial liability.
- Accident insurance, for example: personal, compensation for workers.
- Health insurance, for example: personal.
- Travel insurance.
- Credit insurance.
- Other insurances such as aviation, marine, legal protection, etc.

We have been fortunate enough to be given insurance data from a Norwegian non-life insurance company[a]. The dataset contains car and fire insurance data which can be found in Appendix B. The fire insurance data only contains fire damages on villas while the car insurance data contains data on personal injuries caused in car related accidents. These datasets are both RBNS datasets.

1.2 Course of events

Every day there are hundreds of accidents and the chances are that most of us will, at one point or another, experience an incident where an insurance company is involved. We will now present the course of events that are set in motion when an incident happens and an insurance company has to get involved. A typical timeline of events can be seen in Figure 1.1.

First and foremost the policyholder have to be insured against that certain kind of accident, if not the policyholder has to pay the full amount. An insurance contract that specifies what or who is insured, and what it or they are insured against, has to be signed. This provides the insurer with a fixed amount of money, called premiums, and the insured with a financial coverage against random well-specified events. This insurance contract also has to mention in what time period the contract is valid, which depends on what kind of insurance is signed. Property insurance is usually valid for a time period of one year, while life insurance last longer. In property insurance property is usually insured for example: villas, cabins, houses, cars, boats and pets. In life insurance people get insured against for example: death, disability, etc. The right of the insured to collect these amounts, in case the event happens, creates a claim by the insured to the insurer. The amount which the insurer it obligated to pay in a case of a claim is known as the claim amount or the loss amount. The policyholder is not always the one who is insured. It could be that a mother insures her family, and in that case the mother is the policyholder while her family is insured and the insurance company is the insurer.

The reserving problem that arises because the delay between the accident date and the reporting date is know as the IBNR problem, “Incurred, But Not Reported”. The reserving problem that arises because of delay between the reporting date and the claims closing date is called the RBNS problem, “Reported, But Not Settled”. We will go more in-depth in both cases below.

1.2.1 IBNR

If an accident happens it will have to be reported to the insurance company including the date of the accident. With this information the insurance company has to decide if the accident can be linked up to a policy the policyholder is holding. It is important that the policyholder had a valid insurance contract for that specific accident at the time of the accident. The policyholder is not always certain when the accident

[a] The name of the Norwegian non-life insurance company will not be specified as they wish to remain anonymous

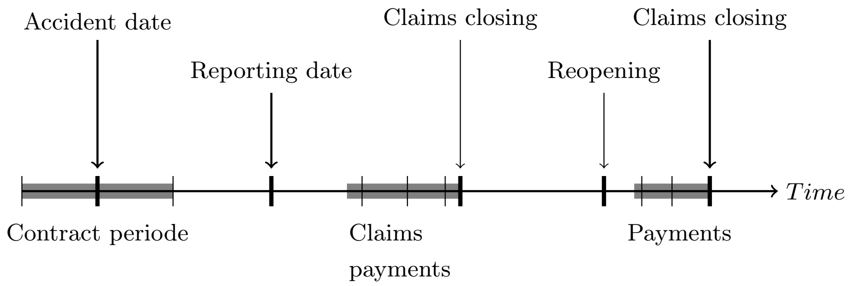

Figure 1.1: *Typical timeline of a non-life insurance claim*

happened. For example if some damage happened to ones cabin, it could take a while before it is noticed. Another example is if a water leak went unnoticed, and later caused mold damage in the house. In such cases the insurance company have to call in a expert to estimate the date of the accident. The insurance company that insured the house during the period when the accident happened will have to cover the damages. These delays are not uncommon, but rather a big part of the daily routine of an insurance company. Delays can vary from hours or days to months or years. One of the reasons for such delays could be as mentioned above, water leaks or damage on ones cabin. Several other examples of IBNR claims are listed below:

- An accident could not be reported right away because it happened during a holiday.
- An accident happened and the policyholder was hospitalized and thus could not report the accident right away.
- A slowly developing occupational disease that was not discovered until several years later.
- A doctor being sued for malpractice because of an operation he or she preformed several years ago.

Each year accidents are reported to the insurance company with a delay, as mentioned above. In all of Europe, insurance companies are obligated to put aside an amount of money to pay all claims for accidents happening during a year. In other words insurance companies have to reserve money for claims they do not know anything about, and that can occur in the future. This is called IBNR reserving.

1.2.2 RBNR

When an accident is reported the insurance company will try to figure out if the accident is something their policy will cover or not. All accidents or claims the insurance company are actively working on are referred to as “open”. When it seems like there will be no more payouts, the case is referred to as “closed”. When small accidents happen, like baggage getting lost in transit, the payouts are minor and the case is closed relatively fast. In cases where there are bigger accidents, like a house burning down, the case will stay open for while. The insurance company have to make an assessment of the scope of the accident and expected cost. They will probably have to do inspections of the accident site to make a full assessment. If a person is injured, a doctor needs to be consulted. In this period there will be no big payouts, just minor payouts to cover assessment expenses, medical consultations and other minor expenses.

The major payouts happen when the damages are evaluated. In the case of a burnt down house the rebuilding will start, or in the case of personal injury, the medical treatment and rehabilitating will commence. In this period the insurance company will not really know the overall cost of the accident, and they will have to appraise it continuously. In this period the case is labelled as “open”, until there are no more payouts. In some cases, given the insured had disability insurance, the case will remain open for the rest of the insured individuals’ life.

A case that gets closed because the insurance company is not expecting there to be any more payouts, can still be re-opened. The policyholder might not be satisfied with the compensation he or she received, or additional information may have surfaced. In situations like these the case might be re-opened and it will have to go through the same stages again. The insurance company will have to re-evaluate to see if there is any basis for the customers dissatisfaction, or to see if the additional information that surfaced provided grounds for more compensation. If it does, the expected costs and payouts have to be re-assessed and the payouts will start again.

The problem that arises because of the delays in settlements are called RBNR problem. Just as with the IBNR, the insurance will have to put aside an amount of money, or reserve, so that they can pay the future payouts for the accidents that have been reported that year. Therefore this is also a big part of the daily routine of an insurance company.

Not all claims begin as an IBNR problem and then become a RBNS problem. Some are only IBNR, and some only RBNS. In some cases the accident is noticed quite early, but the settlements take some time. In other cases it takes some time before the accident is noticed, but then it is settled at once. It is a matter of fact that a lot of non-life insurance company have more RBNS cases then IBNR. The number of IBNR cases are also steadily declining for some products, for example personal injury from car collisions. The reason is the car manufacturers are building better and safer cars, the governments are building better roads and people are driving safer. These are some of the factors that contribute to fewer IBNR cases.

1.3 Reserving future claims and payouts

Estimating IBNR and RBNS reserves is probably one of the most important jobs of an actuary working in an insurance company. These estimates will affect the profitability of a insurance company and bad estimates could have grave consequences for the company. If the actuary over-estimates the reserve it could lead to the insurance company having less money to invest in the market. It could also make it seem like the company is not preforming well, which could lead to them increasing the price of their insurance products. This will not make them popular among their customers. If the actuary under-estimates the reserve it may seem as the company is performing well, and they might decrease the price of their products. This would make them less equipped to tackle unforeseen claims from past accidents which could have grave consequences for the insurance company. The worst case scenario would be that they are insolvent.

As the consequences of over- or underestimating the reserves could be grave, it is important to estimate the necessary reserves as exact as possible. There exist many possible methods for estimating reserves for IBNR and RBNS, like the Chain Ladder method, Bornhuetter-Ferguson method, and others. Both the Chain Ladder method and Bornhuetter-Ferguson method are purely algorithmic methods and uses aggregates to estimate the outstanding liabilities. In the actuarial community there has been some discussions about the convenience of using aggregated data. For presenting the data it is quite suitable, but there could also be loss of information, and in some cases it could lead to poor estimation of the outstanding liabilities. There is a lot of literature that supports using the individual loss data; see Norberg (1989), Norberg (1993), Kaminsky (1987) and Verrall, Nielsen and Jessen (2010). When it comes to the lack of stochasticity in the Chain Ladder method it is shown in Wütherich and Merz (2008) that there are in fact several different stochastic models that justify the Chain Ladder method and the Bornhuetter-Ferguson method. One of the models that leads to the same reserve estimates as the Chain Ladder method is the Poisson model. It should be noted that this revelation was made by the actuaries several years after the algorithm was constructed. Nevertheless the Chain Ladder method and the Bornhuetter-Ferguson are two of the most popular methods of calculating the reserves.

Since none of those who argued for employing the individual loss ever ranked their methods in terms of accuracy, this will be the objective of this thesis. The method that we propose as an alternative to the Chain Ladder method is a approach were we divide the problem into counts and losses and regard the delays in IBNR and RBNS as a multinomial phenomenon governed by delay probabilities. Kenneth Kaminsky was probably the most adamant spokesman for this approach, see Kaminsky (1987), which concerns only the IBNR situation. A similar approach can be done for the RBNS situation, see Verrall et al. (2010). In Bølviken (2015) the author assigns Kaminsky's name to the model and focuses mostly on the IBNR situation. This thesis is inspired by that paper and will adopt the same name for modelling RBNS reserves.

In Bølviken (2015) it is shown that breaking down the problem into counts and losses is always more accurate than using aggregates when estimating the IBNR reserve. In this thesis we will investigate if this holds true for the RBNS as well.

1.4 Objective and outline of the thesis

The objective of this thesis is to tackle the question: Should one use aggregated data, which is what the Chain Ladder method applies, or will dividing the problem into modelling the claim frequency and the claim sizes separately improve the reserve estimate? The mean square error will be used to compare these two models to investigate their uncertainties and bias in the estimation of the RBNS outstanding liabilities. To this end we will perform a data study, numerical study and solve it analytically. Everything but basic probability theory will be explained in the thesis.

Chapter 2 is dedicated to introducing both the Chain Ladder method and the Kaminsky approach and will explain how they can be implemented on both IBNR and RBNS reserve problems. Since the IBNR reserve problem is explained in great detail and solved for an IBNR case in Bølviken (2015), this thesis will mainly concern the RBNS reserve problem. In Chapter the Chain Ladder method and the Kaminsky approach will be implemented on real RBNS data from a Norwegian non-life insurance company. The goal will be to observe and discuss the different obstacles one may encounter when implementing these methods. We will also use bootstrapping to quantify the uncertainty and bias in the estimation of the outstanding liabilities. In Chapter 3 we will implement the methods on a simulated dataset, where we know the underlying situation perfectly to better examine the uncertainty and the bias in the estimation of the reserve. In Chapter 4 we will embark on finding an approximate expression for the uncertainty for both models, and compare them to maybe figure out which model is more accurate. Chapter 5 will present the concluding remarks for this thesis. In Appendix A the various distributions for modelling claim counts and claim sizes that are used in this thesis will be introduced. The fire insurance data and the car insurance data from the Norwegian non-life insurance company is introduced in Appendix B. The computer program that was used in the various simulations and to produce the different plots can be found in Appendix C. The script language that was used in this thesis is R, RStudio Team (2015), and will henceforth not be referenced to throughout the thesis.

Chapter 2

Modelling delay

2.1 Notation

Considering an IBNR and a RBNS situation, we assume that we are at the end of year I, and i, $0 \leq i \leq I$ are historical data going back I+1 years while k, $0 \leq k \leq K$ are the development years. The interpretation of i depends on if it is an IBNR or a RBNS case. If it is an IBNR case, i is denoted as the accident year or occurrence year. When considering a RBNS case it is interpreted as the year the claim was reported. This will be called reported year.

The X_{ik} has the interpretation of the sum of claims that were reported in year i and was settled k years later. In an IBNR case the interpretation is the sum of claims that incurred in year i and was reported k years later. X_{ik} is an observation if $i + k \leq I$. Each of them can be broken down into counts N_{ik} and losses per event $Z_{i,k,1}$, $Z_{i,k,2}$, ... so that:

$$X_{ik} = \sum_{l=1}^{N_{ik}} Z_{i,k,l}. \tag{2.1}$$

It is now possible to present the outstanding loss liabilities R_i as:

$$R_i = X_{i,I-i+1} + \cdots + X_{i,K}, \quad i = 1, \cdots, I, \tag{2.2}$$

which is the amount the insurance company has to reserve against. The issue that will be addressed in this thesis is whether it is better to estimate the outstanding loss liabilities R_i by taking use of the aggregates, or by breaking it down into counts and losses and model them separately.

2.2 Claims development triangles

When working with outstanding loss liabilities one often studies them in so-called claims development triangles, where the insurance claims are separated on two axes as in Figure 2.1. As mentioned above, the most recent accident/reported year is denoted by I while the last development year is denoted by K.

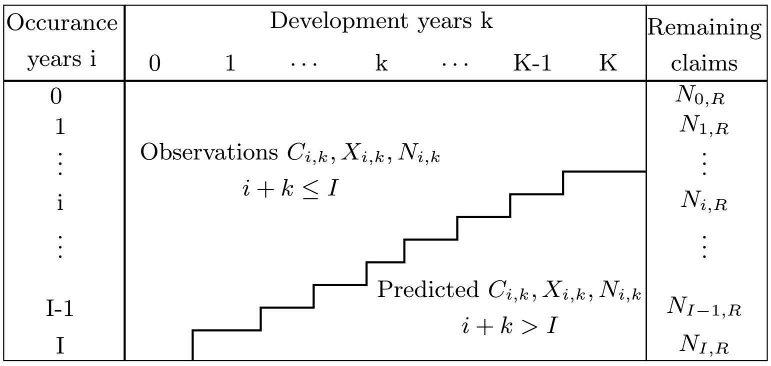

Figure 2.1: *Claims development triangle*

It is worth mentioning that we do not necessarily have to use development years as a measuring unit. Using development periods, where periods can be weeks, months, etc. is another possibility. It really depends on how the insurance company want to utilize the data they have acquired. When it comes to the data we have been given, it is most convenient to use years.

$X_{i,k}$ has the same interpretation as above while $C_{i,k}$ are defined as:

$$C_{i,k} = \sum_{j=0}^{k} X_{i,j} \tag{2.3}$$

which is interpreted as the cumulative claim losses that were reported in year i and were settled at most k years later. This interpretation regards a RBNR case. In an IBNR case the cumulative claim losses are interpreted as the claims that incurred in year i and was reported at most k years later.

Claims $X_{i,k}$ and $C_{i,k}$, as mentioned above, are usually studied in a claims development triangle where the accident/reported years are specified on the y-axis and development years on the x-axis, as in Figure 2.1. At time I the claims development triangle is split into two parts: The upper triangle or trapezoid which shows our historical data, and the lower triangle with the predicted or estimated values of $X_{i,k}$ or $C_{i,k}$.

When working with an IBNR problem, the column "Remaining claims" does not exist. The insurance company do not have any knowledge about the total amount of claims that occurred in accident year i , assuming that year i is not fully developed by the end of year I. In other words, when modelling the lower triangle the claim numbers are independent of each other. In a RBNS case the insurance company knows exactly how many claims were reported in reporting year i . So when predicting the number of claims in the lower triangle, we have to condition on the number of claims that have not been settled by the end of year I. Hence, the last column is of importance in a RBNS case when using the Kaminsky method. The Chain Ladder method has no need for information about the number of claims because it only estimates the accumulated claim sizes as we will see in the next section. This is the main difference between IBNR and RBNS cases.

Most textbooks do not emphasize the ramification of the absolute sizes of I and K in a claims development triangle. There are three possibilities: $K < I$, $K = I$ and $K > I$. The first possibility, where there are more accident/reporting years than development years is not a problem because we have enough historical data that we can use to estimate or predict the future payouts. In this case there will be a upper trapezoid and not a upper triangle of observed information. The second possibility is the same as the first, because there is enough historical data to predict or estimate the future payouts. The third possibility do cause some problems. In this case we have more development years than accident/reporting years, in other words we do not have enough information to estimate or predict future payouts with the models that are presented later on. A part of the solution could be to parametrize the delay probabilities, but in this thesis we will only consider the two first possibilities, because the last one is in itself a master's thesis.

By definition, $X_{i,k}$ and $C_{i,k}$ are observations if $i + k \leq I$. This means that we have to use the observations in the upper triangle/trapezoid,

$$D_I = \{X_{i,k}, C_{i,k}; i + k \leq I, 0 \leq k \leq K\},$$

to estimate or predict the lower triangle $D_I^c = \{X_{i,k}, C_{i,k}; i + k > I, i \leq I, k \leq K\}$.

2.3 The Chain Ladder

The Chain Ladder method is probably one of the most popular ways to estimate reserves. The main reason is the fact that it is distribution-free, in other words non-parametric. It is also known for its simplicity where the basic assumption is that patterns in the claim losses observed in the past will continue in the future (Haavardsson, 2014). This assumption is intuitive and basically says that there exist factors for each development year that describe how the total cumulative claim losses, C_{ij}, will change from one development year to the next.

We will consider Thomas Mack's distribution-free Chain Ladder where there are two embedded assumptions in the Chain Ladder method. The first assumption is a Markov-like assumption that says there exists factors $f_1, \cdots, f_K$ and $l_1, \cdots, l_K$ such that:

$$\mathbb{E}[C_{i,k+1}|C_{i,0},\cdots,C_{i,k}] = C_{i,k}f_{k+1} \quad \textit{and} \quad \mathbb{V}\mathrm{ar}(C_{i,k+1}|C_{i,0},\cdots,C_{i,k}) = C_{i,k}l_{k+1}. \tag{2.4}$$

The second assumption of the Chain Ladder method is that the variables, $C_{i,k}$, from different reported years are independent, i.e.:

$$\{C_{i,0},\cdots,C_{i,I}\}\,,\{C_{j,0},\cdots,C_{j,I}\}\,,\quad i\neq j\,,\ \textit{are independent.} \tag{2.5}$$

These two assumptions are implicitly assumed in the Chain Ladder algorithm. When working with the Chain Ladder method, one usually use the development triangle in Figure 2.1 with $C_{i,k}$'s.

The method is the same for both the IBNR and the RBNS case. Since we are at the end of year I, all the $C_{i,k}$'s which satisfy $i+k\leq I$ are known, i.e. observed data. In the first column, which equals $k=0$, we find the aggregated claims that were reported and settled the same year. The second column equals to those aggregated claim reports that were settled the year they were reported and the year after. With these interpretations in mind, we have that $C_{I,K}$ is the aggregated claim losses that were reported in year I and were settled up to K years later. The interpretation of the IBNR case is similar to the RBNS case, but uses “incurred and reported” instead of “reported and settled”. As of now we will only give the interpretation of the RBNS case as it is equivalent to the IBNR case, except the difference in wording.

To estimate the future cumulative reported claim losses, we will have to take a look at the Markov-like assumption in (2.4) left. To be able to estimate the next cumulative reported claim loss we have to multiply the previous cumulative reported claim loss with a factor f_k. This seems to correspond with what was mentioned earlier, that there exists patterns in how the aggregated reported claim losses evolve from one development year to the next.

A way to estimate these f_k's is by dividing the cumulative reported claim losses up to and including development year k, by the cumulative claim losses up to and including development year k-1, i.e.:

$$\hat{f}_k = \frac{\sum_{i=0}^{I-k} C_{i,k}}{\sum_{i=0}^{I-k} C_{i,k-1}} \quad \textit{for}\ k=1,\cdots,K. \tag{2.6}$$

If we divide the numerator and the denominator by $I-k+1$ we will notice that $\hat{f}_k$ is the average payout after k years divided by the average payout after k-1 years. This makes it an estimate to predict how the future losses will evolve. If we take a closer look at equation (2.6) for $I=2$ we will get:

$$\hat{f}_k = \frac{C_{0,k}+C_{1,k}+C_{2,k}}{C_{0,k-1}+C_{1,k-1}+C_{2,k-1}}.$$

Each $\hat{f}_k$ for $1\leq k\leq I$ is estimated by using as much data as possible from the different claim reported years. Further analysis could be made for $\hat{f}_k$: If $\hat{f}_k\leq 1$ we could conclude that the cumulative reported claim losses would in average decrease from development year k-1 to k . The opposite would apply to $\hat{f}_k \geq 1$, and if $\hat{f}_k = 1$ there would be no change in average. Another interesting aspect of (2.6) seems to be an underlying assumption of the Chain Ladder method: $\hat{f}_k$ for a specific development year is assumed to be the same for all reported claim years. In other words, how the cumulative reported claim settlements evolve from one development year to the next is independent of when they were reported. This coincides with what was mentioned earlier.

When $C_{i,k}$ is known the next one, $C_{i,k+1}$, can be found by multiplying $C_{i,k}$ with $\hat{f}_{k+1}$. $C_{i,k+2}$ can be found by multiplying $C_{i,k+1}$ with $\hat{f}_{k+2}$ and so on. It is therefore possible to write $C_{i,K}$ like:

$$C_{i,K} = C_{i,K-1} f_K = C_{i,K-2}\hat{f}_{K-1}\hat{f}_K = \cdots = C_{i,k}\hat{f}_{k+1}\cdots\hat{f}_{K-1}\hat{f}_K = C_{i,I-i}\prod_{k=I-i+1}^{K}\hat{f}_k.$$

As shown above, this simmers down to multiplying the “last” known observation with the remaining factors. We then have that:

$$C_{i,k} = C_{i,I-i}\prod_{j=I-i+1}^{K}\hat{f}_j \quad for\ k = I-i+1, \cdots, K. \tag{2.7}$$

With (2.6) and (2.7) we have the algorithm that will let us fill in the lower triangle in Figure 2.1.

With the definition of $C_{i,k}$ it is possible to write the outstanding amount (2.2) differently so that it coincides with the Chain Ladder notation.

$$\begin{aligned} R_i^{CL} &= X_{i,I-i+1} + \cdots + X_{i,K} \\ &= C_{i,K} - C_{i,I-i}. \end{aligned} \tag{2.8}$$

We can now use what has been shown above to easily express $C_{i,K}$ with $C_{i,I-i}$, i.e.:

$$\hat{R}_i^{CL} = C_{i,I-i}(\hat{f}_{I-i+1}\hat{f}_{I-i+2}\cdots\hat{f}_{i,K} - 1), \tag{2.9}$$

which has a multiplicative structure. This will be of importance later on.

2.4 Using delay probabilities

Using delay probabilities is probably the most natural way an actuary would tackle a delay problem. The delays could be regarded as a random phenomenon based on probabilities q_k, where q_k is the probability of a claim being settled k years later. We obviously have that $q_0 + \cdots + q_K = 1$ and the process is multinomial. This method is slightly different for IBNR and RBNS cases. If we first describe the RBNS case and let $N_{i,0}, \cdots, N_{i,K}$ be the numbers of claims that arose in year i and were settled $0, \cdots, K$ years later, then $N_{i,0} + \cdots + N_{i,K} = N_i$. We will then have that $N_{i,0}\ldots,N_{i,K}$ given N_i follows the multinomial distribution with probabilities $q_0,\cdots,q_K$ where $N_i=n_i$ is known.

$$P(N_{i,0}=n_{i,0},\cdots,N_{i,K}=n_{i,K}\,|N_i=n_i)=\frac{n_i!}{n_{i,0}!\cdots n_{i,K}!}q_0^{n_{i,0}}\cdots q_K^{n_{i,K}}.$$

From the multinomial distribution we have that the expectation and the variance is:

$$\mathbb{E}[N_{i,k}]=n_iq_k\quad and\quad \mathbb{V}\mathrm{ar}(N_{i,k})=n_iq_k(1-q_k).\tag{2.10}$$

In the IBNR case we make the natural assumption that N_i follows a Poisson distribution with parameter λ_i. Then we have that:

$$\begin{aligned}P(N_{i,0}=n_{i,0},\cdots N_{i,K}=n_{i,K})&=P(N_{i,0}=n_{i,0},\cdots,N_{i,K}=n_{i,K}\,|N_i=n_i)P(N_i=n_i)\\&=\frac{n_i!}{n_{i,0}!\cdots n_{i,K}!}q_0^{n_{i,0}}\cdots q_K^{n_{i,K}}\frac{\lambda_i^{n_i}}{n_i!}e^{\lambda_i}\\&=\frac{q_0^{n_{i,0}}\cdots q_K^{n_{i,K}}}{n_{i,0}!\cdots n_{i,K}!}\lambda_i^{n_i}e^{\lambda_i(q_0+\cdots+q_K)}\\&\overset{\text{a}}{=}\frac{q_0^{n_{i,0}}\cdots q_K^{n_{i,K}}}{n_{i,0}!\cdots n_{i,K}!}(\lambda_i^{n_{i,0}}e^{-q_0\lambda_i})\cdots(\lambda_i^{n_{i,K}}e^{-q_K\lambda_i})\\&=\prod_{k=0}^{K}\frac{(q_k\lambda_i)^{n_{i,k}}}{n_{i,k}!}e^{-q_k\lambda_i}.\end{aligned}$$

The set $\{N_{i,k}\}_{k=0}^{I}$ is stochastically independent with

$$N_{i,k}\sim Poisson(\lambda_{i,k})\;\;where\;\;\lambda_{i,k}=\lambda_iq_k.\tag{2.11}$$

As mentioned earlier, Kaminsky has probably been the most vocal advocate of using delay probabilities and modelling claim numbers and loss separately. From here on and throughout the thesis his name will be assigned to the method of using delay probabilities for calculating both IBNR and RBNS reserves.

2.5 Implementing the Kaminsky approach

When it comes to the IBNR case, fitting the Kaminsky model is not hard. Since the $\lambda_{i,k}$ is in a multiplicative form in (2.11) so that:

$$log(\lambda_{i,k})=log(\lambda_i)+log(q_k).$$

This is a log-linear Poission regression problem and can easily be fitted by standard GLM software where all the parameters will be estimated. This is discussed in detail in Bølviken (2014) and Bølviken (2015), while in de Jong and Heller (2008) the GLM process is described in detail. Since there is a lot of literature on this subject, we are mainly going to focus on the RBNS case.

[a] Since $q_0+\cdots+q_I=1$ and that $n_i=n_{i,0}+\cdots+n_{i,I}$

2.5.1 Delay probabilities

The delay probabilities can be found through maximum likelihood estimation. To find them, the Lagrange method were we use that $\sum_k q_k = 1$ will be applied. The log-likelihood is:

$$\begin{aligned} l(q_0, \cdots, q_K) &= \ln\left\{\prod_{i=0}^{I} P(N_{i,0} = n_{i,0}, \cdots, N_{i,I} = n_{i,I}\, | N_i = n_i)\right\} \\ &= \sum_{i=0}^{I}\sum_{k=0}^{K} \ln(n_i!) - \sum_{i=0}^{I}\sum_{k=0}^{K} \ln(n_{i,k}) + \sum_{i=0}^{I}\sum_{k=0}^{K} n_{i,k}\ln(q_k). \end{aligned} \tag{2.12}$$

If we use the Lagrange method with the constraint: $\sum_k q_k = 1$. We then get

$$\mathcal{L}(q_0, \cdots, q_K) = \sum_{i=0}^{I}\sum_{k=0}^{K} \ln(n_i!) - \sum_{i=0}^{I}\sum_{k=0}^{K} \ln(n_{i,k}) + \sum_{i=0}^{I}\sum_{k=0}^{K} n_{i,k}\ln(q_k) - \lambda(1 - \sum_{k=0}^{K} q_k). \tag{2.13}$$

By setting all the derivatives to 0, $\frac{\delta\mathcal{L}}{\delta\lambda} = 1 - \sum_{k=0}^{K} q_k = 0$, which is just the constraint, and also:

$$\begin{aligned} 0 = \frac{\delta\mathcal{L}(q_0, \cdots, q_K)}{\delta q_l} &= \frac{\sum_{i=0}^{I} n_{i,l}}{q_l} - \lambda \\ q_l &= \frac{\sum_{i=0}^{I} n_{i,l}}{\lambda}. \end{aligned} \tag{2.14}$$

To find λ we notice that:

$$q_0 + \cdots q_K = \frac{\sum_{i=0}^{I} n_{i,0} + \cdots + \sum_{i=0}^{I} n_{i,K}}{\lambda} = \frac{\sum_{i=0}^{I}\sum_{k=0}^{K} n_{i,k}}{\lambda} = 1$$

which gives us that $\lambda = \sum_{i=0}^{I}\sum_{k=0}^{K} n_{i,k}$. Then the ML estimator is:

$$\hat{q}_k = \frac{\sum_{i=0}^{I} n_{i,k}}{\sum_{i=0}^{I}\sum_{k=0}^{K} n_{i,k}}. \tag{2.15}$$

The ML estimator is quite intuitive as well since it is just the sum of column for a given development year divided by the total number of claims.

2.5.2 Mean and variance

The delay dependent mean is the average cost of a claim that is settled after k years. It can be found by taking the total amount of claim losses for a certain development year and dividing it by the total number of claims for that same development year, i.e.:

$$\hat{\xi}_k = \frac{\sum_{i=0}^{I-k} X_{i,k}}{\sum_{i=0}^{I-k} n_{i,k}} \quad \textit{for}\ k = 0, \cdots, K. \tag{2.16}$$

One may also notice that as k grows the uncertainty in $\hat{\xi}_k$ will also grow. When $k = K$ there is only one observation to base the estimate for delay dependent mean on. This will be addressed further in the next chapter.

The variance between the individual losses, $Z_{i,k,j}$, is denoted by σ_k^2, and can be found by taking the variance of all claims belonging to development year k . With the dataset that we have been given, calculating the variance is a bit tricky. The information that is available is only the sum of claims for the different combination of reporting and development year. We do not have any information about the individual claims. The datasets can be found in Appendix B.

Two methods were considered for calculating the variance in the individual losses, but only one of them worked. The method that worked based its estimation of the variance in the individual losses by calculating the variance in the average cost after k years for different reporting years and assuming the individual losses to be Gamma distributed. The other method was developed in hopes of it being more accurate than the former method. As it kept giving negative values for the variance it was eventually discarded. Both methods were documented and therefore both will be presented. The working method will be presented below, while the other one will be presented in Section 2.5.4 with a possible explanation as to why it did not work.

Since we do not have the individual losses we do not know the distribution of them. Therefore we will be making the likely assumption that the individual claims are Gamma distributed. The goal will be to estimate α_k which is defined as $\alpha = \xi^2/\sigma^2$. To this end we will introduce $Y_{ik} = \frac{X_{ik}}{n_{ik}} = \frac{1}{n_{ik}}\sum_{j=1}^{n_{ik}} Z_{ikj}$, which is the average claim cost for reporting year i and development year k . These individual claims are Gamma distributed with ξ_k and α_k, i.e: $Z_{ikj} \sim \xi_k Gamma(a_k)$. All the elements in development year k have the same shape factor α_k independent of reporting year i . We then have that:

$$\begin{cases} Y_{ik}|N_{ik} = 0 & \text{if } N_{ik} = 0 \\ Y_{ik}|N_{ik} \sim \xi_k Gamma(N_{ik}a_k) & \text{if } N_{ik} > 0. \end{cases}$$

By conditioning on $N_{ik} > 0$ we can easily calculate the expectation and the variance of Y_{ik}.

$$\mathbb{E}[Y_{ik}|N_{ik} > 0] = \mathbb{E}[\mathbb{E}[Y_{ik}|N_{ik}, N_{ik} > 0]] = \mathbb{E}[\xi_k] = \xi_k$$

and

$$\begin{aligned} \mathbb{V}\text{ar}(Y_{ik}|N_{ik} > 0) &= \mathbb{V}\text{ar}(\mathbb{E}[Y_{ik}|N_{ik}, N_{ik} > 0]) + \mathbb{E}[\mathbb{V}\text{ar}(Y_{ik}|N_{ik}, N_{ik} > 0)] \\ &= \mathbb{V}\text{ar}(\xi_k) + \frac{\xi_k^2}{\alpha_k}\mathbb{E}[\frac{1}{N_{ik}}|N_{ik} > 0] \\ &= \frac{\xi_k^2}{a_k}\sum_{j=1}^{n_i}\frac{1}{j}\binom{n_i}{j} q_k^j(1-q_k)^{n_i-j} \\ &= \frac{\xi_k^2 f(n_i, q_k)}{a_k} \end{aligned}$$

where $f(n_i, q_k) = \sum_{j=1}^{n_i} \frac{1}{j} bin(n_i, j, q_k)$ and $bin(n_i, j, q_k)$ is the binomial distribution with probability of success q_k. It is also possible to calculate the variance of Y_{ik} by using the empirical formula. By setting these two expressions equal to each other, the shape parameters, α_k's are easily estimated.

$$s_k^2 = \frac{1}{K-1-k} \sum_{i=0}^{I-k} (Y_{i,k} - \hat{\xi}_k)^2 \quad \textit{for } k = 0, \cdots, K-1 \tag{2.17}$$

so that

$$\begin{aligned} s_k^2 &= \mathbb{V}\text{ar}(Y_{ik} | N_{ik} > 0) \\ s_k^2 &= \frac{\xi_k^2 f(n_i, q_k)}{a_k} \\ \hat{a}_k &= \frac{\xi_k^2 f(n_i, q_k)}{s_k^2} \quad \textit{for } k = 0, \cdots, K-1. \end{aligned} \tag{2.18}$$

Again, we have that as k grows the uncertainty in a_k will grow because of lack of information. One may also notice that s_K will always be zero, which is why both s_k and a_k are only defined from $k = 0, \cdots, K-1$.

2.5.3 Modelling the lower triangle

The Kaminsky approach is based on breaking the reserve problem into two parts: modelling claim numbers and modelling claim sizes. Since we have everything we need, $\hat{\xi}_k$ and $\hat{a}_k$, to model the claim sizes we are going to shift our focus to the claim numbers.

When modelling the lower triangle we have to condition on the upper triangle. The good news is that the lower triangle is still multinomial distributed, i.e. we have that $N_{i,K-i+1}$ for $i = 1, \cdots, I$ are multinomial distributed given $N_{i,R}$ for for $i = 1, \cdots, I$. Recall that $N_{i,R}$ are all the claims that have been reported, but not yet settled at time I. We then have to calculate new delay probabilities, $\tilde{q}_{i,k}$, by conditioning them on that the claim is settled for a development year $k > K - i$, i.e.:

$$\begin{aligned} \tilde{q}_{i,k} = P(k = j | k > K - i) &= \frac{P(k = j, k > K - i)}{P(k > K - i)} \\ &= \begin{cases} \frac{P(k=j)}{P(k>K-i)}, & \text{if } j > K - i \\ 0, & \text{if } j \leq K - i \end{cases} \\ &= \frac{P(k = j)}{1 - P(k \leq K - i)}, \quad \text{if } j > K - i \\ &= \frac{q_j}{1 - \sum_{l=o}^{K-i} q_l}, \quad \text{if } j > K - i. \end{aligned}$$

With these "new" delay probabilities is is possible to model the lower triangle. Notice that if we assume, as we will, that all claims that were reported in year i will be settled within K years. Then we have that for $i = 1$ the probability is $\tilde{q}_{1,K} = 1$ and for $i = 2$ it all collapses into a binomial trial. By combining this with the modelling of claim sizes the lower triangle is easy to predict.

2.5.4 An alternative method for calculating variance

An alternative method to calculate the individual claim variance, that is not used, will now be presented. The method is quite intuitive and logical, but it did not work with the dataset that was given for this thesis. This method was developed because it was thought that it would give a better estimate for the variance in the individual losses as it did not assume any underlying distribution for the $Z_{i,k,j}$'s.

We start by defining $Y_{ik} = \frac{X_{ik}}{n_{ik}}$, which is the average cost of a claim in reporting year i and development year k. We also define $Q_k = \sum_{i=0}^{I-k} n_{ik}(Y_{ik} - \hat{\xi}_k)^2$.

$$\begin{aligned}
Q_k &= \sum_{i=0}^{I-k} n_{ik}(Y_{ik} - \hat{\xi}_k)^2 \\
&= \sum_{i=0}^{I-k} n_{ik}(Y_{ik} - \xi_k - (\hat{\xi}_k - \xi_k))^2 \\
&= \sum_{i=0}^{I-k} n_{ik}(Y_{ik} - \xi_k)^2 + \sum_{i=0}^{I-k} n_{ik}(\hat{\xi}_k - \xi_k)^2 - 2\sum_{i=0}^{I-k} n_{ik}(Y_{ik} - \xi_k)(\hat{\xi}_k - \xi_k) \\
&= \sum_{i=0}^{I-k} n_{ik}(Y_{ik} - \xi_k)^2 - \sum_{i=0}^{I-k} n_{ik}(\hat{\xi}_k - \xi_k)^2. \qquad (2.19)
\end{aligned}$$

Here we have used that $\sum_{i=0}^{I-k} n_{ik}Y_{ik} = \sum_{i=0}^{I-k} X_{ik} = \frac{\sum_{i=0}^{I-k} n_{ik} \sum_{i=0}^{I-k} X_{ik}}{\sum_{i=0}^{I-k} n_{ik}} = \hat{\xi}_k \sum_{i=0}^{I-k} n_{ik}$. We define the first expression in $\mathbb{E}[Q_k]$ for I and the last one II.

$$\begin{aligned}
I : \mathbb{E}[\sum_{i=0}^{I-k} n_{ik}(Y_{ik} - \xi_k)^2] &= \sum_{i=0}^{I-k} n_{ik}[\mathbb{V}\text{ar}(Y_{ik} - \xi_k) + (\mathbb{E}[Y_{ik} - \xi_k])^2] \\
&= \sum_{i=0}^{I-k} n_{ik}[\mathbb{V}\text{ar}(\frac{X_{ik}}{n_{ik}}) + (\mathbb{E}[\frac{X_{ik}}{n_{ik}}] - \xi_k)^2] \\
&= \sum_{i=0}^{I-k} n_{ik}[\frac{1}{n_{ik}^2} \mathbb{V}\text{ar}(X_{ik}) + (\frac{1}{n_{ik}} \mathbb{E}[X_{ik}] - \xi_k)^2] \\
&= \sum_{i=0}^{I-k} n_{ik}[\frac{n_i q_k[\xi_k^2(1 - q_k) + \sigma_k^2]}{n_{ik}^2} + \xi_k^2(\frac{n_i q_k}{n_{ik}} - 1)^2] \qquad (2.20)
\end{aligned}$$

and

$$
\begin{aligned}
II : \mathbb{E}[\sum_{i=0}^{I} n_{ik}(\hat{\xi}_k - \xi_k)^2] &= \sum_{i=0}^{I-k} n_{ik}[\mathbb{V}\text{ar}(\hat{\xi}_k - \xi_k) + (E[\hat{\xi}_k - \xi_k])^2] \\
&= \sum_{i=0}^{I-k} n_{ik}\,\mathbb{V}\text{ar}(\hat{\xi}_k) \\
&= \sum_{i=0}^{I-k} n_{ik}\frac{\sum_{i=0}^{I}\mathbb{V}\text{ar}(X_{ik})}{(\sum_{i=0}^{I-k} n_{ik})^2} \\
&= \sum_{i=0}^{I-k} n_{ik}\frac{\sum_{i=0}^{I-k} n_i q_k[\xi_k^2(1-q_k)+\sigma_k^2]}{(\sum_{i=0}^{I-k} n_{ik})^2} \qquad (2.21)
\end{aligned}
$$

where we have used the expectation and the variance of X_{ik} which are calculated in (2.25) and (4.10) respectively. We have then have that:

$$\mathbb{E}[Q_k] = \sum_{i=0}^{I-k} n_{ik}[\hat{q}_k[\xi_k^2(1-q_k)+\sigma_k^2](\frac{n_i}{n_{ik}^2} - \frac{\sum_{i=0}^{I-k} n_i}{(\sum_{i=0}^{I-k} n_{ik})^2}) + \xi_k^2(\frac{n_i q_k}{n_{ik}} - 1)^2]. \qquad (2.22)$$

$\mathbb{E}[Q_k]$ can easily be calculated and everything on the right hand side is known except for σ_k^2. ξ_k can be estimated through $\hat{\xi}_k$. By solving equation (2.22) for σ_k^2 we have an estimate for the variance in the individual losses.

$$\hat{\sigma}_k^2 = \frac{\mathbb{E}[Q_k] + \hat{\xi}_k^2\sum_i n_{ik}((\frac{\sum_i n_i}{\sum_i n_{ik}}\hat{q}_k - 1)^2 - (\frac{n_i\hat{q}_k}{n_{ik}} - 1)^2)}{\hat{q}_k(\sum_i \frac{n_i}{n_{ik}} - \frac{\sum_{i=0}^{I-k} n_i}{\sum_{i=0}^{I-k} n_{ik}})} - \hat{\xi}_k^2(1-\hat{q}_k). \qquad (2.23)$$

The problem with this method is that it kept giving negative values for some of the variances for some k's. Our understanding is that the variance between the individual claims became overshadowed by the variance between the $X_{i,k}$'s. In other words, $\mathbb{V}\text{ar}(X_{i,k}) = n_i q_k[\xi_k^2(1-q_k)+\sigma_k^2] \approx n_i q_k \xi_k^2(1-q_k)$ where as mentioned, σ_k^2 is the variance between the individual claims. When this method did not work, equation (2.17) was used instead to estimate the shape parameter.

2.6 Outstanding loss liabilities

When it comes to estimating the outstanding amount, we can predict $N_{i,I-i+1},\cdots,N_{i,K}$ through their expectations. We have to combine this with the model for the claim, which depends on how long it has taken to report or settle them. We can observe this in how $X_{i,k}$ is constructed. The expectation of $X_{i,k}$ is:

$$\textit{IBNR:}\quad \mathbb{E}[X_{i,k}] = \mathbb{E}[\mathbb{E}[X_{i,k}|N_{i,k}]] = \mathbb{E}[\xi_k N_{i,k}] = \xi_k\lambda_i q_k \qquad (2.24)$$

$$\textit{RBNS:}\quad \mathbb{E}[X_{i,k}] = \mathbb{E}[\mathbb{E}[X_{i,k}|N_{i,k}]] = \mathbb{E}[\xi_k N_{i,k}] = \xi_k n_i q_k \qquad (2.25)$$

where ξ_k is just a delay-dependent mean. It is now possible to find the expectation of R_i as defined in (2.2):

$$\textit{IBNR:}\quad \mathbb{E}[R_i] = \lambda_i(q_{I-i+1}\xi_{I-i+1} + \cdots + q_K\xi_K). \tag{2.26}$$

$$\textit{RBNS:}\quad \mathbb{E}[R_i] = n_i(q_{I-i+1}\xi_{I-i+1} + \cdots + q_K\xi_K). \tag{2.27}$$

With the estimates $\hat{\lambda}_i$, $\hat{\xi}_k$ and $\hat{q}_k$, the Kaminsky prediction for the outstanding amount becomes:

$$\textit{IBNR:}\quad \hat{R}_i^{Ka} = \hat{\lambda}_i(\hat{q}_{I-i+1}\hat{\xi}_{I-i+1} + \cdots + \hat{q}_K\hat{\xi}_K). \tag{2.28}$$

$$\textit{RBNS:}\quad \hat{R}_i^{Ka} = n_i(\hat{q}_{I-i+1}\hat{\xi}_{I-i+1} + \cdots + \hat{q}_K\hat{\xi}_K). \tag{2.29}$$

It should be noted that the Kaminsky method has an additive structure, as one may see above. $\hat{\xi}_k$ can be estimated in various ways, but in this thesis it is estimated by taking the average of all past claims that were settled k years later, as in equation (2.16). In Chapter 4 when we embark of finding approximate expressions for the Kaminsky and Chain Ladder uncertainty, this way of estimating $\hat{\xi}_k$ will be convenient.

2.7 Method for comparing the two models

The method we decide to use to compare both models is the mean square error, MSE, which is defined as $\mathbb{E}[(\hat{\theta}-\theta)^2]$. For more on MSE see Devore and Berk (2007). This method was chosen because the Chain Ladder method has a multiplicative structure while the Kaminsky approach has an additive structure. This will be addressed further in Chapter 4.

$$\begin{aligned}\mathbb{E}[(\hat{\theta}-\theta)^2] &= \mathbb{V}\text{ar}(\hat{\theta}) + [\mathbb{E}[\hat{\theta}] - \theta]^2 \\ &= \textit{Variance of estimator} \;\; + \;\; [\textit{bias}]^2\end{aligned}$$

An estimator is unbiased if the bias is equal to zero, i.e. $\mathbb{E}[\hat{\theta}] = \theta$.

The MSE informs us about the balance between the uncertainty of the estimator and how well it estimates the target, θ. Focusing only on either the uncertainty or the bias will not give us any valuable information. For example: If the objective is to have an uncertainty equal to zero, we can choose $\hat{\theta}$ to be equal to a constant. But, this does not guarantee that the $\hat{\theta}$ is a good estimate for θ. To ensure that $\hat{\theta}$ is a good estimate, we would have to look at both the uncertainty and the bias. When comparing the Chain Ladder method and the Kaminsky approach we will investigate the estimation of the outstanding liabilities through analysing the balance between the uncertainty and the bias, as this will show which model is superior.

In the next chapter the Chain Ladder method and the Kaminsky approach will be implemented on both the real data from a Norwegian non-life insurance company and a simulated dataset. When implementing the Kaminsky approach, the "recipe" described in Section 2.5 will be used to estimate the various parameters to predict the lower triangle. We will also try to figure out which parameters seem to affect the uncertainty and the bias in estimates for the outstanding liabilities.

Chapter 3

Data study

3.1 Data

To compare the two models we will use the car and fire insurance data from the Norwegian non-life insurance company. In the car insurance data there were 1504 incidents while in the fire insurance data there were 2963 incidents. In the case of car insurance we received information dating back to 2009, while the fire insurance information covers the period 2010 to 2015. The number of policies have been increasing by about an average of 5 000 and 10 000 each year for car and fire insurance respectively. As mentioned earlier, when a claim is reported to the insurance company and is valid for one of the policies, the policyholder will not necessarily get a lump sum. The insurance company will not pay everything at once but rather small payments until there are no more payouts. In this dataset we have defined “settlement” as the last payout to the policyholder, in other words the year of the case being “closed”. More information about the dataset can be found in Appendix B on page 57.

Table 3.1: *Number of fire insurance claims that were reported and settled with delay*

	0	1	2	3	4	5	Not Yet Settled	Total(n_i)
2010	212	92	26	8	5	1	2	346
2011	274	105	15	10	4		3	411
2012	269	111	20	9			7	416
2013	319	110	13				9	451
2014	599	198					34	831
2015	378						130	508

The table above presents the number of claims that were settled. The rows are the reported years and the columns are the delays, also known as development years. The column titled “Not Yet Settled” includes the claims that have not yet been settled. This means that 130 claims out of all 508 claims that were reported in year 2015 have not yet been settled at the time we were given the dataset. The last column is the total amount of claims that were reported for each year. It is reasonable that there is a steady increase in claims that have not yet been settled from the different reported years. Remember the interpretation of development triangle, for example: of all the claims that were reported in year 2010, 212 of them were settled the same year, while 92 of them were settled the year after, in 2011. The highlighted gray diagonal are all the claims settled in 2015. One would then expect there to be quite a few claims not yet settled out of the claims reported in 2015.

The data for the car insurance is presented below in the same way as Table 3.1, with the same interpretation.

Table 3.2: *Number of car insurance claims that were reported and settled with delay*

	0	1	2	3	4	5	6	Not Yet Settled	Total(n_i)
2009	12	14	17	10	9	4	4	7	77
2010	32	70	14	14	9	2		5	146
2011	60	51	22	16	8			16	173
2012	77	83	23	13				20	216
2013	65	101	22					35	223
2014	150	148						80	378
2015	108							183	291

Comparing the table above with Table 3.1 we notice that for the fire insurance, most of the claims are settled the year they were reported. For the car insurance, almost the same amount of claims are settled the two first years after the claims were reported. In some cases it might take more time to recover from a personal injury caused by a car accident, than it takes to settle a fire insurance claim. This could be the reason why the settlement of car insurance claims are dragged out over a longer period.

It will be assumed that all claims that were reported after 2009 and 2010 for car and fire insurance respectively, will be settled within their respectively maximum delays of 5 and 6 years.

3.1.1 Delay-dependent mean

The delay-dependent mean, ξ_k, is the average cost for a claim after k years. It is also needed when computing the the lower triangle, D_I^c with the Kaminsky method, but it also holds some interesting information about the dataset. An estimate of the delay-dependent mean for each individual claim can be found by taking the total amount of claim losses for a certain development year and dividing it by the total number of claims for that development year as seen in equation (2.16).

The table below has the delay-dependent mean for both the car and fire insurance. A rather interesting aspect of this table is that it seems like the most expensive claims are settled 4 years after being reported. Table 3.3 also indicates that the fire insurance claims are more expensive for the insurance company than car insurance claims, which is somewhat intuitive. Villas are seemingly often more expensive than injuries related to car accidents.

Table 3.3: *The delay-dependent mean for the individual car and fire losses given in million NOK*

	$\hat{\xi}_0$	$\hat{\xi}_1$	$\hat{\xi}_2$	$\hat{\xi}_3$	$\hat{\xi}_4$	$\hat{\xi}_5$	$\hat{\xi}_6$
Fire	0.03	0.20	1.50	2.27	3.17	0.06	
Car	0.01	0.02	0.19	0.29	0.35	0.25	0.09

There is considerable uncertainty with high delays because of the lack of information for high k's. These errors have limited effect on the projection in the Kaminsky method because the delay probabilities are quite small, which can be found in Table 3.10.

In the next two sections we will implement the Chain Ladder method and the Kaminsky approach on these datasets to see what kind of obstacles that can occur. The results for the outstanding liabilities will be presented in Table 3.15 and 3.16 where bootstrapping has been used to obtain the final estimates. The R-codes for the implementation of these models can be found in Appendix C.

3.2 Chain Ladder

3.2.1 Fire insurance data

Table 3.4: *Cumulative payouts/settlement (in million NOK) in fire insurance presented as a run-off triangle*

	0	1	2	3	4	5
2010	4.85	17.71	52.56	81.15	103.39	103.45
2011	8.04	30.51	50.25	72.13	78.40	
2012	7.28	28.72	64.02	74.87		
2013	10.35	52.88	73.76			
2014	11.38	34.07				
2015	9.41					

We start by implementing the Chain Ladder method on the fire insurance data presented above. For more information about the dataset, consult Appendix B. We present the cumulative payouts/settlements in a run-off triangle fashion. We see that the Table 3.4 is equivalent to Figure 2.1 with $C_{i,k}$.

The cumulative settlements are the total amount of claims settled up to that development year, which is the sum of the incremental settlements to that date. By definition, we then have that all the elements on the outer diagonal are equal to the total amount settled up to that date for each reported year. From looking at the run-off triangle in Table 3.4, we can see that the development years seem to develop in the same way independent of when the accident was reported. In other words all the elements in a column, development year, behave in the same way. They all increase with about the same amount. This does indeed satisfy the assumptions in the Chain Ladder model.

By using the equation (2.6) we can find the Chain Ladder factors that describe how the development years change from one development year to the next. Below we have Table 3.5 with the Chain Ladder factors.

Table 3.5: *Chain Ladder estimates for the development factors for the fire insurance data*

k	1	2	3	4	5
$\hat{f}_k$	3.91	1.85	1.37	1.19	1.00

By looking at these factors we see that all of them are either equal to or larger than 1 when taking two decimals into account. The factors also steadily decrease, starting at 3.19 and gradually declining towards 1.00. This was expected, as we observed that most of the claims would be settled within a short period of time.

From a statistical point of view, the factor estimated for development year 1 is more reliable than the estimates for the other development years, especially the last one. If we look at equation (2.6) and Table 3.4 we see that the reason is because more observations are used to estimate $\hat{f}_1$ compared to $\hat{f}_5$.

Table 3.6: *Cumulated payouts/settlements (in million NOK) for fire insurance*

	0	1	2	3	4	5
2010	4.85	17.71	52.56	81.15	103.39	103.45
2011	8.04	30.51	50.25	72.13	78.40	78.44
2012	7.28	28.72	64.02	74.87	88.79	88.84
2013	10.35	52.88	73.76	100.87	119.63	119.69
2014	11.38	34.07	63.15	86.36	102.42	102.47
2015	9.41	36.78	68.16	93.21	110.55	110.61

Above we have used the factors in Table 3.5 to find the missing lower triangle in Table 3.4. We have used the algorithm described in Chapter 2. Looking at the grey triangle in Table 3.6 we see that they behave the same way as the factors, which is to be expected if we take a look at equation (2.7). The equation tells us to multiply the last known observation with the remaining Chain Ladder factors. When programming we are using more than two decimals on the development factors, which is why there is a relatively small change from development year 4 to 5.

Now that we have computed the lower triangle we can easily find the outstanding loss liabilities. We take the elements in the last column in Table 3.6 and subtract the last know observations for the respective accident years, as in equation (2.8). Another possibility is to use equation (2.9). As mentioned the results of this will be presented later on.

3.2.2 Car insurance data

We have the settlements for the car insurance from year 2009 to 2015. Notice that we have a development triangle with the settlements, X_{ik}, and not the cumulative settlements, C_{ik}. At first glance everything seems to be fine, but with a closer look we notice that element $X_{2013,0} = 1181876$ is an outlier compared to the others. This can cause problems. A condition of the Chain Ladder model is that what happens one year, will happen in the next year as well. In other words, there is no room for outliers in the Cain Ladder model. This outlier clearly does not satisfy the condition above.

Table 3.7: *Payouts/settlements (in million NOK) for car insurance presented in a run-off triangle*

	0	1	2	3	4	5	6
2009	0.05	0.16	8.95	3.64	2.39	0.78	0.36
2010	0.06	0.50	2.68	2.47	3.57	0.73	
2011	0.26	0.86	5.32	4.15	3.13		
2012	0.38	3.12	4.86	4.49			
2013	1.18	1.56	4.65				
2014	0.55	2.94					
2015	0.61						

We have that 65 claims were reported and settled in 2013. In 2014, 150 claims were reported and settled. This means that the average cost of each of those 150 claims was around 4 000 NOK, while the 65 claims that were settled in 2013 cost around 18 000 NOK each in average.

It could be that the road was quite slippery in 2013, which caused a chain collision and therefore the amount $X_{2013,0}$ consist of several middle sized claims. It could also be that there is an individual outlier, i.e. there is a single large claim which is the reason why $X_{2013,0}$ is so big.

An actuary in the industry(email correspondence with an actuary from DNB) explained that software usually uses a truncating method to deal with problems such as these. They also use several other methods to compute an outstanding loss estimate, so they do not solely rely on the Chain Ladder method. Another method that can be used is the one proposed by Weindorfer (2012) where he finds the single claim which is the reason for the amount to be large. Then he preforms the Chain Ladder method without that claim. Here he acts as if the outlier claim did not happen, since it is "unnatural".

The actuary proposed another method where we take the outlier claim into account. First start by removing the claim or the claims that generate high values in $C_{2013,0}$. Then preform the Chain Ladder method on the dataset without the outlier/outliers. Using the Chain Ladder factors acquired from the dataset without the outlier/outliers, one can preform a Chain Ladder method on the outliers and adding the reserves for both with and without the outliers to achieve a total reserve estimate.

We cannot perform any of these methods because we do not have the data necessary to do so. If we take a look at the dataset in Appendix B we do not have information about the individual claims, but rather the aggregated ones for certain development and reported years. Therefore we cannot simply find an outlier claim or claims. This also shows a unfavourable side of the Chain Ladder method which runs on the conditions that there cannot be any outliers. While in real life there is always a possibility for a "surprise" chain collision, or a single expensive claim. We will proceed without any modifications, because there was only one outlier cell. This means that it has a very limited effect on the result. The biggest effect will be through the estimation of $\hat{f}_1$.

Table 3.8: *Chain Ladder estimates for the development factor for car insurance data*

k	1	2	3	4	5	6
$\hat{f}_k$	4.68	3.26	1.77	1.43	1.09	1.04

As in the fire insurance case, all the Chain Ladder factors are greater than 1 and steadily decrease from 4.68 to 1.04. We see that the factors for development year 1 are quite a bit higher for the car insurance data compared to the same development years in the fire insurance. We have to remember that the factor $\hat{f}_1$ describes how the accumulated claim settlements evolve from development year 0 to 1. If we look at Table 3.2, we see that unlike the fire insurance data there are more claims that are settled the year after they get reported than the amount that are settled the same year. In Table 3.3 we also see that the average claim cost rises from development year 0 to 1. It is therefore reasonable that $\hat{f}_1$ is quite big.

Another reason why $\hat{f}_1$ is big, is because of the outlier. How does this affect the the outstanding loss? It only affects the 2014 reserve estimate because it is the only one that requires $\hat{f}_1$. This causes the outlier to have a limited effect. Remember that the Chain Ladder method does not use the claim numbers, only the accumulated losses. The Chain Ladder factors, $\hat{f}_k$, picks up on the trend of how the cost varies from one development year to the next. The estimates in the context of the claim numbers can be surprising. Notice that for 2014 there are only 80 claims left. These 80 claims seem to cost around 28-29 million NOK while the 298 claims cost around 4 million NOK. It should be noted that the final estimate, $C_{2014,6}$, does not seem to be out of this world. It is a realistic estimate compared to the others and the fact that there were 378 claims reported that years.

Table 3.9: *Cumulative payouts/settlements (in million NOK) for fire insurance in a run-off triangle*

	0	1	2	3	4	5	6
2009	0.05	0.21	1.10	4.75	7.13	7.92	8.28
2010	0.06	0.56	3.24	5.71	9.28	10.01	10.47
2011	0.26	1.12	6.45	10.59	13.72	14.99	15.67
2012	0.38	3.51	8.37	12.86	18.40	20.10	21.02
2013	1.18	2.75	7.39	13.09	18.73	20.46	21.40
2014	0.55	3.49	11.38	20.14	28.83	31.45	32.93
2015	0.61	2.84	9.27	16.40	23.48	25.65	26.82

3.3 The Kaminsky method

3.3.1 Delay probabilities

The maximum likelihood (ML) estimator obtained in Section 2.5,

$$\hat{q}_k = \frac{\sum_{i=0}^{I-k} n_{i,k}}{\sum_{i=0}^{I} \sum_{k'=0}^{K-i} n_{i,k'}},$$

is a quite natural estimate for q_k. If we take a look at either Table 3.1 or Table 3.2 we see that q_k is estimated by summing the column for a given k , development year, and dividing it by the sum of all the claim numbers. The delay probabilities are based on the upper triangle, D_I. One may also notice that the uncertainty grows with higher lags, as less information is used to estimate the delay probabilities for high k 's.

Table 3.10: *The delay probabilities for the car and fire insurance modelling*

	$\hat{q}_0$	$\hat{q}_1$	$\hat{q}_2$	$\hat{q}_3$	$\hat{q}_4$	$\hat{q}_5$	$\hat{q}_6$
Fire	0.6922	0.2079	0.0250	0.0091	0.0030	0.0003	
Car	0.3351	0.3105	0.0652	0.0352	0.0173	0.0040	0.0027

We see that in the case of fire insurance most of the claims, around 70%, are settled the year they are reported. These are also the cheapest claims according to the Table 3.3. The number of claims that get settled gradually decline as the development years get higher. We also notice that the most expensive development year only consist of 0.3% of the total claims that were reported. With the car insurance we see that almost the same amount of claims were settled the year they were reported and the year after. The most expensive development year only has 1.74% of the total amount of claims that were reported. The most interesting aspect is that unlike the fire insurance, the car insurance claims are dragged out over longer periods of time. With the fire insurance almost 90% of the claims are settled within the 2 first year after reporting, while only 60% are settled in the same time period for car insurance claims. This may be intuitive since it will most likely take some time before people recover after a car accident, which means that it takes time before a claim is settled. Remember that a case is settled when the insured has received their last compensation payment.

3.3.2 The shape parameter

As specified earlier, because there is no information about the individual claims, the usually easy task of finding the shape parameter becomes a bit more tricky. The method that was described in Section 2.5 is the method that have been used to produce the estimates below.

Table 3.11: *The Gamma shape parameter α for different development years*

	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$
Fire	0.11	0.04	5.13	1.14	1.09	1.09	
Car	0.05	0.08	0.62	1.60	7.38	2.09	2.09

Notice that the two last shape parameters for both fire and car insurance are equal. That is because there is not enough information in development year K to produce an estimate $\hat{\alpha}$, the shape parameter. It is interesting to see how the shape parameter evolves from one development year to the next. One can see that for the 2 first years and the 3 first years for the fire and car insurance respectively, the values are quite low compared to the other. For low values of α, we get heavier tails as seen in Figure 3.1. We are using the parametrization of the Gamma distribution from

Bølviken (2014) which can also be found in Appendix A. The claim size distribution for the 2 first years and the 3 first years for the fire and car insurance respectively are heavy-tailed. One can also see big jumps in $\hat{\alpha}_2$ and $\hat{\alpha}_4$, for fire and car insurance respectively. The high fluctuations in the parameters can be explained by noise in the dataset and lack of information.

Gamma distribution

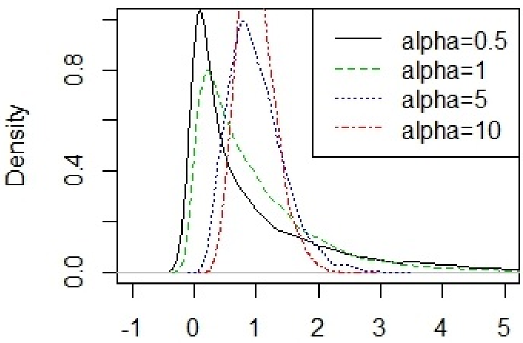

Figure 3.1: *Gamma distribution for different shape parameters α and mean equal to 1*

An interesting question is: How should the α be for different development years? Intuitively, we could assume that all "easy" claims are settled quite early on. One could also assume that these "easy" claims are cheap and therefore easy to settle. As k gets larger, the variance in the individual claims would also become higher because these claims are more complex and cost more to settle which can be confirmed by Table 3.3 where the claim cost grows as k grows. One should then expect the individual claim variance, $\hat{\sigma}_k$, to get higher as k grows, i.e. the tail should become heavier as k grows. If we combine this with Table 3.3 and use the standard deviation for the Gamma distribution: $\sigma = \frac{\xi}{\sqrt{\alpha}}$, we see that the standard deviation grows with k . Even though the shape parameter seems to show the opposite, combining the parameter with $\hat{\xi}_k$'s show us that for high k values the distribution is more heavy-tailed. We also have a somewhat rough estimate of $\hat{\sigma}_k$, s_k that was derived in Section 2.5, equation (2.17) which gives us the estimates below:

Table 3.12: *Emperical standard deviation given in millions*

	s_0	s_1	s_2	s_3	s_4	s_5	s_6
Fire	0.005	0.107	0.217	1.192	2.048	NaN	
Car	0.005	0.011	0.075	0.087	0.074	0.125	NaN

As already mentioned, these estimates are very rough because there is not enough information. They will still help us to paint the picture that the individual variance grows as k gets larger, as you can see in Table 3.12. This give a heavier tail for high k 's. It is hard to determine anything without the information about the individual claims. It may also depend on what product it is: if it is a boat, travel or any other kind of insurance. For some products, the claim size distribution is heavy-tailed for small k's and not for large k's. This does not seem to be the case in our dataset.

3.3.3 Fire and car insurance

By using the estimated delay-dependent means, the delay probabilities and the shape parameter we can predict the lower triangle. Below we have a result of a single iteration of the Kaminsky approach. These results are presented for illustrative purposes only and are not the final result of the Kaminsky approach.

Table 3.13: *Cumulative payouts/settlement (in million NOK) for fire insurance*

	0	1	2	3	4	5
2010	4.85	17.71	52.56	81.15	103.39	103.45
2011	8.04	30.51	50.25	72.13	78.40	78.54
2012	7.28	28.71	64.02	74.87	92.52	92.59
2013	10.35	52.88	73.76	84.76	87.70	87.70
2014	11.38	34.07	176.05	217.15	251.80	251.80
2015	9.41	9.85	116.70	123.33	129.55	129.55

What is interesting with this method of modelling the lower triangle is that it will sometimes predict that there will be no claims settled for some certain combination of the reporting and development year. This possibility is absent from the Chain Ladder method, because it uses the development factors. If $f_k = 1$ the Chain Ladder method will assume that no claims were settled k years later for all reporting years i. This could be seen as a weakness in the Chain Ladder method and a strength in modelling both claim frequency and claim sizes separately and combining them afterwards.

Table 3.14: *Cumulative payouts/settlement (in million NOK) for car insurance*

	0	1	2	3	4	5	6
2009	0.05	0.21	1.10	4.75	7.13	7.92	8.28
2010	0.06	0.56	3.24	5.71	9.28	10.01	11.22
2011	0.26	1.12	6.45	10.59	13.72	18.79	19.61
2012	0.38	3.51	8.37	12.86	53.08	55.55	55.77
2013	1.18	2.75	7.39	14.76	34.25	36.22	37.04
2014	0.55	3.49	8.89	23.74	38.59	40.70	40.70
2015	0.61	0.90	3.94	6.10	22.74	24.12	24.12

As mentioned above, the two tables are only a single iteration of the modelling. To compare the Chain Ladder method and the Kaminsky approach we are going the use bootstrapping, see Efron and Tibshirani (1993) and Devore and Berk (2007). With this method it is possible to see how well the Chain Ladder method and the Kaminsky approach estimate the reserve for different reporting years.

3.4 Comparing the data study results

The method used to estimate outstanding liabilities is called parametric bootstrapping. We will use the estimates gathered in Table 3.3 and Table 3.11 to produce a new dataset. Using this new dataset we are going to predict the lower triangle and estimate the reserve using the Chain Ladder method and the Kaminsky approach. This will be done several times until we have 1000 mean reserve estimates for both methods. It is then possible to find the variability in the estimate and hopefully determine which model is preferable. The R-code can be found in Appendix C. Non-parametric bootstrapping could have been used if the individual data was available. We then could have sampled from the individual data for each reporting and development year combination and produce a upper triangle. With the aggregated data this is not possible.

Below we have the bootstrap results for the fire and car insurance. In Table 3.15 we see one of the trends: that the Kaminsky approach keeps underestimating the reserve while the Chain Ladder method overestimates. It also seems like the standard deviation for the Kaminsky approach is somewhat higher than for the Chain Ladder method. We should expect the standard deviation to increase with the reporting years. This is because there is more to predict and fewer constants for high i's.

Table 3.15: *Kaminsky and Chain Ladder projections (in million NOK) for the fire insurance data*

Reporting	Kaminsky			Chain Ladder		
year	Estimate	Bias	Sd	Estimate	Bias	Sd
2011	0.001	-0.007	0.0004	0.008	0.001	0.001
2012	3.894	-0.348	0.209	4.429	0.151	0.137
2013	14.186	-0.401	0.320	15.018	0.421	0.237
2014	59.266	-0.753	0.532	62.534	2.515	0.765
2015	58.852	-0.483	0.541	59.970	0.964	0.496

For 2014 the standard deviation for the Chain Ladder method jumps to 0.765 with a bias of 2.52 million NOK. After running the program several times, and by increasing the portfolio number, it still jumps quite high in 2014. The reason for this sudden jump can be found in Table 3.1 and Table 3.4. In 2014, 599 claims were settled, which is high compared to the others for the same development year. The cost of those 599 claims were 11,38 million NOK, which means that each claim cost about 19 000 NOK in average. If we compare this to 2013 and 2015, they cost around 32

000 and 24 000 NOK in average respectively. In other words, the 599 claims were relatively cheap compared to the others. The effect of the amount of claims that were settled in 2014 is picked up by the delay probability, while the fact that the claims were relatively small is not picked up by the delay-dependent mean. It estimates the average cost of a claim being settled within a year around 30 000 NOK. It should be noted that the σ_k is high for the last k 's and could also be a reason why the bias and standard deviation is high. When performing a parametric bootstrap we get an outlier in this very cell. This affects the rest of the prediction for 2014 in the sense that the Chain Ladder method uses the accumulated data. We see that this problem does not affect the Kaminsky approach, which has a relatively small standard deviation.

This problem is also present for the car insurance data, see Table 3.16, but it is not so evident. It was addressed when implementing the car insurance data for the Chain Ladder method. In 2014 there were 150 claims settled while they only cost 4 000 NOK each in average. The delay-dependent mean estimated the claims to cost around 10 000 each in average if the claim was settled within a year. It should be noted that in both these cases, non-parametric bootstrapping would have been preferable if individual data was available. Since we are sampling from the individual data for the different combinations of reporting and development year, we would not have gotten any outliers.

Table 3.16: *Kaminsky and Chain Ladder projections (in million NOK) for the car insurance data*

Reporting	Kaminsky			Chain Ladder		
year	Estimate	Bias	Sd	Estimate	Bias	Sd
2010	0.011	-0.035	0.002	0.051	0.006	0.002
2011	0.168	-0.113	0.009	0.297	0.016	0.010
2012	1.904	-0.414	0.030	2.114	0.069	0.029
2013	4.807	-0.144	0.049	5.110	0.160	0.060
2014	14.154	-0.245	0.090	14.92	0.523	0.161
2015	13.197	-0.189	0.079	14.133	0.748	0.219

Compared to the previous dataset we see that the standard deviations are almost the same except for the last 3 years. We see that for the last 3 years, the standard deviation is high compared to the Kaminsky standard deviation. The bias is also high for the Chain Ladder for the 2 last reporting years. The variance of the individual losses is somewhat high for the last development years compared to the former, but not as high in the fire insurance data. This could be in play and cause the standard deviation for the Chain Ladder method to be high. This does not seem to affect the results of the Kaminsky approach, as it has a somewhat low bias. The reason could also be that outliers are created when producing the new dataset with the given estimators.

There are some complications here, and different factors are shadowing the true results that we are after. We are going to do a numerical study where we produce our own dataset. We will test how individual claim variance seem to affect the two models and how they perform when we increase the portfolio number.

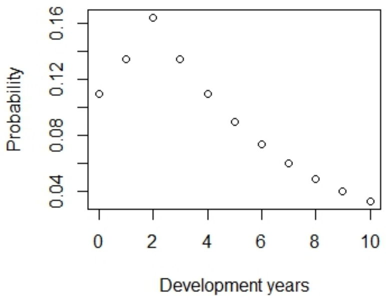

Figure 3.2: *Delay probabilities with a peak at development year 2*

3.5 Numerical study

We will preform a Monte Carlo simulation study with 10 000 simulations. As mentioned above the purpose will be to examine the error while we know perfectly well the underlying situation. The portfolio number will also be increased to see if it has an affect on how these two models preform. We are going to define q_k the same way as Bølviken (2014): $q_k = ce^{-\gamma|k-k_m|}$ for $k = 0, \cdots, K$ where c ensures that $q_0 + \cdots + q_K = 1$. Bølviken (2014) mentions that for the dataset he encountered, most of the claims did not get reported early on, but a few years after they incurred. The dataset he used was an IBNR dataset. In our case, which is a RBNS case, we see that most of claims are settled within a year. For the car insurance we see that the delay probabilities actually plateau for development year 1 and 2 before decreasing rapidly. We have chosen the parameters $\gamma = 0.2$ and $k_m = 2$ which means that the sequence $q_0 + \cdots + q_K$ reaches a peak after 2 years as seen in Figure 3.2. In other words, most of the claims are settled within two years.

$K = I = 10$ means that we have 11 years of historical data and 11 development years. The portfolio number is increased by increasing the number of claims that are reported each reporting year. The number of claims reported was set at 250 000 for $i = 0, \cdots, 10$. The simulation was run twice. Once, when the individual losses were exponentially distributed, which is a special case of the Gamma distribution when the shape parameter $\alpha_k = 1$ for all k's. The second time, it was run when the individual losses were Gamma distributed and $\alpha_k = 0.5$ for $k = 0, \cdots, 10$ to simulate when the claim size distribution is heavy-tailed. It should be mentioned that both of them are heavy-tailed, but the latter one has both a heavier and a longer tail. We want to see how both models fare when the individual claim variance is quite high. In both cases the mean was $\xi_k = 100$ for all k's as if the average of all past losses at delay k was 100 NOK.

By using the parameters mentioned above we simulated a full dataset of claim numbers and corresponding claim amounts. In other words, the upper and lower triangle were simulated. By using the upper triangle, which is the known observations from a reserving point of view, we simulated the lower triangle using both methods. By doing so we could obtain estimates for the outstanding liabilities for both the Chain Ladder method and the Kaminsky approach. This was done 10 000 times. The result of the study are shown in Figure 3.3 while the uncertainty and the bias in their estimates for the exponential losses and the heavy-tailed losses respectively can be found in Table 3.17 and Table 3.18.

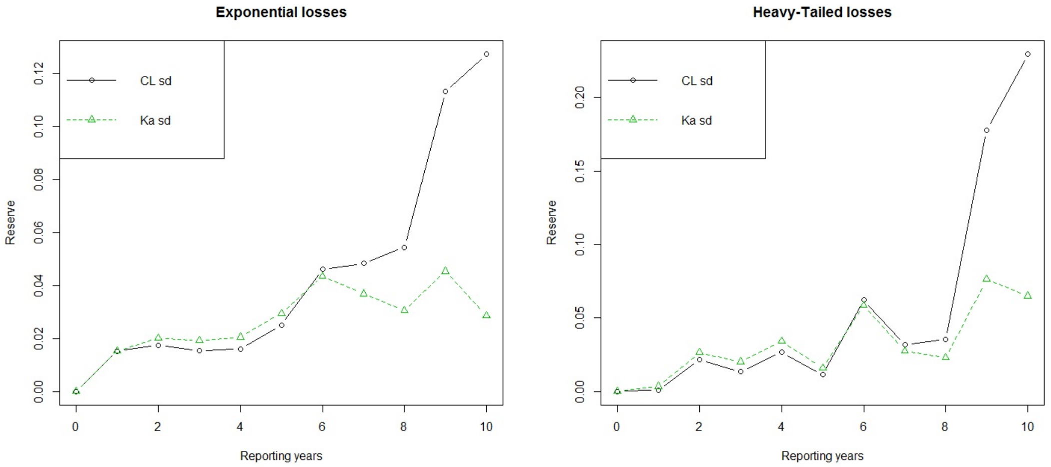

Figure 3.3: *The standard deviation in the estimates for the outstanding liabilities for the Chain Ladder method and the Kaminsky approach*

From the figure above it is clear that from development year 6 and up, the uncertainties in the Kaminsky approach are lower than for the Chain Ladder method. It seems as if the Kaminsky approach tackles increasing the portfolio number and introducing heavy-tailed claim size distribution better. The results from the figure above are reminiscent of and more evident than the results found in the car insurance data in Table 3.16. The uncertainty is bigger and grows more rapidly for the Chain Ladder method compared to the Kaminsky approach. We can take a closer look at the results in Table 3.17 and Table 3.18 for exponential distributed claim losses and heavy-tailed distributed losses respectively.

For the exponential losses, the bias is small and somewhat similar for both the models. One might notice that the bias and the standard deviation for the Chain Ladder method grows for the last 4 years.

The Kaminsky approach seems to be unaffected by the exponential losses. The standard deviation is stable for all reporting years though it is higher than the Chain Ladder method for the first 5 reporting years. The standard deviation for the Kaminsky approach is stable, and therefore the bias is stable as well and does not seem to grow with the reporting year as for the Chain Ladder method.

Table 3.17: *Results from the numerical study with exponential losses*

	Exponential losses					
	Chain Ladder			Kaminsky		
Year	Estimate	Bias	Sd	Mean estimate	Bias	Sd
1	0.661	0.000	0.015	0.661	0.000	0.015
2	1.397	0.000	0.017	1.381	0.001	0.020
3	1.714	0.001	0.015	1.714	0.001	0.019
4	2.069	0.001	0.016	2.068	0.000	0.020
5	4.073	0.000	0.025	4.075	0.002	0.029
6	8.363	-0.003	0.046	5.359	-0.004	0.043
7	8.123	-0.003	0.048	8.129	0.003	0.037
8	6.652	-0.004	0.054	6.655	-0.001	0.030
9	13.537	0.002	0.113	13.534	-0.001	0.045
10	7.272	0.004	0.127	7.263	0.000	0.028

The Kaminsky approach seems to perform a lot better than the Chain Ladder method for heavy-tailed claim loss distribution. In Table 3.18 we can observe that the bias is somewhat close to zero for all reporting years for the Kaminsky approach. Again, we see that the standard deviation is very stable, unlike the Chain Ladder method. The standard deviation seems to grow for the Chain Ladder method as the reporting years move toward reporting year 10. One should notice that the bias is quite low for the Chain Ladder method, but not as low as for the Kaminsky approach.

Table 3.18: *Results from the numerical study with heavy-tailed losses*

	Heavy-Tailed losses					
	Chain Ladder			Kaminsky		
Year	Estimate	Bias	Sd	Mean estimate	Bias	Sd
1	0.036	0.000	0.001	0.036	0.000	0.003
2	1.401	0.000	0.026	1.401	-0.001	0.026
3	1.083	0.001	0.013	1.082	0.000	0.020
4	2.898	0.001	0.027	2.897	0.000	0.034
5	0.904	0.000	0.011	0.904	0.000	0.016
6	8.928	0.001	0.062	8.925	0.000	0.058
7	2.794	0.000	0.032	2.794	0.000	0.028
8	2.068	0.000	0.035	2.067	0.000	0.023
9	17.712	0.000	0.178	18.712	0.000	0.076
10	14.447	0.004	0.230	14.446	0.003	0.065

It seems as if the Kaminsky approach is performing better than the Chain Ladder method under these conditions. While the distinction between these two models uncertainty in their predictions are quite on par until reporting year 6. The question is if this distinction will be even more clear if we further increase the number of policies? Will the difference between the uncertainties and the bias for the corresponding methods grow with more volatile loss distributions? These questions will be answered in the next chapter, where we try to acquire an approximate expression for these uncertainties when the number of policies increase toward infinity.

Chapter 4

Error analysis for RBNR

4.1 The approach

In the numerical study we observed that when the number of policies in a portfolio were increased, the variance in individual losses had a big impact on the uncertainty in the prediction of the outstanding liabilities for the Chain Ladder method. When compared to the Chain Ladder method, the uncertainty in the Kaminsky prediction of the outstanding liabilities seemed to hardly be affected at all by the individual claim variance. The bias was also smaller for the Kaminsky approach compared to the the Chain Ladder method. In this chapter both models will be compared analytically. We will investigate if it is a general rule that the Kaminsky approach is superior to the Chain Ladder method. In other words, we will investigate if modelling the claim frequency and claim size separately improves the accuracy of the estimate of outstanding liabilities.

Mack (1993) introduced the Markov-like assumptions, which can be found in the equations in (2.4), to be able to quantify the Chain Ladder error. The Chain Ladder method has a multiplicative structure as seen in equation (2.9), while the Kaminsky approach has an additive structure as seen in equation (2.28). The way Mack (1993) quantified the error will not suffice, due to the additive structure of the Kaminsky approach.

This is the reason behind choosing the mean square error (MSE) to compare the two models. We have that $\hat{R}_i$ fitted on past observations is independent of the future R_i. This gives us:

$$\mathbb{E}[(\hat{R}_i - R_i)^2] = \mathbb{V}\mathrm{ar}(\hat{R}_i) + \mathbb{V}\mathrm{ar}(R_i) + (\mathbb{E}[\hat{R}_i] - \mathbb{E}[R_i])^2$$

where $\mathbb{V}\mathrm{ar}(R_i)$ is unaffected by how we perform the RBNS reserve calculations and will be called the unpredictable error. The mathematical expression for $\mathbb{V}\mathrm{ar}(R_i)$ can be found below. The calculations for variance and the covariance in the expression can be found in Section 4.3.

$$\begin{aligned}\mathbb{V}\text{ar}(R_i) &= \mathbb{V}\text{ar}(\sum_{k=I-i+1}^{K} X_{i,k}) \\ &= \sum_{k=I-i+1}^{K} \mathbb{V}\text{ar}(X_{i,k}) + \sum_{k=I-i+1}^{K} \sum_{k\neq k'} cov(X_{i,k}, X_{i,k'}) \\ &= \sum_{k=I-i+1}^{K} (q_k\xi_k^2 + q_k\sigma_k^2)n_i - (\sum_{k=I-i+1}^{K} q_k\xi_k)^2 n_i\end{aligned}$$

It is already known that the estimate for the outstanding liabilities in the Kaminsky approach is an unbiased estimator for the outstanding liabilities. It is also shown in Section 4.3 that for large portfolio $\mathbb{E}[R_i^{CL}] \approx E[R_i]$. In other words the bias term, $\mathbb{E}[\hat{R}_i] - \mathbb{E}[R_i]$, will become zero when the number of policies in a portfolio increase toward infinity. Since $\mathbb{V}\text{ar}(R_i)$ is unaffected by how we perform the modelling of the reserve, our main concern will be to compare the uncertainties in the prediction of the outstanding liabilities for both approaches. It is impossible to obtain a closed and exact formula for these uncertainties. Because of this we will try to approximate them for large portfolios, i.e. we will let $n_i \to \infty$.

4.2 Large portfolio approximation

One knows from the central limit theorem that both R_i and $\hat{R}_i$ becomes normally distributed for a large portfolio. In this section the key results will be presented, while the lengthy calculations will be presented in section 4.3.

For the Chain Ladder we introduce:

$$a_k = \sum_{l=0}^{k} q_l\xi_l\,, \quad b_k = \sum_{l=0}^{k} q_l[\xi_l^2 + \sigma_l^2] \tag{4.1}$$

and

$$c_k = b_{k-1}(\frac{a_K}{a_{k-1}} - 1)^2 - b_k(\frac{a_K}{a_k} - 1)^2\,, \quad d_k = b_k - a_k^2 \tag{4.2}$$

for $k = 0, \cdots, K$. We then have that $\mathbb{V}\text{ar}(\hat{R}_i^{cl})$ approximately becomes:

$$\widetilde{\mathbb{V}\text{ar}}(\hat{R}_i^{CL}) = n_i^2 \sum_{k=I-i+1}^{K} \eta_k^{-1}(c_k + b_k - b_{k-1}) + n_i d_{I-i}(\frac{a_K}{a_{I-i}} - 1)^2. \tag{4.3}$$

The stochastic remainder term $o(\sqrt{n_i})$ is, as one may see in equation (4.31) for $\mathbb{V}\text{ar}(\hat{R}_i^{cl})$ in Section 4.3, dominated by $\sqrt{n_i}$ in the sense that $\frac{o(\sqrt{n_i})}{\sqrt{n_i}} \to 0$ in some stochastic limit as $n_i \to \infty$. There will also be other remainder terms in the calculations in section 4.3, but they will be treated informally. For example, through the limit process where $n_0 = \omega_0\bar{n}, \dots n_K = \omega_K\bar{n}$ where $\bar{n} \to \infty$. Here we have that $\omega_0, \dots, \omega_K$ are fixed and positive weights such that $\omega_0 + \omega_1 + \cdots + \omega_K = 1$. It gives us that $\widetilde{\mathbb{V}\mathrm{ar}}(\hat{R}_i^{cl}) / \mathbb{V}\mathrm{ar}(\hat{R}_i^{cl}) \to 1$ as $n_i \to \infty$.

The variance between the individual losses, $Z_{i,k}$, is denoted by σ_k^2. Both Bølviken (2015) and Wütherich and Merz (2008) discusses the connection between the Poisson version of the Burnhuetter-Ferguson method and the Chain Ladder method. It is shown that both methods are equivalent. For a simple proof see Bølviken (2015). Bølviken (2015) further discusses the consequence of this equivalence. When there is no uncertainty around the estimation of the delay-dependent means, ξ_k, i.e. when $\sigma_k = 0$ for all k we have that the Kaminsky predictions coincide with the Burnhuetter-Ferguson ones, which again coincide with the Chain Ladder method.

This makes it possible to take a sizeable step towards a similar approximation $\widetilde{\mathbb{V}\mathrm{ar}}(\hat{R}_i^{Ka})$ for the Kaminsky variance. By inserting the new condition, that $\sigma_k = 0$ for all k, in the expressions above we get:

$$b_k^{(0)} = \sum_{l=0}^{k} q_l \xi_l^2\,, \quad c_k^{(0)} = b_{k-1}^{(0)}(\frac{a_K}{a_{k-1}} - 1)^2 - b_k^{(0)}(\frac{a_K}{a_k} - 1)^2 \quad and \quad d_k^{(0)} = b_k^{(0)} - a_k^2 \tag{4.4}$$

for $k = 0, \cdots, K$. When the error in the estimate $\hat{\xi}_k$ is lead through a linearization argument in Section 4.3 we have that:

$$\widetilde{\mathbb{V}\mathrm{ar}}(\hat{R}_i^{Ka}) = n_i^2 \sum_{k=I-i+1}^{K} \eta_k^{-1}(c_k^{(0)} + b_k - b_{k-1}) + n_i d_{I-i}^{(0)}(\frac{a_K}{a_{I-i}} - 1)^2 \tag{4.5}$$

where $\eta_k = \sum_{i=k}^{I} n_i$. Again we have that $\widetilde{\mathbb{V}\mathrm{ar}}(\hat{R}_i^{Ka}) / \mathbb{V}\mathrm{ar}(\hat{R}_i^{Ka}) \to 1$ as $n_i \to \infty$.

Equation (4.3) and (4.5) are very similar, and it is therefore possible to study their differences analytically. By taking the difference between these two equations we get:

$$\begin{aligned}\widetilde{\mathbb{V}\mathrm{ar}}(\hat{R}_i^{CL}) - \widetilde{\mathbb{V}\mathrm{ar}}(\hat{R}_i^{Ka}) &= n_i^2 \sum_{k=I-i+1}^{K} \eta_k^{-1}(c_k + b_k - b_{k-1}) + n_i d_{I-1}(\frac{a_K}{a_{I-i}} - 1)^2 \\ &\quad - n_i^2 \sum_{k=I-i+1}^{K} \eta_k^{-1}(c_k^{(0)} + b_k - b_{k-1}) - n_i d_{I-i}^{(0)}(\frac{a_K}{a_{I-i}} - 1)^2 \\ &= n_i(d_{I-i} - d_{I-i}^{(0)})(\frac{a_K}{a_{I-i}} - 1)^2 + n_i^2 \sum_{k=I-i+1}^{K} \eta_k^{-1}(c_k - c_k^{(0)}) \\ &= A_{1,i} + A_{2,i} \end{aligned} \tag{4.6}$$

where

$$A_{1,i} = n_i(d_{I-i} - d_{I-i}^{(0)})(\frac{a_K}{a_{I-i}} - 1)^2 \quad and \quad A_{2,i} = n_i^2 \sum_{k=I-i+1}^{K} \eta_k^{-1}(c_k - c_k^{(0)}). \tag{4.7}$$

Starting with $A_{1,i}$, because it is the easiest term to work with, we see that $A_{1,i} \geq 0$ since $d_k^{(0)} \leq d_k$ because $b_k^{(0)} \leq b_k$. $A_{2,i}$ is a bit more complicated to work with, but by inserting the definition of c_k and $c_k^{(0)}$ into $A_{2,i}$ we get that:

$$\begin{aligned}
A_{2,i} &= n_i^2 \sum_{k=I-i+1}^{K} \eta_k^{-1}(c_k - c_k^{(0)}) \\
&= n_i^2 \sum_{k=I-i+1}^{K} \eta_k^{-1}(b_{k-1}(\frac{a_K}{a_{k-1}} - 1)^2 - b_k(\frac{a_K}{a_k} - 1)^2 - b_{k-1}^{(0)}(\frac{a_K}{a_{k-1}} - 1)^2 + b_k^{(0)}(\frac{a_K}{a_k} - 1)^2) \\
&= n_i^2 \sum_{k=I-i+1}^{K} \eta_k^{-1}(b_{k-1} - b_{k-1}^{(0)})(\frac{a_K}{a_{k-1}} - 1)^2 - n_i^2 \sum_{k=I-i+1}^{K} \eta_k^{-1}(b_k - b_k^{(0)})(\frac{a_K}{a_k} - 1)^2 \\
&= n_i^2 \eta_{I-i+1}^{-1}(b_{I-i} - b_{I-i}^{(0)})(\frac{a_K}{a_{I-i}} - 1)^2 + n_i^2 \sum_{k=I-i+2}^{K} \eta_k^{-1}(b_{k-1} - b_{k-1}^{(0)})(\frac{a_K}{a_{k-1}} - 1)^2 \\
&- n_i^2 \sum_{k=I-i+1}^{K} \eta_k^{-1}(b_k - b_k^{(0)})(\frac{a_K}{a_k} - 1)^2 \\
&\overset{\text{a}}{=} n_i^2 \eta_{I-i+1}^{-1}(b_{I-i} - b_{I-i}^{(0)})(\frac{a_K}{a_{I-i}} - 1)^2 + n_i^2 \sum_{k=I-i+1}^{K-1} \eta_{k+1}^{-1}(b_k - b_k^{(0)})(\frac{a_K}{a_k} - 1)^2 \\
&- n_i^2 \sum_{k=I-i+1}^{K-1} \eta_k^{-1}(b_k - b_k^{(0)})(\frac{a_K}{a_k} - 1)^2 \\
&= n_i^2 \eta_{I-i+1}^{-1}(b_{I-i} - b_{I-i}^{(0)})(\frac{a_K}{a_{I-i}} - 1)^2 + n_i^2 \sum_{k=I-i+1}^{K-1} (b_k - b_k^{(0)})(\frac{a_K}{a_k} - 1)^2(\eta_{k+1}^{-1} - \eta_k^{-1})
\end{aligned} \tag{4.8}$$

again by the virtue of $b_k^{(0)} \leq b_k$ and $\eta_{k+1} < \eta_k$ we have that $A_{2,i} \geq 0$. It follows that $\widetilde{\mathbb{V}\mathrm{ar}}(\hat{R}_i^{Ka}) \leq \widetilde{\mathbb{V}\mathrm{ar}}(\hat{R}_i^{CL})$. It is shown here that the Kaminsky approach is more accurate in estimating the outstanding liabilities than the Chain Ladder method for large portfolios regardless of claim frequency and claim size distribution. This result coincides with the results from the the numerical study and the data study for the car insurance data.

It would be interesting to calculate $A_{1,i}$ and $A_{2,i}$ when the parameters are the same as in the numerical study where $\xi_k = \xi$ and $\sigma_k = \sigma$ for all k . In other words, when the average of all past losses per event and when the variation is the same independent of development year k , then $a_k = \sum_{l=0}^{k} q_l \xi = Q_k \xi$,

$$\begin{aligned}
d_i - d_i^{(0)} &= b_i - b_i^{(0)} \\
&= Q_k(\xi^2 + \sigma^2) - Q_k \xi^2 \\
&= Q_k \sigma^2
\end{aligned}$$

[a] Here we have used that $\sum_{k=I-i+1}^{K} \eta_k^{-1}(b_k - b_k^{(0)})(\frac{a_K}{a_k} - 1)^2 = \sum_{k=I-i+1}^{K-1} \eta_k^{-1}(b_k - b_k^{(0)})(\frac{a_K}{a_k} - 1)^2$ since the last term in the sum will always be zero.

and

$$\begin{aligned}
c_k - c_k^{(0)} &= (b_{k-1} - b_{k-1}^{(0)})(\frac{a_K}{a_{k-1}} - 1)^2 - (b_k - b_k^{(0)})(\frac{a_K}{a_k} - 1)^2 \\
&= Q_{k-1}\sigma^2(\frac{Q_K\xi}{Q_{k-1}\xi} - 1)^2 - Q_k\sigma^2(\frac{Q_K\xi}{Q_k\xi} - 1)^2 \\
&\overset{\text{b}}{=} \sigma^2(\frac{1}{Q_{k-1}} - \frac{1}{Q_k} + Q_{k-1} - Q_k) \\
&= \sigma^2(\frac{Q_k - Q_{k-1}}{Q_{k-1}Q_k} + Q_{k-1} - Q_k) \\
&= q_k(\frac{1}{Q_{k-1}Q_k} - 1)\sigma^2
\end{aligned}$$

where $Q_k = q_0 + \cdots q_k$ is the distribution function of the delay. It follows then that

$$A_{1,i} = n_i Q_{I-i}(\frac{1}{Q_{I-i}} - 1)^2\sigma^2 \quad \textit{and} \quad A_{2,i} = n_i^2 \sum_{k=I-i+1}^{K} \eta_k^{-1} q_k(\frac{1}{Q_{k-1}Q_k} - 1)\sigma^2. \tag{4.9}$$

The beauty of the equations in (4.9) is that both $A_{1,i}$ and $A_{2,i}$ are proportional to σ^2. This shows us that the more volatile the loss distribution is, the more superior the Kaminsky approach is compared to the Chain Ladder method. When the variation in claims per event is large it is far better to solve the problem by breaking the predictions down into claim losses and claim counts. The expressions for $A_{1,i}$ and $A_{2,i}$ and the difference in $d_i - d^{(0)}$ shows that this is a general phenomenon. This was also observed in the numerical study that it seem as if the Kaminsky approach fare for more volatile claim size distribution.

4.3 Mathematical arguments

In this section we will thoroughly go through the lengthy calculations that were skipped in the preceding section. First, we will tackle the Chain Ladder asymptotics. The section below is divided into three parts: *Preliminaries, the Chain Ladder coefficients* and *mean and variance of* $\hat{R}_i^{CL}$.

In the first part we will derive some preliminary results needed for the two other parts. In the second part our goal will be to derive an expression for the $\hat{f}_k$. Using the results from the two previous parts, we will derive an expression for the mean and the variance of outstanding losses for the Chain Ladder method in the third and last part. In section 4.3.2 we will find an expression for the variance of the outstanding losses for the Kaminsky approach.

[b] Here we have used that Q_K by definition is equal to 1 and rewritten the quadratic terms into standard forms

4.3.1 The Chain Ladder asymptotics

Preliminaries

Recall that each aggregate $X_{i,k}$ becomes normally distributed as $n_i \to \infty$ which is a consequence of the central limit theorem. It is also known that each $N_{i,k}$ is multinomial given $N_i = n_i$ with expectation and variance as given in (2.10). Using these results we can also derive the expectation, as given in (2.25), variance and covariance for $X_{i,k}$:

$$\begin{aligned}\mathbb{V}\mathrm{ar}(X_{i,k}) &= \mathbb{V}\mathrm{ar}(\mathbb{E}[X_{i,k}|N_{i,k}]) + \mathbb{E}[\mathbb{V}\mathrm{ar}(X_{i,k}|N_{i,k})] \\ &= \mathbb{V}\mathrm{ar}(\xi_k N_{i,k}) + \mathbb{E}[\sigma_k^2 N_{i,k}] \\ &= \xi_k^2 n_i q_k(1-q_k) + \sigma_k^2 n_i q_k \\ &= n_i q_k[\xi_k^2(1-q_k) + \sigma_k^2]. \end{aligned} \tag{4.10}$$

With this result we can find the covariance between $X_{i,k}$ and $X_{i,l}$. We will have to use that $cov(N_{i,k}, N_{i,l}) = -n_i q_k q_l$ since $n_{i,k}$'s are multinomial distributed.

$$\begin{aligned}cov(X_{i,k}, X_{i,l}) &= cov(\mathbb{E}[X_{i,k}|N_{i,l}]\,\mathbb{E}[X_{i,l}|N_{i,l}]) + \mathbb{E}[cov(X_{i,k}, X_{i,l}|N_{i,k}N_{i,l})] \\ &\overset{\mathrm{c}}{=} cov(\xi_k N_{i,k}, \xi_l N_{i,l}) + \mathbb{E}[0] \\ &= \xi_k \xi_l cov(N_{i,k}, N_{i,l}) \\ &= -\xi_k \xi_l n_i q_k q_l \qquad \text{if } k \neq l. \end{aligned} \tag{4.11}$$

Now that we know the expectation, variance and the covariance of $X_{i,k}$ we can compute the expectation and the variance for the cumulative losses $C_{i,k} = X_{i,0} + \cdots + X_{i,k}$.

$$\mathbb{E}[C_{i,k}] = \mathbb{E}[X_{i,0} + \cdots + X_{i,k}] = (q_0\xi_0 + \cdots q_k\xi_k)n_i = a_k n_i \tag{4.12}$$

[c] The covariance is 0 in RBNS since the counts $N_{i,k}$ are observed the sizes of the claims are independent from year to another

and

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(C_{i,k}) &= \sum_{j=0}^{k}\mathbb{V}\mathrm{ar}(X_{i,j}) + \sum_{j=0}^{k}\sum_{j'\neq j} cov(X_{i,j}, X_{i,j'}) \\
&= \sum_{j=0}^{k}(q_j(1-q_j)\xi_j^2 + q_j\sigma_j^2)n_i - \sum_{j=0}^{k}\sum_{j'\neq j} q_j q_{j'}\xi_j\xi_{j'}n_i \\
&= \sum_{j=0}^{k}(q_j\xi_j^2 + q_j\sigma_j^2)n_i - (\sum_{j=0}^{k}\sum_{j'\neq j, j'=0}^{k} q_j q_{j'}\xi_j\xi_{j'} + \sum_{j=0}^{k} q_j^2\xi_j^2)n_i \\
&= \sum_{j=0}^{k}(q_j\xi_j^2 + q_j\sigma_j^2)n_i - \sum_{j=0}^{k}\sum_{j'=0}^{k} q_j q_{j'}\xi_j\xi_{j'}n_i \\
&= b_k n_i - (\sum_{j=0}^{k} q_j\xi_j)^2 n_i \\
&= (b_k - a_k^2)n_i \\
&= d_k n_i \qquad (4.13)
\end{aligned}$$

where a_k, b_k and d_k were defined in (4.1) and (4.2) right. The variance between the individual losses, $Z_{i,k}$, is denoted by σ_k^2.

With the acquired expressions for both the expectation and the variance of $C_{i,k}$ we can express the cumulative losses $C_{i,k}$ in (2.3) as such:

$$C_{i,k} = n_i a_k + \sqrt{n_i d_k}\epsilon_{i,k} + o(\sqrt{n_i}). \qquad (4.14)$$

As mentioned earlier, the stochastic remainder term $o(\sqrt{n_i})$ is dominated by $\sqrt{n_i}$ in the sense that $\frac{o(\sqrt{n_i})}{\sqrt{n_i}} \to 0$ in some stochastic limit as $n_i \to \infty$. There will be other remainder terms below, but we will treat them informally. For example, through the limit process where $n_0 = \omega_0\bar{n}, \ldots n_K = \omega_K\bar{n}$ where $\bar{n} \to \infty$. Again, we have that $\omega_0, \ldots, \omega_K$ are fixed and positive weights such that $\omega_0 + \omega_1 + \cdots + \omega_K = 1$.

The Chain Ladder coefficients

All the $\epsilon_{i,k}$ in (4.14) are N(0,1) and are independent between accident years i , but the aggregates $X_{i,k}$'s are correlated in $C_{i,k}$. Before we can calculate $cor(C_{i,k}, C_{i,l}) = cor(\epsilon_{i,k}, \epsilon_{i,l})$ we need to calculate the covariance between $C_{i,k}$ and $C_{i,l}$.

Suppose that $k \leq l$, then:

$$\begin{aligned}
cov(C_{i,k}, C_{i,l}) &= \sum_{j=0}^{k}\sum_{j'=0}^{l} cov(X_{i,j}, X_{i,j'}) \\
&= \sum_{j=0}^{k} \mathbb{V}\mathrm{ar}(X_{i,j}) + \sum_{j=0}^{k}\sum_{j'\neq j, j'=0}^{l} cov(X_{i,j}, X_{i,j'}) \\
&= \sum_{j=0}^{k}(q_j(1-q_j)\xi_j^2 + q_j\sigma_j^2)n_i - \sum_{j=0}^{k}\sum_{j'\neq j, j'=0}^{l} q_j q_{j'}\xi_j\xi_{j'}n_i \\
&= \sum_{j=0}^{k}(q_j\xi_j^2 + q_j\sigma_j^2)n_i - (\sum_{j=0}^{k}\sum_{j'\neq j, j'=0}^{l} q_j q_{j'}\xi_j\xi_{j'} + \sum_{j=0}^{k} q_j^2\xi_j^2)n_i \\
&= b_k n_i - (\sum_{j=0}^{k} q_j\xi_j)(\sum_{j'=0}^{l} q_{j'}\xi_{j'})n_i \\
&= (b_k - a_k a_l)n_i. \qquad (4.15)
\end{aligned}$$

With the expression for the covariance, we can calculate the correlation between the aggregates $C_{i,k}$'s.

$$\begin{aligned}
\rho_{k,l}^{\epsilon} = cor(\epsilon_{i,k}, \epsilon_{i,l}) &= \frac{cov(C_{i,k}, C_{i,l})}{\sqrt{\mathbb{V}\mathrm{ar}(C_{i,k})\,\mathbb{V}\mathrm{ar}(C_{i,l})}} \\
&= \frac{(b_k - a_k a_l)}{\sqrt{d_k d_l}}.
\end{aligned}$$

We then have that:

$$\rho_{k,l}^{\epsilon} = \begin{cases} \frac{b_k - a_k a_l}{\sqrt{d_k d_l}} & \text{if } k \leq l \\ \frac{b_l - a_l a_k}{\sqrt{d_l d_k}} & \text{if } l > k \end{cases} \qquad (4.16)$$

$\rho_{k,l}^{\epsilon}$ will be useful later on, but now we can find an expression for $\hat{f}_k$ by inserting the expression for $C_{i,k}$ in (4.14) into (2.6). We then get:

$$\hat{f}_k = \frac{\sum_{i=k}^{I}[n_i a_k + \sqrt{n_i d_k}\epsilon_{i,k} + o(\sqrt{\bar{n}})]}{\sum_{i=k}^{I}[n_i a_{k-1} + \sqrt{n_i d_{k-1}}\epsilon_{i,k-1} + o(\sqrt{\bar{n}})]}$$

for $k = 1, \cdots, I$. With further manipulation we get:

$$\hat{f}_k = \frac{a_k + \eta_k^{-\frac{1}{2}} d_k^{-\frac{1}{2}}\delta_{1,k} + o(\bar{n}^{-\frac{1}{2}})}{a_{k-1} + \eta_k^{-\frac{1}{2}} d_{k-1}^{-\frac{1}{2}}\delta_{2,k} + o(\bar{n}^{-\frac{1}{2}})} \qquad (4.17)$$

where

$$\eta_k = \sum_{i=k}^{I} n_i\,, \quad \delta_{1,k} = \sum_{i=k}^{I}\frac{n_i^{\frac{1}{2}}}{\eta_k^{\frac{1}{2}}}\epsilon_{i,k}\,, \quad and \quad \delta_{2,k} = \sum_{i=k}^{I}\frac{n_i^{\frac{1}{2}}}{\eta_k^{\frac{1}{2}}}\epsilon_{i,k-1}. \qquad (4.18)$$

By taking a closer look at $\delta_{1,k}$ and $\delta_{2,k}$, we see that since $\epsilon_{1,k}$'s are N(0,1) distributed,

$$\mathbb{E}[\delta_{1,k}] = \mathbb{E}[\delta_{2,k}] = \sum_{i=k}^{I} \frac{n_i^{\frac{1}{2}}}{\eta_k^{\frac{1}{2}}} \mathbb{E}[\epsilon_{i,k-1}] = 0$$
$$\mathbb{V}\text{ar}(\delta_{1,k}) = \mathbb{V}\text{ar}(\delta_{2,k}) = \sum_{i=k}^{I} \frac{n_i}{\eta_k} \mathbb{V}\text{ar}(\epsilon_{i,k-1}) = \frac{\eta_k}{\eta_k} = 1$$

$\delta_{1,k}$ and $\delta_{2,k}$ are also standard normal. It is now possible calculate the correlation between $\delta_{i,k}$ and $\delta_{2,k}$.

$$\begin{aligned} cor(\delta_{1,k}, \delta_{1,l}) &= \frac{\mathbb{E}[\delta_{1,k}\delta_{1,l}] - \mathbb{E}[\delta_{1,k}]\,\mathbb{E}[\delta_{1,l}]}{\sqrt{\mathbb{V}\text{ar}(\delta_{1,k})\,\mathbb{V}\text{ar}(\delta_{1,l})}} \\ &= \mathbb{E}[\delta_{1,k}\delta_{1,l}] \\ &= \eta_k^{-\frac{1}{2}}\eta_l^{-\frac{1}{2}}\,\mathbb{E}\left[\sum_{i=k+1}^{I}\sum_{i'=l+1}^{I} n_i^{\frac{1}{2}} n_{i'}^{\frac{1}{2}} \epsilon_{i,k}\epsilon_{i',l}\right] \\ &= \eta_k^{-\frac{1}{2}}\eta_l^{-\frac{1}{2}} \sum_{i=max(k,l)+1}^{I} n_i\,\mathbb{E}[\epsilon_{i,k}\epsilon_{i,l}] \\ &= \eta_k^{-\frac{1}{2}}\eta_l^{-\frac{1}{2}} \sum_{i=max(k,l)+1}^{I} n_i cor(\epsilon_{i,k}, \epsilon_{i,l}) \\ &= \rho^n_{k,l}\rho^\epsilon_{k,l}. \end{aligned}$$

Where we have defined $\rho^n_{k,l}$ to be:

$$\rho^n_{k,l} = \begin{cases} \sqrt{\frac{\eta_l}{\eta_k}} & \text{if } k \le l \\ \sqrt{\frac{\eta_k}{\eta_l}} & \text{if } l > k \end{cases} \tag{4.19}$$

When transitioning from the third to the fourth equality sign, we have used that $\mathbb{E}[\epsilon_{i,k}\epsilon_{i',l}] = 0$ when $i \neq i'$. We have also used that $E[\epsilon_{i,k}\epsilon_{i,l}] = cor(\epsilon_{i,k}, \epsilon_{i,l})$ since, by definition, the expectations and the variances are equal to zero in the transition from the fourth to the fifth equality sign. In summary, the other three correlations will be:

$$cor(\delta_{1,k}, \delta_{1,k}) = \rho^n_{k,l}\rho^\epsilon_{k,l}, \quad cor(\delta_{1,k}, \delta_{2,l}) = \rho^n_{k,l}\rho^\epsilon_{k,l-1} \quad and \quad cor(\delta_{2,k}, \delta_{2,l}) = \rho^n_{k,l}\rho^\epsilon_{k-1,l-1}. \tag{4.20}$$

The expression for $\hat{f}_k$ in (4.17) is hard to work with, and a more compliant expression is needed. One way to get a more compliant expression is to approximate $\hat{f}_k$ through a Taylor expansion centred in $\eta_k^{-\frac{1}{2}}$. This is the same as noticing that (4.17), with minor manipulation, can be expanded into a the geometric series. By only using the linear part of the expansion we get the following expression:

$$
\begin{aligned}
\hat{f}_k &= \frac{\frac{a_k}{a_{k-1}} + \frac{\eta_k^{-\frac{1}{2}} d_k^{\frac{1}{2}} \delta_{1,k}}{a_{k-1}} + o(\bar{n}^{-\frac{1}{2}})}{1 + \frac{\eta_k^{-\frac{1}{2}} d_{k-1}^{\frac{1}{2}} \delta_{2,k}}{a_{k-1}} + o(\bar{n}^{-\frac{1}{2}})} \\
&= \left(\frac{a_k}{a_{k-1}} + \frac{\eta_k^{-\frac{1}{2}} d_k^{\frac{1}{2}} \delta_{1,k}}{a_{k-1}} + o(\bar{n}^{-\frac{1}{2}}) \right) \left(1 - \frac{\eta_k^{-\frac{1}{2}} d_{k-1}^{\frac{1}{2}} \delta_{2,k}}{a_{k-1}} + o(\bar{n}^{-\frac{1}{2}}) \right) \\
&= \frac{a_k}{a_{k-1}} \left(1 + \eta_k^{-\frac{1}{2}} \frac{d_k^{\frac{1}{2}}}{a_k} \delta_{1,k} - \eta_k^{-\frac{1}{2}} \frac{d_{k-1}^{\frac{1}{2}}}{a_{k-1}} \delta_{1,k} - \eta_k^{-\frac{1}{2}} \eta_k^{-\frac{1}{2}} \frac{d_{k-1}^{\frac{1}{2}}}{a_{k-1} a_k} \delta_{1,k} \delta_{2,k} \right) + o(\bar{n}^{-\frac{1}{2}}) \\
&\overset{\text{d}}{=} \frac{a_k}{a_{k-1}} \left(1 + \eta_k^{-\frac{1}{2}} \left(\frac{d_k^{\frac{1}{2}}}{a_k} \delta_{1,k} - \frac{d_{k-1}^{\frac{1}{2}}}{a_{k-1}} \delta_{2,k} \right) \right) + o(\bar{n}^{-\frac{1}{2}}) \\
&= \frac{a_k}{a_{k-1}} \left(1 + \eta_k^{-\frac{1}{2}} Y_k \right) + o(\bar{n}^{-\frac{1}{2}}). \qquad (4.21)
\end{aligned}
$$

Where we have defined Y_k to be:

$$
Y_k = \frac{d_k^{\frac{1}{2}}}{a_k} \delta_{1,k} - \frac{d_{k-1}^{\frac{1}{2}}}{a_{k-1}} \delta_{2,k}. \qquad (4.22)
$$

There is a need to calculate the variance and the covariance of $Y_1, \cdots, Y_K$. Here we will need the results in (4.16), (4.19) and (4.20) to calculate the variance and the covariance.

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}(Y_k) &= \mathbb{V}\mathrm{ar}\Big(\frac{d_k^{\frac{1}{2}}}{a_k} \delta_{1,k} - \frac{d_{k-1}^{\frac{1}{2}}}{a_{k-1}} \delta_{2,k}\Big) \\
&= \frac{d_k}{a_k^2} \, \mathbb{V}\mathrm{ar}(\delta_{1,k}) + \frac{d_{k-1}}{a_{k-1}^2} \, \mathbb{V}\mathrm{ar}(\delta_{2,k}) - 2 \frac{d_k^{\frac{1}{2}} d_{k-1}^{\frac{1}{2}}}{a_k a_{k-1}} cov(\delta_{1,k}, \delta_{2,k}) \\
&= \frac{d_k}{a_k^2} + \frac{d_{k-1}}{a_{k-1}^2} - 2 \frac{d_k^{\frac{1}{2}} d_{k-1}^{\frac{1}{2}}}{a_k a_{k-1}} \rho^n_{k,k} \rho^\epsilon_{k,k-1} \\
&= \frac{d_k}{a_k^2} + \frac{d_{k-1}}{a_{k-1}^2} - 2 \frac{d_k^{\frac{1}{2}} d_{k-1}^{\frac{1}{2}}}{a_k a_{k-1}} \sqrt{\frac{\eta_k}{\eta_k}} \frac{(b_{k-1} - a_{k-1} a_k)}{\sqrt{d_k d_{k-1}}} \\
&= \frac{d_k}{a_k^2} + \frac{d_{k-1}}{a_{k-1}^2} - 2 \frac{b_{k-1} - a_{k-1} a_k}{a_k a_{k-1}} \\
&= \frac{b_k}{a_k^2} + 1 + \frac{b_{k-1}}{a_{k-1}^2} + 1 - 2 \frac{b_{k-1}}{a_k a_{k-1}} - 2 \\
&= \frac{b_k}{a_k^2} + \frac{b_{k-1}}{a_{k-1}^2} - 2 \frac{b_{k-1}}{a_k a_{k-1}}. \qquad (4.23)
\end{aligned}
$$

[d] Here we have that the last term gets eaten by the remainder term since η_k^{-1} is a very small number

When calculating the covariance we are going to assume that $k \leq l$

$$
\begin{aligned}
cov(Y_k, Y_l) =& cov(\frac{d_k^{\frac{1}{2}}}{a_k}\delta_{1,k} - \frac{d_{k-1}^{\frac{1}{2}}}{a_{k-1}}\delta_{2,k}, \frac{d_l^{\frac{1}{2}}}{a_l}\delta_{1,l} - \frac{d_{l-1}^{\frac{1}{2}}}{a_{l-1}}\delta_{2,l})\\
=& \frac{d_k^{\frac{1}{2}} d_l^{\frac{1}{2}}}{a_k a_l}\rho^n_{k,l}\rho^\epsilon_{k,l} - \frac{d_k^{\frac{1}{2}} d_{l-1}^{\frac{1}{2}}}{a_k a_{l-1}}\rho^n_{k,l}\rho^\epsilon_{k,l-1} - \frac{d_{k-1}^{\frac{1}{2}} d_l^{\frac{1}{2}}}{a_{k-1} a_l}\rho^n_{k,l}\rho^\epsilon_{k-1,l} + \frac{d_{k-1}^{\frac{1}{2}} d_{l-1}^{\frac{1}{2}}}{a_{k-1} a_{l-1}}\rho^n_{k,l}\rho^\epsilon_{k-1,l-1}\\
=& \rho^n_{k,l}(\frac{d_k^{\frac{1}{2}} d_l^{\frac{1}{2}}}{a_k a_l}\frac{(b_k - a_k a_l)}{\sqrt{d_k d_l}} - \frac{d_k^{\frac{1}{2}} d_{l-1}^{\frac{1}{2}}}{a_k a_{l-1}}\frac{(b_k - a_k a_{l-1})}{\sqrt{d_k d_{l-1}}} - \frac{d_{k-1}^{\frac{1}{2}} d_l^{\frac{1}{2}}}{a_{k-1} a_l}\frac{(b_{k-1} - a_{k-1} a_l)}{\sqrt{d_{k-1} d_l}}\\
&+ \frac{d_{k-1}^{\frac{1}{2}} d_{l-1}^{\frac{1}{2}}}{a_{k-1} a_{l-1}}\frac{(b_{k-1} - a_{k-1} a_{l-1})}{\sqrt{d_{k-1} d_{l-1}}})\\
=& \rho^n_{k,l}(\frac{b_k - a_k a_l}{a_k a_l} - \frac{b_k - a_k a_{l-1}}{a_{k-1} a_l} - \frac{b_{k-1} - a_{k-1} a_l}{a_{k-1} a_l} + \frac{b_{k-1} - a_{k-1} a_{l-1}}{a_{k-1} a_{l-1}})\\
=& \rho^n_{k,l}(\frac{b_k}{a_k a_l} - \frac{b_k}{a_{k-1} a_l} - \frac{b_{k-1}}{a_{k-1} a_l} + \frac{b_{k-1}}{a_{k-1} a_{l-1}})\\
=& \rho^n_{k,l}(\frac{b_k}{a_k} - \frac{b_{k-1}}{a_{k-1}})(\frac{1}{a_l} - \frac{1}{a_{l-1}}), \quad for\ k \leq l.
\end{aligned} \tag{4.24}
$$

Mean and variance of $\hat{R}_i^{CL}$

It is now possible to find the expression for $\hat{R}_i^{CL}$, but first we need to find the expression for the growth factor $\hat{\alpha}_i = \hat{f}_{I-i+1} \cdots \hat{f}_K$. From (4.21) we get:

$$
\hat{\alpha}_i = (\frac{a_{I-i+1}}{a_{I-i}}(1 + \eta_{I-i+1}^{-\frac{1}{2}} Y_{I-i+1}) + o(\bar{n}^{-\frac{1}{2}})) \cdots (\frac{a_K}{a_{K-1}}(1 + \eta_K^{-\frac{1}{2}} Y_K) + o(\bar{n}^{-\frac{1}{2}})).
$$

After multiplying and excluding all the cross-terms, since they are small in order, we get:

$$
\begin{aligned}
\hat{\alpha}_i &= (\prod_{k=I-i+1}^{K} \frac{a_k}{a_{k-1}})(1 + \sum_{k=I-i+1}^{K} \eta_k^{-\frac{1}{2}} Y_k + o(\bar{n}^{-\frac{1}{2}}))\\
&= \frac{a_K}{a_{I-i}} + \frac{a_K}{a_{I-i}} \sum_{k=I-i+1}^{K} \eta_k^{-\frac{1}{2}} Y_k + o(\bar{n}^{-\frac{1}{2}})
\end{aligned} \tag{4.25}
$$

The linearization of $\hat{R}_i^{CL} = (\hat{\alpha}_i - 1)C_{i,I-i}$ can now be obtained by setting $C_{i,I-i}$ equal to the two first terms in (4.14) when $k = I - i$. This yields:

$$
\begin{aligned}
\hat{R}_i^{CL} =& n_i(a_K - a_{I-i}) + n_i a_K \sum_{k=I-i+1}^{K} \eta_k^{-\frac{1}{2}} Y_k + (\frac{a_K}{a_{I-i}} - 1)\sqrt{n_i d_{I-i}}\epsilon_{i,I-i}\\
&+ \sqrt{n_i d_{I-i}}\frac{a_k}{a_{I-i}}\epsilon_{i,I-i} \sum_{k=I-i+1}^{K} \eta_k^{-\frac{1}{2}} Y_k + o(\bar{n}^{\frac{1}{2}})
\end{aligned}
$$

The two last terms before $o(\bar{n}^{\frac{1}{2}})$ is of a lower order and can be dumped into the discrepancy. We then have that:

$$\hat{R}_i^{CL} = n_i(a_K - a_{I-i}) + n_i a_K \sum_{k=I-i+1}^{K} \eta_k^{-\frac{1}{2}} Y_k + (\frac{a_K}{a_{I-i}} - 1)\sqrt{n_i d_{I-i}}\epsilon_{i,I-i} + o(\bar{n}^{\frac{1}{2}}) \tag{4.26}$$

The expectation can easily be calculated since $Y_{I-i+1}, \cdots, Y_K$ and $\epsilon_{i,I-i}$ are zero-mean so that:

$$\begin{aligned}\mathbb{E}[\hat{R}_i^{CL}] &= n_i(a_K - a_{I-i}) + o(\bar{n}^{\frac{1}{2}}) \\ &= n_i(\sum_{l=0}^{K} q_l\xi_l - \sum_{l=0}^{I-i} q_l\xi_l) + o(\bar{n}^{\frac{1}{2}}) \\ &= n_i(\sum_{l=I-i+1}^{K} q_l\xi_l) + o(\bar{n}^{\frac{1}{2}}) \\ &= \mathbb{E}[R_i] + o(\bar{n}^{\frac{1}{2}}).\end{aligned} \tag{4.27}$$

The result of this equality is that the bias term in the MSE becomes zero as the number of policies in a portfolio increases towards infinity. To calculate the variance of $\hat{R}_i^{CL}$, it will be wise to focus on Y_k-term first, because this will cause the biggest problems.

$$\mathbb{V}\text{ar}(\sum_{k=I-i+1}^{K} \eta_k^{-\frac{1}{2}} Y_k) = \sum_{k=I-i+1}^{K} \eta_k^{-1}\,\mathbb{V}\text{ar}(Y_k) + 2\sum_{k=I-i+1}^{K}\sum_{l=k+1}^{K} \eta_k^{-\frac{1}{2}}\eta_l^{-\frac{1}{2}} cov(Y_k, Y_l) \tag{4.28}$$

By defining the first and the last term in (4.28), V1 and V2 respectively, we get:

$$V1 = \sum_{k=I-i+1}^{K} \eta_k^{-1}(\frac{b_k}{a_k^2} + \frac{b_{k-1}}{a_{k-1}^2} - 2\frac{b_{k-1}}{a_k a_{k-1}}) \tag{4.29}$$

and

$$\begin{aligned}V2 &= 2\sum_{k=I-i+1}^{K}\sum_{l=k+1}^{K} \eta_k^{-\frac{1}{2}}\eta_l^{-\frac{1}{2}}\rho_{k,l}^n(\frac{b_k}{a_k} - \frac{b_{k-1}}{a_{k-1}})(\frac{1}{a_l} - \frac{1}{a_{l-1}}) \\ &= 2\sum_{k=I-i+1}^{K}\sum_{l=k+1}^{K} \eta_k^{-\frac{1}{2}}\eta_l^{-\frac{1}{2}}\sqrt{\frac{\eta_l}{\eta_k}}(\frac{b_k}{a_k} - \frac{b_{k-1}}{a_{k-1}})(\frac{1}{a_l} - \frac{1}{a_{l-1}}) \\ &= 2\sum_{k=I-i+1}^{K} \eta_k^{-1}(\frac{b_k}{a_k} - \frac{b_{k-1}}{a_{k-1}})\sum_{l=k+1}^{K}(\frac{1}{a_l} - \frac{1}{a_{l-1}}) \\ &= 2\sum_{k=I-i+1}^{K} \eta_k^{-1}(\frac{b_k}{a_k} - \frac{b_{k-1}}{a_{k-1}})(\frac{1}{a_K} - \frac{1}{a_k}).\end{aligned}$$

By adding V1 and V2 we obtain the variance expression for (4.28):

$$\begin{aligned}\mathbb{V}\text{ar}(\sum_{k=I-i+1}^{K}\eta_k^{-\frac{1}{2}}Y_k) &= \sum_{k=I-i+1}^{K}\eta_k^{-1}((\frac{b_k}{a_k^2}+\frac{b_{k-1}}{a_{k-1}^2}-2\frac{b_{k-1}}{a_ka_{k-1}})+2(\frac{b_k}{a_k}-\frac{b_{k-1}}{a_{k-1}})(\frac{1}{a_K}-\frac{1}{a_k}))\\ &= \sum_{k=I-i+1}^{K}\eta_k^{-1}((\frac{b_k}{a_k^2}+\frac{b_{k-1}}{a_{k-1}^2}-2\frac{b_{k-1}}{a_ka_{k-1}})+2(\frac{b_k}{a_ka_K}-\frac{b_k}{a_k^2}-\frac{b_{k-1}}{a_{k-1}a_K}+\frac{b_{k-1}}{a_{k-1}a_k}))\\ &= \sum_{k=I-i+1}^{K}\eta_k^{-1}(-\frac{b_k}{a_k^2}+2(\frac{b_k}{a_ka_K}-\frac{b_{k-1}}{a_{k-1}a_K})+\frac{b_{k-1}}{a_{k-1}^2}). \qquad (4.30)\end{aligned}$$

The variance of the outstanding losses for the Chain Ladder method, $\hat{R}_i^{CL}$, can now be calculated. By taking advantage of the fact that $Y_{I-i+1},\cdots Y_K$ of the future and $\epsilon_{i,I-i}$ of the past are stochastically independent and that the variance of $\epsilon_{i,I-i}$ is equal to 1 from the definition of the $\epsilon_{i,k}$'s, we have that:

$$\begin{aligned}\mathbb{V}\text{ar}(\hat{R}_i^{CL}) =&(n_ia_K)^2\sum_{k=I-i+1}^{K}\eta_k^{-1}(-\frac{b_k}{a_k^2}+2(\frac{b_k}{a_ka_K}-\frac{b_{k-1}}{a_{k-1}a_K})+\frac{b_{k-1}}{a_{k-1}^2})+n_id_{I-i}(\frac{a_K}{a_{I-i}}-1)^2\\ &+o(\bar{n}^{\frac{1}{2}})\\ =&n_i^2\sum_{k=I-i+1}^{K}\eta_k^{-1}(-\frac{a_K^2b_k}{a_k^2}+2(\frac{a_Kb_k}{a_k}-\frac{a_Kb_{k-1}}{a_{k-1}})+\frac{a_K^2b_{k-1}}{a_{k-1}^2})+n_id_{I-i}(\frac{a_K}{a_{I-i}}-1)^2+o(\bar{n}^{\frac{1}{2}})\\ =&n_i^2\sum_{k=I-i+1}^{K}\eta_k^{-1}((\frac{a_K^2b_{k-1}}{a_{k-1}^2}-2\frac{a_Kb_{k-1}}{a_{k-1}}+b_{k-1})-(\frac{a_K^2b_k}{a_k^2}-\frac{a_Kb_k}{a_k}+b_k)+b_k-b_{k-1})\\ &+n_id_{I-i}(\frac{a_K}{a_{I-i}}-1)^2+o(\bar{n}^{\frac{1}{2}})\\ =&n_i^2\sum_{k=I-i+1}^{K}\eta_k^{-1}(b_{k-1}(\frac{a_K}{a_{k-1}}-1)^2-b_k(\frac{a_K}{a_k}-1)^2+b_k-b_{k-1})+n_id_{I-i}(\frac{a_K}{a_{I-i}}-1)^2\\ &+o(\bar{n}^{\frac{1}{2}})\\ =&n_i^2\sum_{k=I-i+1}^{K}\eta_k^{-1}(c_k+b_k-b_{k-1})+n_id_{I-i}(\frac{a_K}{a_{I-i}}-1)^2+o(\bar{n}^{\frac{1}{2}}) \qquad (4.31)\end{aligned}$$

where c_k is defined as in (4.2) left. This is the approximation for the variance of $\hat{R}_i^{CL}$ in Section 4.

4.3.2 Kaminsky asymptotics

To verify the approximate variance of the outstanding losses for the Kaminsky approach we will start at a natural point:

$$\hat{R}_i^{Ka}-\mathbb{E}[R_i]=n_i\sum_{k=I-i+1}^{K}\hat{q}_k\hat{\xi}_k-n_i\sum_{k=I-i+1}^{K}q_k\xi_k. \qquad (4.32)$$

By implementing some further manipulation by adding and subtracting a $n_i \sum_{k=I-i+1}^{K} \hat{q}_k \xi_k$ we get:

$$\begin{aligned}
\hat{R}_i^{Ka} - \mathbb{E}[R_i] &= \sum_{k=I-i+1}^{K} [n_i \hat{q}_k \hat{\xi}_k - n_i q_k \xi_k + n_i \hat{q}_k \xi_k - n_i \hat{q}_k \xi_k] \\
&= \sum_{k=I-i+1}^{K} [n_i \hat{q}_k \xi_k - n_i q_k \xi_k + n_i \hat{q}_k (\hat{\xi}_k - \xi_k)] \\
&= n_i \sum_{k=I-i+1}^{K} \xi_k (\hat{q}_k - q_k) + n_i \sum_{k=I-i+1}^{K} \hat{q}_k (\hat{\xi}_k - \xi_k).
\end{aligned}$$

By replacing $\hat{q}_k$ by its true value in the last sum and lump all the discrepancy into the reminder term, we have that:

$$\hat{R}_i^{Ka} - \mathbb{E}[R_i] = B_{1,i} + B_{2,i} + o(\bar{n}^{-\frac{1}{2}}) \tag{4.33}$$

where

$$B_{1,i} = n_i \sum_{k=I-i+1}^{K} \xi_k (\hat{q}_k - q_k) \quad \textit{and} \quad B_{2,i} = n_i \sum_{k=I-i+1}^{K} q_k (\hat{\xi}_k - \xi_k). \tag{4.34}$$

We can also notice that $B_{1,i}$ is the error of the Kaminsky method when the ξ_k are fixed for a development year k without any randomness. This is very important and can be taken advantage of. As mentioned earlier, the connection between the Poisson version of the Burnhuetter-Ferguson method and the Chain Ladder method is discussed in both Bølviken (2015) and Wütherich and Merz (2008). It is shown that both methods are equivalent. The consequence of this equivalence is that when there is no uncertainty around the estimation of the delay-dependent means, ξ_k, i.e. when $\sigma_k = 0$ for all k we have that the Kaminsky predictions coincide with the Chain Ladder method.

This makes it possible to obtain $\mathbb{V}\text{ar}(B_{1,i})$ by setting $\sigma_0 = \sigma_1 = \cdots = \sigma_K = 0$ into the Chain Ladder variance in (4.3) which yields:

$$\mathbb{V}\text{ar}(B_{1,i}) = n_i^2 \sum_{k=I-i+1}^{K} \eta_k^{-1} (c_k^{(0)} + b_k^{(0)} - b_{k-1}^{(0)}) + n_i d_{I-i}^{(0)} (\frac{a_K}{a_{I-i}} - 1)^2 \tag{4.35}$$

where $b_k^{(0)}$, $c_k^{(0)}$ and $d_k^{(0)}$ are defined as in (4.4). To calculate the variance of $B_{2,i}$, remember that $\hat{\xi}_k$ is the average loss when delayed k years. The observation behind is $N_{kk}+, \cdots, +N_{I,k}$ with expected value $(n_k + \cdots + n_I) q_k$. This gives us:

$$\mathbb{V}\text{ar}(\hat{\xi}_k) = \frac{\sigma_k^2}{(n_k + \cdots + n_K) q_k} + o(\bar{n}^{-1})$$

so that

$$\mathbb{V}\text{ar}(B_{2,i}) = n_i^2 \sum_{k=I-i+1}^{K} q_k^2 \, \mathbb{V}\text{ar}(\hat{\xi}_k) = n_i^2 \sum_{k=I-i+1}^{K} \eta_k^{-1} q_k \sigma_k^2 + o(\bar{n}). \tag{4.36}$$

The final step is quite easy because $B_{1,i}$ and $B_{2,i}$ are independent, $\mathbb{V}\text{ar}(\hat{R}_i^{Ka})$ can easily be calculated:

$$\begin{aligned}\mathbb{V}\text{ar}(\hat{R}_i^{Ka}) &= n_i^2 \sum_{k=I-i+1}^{K} \eta_k^{-1}(c_k^{(0)} + b_k^{(0)} - b_{k-1}^{(0)} + q_k\sigma_k^2) + n_i d_{I-i}^{(0)}(\frac{a_K}{a_{I-i}} - 1)^2 + o(\bar{n}) \\ &= n_i^2 \sum_{k=I-i+1}^{K} \eta_k^{-1}(c_k^{(0)} + b_k - b_{k-1}) + n_i d_{I-i}^{(0)}(\frac{a_K}{a_{I-i}} - 1)^2 + o(\bar{n}) \qquad (4.37)\end{aligned}$$

which is the approximation (4.5). This concludes the lengthy calculations that were used to show that the Kaminsky approach is more accurate than the Chain Ladder method in the preceding section. It was also shown that breaking the problem into counts and sizes is a more superior method the more volatile the claim size distribution is.

Chapter 5

Concluding remarks

The objective of this thesis was to investigate which model would be more accurate to estimate the outstanding liabilities. Should insurance companies use aggregated data to estimate the outstanding liabilities, or should they divide the problem into modelling the claim frequency and the claim sizes separately? From the results in Chapter 3 and 4 it seems as it would be advantageous to model the claim counts and the claim sizes separately to estimate the outstanding liabilities.

In Chapter 2 we introduced the Chain Ladder method and the Kaminsky approach, and wanted to find out how the insurance company could go forth to implement them. One of the underlying assumptions of the Chain Ladder method is that there are no “unnatural” claims. Every year develops the same way, as if there is a pattern that will continue in the future. This will not always be true. “Unnatural” claims like natural disasters, big chain collisions and ect. have to be handled separately to predict the reserves. Different methods of handling outliers were described in Chapter 3. The stochasticity of the Kaminsky approach allows it to deal with outlying claim or claims with no additional effort.

When implementing the methods on real data from a Norwegian non-life insurance company in Chapter 3, we observed that for the fire insurance data the Chain Ladder method was affected by the outlier in reporting year 2014. The standard deviation and the bias were both larger for the Chain Ladder method than for the Kaminsky approach for the 2 last reporting years. The results from the bootstrapping simulation on the car insurance data seemed more reliable in the sense that it was not affected by outliers. The Kaminsky standard deviations and the bias were more stable than for the Chain Ladder method. It should be noted that the bias and the standard deviation were somewhat higher for the Kaminsky approach for the first 3 and 4 reporting years.

In the numerical study, the number of policies in a portfolio were increased and the exponential losses and heavy-tailed losses were introduced. The standard deviation and the bias for the Kaminsky apporach were very stable compared to the Chain Ladder ones. It was also clear that when the more volatile the loss distribution was more superior was the Kaminsky approach. This was confirmed in the large portfolio approximation in Chapter 4. It was shown that when the delay-dependent mean and the variance in the individual losses were the same for all delays, the difference between the Chain Ladder approximation and the Kaminsky approximation was proportional to the variance in the individual losses.

This thesis has shown that it is more advantageous to model the counts and the sizes separately in a RBNS case than by using aggregates, especially when the loss distribution is volatile. This is also true for the IBNR case as shown in Bølviken (2015). Verrall et al. (2010) discusses how insurance companies do not tend to use individual data as it can be hard to utilize and computationally difficult. This is also one of the reasons why the Chain Ladder method is popular. It should be noted that since there was no available data on individual losses, the Kaminsky approach was implemented using aggregated data to estimate parameters on an individual level. The method fared very well. If individual data would have been available, the Kaminsky approach would have been even more accurate as the parameters would have been estimated more accurately. Another argument for using the Chain Ladder method is that the method is distribution-free, i.e. non-parametric. This is not necessarily a strength, but can be seen as a weakness in the sense that the model treats all situations equally. This has been pointed out several times in the thesis, and it is not a realistic assumption to make because there is always a possibility for an outlier. The Chain Ladder method is very sensitive when it comes to for example: small changes in portfolio as strong growth that can influence the observed history and changes in product and/ or assessments of claims. The Kaminsky approach can be affected by these examples as well, but the model is more adaptable for these scenarios. The Kaminsky approach requires no specific loss distribution, and is parametric in the sense that it has a Poisson or multinomial basis depending on whether it is an IBNR or a RBNS case. The flexibility in the Kaminsky approach makes it a valuable resource for an actuary that is estimating outstanding liabilities in an insurance company.

Appendix A

Distributions

A.1 Claim number distribution

A.1.1 Poisson distribution

The Poisson distribution is qualified to model claim numbers. The proof is shown in Bølviken (2014). The Poisson distribution:

$$P(N = n) = \frac{\lambda^n}{n!} e^{-\lambda}, \ for\ n = 0, 1, \cdots$$

with

$$\mathbb{E}[N] = \lambda \quad and \quad \mathbb{V}\text{ar}(N) = \lambda$$

The parameters are defined as:

$$\lambda = \mu T \quad and \quad \lambda = J\mu T$$

on policy level and portfolio level respectively. μ is the intensity while J is the number of policies and T is the exposure. This makes μT the frequency. When modelling with delay, a possibility as mentioned Chapter 2 is:

$$N_{i,k} \sim Poisson(\lambda_{i,k})$$

where $\lambda_{i,k} = \lambda_i q_k$ where $\lambda_i = A_i \mu$ and A_i is the portfolio value in year i .

A.1.2 Multinomial distribution

The multinomial distribution is a generalized binomial distribution. There exists $K + 1$ categories where each category is assigned a fixed probability q_k of success. In our case we assume the categories to be development years and the q_k's to be the delay probability. Chapter 2 addresses how the claim numbers are governed by delay-probabilities $q_0, \cdots, q_K$ in-depth. The formula is:

$$P(X_1 = x_1, \cdots, X_K = x_K \,|\, \sum_{k=0}^{K} x_k = n) = \frac{n!}{x_1! \cdots x_K!} q_1^{x_1} \cdots q_K^{x_K}$$

where the expectation, the variance and the covariance is:

$$\mathbb{E}[X_k] = nq_k \;, \mathbb{V}\mathrm{ar}(X_k) = nq_k(1 - q_k) \;\; and \;\; Cov(X_k, X_l) = nq_k q_l$$

for $k \neq l$. For the RBNS case, where the goal is to model the claim numbers for the different development years and reporting years, we have:

$$P(N_{i,0} = n_{i,0}, \cdots, N_{i,K} = n_{i,K} \,|N_i = n_i) = \frac{n_i!}{n_{i,0}! \cdots n_{i,K}!} q_0^{n_{i,0}} \cdots q_K^{n_{i,K}},$$

where $N_{i,0} + \cdots + N_{i,K} = N_i$ and multinomial distributed with probabilities $q_0, \cdots, q_K$ where $N_i = n_i$ is known.

A.2 Claim size distributions

A.2.1 Gamma distribution

The Gamma distribution is often used to simulate claim sizes. One of the reasons for this is that its shape is flexible and this makes it useful in many contexts. The loss Z can be Gamma distributed. In Bølviken (2014), when loss is Gamma distributed, Z is defined as $Z = \xi G$ where G$\sim$ Gamma(α) is called the standard Gamma with mean one and shape α. We then have that:

$$\mathbb{E}[Z] = \xi \;\; and \;\; \mathbb{V}\mathrm{ar}(Z) = \frac{\xi^2}{\alpha}.$$

When $\alpha \to \infty$ the Gamma variables become normal and the standard deviation$\to 0$. The smaller the α becomes, the heavier the tail will become. The density function is:

$$f(x) = \frac{(\frac{\alpha}{\xi})^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{\alpha x}{\xi}}, \;\; x > 0 \;\; where \; \Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

The density function to a standard Gamma is equal to the one above, but with $\xi = 1$.

A.2.2 Exponential distribution

This distribution is a special case only when the shape parameter is equal to 1. It is a somewhat heavy-tailed distribution. The density function is:

$$f(x) = \frac{1}{\xi} e^{-\frac{x}{\xi}} = Gamma(\xi, \alpha = 1).$$

The expectation and variance is:

$$\mathbb{E}[X] = \xi \quad and \quad \mathbb{V}\mathrm{ar}(Z) = \xi^2.$$

Appendix B

Data

The data was given by a Norwegian non-life insurance company. The data for fire insurance and car insurance can be found below. For the fire insurance, there is data from 2010 to 2015 and it contains data on fire damage on villas. For the car insurance there is data from 2009 to 2015 and it contains data on personal injury in car related accidents. The data contains:

Reported Years: These are the years the claims were reported to the insurance company.

Development Years: These are also known as delay years or lags. They indicate how many years it takes for a claim to be settled, counting from the reporting year. The claims that have not yet been settled are set to 2020. In other words, if reported year + development year = 2020 then they have not yet been settled.

Number of Claims: These are the total number of claims for the combination of reported year and development year that have been settled.

Payouts: The total amount the insurance company had to pay out for the different combination of reported and development year.

B.1 Car insurance

Table B.1: *Car insurance data*

Reported Year	Development Year	Number of Claims	Payout
2009	0	12	-48064.0
2009	1	14	-161108.0
2009	2	17	-894886.0
2009	3	10	-3641613.1
2009	4	9	-2387052.0
2009	5	4	-784852.0
2009	6	4	-362340.0
2009	11	7	-10499516.2
2010	0	32	-59994.0
2010	1	70	-502431.9
2010	2	14	-2675988.0
2010	3	14	-2474542.0
2010	4	9	-3571742.0
2010	5	2	-729123.8
2010	10	5	-632908.9
2011	0	60	-258203.0
2011	1	51	-864706.8
2011	2	22	-5322439.2
2011	3	16	-4147862.9
2011	4	8	-3127485.6
2011	9	16	-2054619.2
2012	0	77	-384433.0
2012	1	83	-3124323.2
2012	2	23	-4860033.0
2012	3	13	-4486513.0
2012	8	20	-3734359.8
2013	0	65	-1181876.5
2013	1	101	-1564453.0
2013	2	22	-4647047.8
2013	7	35	-4599373.1
2014	0	150	-554078.0
2014	1	148	-2938129.4
2014	6	80	-2520063.6
2015	0	108	-607523.4
2015	5	183	-914747.9

B.2 Fire insurance

Table B.2: *Fire insurance data*

Reported Year	Development Year	Number of Claims	Payout
2010	0	212	-4853226
2010	1	92	-12858772
2010	2	26	-34847703
2010	3	8	-28592712
2010	4	5	-22237291
2010	5	1	-55505
2010	10	2	-2041059
2011	0	274	-8038561
2011	1	105	-22469769
2011	2	15	-19741824
2011	3	10	-21876168
2011	4	4	-6275362
2011	9	3	-12112507
2012	0	269	-7283445
2012	1	111	-21434904
2012	2	20	-35305370
2012	3	9	-10841351
2012	8	7	-18209330
2013	0	319	-10352733
2013	1	110	-42526638
2013	2	13	-20881159
2013	7	9	-28143883
2014	0	599	-11381789
2014	1	198	-22691535
2014	6	34	-53914940
2015	0	378	-9404576
2015	5	130	-18332702

Appendix C

R-code

C.1 Sorting the data

Here we are sorting the fire insurance data and making a run-off triangle with incremental losses. For the fire insurance data we will use the data from reporting year 2010 to 2015.

```r
##########################
#######Brann############
##########################

Brann    = read.table("SumBrann.txt", skip="1") #Reading the data, skipping the first row

BMelAr  = Brann[,1]  #reported year
BAvv  = Brann[,2]    #development years
BAnS  = Brann[,3]    #number of accidents
BSkae = Brann[,4]    #accident estimate
BRegg = Brann[,5]    #Regress
BUtBe = Brann[,6]    #settlements
BRBNS = Brann[,7]    #RBNS

BrannTable = cbind(BMelAr, BAvv, BAnS, BUtBe)
cat("\n\n")
BrannTable

b = max(which(BrannTable[,1] %in% 2009)) #finding last index where 2009 appears

BrannTable2010 = BrannTable[-(1:b),] #making a matrix with only SkadeAr 2010 and up and avv and utbetaltbelop

#Elementing data were claims where settled in 2020. They have not yet been settled.

MelAr2020 = c()
Avv2020    = c()
AnS2020    = c()
bindex     = c()
h=1
for (b in 1:length(BrannTable2010[,1]))
  {
  #Data is given in such way that the reported claims in year x that hasn't
been settled yet are "Settled"
  #in 2020
  if(BrannTable2010[b,1] + BrannTable2010[b,2] == 2020){
    MelAr2020[h] = BrannTable2010[b,1]; Avv2020[h]= BrannTable2010[b,2]; AnS2020[h]=  BrannTable2010[b,3]
    bindex[h] = b
    h = h+1
  }
}
BrannTableUse = BrannTable2010[-bindex, ] #Table from 2010 and up to 2015 with data of settlements up to 2015
Table2020     = cbind(MelAr2020, AnS2020) #contains the the claims that have not been reported yet.
\\
#################################
#Upper-Triangle Matrix##########
#################################
#Making a empty Matrix
AccidentYear = unique(BrannTableUse[,1])
DevelopmentYear = unique(BrannTableUse[,2])
UpperTriangle = matrix(rep(NA),length(AccidentYear),length(AccidentYear))
rownames(UpperTriangle)=AccidentYear
colnames(UpperTriangle)=DevelopmentYear
a = 1;b=1
for (i in AccidentYear){
  c = which(BrannTableUse[,1] %in% i) #finding indexs.
  b=1
  for (j in c){
    UpperTriangle[a,b] = (-1)*BrannTableUse[j,4] #Making settlements positive, and adding them to the
    #run-off-triangle
    b = b+1 #changing colomn
  }
  a = a+1 #changing rows
}
cat("\n\n\n Upper Triangle \n\n\n")
```

For the car insurance data we will use information from reporting year 2009 to 2015.

```
##########################
#######Bil###############
##########################

Bil    = read.table("SumBil.txt", skip="1") #Reading the data, skipping the first row

BiMelAr = Bil[,1] #reported year
BiAvv   = Bil[,2]   #development years
BiAnS   = Bil[,3]   #number of accidents
BiSkae  = Bil[,4]   #accident estimate
BiRegg  = Bil[,5]   #Regress
BiUtBe  = Bil[,6]   #settlements
BiRBNS  = Bil[,7]   #RBNS

BilTable = cbind(BiMelAr, BiAvv, BiAnS, BiUtBe)
cat("\n\n")
BilTable

b = max(which(BilTable[,1] %in% 2008)) #finding last index where 2008 appears

BilTable2009 = BilTable[-(1:b),] #making a matrix with only SkadeAr 2009 and up and avv and utbetaltbelop

#Elementing data were claims where settled in 2020. They have not yet been settled.
```

```

BilMelAr2020 = c()
BilAvv2020   = c()
BilAnS2020   = c()
Bilindex     = c()
h=1
for (b in 1:length(BilTable2009[,1]))
{
  #Data is given in such way that the reported claims in year x that hasn't been settled yet are "Settled"
  #in 2020
  if(BilTable2009[b,1] + BilTable2009[b,2] == 2020){
    BilMelAr2020[h] = BilTable2009[b,1]; BilAvv2020[h]= BilTable2009[b,2]; BilAnS2020[h]=  BilTable2009[b,3]
    Bilindex[h] = b
    h = h+1
  }
}
BilTableUse = BilTable2009[-Bilindex, ] #Table from 2010 and up to 2015 with data of settlements up to 2015
BilTable2020     = cbind(BilMelAr2020, BilAnS2020)

################################
#Upper-Triangle Matrix##########
################################
#Making a empty Matrix
AccidentYearBil = unique(BilTableUse[,1])
DevelopmentYearBil = unique(BilTableUse[,2])
UpperTriangleBil = matrix(rep(NA),length(AccidentYearBil),length(AccidentYearBil))
rownames(UpperTriangleBil)=AccidentYearBil
colnames(UpperTriangleBil)=DevelopmentYearBil
a = 1;b=1
for (i in AccidentYearBil){
  c = which(BilTableUse[,1] %in% i) #finding indexs.
  b=1
  for (j in c){
    UpperTriangleBil[a,b] = (-1)*BilTableUse[j,4] #Making settlements positive, and adding them to the
    #run-off-triangle
    b = b+1 #changing colomn
  }
  a = a+1 #changing rows
}
cat("\n\n\n Upper Triangle \n\n\n")
```

C.2 R-code for Section 3.1

This is how Table 3.1 and Table 3.2 were made. In addition, these matrices were used in the Kaminsky approach as well.

```
######Fire##########
####################
#Number of accidents#
####################
AccidentYear = unique(BrannTableUse[,1])
DevelopmentYear = unique(BrannTableUse[,2])
UpperTriangleN = matrix(rep(NA),length(AccidentYear),length(AccidentYear))
rownames(UpperTriangleN)=AccidentYear
colnames(UpperTriangleN)=DevelopmentYear
```

```

a = 1;b=1
for (i in AccidentYear){
  c = which(BrannTableUse[,1] %in% i) #finding indexs.
  b=1
  for (j in c){
    UpperTriangleN[a,b] = BrannTableUse[j,3]
    #run-off-triangle with accident numbers
    b = b+1 #changing colomn
  }
  a = a+1 #changing rows
}
UpperTriangleN[is.na(UpperTriangleN)] = 0

UpperTriangleN

#######Car##########
####################
#Number of accidents#
####################
AccidentYearBil = unique(BilTableUse[,1])
DevelopmentYearBil = unique(BilTableUse[,2])
UpperTriangleNBil = matrix(rep(NA),length(AccidentYearBil),length(
    AccidentYearBil))
rownames(UpperTriangleNBil)=AccidentYearBil
colnames(UpperTriangleNBil)=DevelopmentYearBil

a = 1;b=1
for (i in AccidentYearBil){
  c = which(BilTableUse[,1] %in% i) #finding indexs.
  b=1
  for (j in c){
    UpperTriangleNBil[a,b] = BilTableUse[j,3]
    #run-off-triangle with accident numbers
    b = b+1 #changing colomn
  }
  a = a+1 #changing rows
}
UpperTriangleNBil[is.na(UpperTriangleNBil)] = 0

UpperTriangleNBil
```

The delay-dependent mean for the fire and car insurance data was calculated, and this was also used to implement the Kaminsky approach.

```
##########################
#####average#############
##########################
UpperTriangleA = UpperTriangle

xik = c()


UpperTriangleA[is.na(UpperTriangleA)] = 0

for (i in 1:length(AccidentYear)){
  xik[i] = sum(UpperTriangleA[,i])/sum(UpperTriangleN[,i])
}

plot(0:5, xik)

########Car###########
##########################
#####average#############
##########################
```

```
UpperTriangleABil = UpperTriangleBil

xikk = c()


UpperTriangleABil[is.na(UpperTriangleABil)] = 0

for (i in 1:length(AccidentYearBil)){
  xikk[i] = sum(UpperTriangleABil[,i])/sum(UpperTriangleNBil[,i])
}

plot(0:6, xikk)
```

C.3 R-code for Section 3.2

The code below implements the Chain Ladder method on both the fire and car insurance data as seen in Section 3.2.1 and 3.2.2. A run-off triangle with incremental losses that was obtained in C.1 will be used to move forward with the Chain Ladder method.

Fire insurance data:

```
##Cumulated Upper Triangle

CumulatedTriangle = t(apply(UpperTriangle, 1, cumsum))
cat("\n\n\n Cumulated Upper Triangle \n\n\n")
CumulatedTriangle # cumulated run-off-triangle

J = length(DevelopmentYear)

CLMest = rep(0,J-1) #vector of zeros with length J-1, we dont have a estimate for development year 0

for (i in 2:J){
  CLMest[i-1] = sum(CumulatedTriangle[1:(J-i+1),i])/sum(CumulatedTriangle[1:(J-i+1),i-1])
}
cat("\n\n\n Chain-Ladder Estimates \n\n\n")
CLMest

cumulatedLowerTriangle = matrix(rep(NA),length(AccidentYear), length(AccidentYear))

for (j in 2:J){
  for (i in 1:J){
    if (i>J-j+1){
      cumulatedLowerTriangle[j,i] = CumulatedTriangle[j, J-j+1]*prod(CLMest[(J-j+1):(i-1)])
    }
  }
}

cat("\n\n\n Estimated Values \n\n\n")
cumulatedLowerTriangle

CumulatedTriangle[is.na(CumulatedTriangle)] = 0
cumulatedLowerTriangle[is.na(cumulatedLowerTriangle)] = 0

cat("\n\n\n The Full Triangle \n\n\n")
```

```
FullTriangle = CumulatedTriangle + cumulatedLowerTriangle
FullTriangle
```

Car insurance data:

```
##Cumulated Upper Triangle

CumulatedTriangleBil = t(apply(UpperTriangleBil, 1, cumsum))
cat("\n\n\n Cumulated Upper Triangle \n\n\n")
CumulatedTriangleBil # cumulated run-off-triangle

J = length(DevelopmentYearBil)

CLMest = rep(0,J-1) #vector of zeros with length J-1, we dont have a estimate for development year 0

for (i in 2:J){
  CLMest[i-1] = sum(CumulatedTriangleBil[1:(J-i+1),i])/sum(CumulatedTriangleBil[1:(J-i+1),i-1])
}
cat("\n\n\n Chain-Ladder Estimates \n\n\n")
CLMest

cumulatedLowerTriangleBil = matrix(rep(NA),length(AccidentYearBil), length(AccidentYearBil))

for (j in 2:J){
  for (i in 1:J){
    if (i>J-j+1){
      cumulatedLowerTriangleBil[j,i] = CumulatedTriangleBil[j, J-j+1]*prod(CLMest[(J-j+1):(i-1)])
    }
  }
}

cat("\n\n\n Estimated Values \n\n\n")
cumulatedLowerTriangleBil

CumulatedTriangleBil[is.na(CumulatedTriangleBil)] = 0
cumulatedLowerTriangleBil[is.na(cumulatedLowerTriangleBil)] = 0

cat("\n\n\n The Full Triangle \n\n\n")
FullTriangleBil = CumulatedTriangleBil + cumulatedLowerTriangleBil
FullTriangleBil
```

C.4 R-code for Section 3.3

The code for calculating the delay portabilities for both the fire and the car insurance data can be found below.

```
#########Fire#############
##########################
#########q's##############
##########################
TotalBrann = sum(UpperTriangleN) + sum(Table2020[,2])
q          = c()

```

```r
for ( i in 1:length(AccidentYear)){
  q[i] = sum(UpperTriangleN[,i])/TotalBrann

}
#########Car##############
##########################
#########q's##############
##########################
TotalBil = sum(UpperTriangleNBil) + sum(BilTable2020[,2])
qbil          = c()

for ( i in 1:length(AccidentYearBil)){
  qbil[i] = sum(UpperTriangleNBil[,i])/TotalBil

}
```

The R-code below was run to find the shape parameter:

```r
#######Fire##############
##########################
########Y_ik##############
##########################

UpperTriangleY = UpperTriangle

UpperTriangleY[is.na(UpperTriangleY)] = 0

TikMatrix = UpperTriangleY/UpperTriangleN
TikMatrix[is.nan(TikMatrix)]=0  #Matrix with T_ik's, which are equal to Y_ik

##########################
#######Finding alpha#####
##########################
nis = rowSums(UpperTriangleN) + Table2020[,2] # The total amount of claims that were reported from year 2009 to 2015
EmpVar = c()

for (i in 1:length(AccidentYear)){
          a = TikMatrix[,i][TikMatrix[,i] !=0]
  EmpVar[i] = (1/(length(a)-1))*sum((a - xik[i])^2) # Empirical variance
}

a_k = c()
a = c()
b = c()

for (i in 1:length(AccidentYear)){
  j = 1:nis[i]
  a[i] = xik[i]^2
  b[i] = sum((1/j)*dbinom(j, nis[i],q[i]))
  a_k[i] = (a[i]/EmpVar[i])*b[i]
}

a_k[is.na(a_k)] = 0
a_k[length(AccidentYear)]= a_k[length(AccidentYear)-1] # the shape parameter with the last element equal to the element before.

##########################
########Y_IK##############
##########################
Yikbil = UpperTriangleABil/UpperTriangleNBil

Yikbil[is.nan(Yikbil)] = 0

###########Car###########
```

```
###########################
#######alpha og varians##
###########################

nisbil = rowSums(UpperTriangleNBil) + BilTable2020[,2]
EmpVarbil = c()

for (i in 1:length(AccidentYearBil)){
  a = Yikbil[,i][Yikbil[,i] !=0]
  EmpVarbil[i] = (1/(length(a)-1))*sum((a - xikk[i])^2)
}

a_kbil = c()
abil = c()
bbil = c()

for (i in 1:length(AccidentYearBil)){
  j = 1:nisbil[i]
  abil[i] = xikk[i]^2
  bbil[i] = sum((1/j)*dbinom(j, nisbil[i],qbil[i]))
  a_kbil[i] = (abil[i]/EmpVarbil[i])*bbil[i]
}

a_kbil[is.na(a_kbil)] = 0
a_kbil[length(AccidentYearBil)]= a_kbil[length(AccidentYearBil)-1]
```

These are the different plots of the Gamma distribution for different shape parameters:

```
#######Gammaplot########
z = runif(10000, 0, 2)

Gamma1 = rgamma(z, 0.5)/0.5
Gamma2 = rgamma(z, 1)/1
Gamma3 = rgamma(z, 5)/5
Gamma4 = rgamma(z, 10)/10



plot(density(Gamma1), xlim=c(-1,5), ylim=c(0,1), main="Gamma distribution")
legend("topright", c("alpha=0.5", "alpha=1", "alpha=5", "alpha=10"), lty= 1:4, col=c(1, 3, 4,10))
lines(density(Gamma2), lty=2, col=3)
lines(density(Gamma3), lty=3, col=4)
lines(density(Gamma4),lty=4, col=10)
```

Below we are going to implement the Kaminsky approach using the delay-dependent mean, the delay probabilities, run-off triangle with incremental claims and claim numbers and the shape parameter.

Fire insurance data:

```
###########################
#####Simulering###########
###########################

q_tilde = matrix(0, nrow =length(AccidentYear), ncol = length(AccidentYear)) # empty matrix for the new delay probabilities.

```

```r
for (i in 2:length(AccidentYear)){
  c = which(UpperTriangleN[i,] %in% 0)
  for (k in c){
    q_tilde[i,k] = q[k]/(1-sum(q[1:(length(AccidentYear)-i+1)]))#new delay probabilities conditioned on the upper triangle.
  }
}

LowerN = matrix(0, nrow =length(AccidentYear), ncol = length(AccidentYear)) # empty matrix for the lower triangle

LO = Table2020[,2] #Leftover, claims that have not yet been settled by the time we got the dataset

b = length(AccidentYear)
for (i in 2:length(AccidentYear)){
  if(i == 2){
    LowerN[i,length(AccidentYear)] = rbinom(1, LO[i], q_tilde[i,][q_tilde[i,] !=0] )
  }
  LowerN[i, b:length(AccidentYear)] = rmultinom(1, LO[i], q_tilde[i,][q_tilde[i,] !=0] )
  b = b-1
}

N = UpperTriangleN + LowerN #combining both matrices

LowerClaim = matrix(0, nrow =length(AccidentYear), ncol = length(AccidentYear)) # empty matrix for the lower triangle of incremental claims

for (i in 2:length(AccidentYear)){
  c = which(UpperTriangleN[i,] %in% 0)
  for (k in c){
    LowerClaim[i,k] = sum(rgamma(LowerN[i,k], a_k[k])*xik[k])
  }
}

UpperTriangle[is.na(UpperTriangle)]= 0

Claims = UpperTriangle + LowerClaim #combining the matrices

CumulatedClaim = t(apply(Claims, 1, cumsum))aggregating to compare with the chain ladder ones
```

Car insurance data:

```r
##########################
#####Simulering###########
##########################

q_tildebil = matrix(0, nrow =length(AccidentYearBil), ncol = length(AccidentYearBil))

for (i in 2:length(AccidentYearBil)){
  c = which(UpperTriangleNBil[i,] %in% 0)
  for (k in c){
    q_tildebil[i,k] = qbil[k]/(1-sum(qbil[1:(length(AccidentYearBil)-i+1)]))
  }
}

LowerNbil = matrix(0, nrow =length(AccidentYearBil), ncol = length(AccidentYearBil))

LObil = BilTable2020[,2] #Leftover

```

```
b = length(AccidentYearBil)
for (i in 2:length(AccidentYearBil)){
  if(i == 2){
    LowerNbil[i,length(AccidentYearBil)] = rbinom(1, LObil[i], q_tildebil[i,][q_tildebil[i,] !=0] )
  }
  LowerNbil[i, b:length(AccidentYearBil)] = rmultinom(1, LObil[i], q_tildebil[i,][q_tildebil[i,] !=0] )
  b = b-1
}

Nbil = UpperTriangleNBil + LowerNbil

LowerClaimbil = matrix(0, nrow =length(AccidentYearBil), ncol = length(AccidentYearBil))

for (i in 2:length(AccidentYearBil)){
  c = which(UpperTriangleNBil[i,] %in% 0)
  for (k in c){
    LowerClaimbil[i,k] = sum(rgamma(LowerNbil[i,k], a_kbil[k])*xikk[k])
  }
}

UpperTriangleBil[is.na(UpperTriangleBil)]= 0

Claimsbil = UpperTriangleBil + LowerClaimbil

CumulatedClaimbil = t(apply(Claimsbil, 1, cumsum))
```

C.5 R-code for Section 3.4

In Section 3.4 we did a bootstrap simulation with the parameters estimated when implementing the Kaminsky approach.

```
##Parametric Bootstrap
B=1000 #number of bootstrap simulations
O=100  #number of reserve estimation per b in 1 to 1000

#Getting data from the data study
q_fire = c(0.6922038475, 0.2078974013, 0.0249746878, 0.0091123861, 0.0030374620, 0.0003374958)
xik_fire = c(25019.18,  198022.11, 1496973.73, 2270749.29, 3168072.58, 55505.00)
a_fire = c(0.11003951, 0.04011376, 5.13440851, 1.14195556, 1.09100836, 1.09100836)
ni_fire = c(346,  411,  416,  451,  831,  508)

q_car = c(0.335106383, 0.310505319, 0.065159574, 0.035239362, 0.017287234, 0.003989362, 0.002659574)
xik_car = c(6139.23,  19604.18, 187759.12, 278311.91, 349472.30, 252329.29, 90585.00)
a_car = c(0.05039394, 0.07562587, 0.61855204, 1.59851134, 7.38483021, 2.09441464, 2.09441464)
ni_car = c(77,  146,  173,  216,  223,  378,  291 )
##################################

#by changing the parameters below we can implement the bootstrap simulation for both datasets.
q =q_fire
xi= xik_fire
```

```
a = a_fire
ni= ni_fire

n = length(ni)

#making empty matrices
BMC = matrix(0, n, B); BCL=matrix(0, n, B); BKa=matrix(0, n, B)
OMC = matrix(0, n, O); OCL=matrix(0, n, O); OKa=matrix(0, n, O)

#Bootstrapping
for (b in 1:B){
  for(o in 1:O){
    z = matrix(0, n, n); N = matrix(0, n, n) #empty matrices

    ####Modeling z's and n's######

    for (i in 1:n){
      N[i,] = t(rmultinom(1, ni[i], q)) #simulating claim counts, one row at the time
    }

    for (i in 1:n){
      for (k in 1:n){
        z[i,k] = sum(rgamma(N[i,k], a[k],a[k])*xi[k]) #simulating claim sizes, one cell at the time
      }
    }

    #######Finding the "true" reserve#########
    cumMC = t(apply(z, 1, cumsum))

    for (i in (1:n)){
      OMC[i,o] = cumMC[i,n] - cumMC[i, n+1-i] #finding the "true" reserves.
    }

    #######Making matrices with "known" information#########
    h = n;d = n
    knownN = N
    knownz = z
    for (i in (2:n)){
      for ( k in d:h){
        knownN[i,k] = 0
        knownz[i,k] = 0
      }
      d = d-1
    }
    ######Chain Ladder reserves############
    CLest = rep(0,n)

    cumCL = t(apply(knownz, 1, cumsum))

    for (i in 2:n){
      CLest[i-1] = sum(cumCL[1:(n-i+1),i])/sum(cumCL[1:(n-i+1),i-1])#estimating the f_k's
    }

    for (g in 2:n){
      for (i in 1:n){
        if (i>n-g+1){
          cumCL[g,i] = cumCL[g, n-g+1]*prod(CLest[(n-g+1):(i-1)])#predicting
        }
      }
    }

    for (i in (1:n)){
      OCL[i,o] = cumCL[i,n] - cumCL[i, n+1-i] #finding the "true" reserves.
    }

    #########Kaminsky method##########

    Lo = rowSums(N) - rowSums(knownN) #Leftover
    nis = rowSums(N)

    xik = c()

    for (i in 1:n){
      xik[i] = sum(knownz[,i])/sum(knownN[,i])
    }
    xik[is.nan(xik)] = 0

    q_tilde = matrix(0, n, n)

    x = n

    for (i in (2:n)){
      for (k in x:n){
        q_tilde[i,k] = q[k]/(1-sum(q[1:(n-i+1)])) #making new q's given LO
      }
      x = x-1
    }

    LowerN = matrix(0, n, n)

    j = n
    for (i in 2:n){
      LowerN[i, j:n] = rmultinom(1, Lo[i], q_tilde[i,][q_tilde[i,] !=0])#predicting N's
      j = j-1
    }

    LowerClaim = matrix(0, n,n)

    x = n
    for (i in (2:n)){
      for (k in x:n){
        LowerClaim[i,k] = sum(rgamma(LowerN[i,k], a[k], a[k])*xik[k])#simulating claims
      }
      x=x-1
    }
    KaC = LowerClaim + knownz

    cumKa = t(apply(KaC, 1, cumsum))

    for (i in 1:n){
      OKa[i,o] = cumKa[i,n] - cumKa[i, n-i+1]
    }

  }

  for ( i in 1:n){
    BMC[i,b] = mean(OMC[i,])
    BCL[i,b] = mean(OCL[i,])
    BKa[i,b] = mean(OKa[i,])
  }
  print(b)
}

BootMC = c(); BootSdMC = c()
BootCL = c(); BootSdCL = c()
BootKa = c(); BootSdKa = c()
for (i in 1:n){
  BootMC[i]   = mean(BMC[i,])/1000000 #finding the mean
  BootSdMC[i] = sd(BMC[i,])/1000000   standard deviation
  BootCL[i]   = mean(BCL[i,])/1000000
  BootSdCL[i] = sd(BCL[i,])/1000000
  BootKa[i]     = mean(BKa[i,])/1000000
  BootSdKa[i]  = sd(BKa[i,])/1000000
}
#different ways of looking at the data
cbind(BootMC, BootCL, BootKa)
cbind(BootCL/BootMC, BootKa/BootMC)
cbind(BootSdMC, BootSdCL, BootSdKa)
cbind(BootSdCL/BootSdMC, BootSdKa/BootSdMC)
```

C.6 R-code for Section 3.5

In Section 3.5 we did a Monte Carlo simulation for two different loss distributions as well as increasing the number of policies in a portfolio. We will start with the plots of the new delay probabilities using the formula described in Section 3.5:

```
#MOnte Carlo simulasjon
m = 1000
L = 10  #pluss one for the left over
lm = 3

gamma = 0.2
q = exp(-gamma*abs(0:L-lm)) ; q = q/sum(q) #the probabilities.
plot(0:L, q, xlab = "Development years", ylab = "Probability") #plotting the probability

MCR = matrix(0, L+1, m) ; CLR = matrix(0, L+1, m); KaR = matrix(0, L+1, m) #empty matrices

Ni = sample(250000, L+1, replace = TRUE)
xi = rep(100,L+1) #delay dependet means
a = sample(1, L+1, replace = TRUE) #shape parameter 1 or 0.5


for (j in 1:m){
  z = matrix(0, L+1, L+1); n = matrix(0, L+1, L+1) #empty matrices

#######Modelling z's and n's########
  for (i in (1:(L+1))){
    n[i,] = t(rmultinom(1, Ni[i], q)) #simulating claim counts, one row at the time
  }

  for ( i in (1:(L+1))){
    for(k in (1:(L+1))){
      z[i,k] = sum(rgamma(n[i,k], a[k],a[k])*xi[k]) #simulating claim sizes, one cell at the time
    }
  }
  #######Finding the "true" reserve#########
  cumMC = t(apply(z, 1, cumsum))

  for (i in (1:(L+1))){
    MCR[i,j] = cumMC[i,L+1] - cumMC[i, L+1+1-i] #finding the "true" reserves.
  }

  #######Making matrices with "known" information#########
  b = L+1;d = L+1
  knownn = n
  knownz = z
  for (i in (2:(L+1))){
```

```r
  for ( k in d:b){
    knownn[i,k] = 0
    knownz[i,k] = 0
  }
  d = d-1
}
######Finding the CL reserve##############
CLest = rep(0,L)

cumCL = t(apply(knownz, 1, cumsum))

for (i in 2:(L+1)){
  CLest[i-1] = sum(cumCL[1:(L+1-i+1),i])/sum(cumCL[1:(L+1-i+1),i-1])# estimating the f_k's
}

for (g in 2:(L+1)){
  for (i in 1:(L+1)){
    if (i>L+1-g+1){
      cumCL[g,i] = cumCL[g, L+1-g+1]*prod(CLest[(L+1-g+1):(i-1)])#predicting
    }
  }
}
for (i in (1:(L+1))){
  CLR[i,j] = cumCL[i,L+1] - cumCL[i, L+1+1-i] #finding the CL reserves.
}

#########Kaminsky method##########

Lo = rowSums(n) - rowSums(knownn) #Leftover
nis = rowSums(n)

xik = c()

for (i in 1:(L+1)){
  xik[i] = sum(knownz[,i])/sum(knownn[,i])
}


q_tilde = matrix(0, L+1, L+1)

for (i in (2:(L+1))){
  c = which(knownn[i,] %in% 0)
  for (k in c){
    q_tilde[i,k] = q[k]/(1-sum(q[1:(L+1-i+1)])) #making new q's given LO
  }
}

LowerN = matrix(0, L+1, L+1)

b = L+1
for (i in 2:(L+1)){
  LowerN[i, b:(L+1)] = rmultinom(1, Lo[i], q_tilde[i,][q_tilde[i,] !=0])# predicting N's
  b = b-1
}

LowerClaim = matrix(0, L+1,L+1)

for (i in (2:(L+1))){
  c = which(knownn[i,] %in% 0)
  for (k in c){
    LowerClaim[i,k] = sum(rgamma(LowerN[i,k], a[k], a[k])*xik[k])#simulating claims
  }
}
KaC = LowerClaim + knownz

```

```
  cumKa = t(apply(KaC, 1, cumsum))

  for (i in 1:(L+1)){
    KaR[i,j] = cumKa[i,L+1] - cumKa[i, L+1-i+1]
  }

  print(j)
}

RMC = c(); VarRMC = c()
RCL = c(); VarRCL = c()
RKa = c(); VarRKa = c()
for (i in 1:(L+1)){
  RMC[i]    = mean(MCR[i,])/1000000
  VarRMC[i] = sd(MCR[i,])/1000000
  RCL[i] = mean(CLR[i,])/1000000
  VarRCL[i] = sd(CLR[i,])/1000000
  RKa[i] = mean(KaR[i,])/1000000
  VarRKa[i] = sd(KaR[i,])/1000000
}
#different ways of looking at the data
cbind(RMC, RCL, RKa)
cbind(RCL-RMC, RKa-RMC)
cbind(RCL/RMC, RKa/RMC)
cbind(VarRMC, VarRCL, VarRKa)
cbind(VarRCL/VarRMC, VarRKa/VarRMC)

#mean reserve
a1 = RMC
b= RCL
c = RKa

xrange = range(0:L)
yrange = range(RKa)

plot(xrange, yrange, type="n", xlab="Reporting years",
     ylab="Reserve" )
colors <- rainbow(L)
linetype <- c(0:L)
plotchar <- seq(18,18+L,1)

lines(0:L, a1, type="b", lty=1, col=1, lwd=1.5, pch=1)
lines(0:L, b, type="b", lty=2, col=3, lwd=1.5, pch=2)
legend("bottomright", c( "CL", "Ka"), lty= 1:2, col=c(1, 3),pch=c(1,2))

aa = VarRMC
bb= VarRCL
cc = VarRKa

xxrange = range(0:L)
yyrange = range(VarRCL)

plot(xxrange, yyrange, type="n", xlab="Reporting years",
     ylab="Reserve" )
colors <- rainbow(L)
linetype <- c(0:L)
plotchar <- seq(18,18+L,1)

lines(0:L, aa, type="b", lty=1, col=1, lwd=1.5, pch=1)
lines(0:L, bb, type="b", lty=2, col=3, lwd=1.5, pch=2)
title("Exponential losses")
legend("topleft", c( "CL sd", "Ka sd"), lty= 1:2, col=c(1, 3 ),pch=c(1,2))

print(a)
```

Bibliography

Bølviken, E. (2014). *Computation and Modelling in Insurance and Finance.* Cambridge University Press, Cambridge.

Bølviken, E. (2015). Accuracy of claim reserving. Working paper by this date.

de Jong, P. and Heller, G. Z. (2008). *Generalized linear models for insurance data.* Cambrigde University Press, Cambridge.

Devore, J. L. and Berk, K. N. (2007). *Modern Mathematical Statistics with Apllications.* Thomson Brooks/Cole, Belmont CA.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman & Hall/CRC, Boca Raton, FL.

Haavardsson, N. F. (2014). STK4540 notes from lecture on reserving.

Kaminsky, K. S. (1987). Prediction of ibnr claim counts by modelling the distribution of report lags. *Insurance: Mathematics and Economics*, 6.

Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *Astin Bulletin*, 23(2).

Norberg, R. (1989). A contribution of modelling ibnr claims. *Scandinavian Actuarial Journal.*

Norberg, R. (1993). Prediction of outstanding liabilities in non-life insurance. *Astin Bulletin*, 23.

RStudio Team (2015). *RStudio: Integrated Development Environment for R.* RStudio, Inc., Boston, MA.

Verrall, R., Nielsen, J. P., and Jessen, A. H. (2010). Prediction of rbns and ibnr claims using claim amounts and claim counts. *Astin Bulletin*, 40(2).

Weindorfer, B. (2012). A practical guide to the use of the chain-ladder method for determining technical provisions for outstanding reported claims in non-life insurance. *Working Paper Series.*

Wütherich, M. V. and Merz, M. (2008). *Stochastic claims reserving methods in insurance.* Wiley Finance, West Sussex, England.