How big is large?

a study of the limit for large insurance claims in case reserves.

Karl-Sebastian Lindblad

February 15, 2011

Abstract

A company issuing an insurance will provide, in return for a monetary premium, acceptance of the liability to make certain payments to the insured person or company if some beforehand specified event occurs. There will always be a delay between occurrence of this event and actual payment from the insurance company. It is therefore necessary for the company to put aside money for this liability. This money is called the reserve. When a claim is reported, a claim handler will make an estimate of how much the company will have to pay to the claimant. This amount is booked as a liability. This type of reserve is called; "case reserve". When making the estimate, the claim handler has the option of giving the claim a standard reserve or a manual reserve. A standard reserve is a statistically calculated amount based on historical claim costs. This type of reserve is more often used in small claims. A manual reserve is a reserve subjectively decided by the claim handler. This type of reserve is more often used in large claims. This thesis propose a theory to model and calculate an optimal limit above which a claim should be considered large. An application of the method is also applied to some different types of claims.

Keywords: Insurance claims, Monte Carlo simulation, large claims, small claims, case reserve, distributions for insurance claims, general insurance, non-life insurance

Acknowledgements

I would like to thank Charlotta Linse for listening and tossing ideas. I would also like to thank my commissioner at If, Helge Blaker and my tutor at KTH Harald Lang for useful and very interesting discussions and comments.

Stockholm, February 2011

Kalle Lindblad

Contents

Chapter 1

Introduction

A company issuing an insurance will provide, in return for a monetary premium, acceptance of the liability to make certain payments to the insured person or company if some beforehand predefined event occurs.

Both the amount to be paid and the actual occurrence of the event can be modeled as random variables. If the event actually occurs and is reported to the insurance company, this type of liability becomes an insurance claim. In general there is a delay between the specified event occurring and the insurers actual payment and settlement of the claim.

One reason for this is that there is usually some kind of reporting delay between occurrence of a claim and the time for reporting it to the insurance company. There is also a settlement delay. This means that it takes time to evaluate the final settlement. It could for instance be difficult to establish the actual cost of rebuilding something. Estimating the payment for someone being injured can take a long time and in some cases the actual injury will not be noticed until years after the actual occurrence of an accident. In some claim cases, the claim has to be settled in court. The point is that a settlement can take time, sometimes days and sometimes years. [3]

The term "claims reserving" means that an insurance company has to put sufficient provisions aside, so that it is able to settle all the claims that are caused by its insurance contracts up until today. This means that the company has to put aside money for claims that have been reported and also for events that have occurred but that have not yet been reported. The latter is usually called IBNR reserve, meaning reserves for claims that are Incurred But Not Reported.

Since no one knows for sure how many claimants that will report an accident that has occurred in the past, the insurance company has to make

a statistical estimate. There are several methods to do this. However, this type of reserving will not be covered in the thesis.

This thesis will focus on claims that have been reported to the insurance company, since they are all connected to a specific insurance case. The reserves from this type of claim is called case reserves. [10]

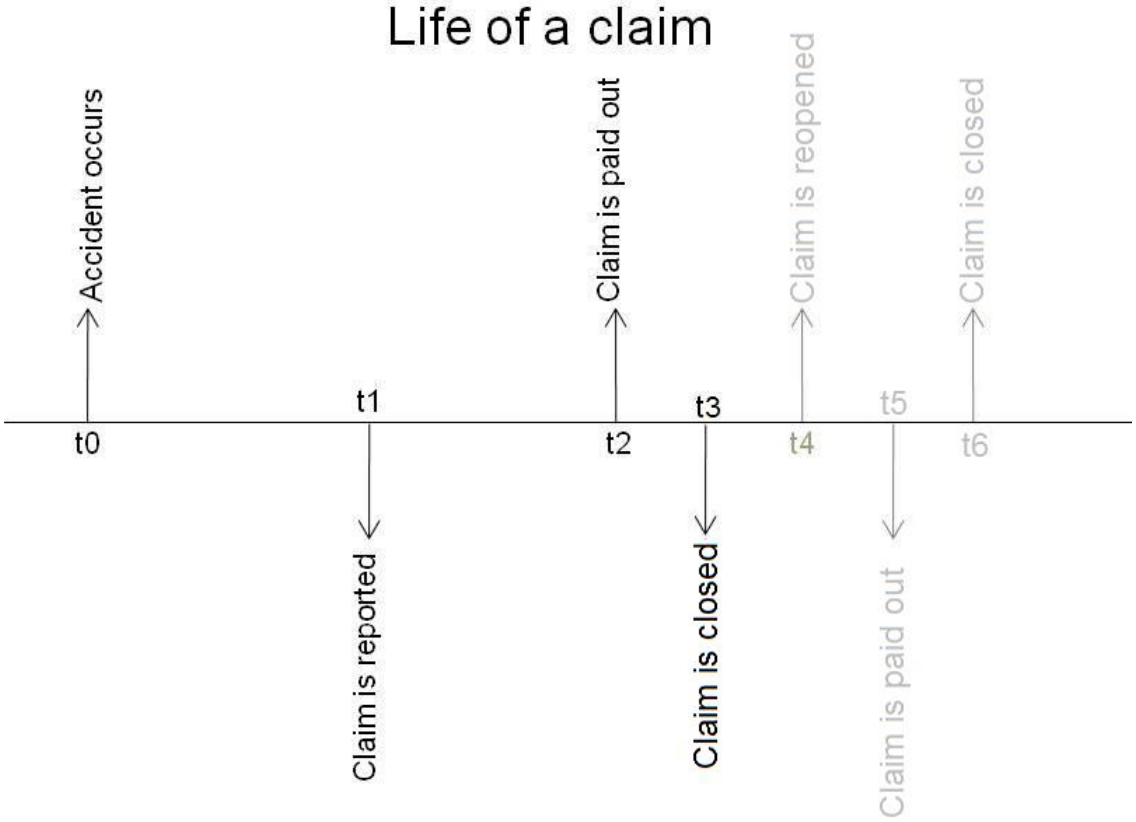

Figure 1.1: There is usually some time before occurance of accident and reporting of a claim. There is also usually a delay between reporting and actual payment. This is a simplified version since the claim is assumed to be paid out once in a lump sum. For example, sometimes several claim payments are paid out over a period of time. After the claim has been paid out in full, it is usually considered to be closed. However, it happens that old claims are reopened due to new circumstances. This would in turn lead to new payments.

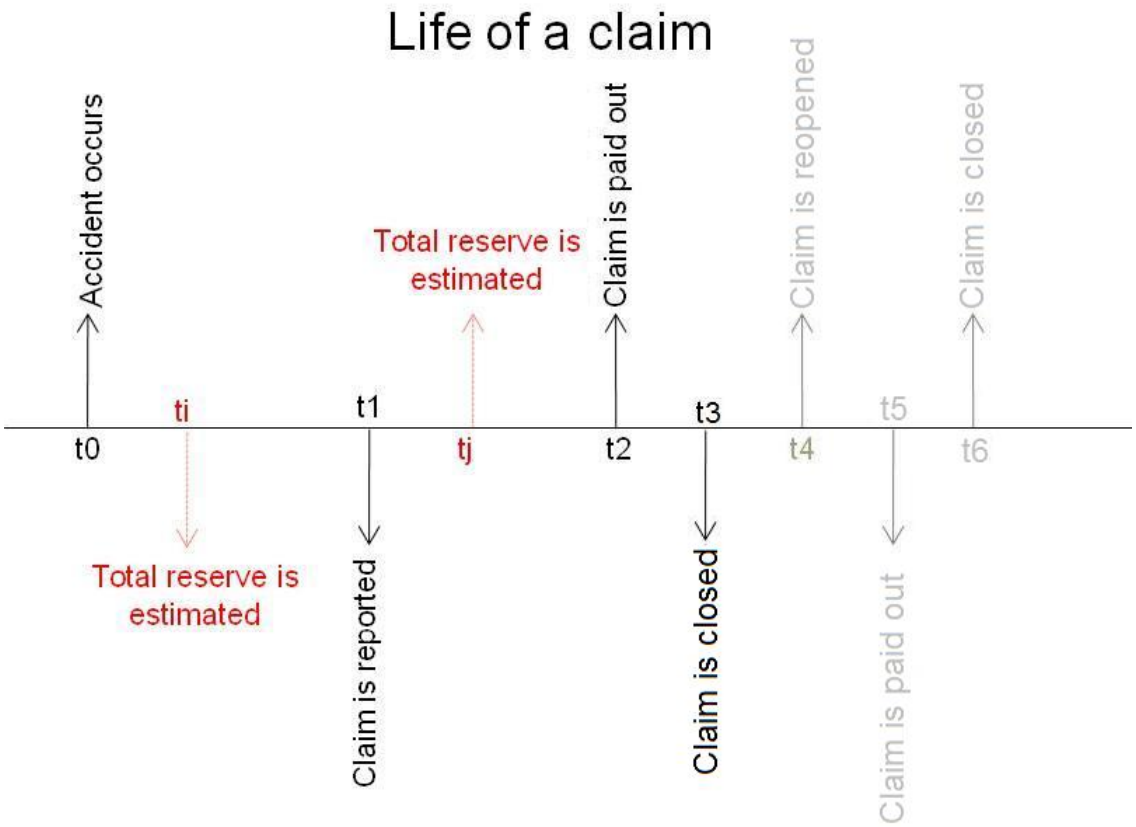

Figure 1.2: Between time t_0 and t_1, at time t_i, the accident has occurred but it has not yet been reported. When estimating the proper reserve of these "future claims", several statistical methods can be used. This type of reserve is called "IBNR reserve". If the claim is between time t_1 and t_2, at t_j, the claim has been reported but not yet paid out. An estimate of how much the accident will cost is made. This estimate is then added to the total reserve. The reserve coming from reported but not paid claims is called "case reserve".

Chapter 2

The Problem

2.1 Background

The reserve of a claim should be equal to the expected remaining payments of the claim. A claim is reported to a claims handler.

In the insurance system, the reserve of a claim can be either manual or standard. Manual means that the handler will estimate the size of the claim and enter a reserve into the system. Standard means that the reserve is calculated automatically. The calculation would for example be dependent on which line of business the claim falls in. It could also depend on whether the customer is a private person or company. In other words, a single claim can vary quite a lot in size, ranging from 0 to a billion SEK.

Typically one would want to put a standard reserve for smaller claims and a manual reserve for larger claims. This would minimize the error and create a reliable reserve estimate. This makes sense since there will be a large number of small claims. They will also be of a more standard nature, for example a broken window, a burnt out oven or a bicycle theft. This makes good parameters for making a statistical estimate. Larger claims, on the other hand, will not be as plentiful, they will also, most of the time be of a unique nature. This means that they will most likely need an individual estimation to be accurate.

There are several reasons to make automatic statistical estimates of reserves. Having an automatic calculation will obviously minimize the work of claims handlers. Since a single small claim does not affect the whole result of the company it is considered more accurate to view many small claims in terms of statistics. Having a standard reserve will also make the claims reserve predictable. This is because the model for calculating is specified beforehand and the number of claims received from year to year is quite

similar. [10]

When a claim handler receives a claim he or she first of all has to decide if the claim will be large and need individual treatment or if it is small and get a standard reserve. There has to be some limit above which a claim would be considered "large" and receive a manually estimated reserve.

2.2 Aim of thesis

The purpose of this Master Thesis is to describe theory and methods for modeling the optimal large claims limit. It aims at finding a procedure for estimating this limit. Several modeling issues must therefore be dealt with. One must define what an optimal limit really is. A definition can for instance be made using loss functions.

The actions of the claim handlers should be described in a model using historical data where both the estimation of the claim and the actual payment of the claim is known. This way a model can be built around how "correct" the claims handlers actually were.

One must also model the standard reserve process. In this process historical data are also known. This way, one can build a model around how "correct" the automatic process has been.

Another issue is how one should model the claims. Using historical data, a good approach would be to assume a distribution. Perhaps combining different distributions to catch large tail events.

In addition, one question deals with the problem of knowing at which level a standard reserve should be calculated. Should one reserve be calculated, for instance, for all motor claims? Should one reserve be calculated for each subset of motor claims, for example motor-hull claims? Or maybe there is no point in actually differentiating between claim types at all?

2.3 Outline

The disposition of this paper is as follows. The first chapter gives an introduction to the topic. Chapter two presents the background of the problem. It also presents the aim and the outline of the thesis. Chapter three will take care of assumptions, definitions and modeling of the problems. Chapter four will give a few solution approaches on how to solve the problem presented.

The fifth chapter deals with the results. Chapter six contains a concluding discussion together with suggestions for further research.

Chapter 3

Theory

3.1 Assumptions

In order to make a proper model, it is necessary to make some simplifications regarding the data used. When discussing data, the author generally mean data regarding claims of an insurance company. The claims can be of different types. For example house fire claims, theft claims, car claims, personal injury claims and so on. Different types of claims behave differently. Some take a long time to settle. Others result in annual payments that go on as long as the claimant lives. Some types of claims are frequently reopened. To be able to make a reasonable model one must assume and use data that are reported and settled quite quickly. Inflation and discounting between occurrence and payout is therefore ignored. For simplification it is assumed that the total payment of a claim is always paid out as a single lump sum. It would also be required that data are not frequently reopened. It will be assumed that a closed claim will remain closed. When experimenting with the presented models, one must have these assumptions in mind when choosing what type of data to use.

An assumption is done regarding the claims handlers. When first receiving a claim, they must decide whether the claim should be considered large or small. This decision is made without detailed knowledge of the actual claim. They must use their experience and give a ballpark figure of where the estimate will land. It is assumed that the claims handlers will get this right every time.

3.2 Defining an optimal limit

One can define an optimal limit in many ways. There is very little literature on the subject of the large claims limit. This means that the definitions made here are perhaps some out of many possible. The total number of reported

claims over a period of time is n. The reserve R is defined as the sum of all estimations of reported but not paid claims r_i.

$$R = \sum_{i=1}^{n} r_i \tag{3.1}$$

The total actual paid amount S is the sum of all claim payments s_i.

$$S = \sum_{i=1}^{n} s_i \tag{3.2}$$

Here one wants to define some kind of "loss function" that can later be minimized. A loss function can be described as a function that measures some kind of loss. For example a model of how much error there will be in a reserve estimation. On first thought, one might think of a loss function that looks like $R - S$. Meaning that the optimal reserve estimation would be as close as possible to the actual future payments.

$$\Theta = R - S = \sum_{i=1}^{n} r_i - s_i = \sum_{i=1}^{n} \theta_i = 0 \tag{3.3}$$

This is an easy way of measuring the "loss". It is also quite easy to give the "loss" an interpretation. In this case, it would be interpreted as how wrong the estimation of the reserve is in general. It would also be easy to perhaps apply some kind of Value at Risk model to this definition.

In reserving, it is perhaps worse to have a negative reserve error. One could for example look at the risk of having a negative reserve error and decide upon a claims limit that would give 95% probability of having a positive reserve error.

Every estimation r_i corresponds to an actual payment s_i. To make the model more stable, it would perhaps be desirable to minimize $|r_i - s_i| = |\theta_i|$. In other words, to make each estimation come as close as possible to its actual payment. Minimizing only $|\Theta| = |R - S|$ would mean that the model could become dependent on a few large negative and positive estimation errors that neutralize each other. Trying to instead minimize $|\theta_i|$ would mean that the model is less dependent on single large estimation errors.

An optimal limit for large claims would therefore be a limit that minimizes all $|\theta_i|$. This means the limit would be such that: $\sum_{i=1}^{n} |\theta_i|$ is minimized. The equivalent would be to say: find a limit u that fulfills the following loss function:

$$\min \sum_{i=1}^{n} |\theta_i(u)| = \min \sum_{i=1}^{n} |r_i(u) - s_i| \tag{3.4}$$

Where $r_i(u) = I(s_i < u) \cdot r_i^{Standard} + I(s_i > u) \cdot r_i^{Manual}$. $I(.)$ is the identity function. s_i, n and r_i can be considered to be stochastic random variables. The modeling of s_i and n is discussed in section 3.5, while the modeling of r_i is discussed in section 3.3 and 3.4.

There exist other options for loss functions. One of the most widely used is perhaps the quadratic loss function. Find a limit u that satisfies:

$$\min \sum_{i=1}^{n} \theta_i(u)^2 = \min \sum_{i=1}^{n} (r_i(u) - s_i)^2 \tag{3.5}$$

This type of loss function will treat negative and positive errors equally. In claims reserving, one could imagine that it would be worse to reserve to little. If an insurance company does not have enough money to pay for its liabilities, it is probably in big trouble. If too much money is reserved the insurance company might have a slightly worse result. The latter alternative would certainly be preferred. One option would be to use the LINEX loss function [11]. Find the u that satisfies:

$$\min \sum_{i=1}^{n} \exp(-\beta \cdot \theta_i(u)) + \beta \cdot \theta_i(u) + 1 =$$

$$= \min \sum_{i=1}^{n} \exp(-\beta \cdot (r_i(u) - s_i)) + \beta \cdot (r_i(u) - s_i) + 1 \tag{3.6}$$

Where β is some constant > 0. This type of loss function will punish negative estimation errors exponentially and only punish positive estimation errors linearly. The LINEX function will look quite similar to the quadratic loss function for small losses since $e^x - x - 1$ is proportional to $\frac{x^2}{2}$. The big difference is that the LINEX function can be positively or negatively skewed.

Another, a bit simpler loss function would be. Find u that satisfies:

$$\min \sum_{i=1}^{n} I(\theta_i > 0) \cdot k_1 \cdot |\theta_i(u)| + I(\theta_i < 0) \cdot k_2 \cdot |\theta_i(u)| =$$

$$\min \sum_{i=1}^{n} I(r_i(u) - s_i > 0) \cdot k_1 \cdot |r_i(u) - s_i| +$$

$$+ I(r_i(u) - s_i < 0) \cdot k_2 \cdot |r_i(u) - s_i| \tag{3.7}$$

Where k_1 and k_2 are weight constants. The case where k_1 and k_2 are equal to one would be the same case as the absolute value loss function

described earlier. With this equation one can control how much to value positive and negative errors. In the case of reserving one could argue that the only error that should be punished is when the reserve is too small. In that case k_1 should be put to zero and k_2 can be put to any positive number.

As argued earlier, the standard reserve can be treated with statistical methods. The error coming from this kind of estimation would sometimes be larger than zero and sometimes smaller. In a perfect world, the total error would even out in the end and become zero. It would perhaps therefore be logical to treat the error coming from these claims, not as individual errors, but as one error. This would result in a total estimation error that would oscillate around zero. As the large claims limit increases, it will result in instability and an increase in variance.

The aim of the manual reserve is however to treat every claim individually and try to come as close as possible to the actual claim cost. This means one perhaps should treat these errors individually. A loss function could look like:

Find u that satisfies:

$$\min \sum_{i=1}^{n} I(s_i < u) \cdot |r_i^{Standard} - s_i| + I(s_i > u) \cdot |r_i^{Manual} - s_i| \tag{3.8}$$

3.3 Modeling the actions of the claims handlers

Modeling the actions of humans is a very tricky business. Here, the process of reserving a manual claim will be considered stochastic. I have not found any literature describing this type of modeling. This means that the theory for this part has been developed solely for this thesis.

It would be reasonable to assume that the estimate of the claims handler r_i^{Manual} most of the time will be more or less correct. This means that the model perhaps can be assumed to be normally distributed or Student-t distributed around a mean which is the actual future claim payment s_i.

$$r_i^{Manual} = s_i + N(0, \sigma)$$

or

$$r_i^{Manual} = s_i + t(0, \nu, \sigma) \tag{3.9}$$

Where σ is standard deviation and ν is degrees of freedom of the Student-t distribution.

It would also be logical to assume that the larger a claim actually turns out to be, the harder it is to estimate. This means that a claims handler will more often make a larger estimation error when handling a big claim. At least, the spread of the estimation should be somehow dependent on the size of the claim. As an illustration, when a window is broken on a house, the estimation error of its cost might differ with a few hundred SEK. Whereas if a large villa has been burnt to the ground, the error will likely differ with the size of tens or hundreds of thousand SEK. To incorporate this into the model one can make the following multiplication.

$$r_i^{Manual} = s_i + f(s_i) \cdot N(0, \sigma)$$

or

$$r_i^{Manual} = s_i + f(s_i) \cdot t(0, \nu, \sigma) \tag{3.10}$$

Where $f(s_i)$ is a function of s_i. If $f(s_i)$ should mirror a dependence of the size of s_i, then typically $f(s_i)$ can be modeled as s_i^α. Where α is a constant that is somehow dependent on what type of data are used. This gives:

$$r_i^{Manual} = s_i + s_i^\alpha \cdot N(0, \sigma)$$

or

$$r_i^{Manual} = s_i + s_i^\alpha \cdot t(0, \nu, \sigma) \tag{3.11}$$

When estimating α and the parameters of the normal distribution or student-t distribution one can use historical data, where the estimation of the claims handler is known and where the actual paid amount is also known. In other words r_i^{Manual} and s_i are known.

$$\frac{r_i^{Manual} - s_i}{s_i^\alpha} = N(0, \sigma)$$

or

$$\frac{r_i^{Manual} - s_i}{s_i^\alpha} = t(0, \nu, \sigma) \tag{3.12}$$

α can be fine tuned so that the data looks approximately normal or student-t distributed. The parameters of the chosen distribution can then be estimated. This can be done using for example maximum likelihood, scatter-plots or qq-plots (see later sections).

3.4 Modeling the standard reserve process

The standard reserving method is supposed to make an automatic judgment of how much to reserve, when the claim is considered to be small. In other words; how should one make an estimate when the actual future payment s_i is smaller than the large claims limit u? A logical way to do it would be to use the expected value of s_i:

$$r_i^{Standard} = E[s_i|s_i < u] = \int_{x=0}^{\infty} x \cdot f_{s_i=x|s_i<u}(s_i = x|s_i < u)dx =$$

$$= \int_{x=0}^{u} x \cdot f_{s_i=x|s_i<u}(s_i = x|s_i < u)dx \tag{3.13}$$

To calculate the expected value, one would need to find some distribution for the future payments e.g s_i.

3.5 Modeling the claims data

3.5.1 Distribution of the claims

A model has to be built around the future payment of a claim s_i. s_i is considered to be a random variable belonging to some distribution.

$$s_i \in X \tag{3.14}$$

where X has some distribution.

There are several ways to model this kind of stochastic future payments. All insurance claims are highly positively skewed. A lognormal distribution might be fitting for some types of claims. Typically the ones that tend to have a small upper tail. An example could perhaps be glass damage. The cost of fixing a glass window or wall will most likely never be large compared to claims originating from for example, house fires. [2] On the other hand, claims that have a large upper tail are also possible. The scenario would be that most claims can be found close to zero, but there is the possibility of runaway claims many times larger than the expected value. These claims can be modeled quite accuratly with the generalized pareto distribution or GPD. [8]

Some insurance claims tend to resemble both distributions. Close to zero, they look like the lognormal distribution and in the tail they look like the GPD. In this case Ananda and Cooray [1] suggest using a composite distribution consisting of a log-normal part and a generalized pareto part. A lognormal distribution alone, will not catch the relatively long tail of claims data. While a generalized pareto distribution will not catch the behavior of smaller claims. The combination of the two will simulate the desired behavior. In other words, we want to build a model that has a lognormal distribution until a certain threshold θ. After the threshold it should have a generalized pareto distribution. The composite density function $f(x)$ can be described by:

$$f(x) \begin{cases} cf_1(x), \text{ if } x \in (0, \theta] \\ cf_2(x), \text{ if } x \geq \theta \end{cases} \tag{3.15}$$

Where c is a normalizing constant, $f_1(x)$ has the form of a two-parameter log-normal density and $f_2(x)$ has the form of a two-parameter GPD density.

$$f_1(x) = \frac{(2\pi)^{-1/2}}{x\sigma} \exp[-\frac{1}{2}(\frac{\ln x - \mu}{\sigma})^2], \text{ if } x \in (0, \theta] \tag{3.16}$$

$$f_2(x) = \frac{\alpha\theta^\alpha}{x^{\alpha+1}}, \text{ if } x \geq \theta \tag{3.17}$$

where $\theta, \mu, \sigma, \alpha$ are unknown parameters such that $\theta > 0, \sigma > 0, \alpha > 0$. To make the model realistic we impose continuity and differentiability conditions on θ,

$$f_1(\theta) = f_2(\theta)$$

and

$$f_1'(\theta) = f_2'(\theta) \tag{3.18}$$

where $f'(\theta)$ is the first derivative of $f(x)$ evaluated at θ.
If we impose the conditions of (3.17) on (3.14) we get $\ln(\theta) - \mu = \alpha\sigma^2$ and $\exp(-\alpha^2\sigma^2) = 2\pi\alpha^2\sigma^2$.

Since $\int_0^\infty f(x)dx = 1$, we get $c(\int_0^\theta f_1(x)dx + 1) = 1$. We get

$$\int_0^\theta f(x)dx = \int_0^\theta \frac{(2\pi)^{-1/2}}{x\sigma} \exp[-\frac{1}{2}(\frac{\ln x - \mu}{\sigma})^2]dx$$

$$= \int_{-\infty}^{\frac{\ln(\theta)-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}t^2]dt = \Phi(\frac{\ln(\theta) - \mu}{\sigma}) = \Phi(\alpha\sigma) \tag{3.19}$$

where $\Phi(.)$ is the cumulative distribution function of the standard normal distribution. This gives $c = [1 + \Phi(\alpha\sigma)]^{-1}$

With the result above one can show that the composite density can be reparametrized and re-written as

$$f(x) \begin{cases} \frac{\alpha\theta^\alpha}{(1+\Phi(k))x^{\alpha+1}} \exp[-\frac{\alpha^2}{2k^2} \ln^2(\frac{x}{\theta})], \text{ if } x \in (0, \theta) \\ \frac{\alpha\theta^\alpha}{(1+\Phi(k))x^{\alpha+1}}, \text{ if } x \geq \theta \end{cases} \tag{3.20}$$

where $\Phi(.)$ is the cumulative distribution of the standard normal distribution and k is a known constant. k is given by the solution to $\exp(-k^2) = 2\pi k^2$. This gives $k = 0.372238898$. Here $\alpha\sigma = k$ and $c = 1/(1 + \Phi(k))$. The composite probability density will therefore only have two parameters $\theta > 0$ and $\alpha > 0$. The cumulative distribution function of the composite model is

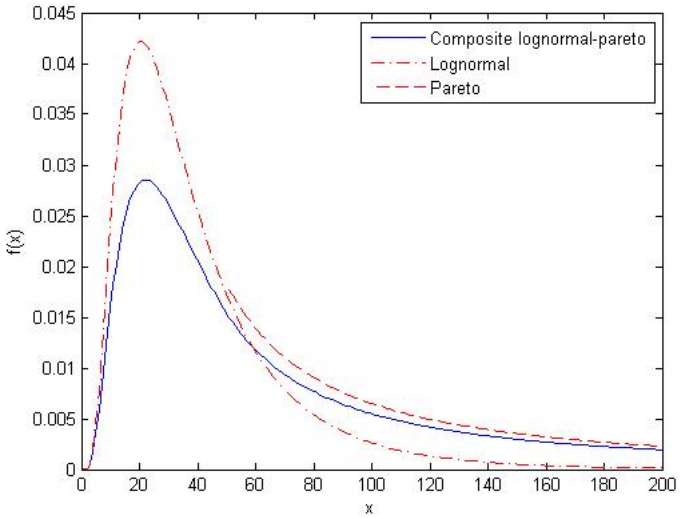

Figure 3.1: The lognormal distribution has a small upper tail, while the pareto distribution has a large one. The composite version has properties from both. In this case $\theta = 50$ and $\alpha = 0.5$

$$F(x) \begin{cases} \frac{1}{(1+\Phi(k))}\Phi((\alpha/k))\ln(x/\theta), \text{ if } x \in (0,\theta) \\ 1 - \frac{1}{(1+\Phi(k))}(\theta/x)^{\alpha}, \text{ if } x \geq \theta \end{cases} \tag{3.21}$$

3.5.2 Number of claims

The number of reported claims during a certain period of time n can be considered to be stochastic. The total number of insured during a period of time is n_{tot}. Define p as the probability for an insured to actually have an accident leading to a claim during this period. Each claim can be considered to be independent of the other. n can then be modeled as Bin(n_{tot},p). [5]

n_{tot} can be considered to be large and p can be considered to be quite small. A well known approximation is then: Bin(n_{tot},p) $\sim$ Po($n_{tot} \cdot p$) [9]. Define $n_{tot} \cdot p$ as λ. This leads to:

$$n \sim Po(\lambda) \tag{3.22}$$

Chapter 4

Simulation approaches

For this chapter, methods in [6], [4] and [7] have been used for parametric and non-parametric simulation. [2], [1] and [8] have been used for distribution estimation techniques such as maximum likelihood, qq-plotting and information regarding heavy-tailed and short-tailed distributions

4.1 Simulating the optimal large claims limit

To investigate the optimal large claims limit u, one has to decide on a loss function. For example, the loss functions of equations 3.4, 3.5,3.6 or 3.7 can be used. Find the limit u that satisfies:

$$\min \sum_{i=1}^{n} |\theta_i(u)| = \min \sum_{i=1}^{n} |r_i(u) - s_i| \tag{4.1}$$

or

$$\min \sum_{i=1}^{n} \exp(-\theta_i(u)) + \theta_i(u) + 1 = \min \sum_{i=1}^{n} \exp(-(r_i(u) - s_i)) + (r_i(u) - s_i) + 1 \tag{4.2}$$

or

minimize

$$\sum_{i=1}^{n} k_1|\theta_i(u)| = \sum_{i=1}^{n} k_1|r_i(u) - s_i| \text{ for } \theta = r_i - s_i > 0$$

$$\sum_{i=1}^{n} k_2|\theta_i(u)| = \sum_{i=1}^{n} k_2|r_i(u) - s_i| \text{ for } \theta = r_i - s_i < 0 \tag{4.3}$$

Where $r_i(u) = I(s_i < u) \cdot r_i^{Standard} + I(s_i > u) \cdot r_i^{Manual}$.
or

$$\min \sum_{i=1}^{n} I(s_i < u) \cdot (r_i^{Standard} - s_i) + I(s_i > u) \cdot |r_i^{Manual} - s_i| \tag{4.4}$$

One way to simulate an approximation of the large claims limit is to measure the loss function for different values of u. u would range from 0 to a suitable large number. For each u, the result of the loss function would represent the "error" of using this u. One could thereafter plot this sum as a function of u to figure out where the minimum loss occurs and to figure out the behavior of the loss function compared to u.

To avoid large random errors, one can simulate many times for every u and take the average of these results.

If this is to be possible, one needs information about $r_i^{standard}$ and r_i^{manual}. To make the right calculations new samples must be generated from historical data. The number of claims received during a certain period has to be simulated. There also has to be a procedure to simulate the actual claims, s_i, from some kind of distribution.

4.2 Simulating the standard reserve

Equation 3.12 gave a model for standard reservation of claims, as follows:

$$r_i^{Standard} = \sum_{x=0}^{u} x \cdot P(s_i = x) \tag{4.5}$$

What one wants is basically the expected value of the claim cost conditioned on the fact that the actual claim cost is smaller than u. This is approximated by taking the average value of the sorted samples up to the point of u. To be more specific, one makes enough samples of the claims distribution and take the average of all claims that are smaller than u.

$$E[s_i | s_i < u] = r_i^{Standard} \approx \frac{1}{N_u} \cdot \sum_{j=0}^{N_u} s_j^{sample} \tag{4.6}$$

where N_u is the number of claims smaller than u and s_j^{sample} is a claim cost smaller than u from the sampled distribution.
To decrease the random error in this procedure one can simulate this many times for the same u and take an average of these simulations.

4.3 Simulating the manual reserve

Equation 3.11 gave a model for manual reservation of claims.

$$\frac{r_i^{Manual} - s_i}{s_i^{\alpha}} = N(0, \sigma)$$

or

$$\frac{r_i^{Manual} - s_i}{s_i^{\alpha}} = t(0, \nu, \sigma) \tag{4.7}$$

Here, one first of all needs to find an α that makes the data look approximately normal or student-t shaped. This can be done using a trial and error approach. One way of doing it would be to look at real historical data coming from claim handlers. The data should contain information about how much money was actually reserved for a claim, r_i^{manual}. The data also should contain information about how much a claim actually ended up costing, s_i. This way one could analyze a histogram of equation 3.11 for different α. Different α could be tried until the plot has a shape that corresponds to a normal or student-t distribution. Then one could simulate the parameters of this distribution. This could easily be done using a standard method such as maximum-likelihood and qq-plotting.

4.4 Simulating the number of claims that should be sampled in each period.

To decide how many samples should be simulated in one time period one could use equation 3.21.

$$n \sim Po(\lambda) \tag{4.8}$$

where $n_{tot} \cdot p$ is λ. n_{tot} is the total number of people covered by this insurance. p is the probability that a policyholder will actually have a claim during the period.

4.5 Simulating the claim payments

Definition of the total payment of a claim in this context would be the total amount paid out on a claim before it is considered to be closed.

4.5.1 Non-parametric sampling

One way of approaching simulation of data is to use non-parametric resampling or non-parametric bootstrapping. It is assumed one has a vector of samples coming from actual data, for example a large number of insurance claims. Instead of looking at the values and assuming a distribution, one

can use the "sample distribution". This means one randomly picks samples from the vector of real data with replacement, until a new vector has been created. By doing a non-parametric resampling of data one can create infinitely many new vectors. This would only be possible if the measured vector is large enough and representable for the whole underlying distribution. A positive aspect of non-parametric resampling is that it is relatively easy and straightforward to do. One does not have to bother with any parameters for any distributions. It can actually also give quite close results to parametric sampling, meaning that in some cases perhaps the easy way is the better way. One of the major negative aspects of non-parametric resampling is that, when resampling, the largest value in any vector can never be larger than the largest of the historical values. If there is a possibility of catastrophically large or small values in the future, meaning larger or smaller than ever recorded, this will not be modeled into the "sample distribution".

4.5.2 Parametric sampling

Another option for simulating the claims would be to assume some kind of distribution for the claims and estimate the parameters. The main benefit of using a parametric sampling is that this kind of sampling catches a greater number of aspects on how the "true" distribution works compared to non-parametric approaches. In this model, for example, values that have previously never occurred can occur. If one is modeling worst case scenarios, this is highly relevant.

What type of distribution I will employ will differ with different types of claims. This has been discussed in earlier chapters. If the claims seem to have a short tail when plotting a histogram, a lognormal distribution might be fitting. If, on the other hand, the claims seem to have a large upper tail a composite lognormal-pareto distribution might be fitting. For fitting the lognormal distribution a simple maximum likelihood estimate of the parameters could give a nice fit to the distribution. A QQ-plot could thereafter be used to fine tune these parameters. When fitting a composite lognormal-pareto distribution, Ananda and Cooray [1] give an algorithm for calculating the maximum likelihood parameters.

Let $X_1, X_2, ..., X_n$ be a random sample from the two-parameter composite lognormal-pareto model described in chapter 3. $x_1 \leq x_2 \leq x_3 \leq ... \leq x_n$ is assumed to be an ordered sample. Suppose the unknown parameter θ is in between the m^{th} observation and the $m+1^{st}$ observation, in other words, $x_m \leq \theta \leq x_{m+1}$. Then the likelihood function is given by

$$L(\alpha, \theta) = C_0 \alpha^n \theta^{n\alpha} \Big(\prod_{i=m+1}^{n} x_i^{-\alpha} \Big) \exp\Big[-\frac{\alpha^2}{2k^2} \sum_{i=1}^{m} ln^2(x_i/\theta) \Big] \qquad (4.9)$$

where $C_0 = 1/[(\prod_{i=1}^{n} x_i)(1 + \Phi(k))^n]$.

The maximum likelihood (ML) estimators of θ and α, $\hat{\theta}_{ML}$ and $\hat{\alpha}_{ML}$ respectively, can be obtained numerically as follows. First, for a given θ, numerically find the value of α that maximizes $L(\alpha, \theta)$. Then by changing θ over the interval $(0, \infty)$, find the values of θ and α that maximizes $L(\alpha, \theta)$. It is important to notice that as θ changes, since $x_m \leq \hat{\theta} \leq x_{m+1}$, the sum in $L(\alpha, \theta)$ should change accordingly. An algorithm for calculating the maximum likelihood shall be presented below.

Step 1. For each m $(m = 1, 2, ..., n-1)$, calculate $\hat{\alpha}_{tem}$ and $\hat{\theta}_{tem}$ as follows:

For $m = 1$,

$$\hat{\alpha}_{tem} = n(\sum_{i=1}^{n} ln(x_i/x_1))^{-1}$$

$$\hat{\theta}_{tem} = x_1 \prod_{i=1}^{n} (x_i/x_1)^{k^2} \tag{4.10}$$

Otherwise,

$$\hat{\alpha}_{tem} = \frac{k^2(n\sum_{i=1}^{m} \ln x_i - m\sum_{i=1}^{n} lnx_i)}{2(m\sum_{i=1}^{m}(\ln x_i)^2 - (\sum_{i=1}^{m} \ln x_i)^2)}$$

$$+\frac{(k^4(n\sum_{i=1}^{m} \ln x_i - m\sum_{i=1}^{n} \ln x_i)^2 + 4mnk^2(m\sum_{i=1}^{m}(\ln x_i)^2 - (\sum_{i=1}^{m} \ln x_i)^2))^{1/2}}{2(m\sum_{i=1}^{m}(\ln x_i)^2 - (\sum_{i=1}^{m} \ln x_i)^2)} \tag{4.11}$$

$$\hat{\theta}_{tem} = (\prod_{i=1}^{m} x_i)^{1/m} \exp(\frac{nk^2}{m\hat{\alpha}_{tem}}) \tag{4.12}$$

If $\hat{\theta}_{tem}$ is in between $x_m \leq \hat{\theta}_{tem} \leq x_{m+1}$, then the ML estimator of α and θ are

$$\hat{\alpha}_{ML} = \hat{\alpha}_{tem}$$

$$\hat{\theta}_{ML} = \hat{\theta}_{tem} \tag{4.13}$$

Step 2. If there is no solution for θ (this means $x_n \leq \hat{\theta}_{tem}$) with the conditions given in Step 1, the ML estimate of α and θ are

$$\hat{\alpha}_{ML} = nk/\sqrt{n\sum_{i=1}^{n}(\ln x_i)^2 - (\sum_{i=1}^{n}\ln x_i)^2}$$

$$\hat{\theta}_{ML} = (\prod_{i=1}^{n} x_i)^{1/n}\exp(\frac{k^2}{\hat{\alpha}_{ML}}) \tag{4.14}$$

Note that if $\hat{\theta}_{tem}$ is closer to x_1 or x_n Pareto or lognormal will respectively be a superior model than the composite lognormal-pareto model. In order to find the ML estimators, one needs to check only $n-1$ intervals.

After the ML estimators have been found, one can use it in combination with QQ-plotting to find a nice estimate.

4.6 Simulation in practice

This will be a walk-through on how to actually simulate a loss function in a suitable program.

Step 1. Calculate α and the distributions to use in order to simulate the manual reserve using historical data. Note that these data must contain information on both how much was paid out for a claim and how much was reserved in the beginning for a claim. The reserve for a manual claim will later be drawn by taking $r_i^{manual} = s_i + s_i^{\alpha} \cdot t(0, \nu, \sigma)$ or $r_i^{manual} = s_i + s_i^{\alpha} \cdot N(0, \sigma)$.

Step 2. Decide on a distribution for the claims and if using parametric sampling, calculate the parameters for the distribution.

Step 3. Start a loop that will repeat once for every value of u (the large claims limit).

Step 4. Calculate the standard reserve value, $r_i^{standard}$ for the current u. This is done using information in 4.2. In order to decrease the random error, one should use a large number of values when sampling from the distribution.

Step 5. Calculate λ and decide on how many claims shall be simulated, n, by sampling from $Po(\lambda)$.

Step 6. Draw a value from the distribution, s_i. If s_i is smaller than the current u, give it the standard reserve $r_i^{standard}$. If it is larger than the

current u give it the manual reserve r_i^{manual}. When the reserve for s_i has been decided, take $\theta_i = (r_i - s_i)$ and calculate the loss function, $g(\theta_i)$, that has been decided on, for example $g(\theta_i) = |\theta_i|$. Do this procedure n times.

Step 7. Sum all $g(\theta_i)$ to create the total "error" for the current value of u.

Step 8. Repeat steps 4, 5, 6 and 7 a large number of times and take the average of these values. This will decrease the random error.

Step 9. Change the value of u and start over from Step 3. In theory one should do this step for $u \to \infty$. In practice however one must decide on a suitably large value for u.

Step 10. Plot the error as a function of u

Methods similar to this one can calculate very difficult stochastic problems. Problems which would be very hard or even impossible to calculate explicitly. The fact that we simulate the event a large number of times will decrease the random effect we are creating every time we draw a sample.

Chapter 5

Results

In this section the theory and modeling is tested using actual data. We want to simulate the claims that are received by the insurance company over the course of a year. We have decided to use fire insurance claims, motor vehicle claims and cargo claims. These were chosen because they intuitively tend to be quite different fields. They were also chosen because the raw data contain sufficiently many manually estimated claims. This means we were able to make a proper model for the claim handlers' distributions. At first, simulation has been done one claim type at a time. Then all claim types are merged together making a mixed simulation. A comparison is made to investigate whether a better result can be achieved by narrowing or widening the definition of a claim type. Ideally one would also like to split each field into different sub-categories. For example, motor could be split into motor-hull damage, etc. Unfortunately, there are not enough large claims on each of these categories to make proper estimates.

5.1 Fire claims

The data used are from fire insurance claims from 2008 and 2009. We have used 2000 (claims) as λ, when calculating the number of claims received each year. This is approximately the number of fire claims received each year. They consist of a matrix. The matrix has two columns. One column contains the original reserve of a single claim, and the other contains the actual payment of the same claim. The claims are considered to be closed and the payment is considered to have been paid out in a lump sum.

5.1.1 Deciding the distribution for claimhandlers

We want to calculate α. Different values of α in equation 3.11 is plotted in a histogram until the graph looks approximately student-t-distributed. For these data $\alpha = 0.5$ made equation 3.11 approximately student-t-distributed.

This can be seen in figure 5.1. With a maximum likelihood estimation of the parameters we find:

$$\mu = 3.86323$$

$$\sigma = 417.396$$

$$\nu = 1.876 \tag{5.1}$$

where μ is the expected value, σ is the standard deviation and ν is the degrees of freedom of the student-t-distribution.

It can also be seen that the qq-plot in figure 5.2 looks nice. The real values seem to be approximately the same distribution as the fitted values.

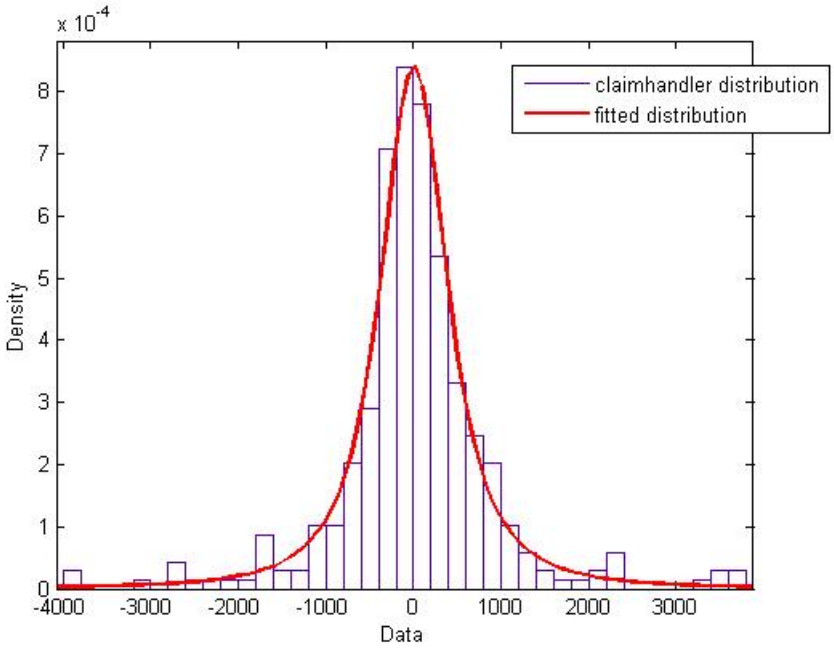

Figure 5.1: With a trial and error approach α was found to be roughly 0.5. This means that $\frac{s_i - r_i}{\sqrt{s_i}}$ is roughly student-t distributed

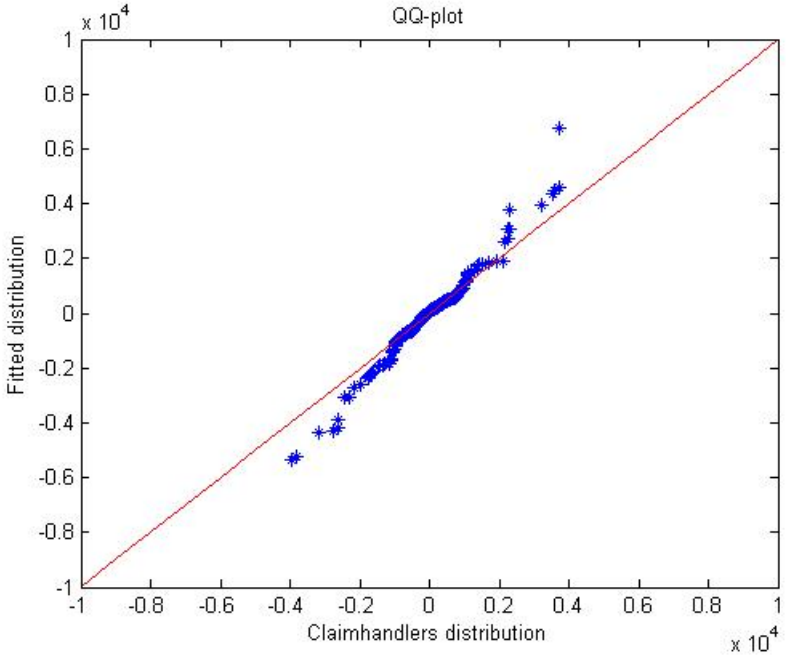

Figure 5.2: The qq-plot seems linear. This means that the fitted distribution has roughly the same distribution as the real values coming from equation 3.11.

5.1.2 Non-parametric bootstrapping

We try to use non-parametric bootstrapping to measure the large claims limit. We create a loop that calculates the mean squared error for different u ranging from 0 to $3,000,000$. For every u a certain number of steps are made. We start with u=0.

First, the standard reserve is calculated by drawing values from the claim payments vector a large number of times. The ones below the large claims limit u is then sorted out. The mean from these values is then calculated as the standard reserve.

Second, we calculate how many samples, n , that will be generated by sampling one value from $Po(\lambda)$. λ is calculated by taking $n_{tot} \cdot p$, where n_{tot} is the total number of insured and p is the probability of actually having a claim. Call the number of claims in the claims vector k. p is approximated to be k/n_{tot}. This means $\lambda = k$.

Third, we draw n random samples from the vector of claim payments.

For every sample s_i, if it is larger than the current u, give it a reserve by drawing a sample from $r_i^{manual} = s_i + \sqrt{s_i} \cdot t(\mu, \sigma, \nu)$. If the sample is smaller than u give it the standard reserve $r_i^{standard}$ calculated earlier.

Fourth, we go through the n samples and give every generated value a reserve. Calculate the loss function for every iteration by using r_i and s_i. In this case we have chosen to use the loss functions of equation 3.4 and equation 3.6 with $k_2 = 0$. The latter loss function will only punish negative reservation errors.

Fifth, do the second, third and fourth step a large number of times and take an average of the result. This reduces the random factor involved.

We enlarge u and do all steps again. We do this procedure until u has reached $3,000,000$. Then we plot the loss function against its corresponding u. This is shown in figures 5.3 and 5.4.

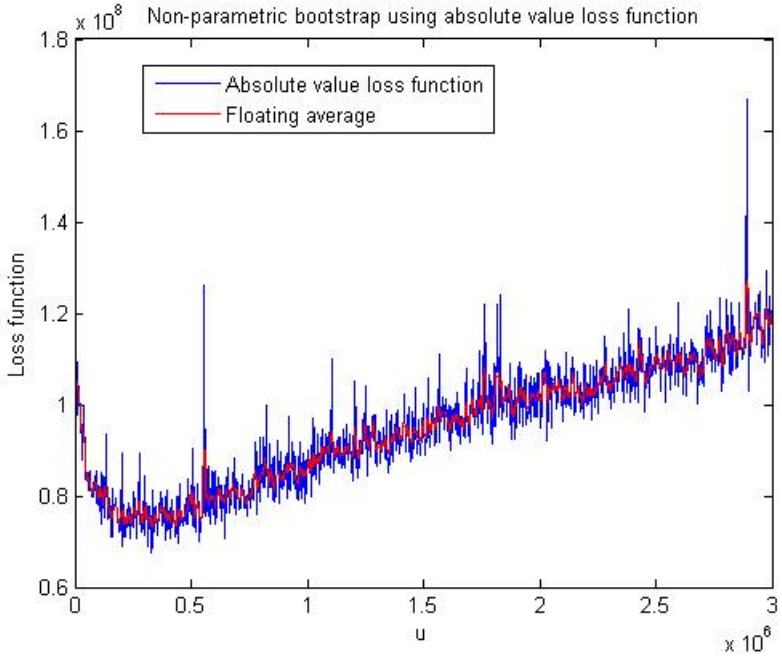

Figure 5.3: A non-parametric bootstrap of the large claims limit compared to the absolute value of the error. To make the graph a bit smoother a floating average has been added.

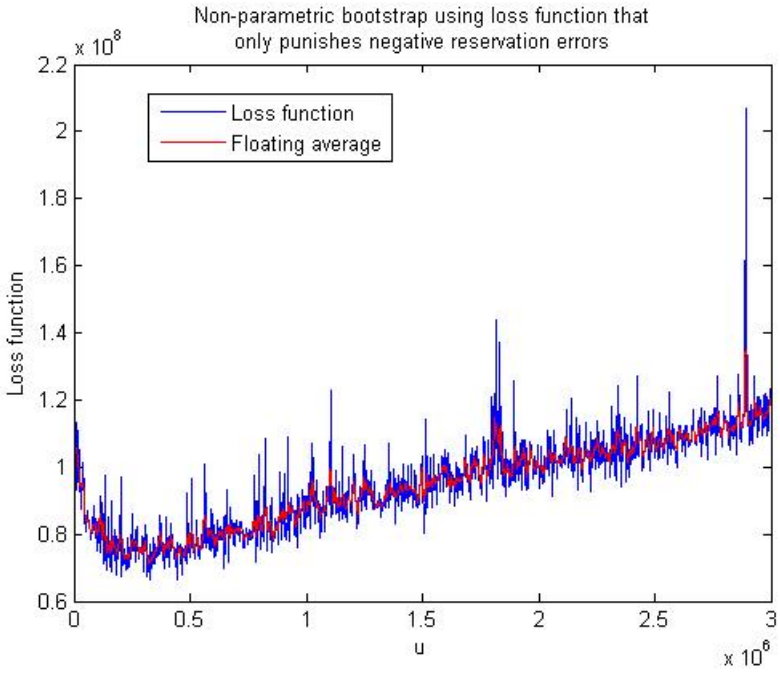

Figure 5.4: A non-parametric bootstrap of the large claims limit compared to the error generated by negative reservation errors. To make the graph a bit smoother a floating average has been added.

5.1.3 Parametric bootstrapping

Here we have to assume and test some distribution. The distribution should be fitting the fire insurance claim payments. Typically we can assume that fire insurance claims look more like a lognormal distribution close to zero and more like a pareto distribution in the tail. Therefore we have assumed a composite lognormal pareto distribution. We tried to fit the parameters to this distribution using maximum likelihood. After some fine-tuning. the qq-plot looks pretty nice. Figure 5.5 looks fairly linear. The parameters for the composite lognormal pareto distribution is estimated to be:

$$\theta = 100,000$$

$$\alpha = 0.7 \tag{5.2}$$

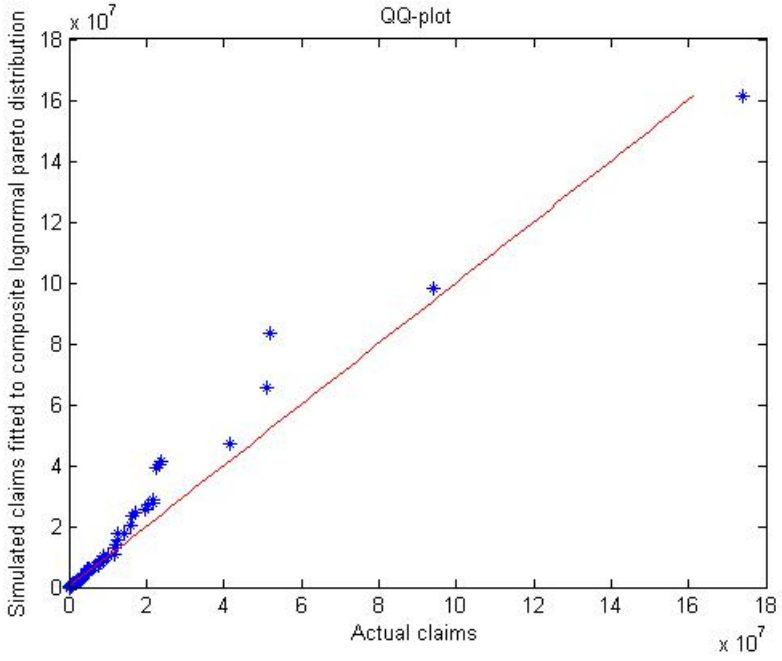

Figure 5.5: A qq-plot showing a fitted composite lognormal pareto distribution against real samples from fire claims. The plot looks linear. This means we can assume that the fitted distribution is roughly the same as the underlying distribution of the Norwegian fire claims.

Next, we simulate the functions compared to the large claims limit by repeating the same steps as in the non-parametric bootstrapping. The difference will be that instead of drawing samples from the actual vector of samples we will generate samples and create a new vector from the fitted distribution and draw samples from that.

For further analysis we also make two plots of the the absolute value loss function against u. One where there are only errors coming from manual reserve and one where there are only errors coming from standard reserve.

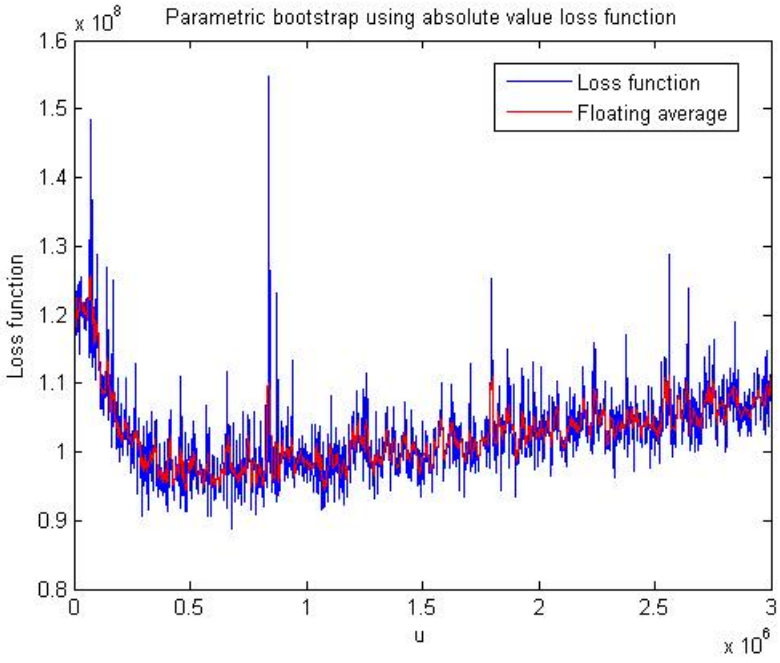

Figure 5.6: A parametric bootstrap of the large claims limit compared to the absoulte value of the error. To make the graph a bit smoother a floating average has been added.

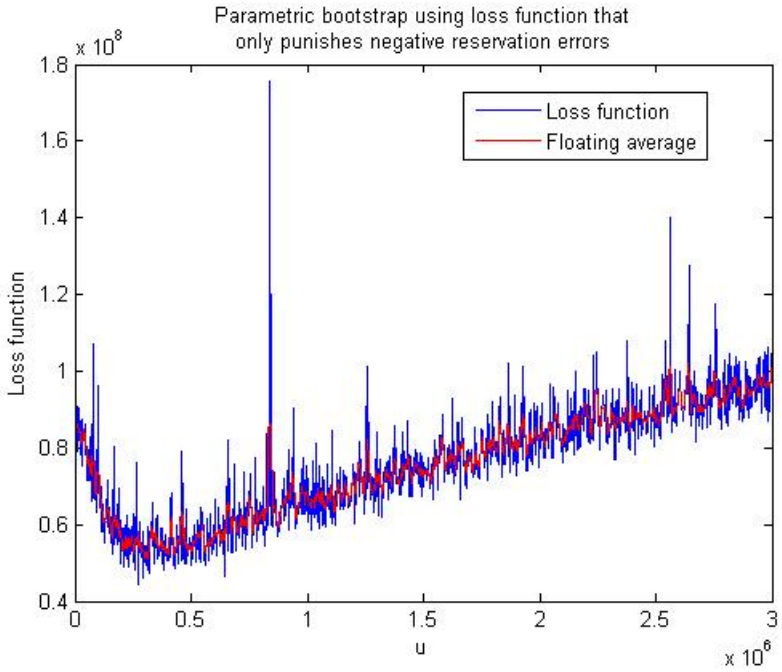

Figure 5.7: A parametric bootstrap of the large claims limit compared to the error generated by negative reservation errors. To make the graph a bit smoother a floating average has been added.

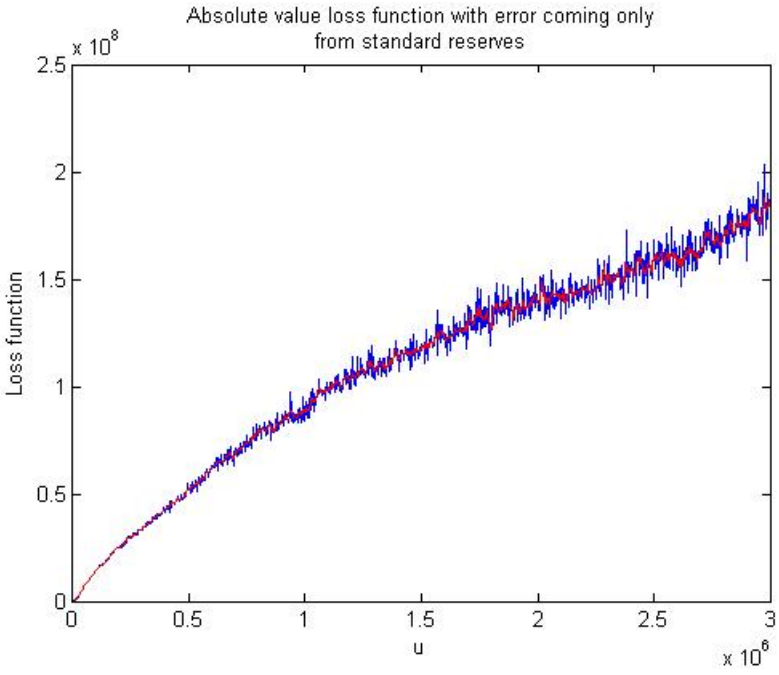

Figure 5.8: A non-parametric bootstrap containing only error coming from standard reserves. Absolute value loss function is used.

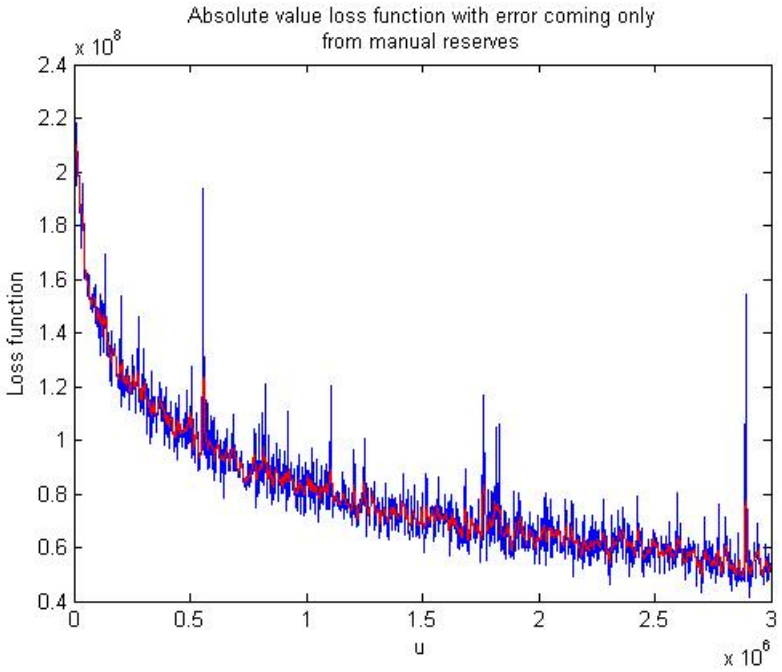

Figure 5.9: A non-parametric bootstrap containing only error coming from manual reserves. Absolute value loss function is used.

5.1.4 Optimal large claims limit

Can we decide on an actual optimal limit? If we look at figure 5.3 and 5.6, they look quite similar. Figure 5.3 has a minimum somewhere between 150,000 and 500,000 while figure 5.6 has a minimum somewhere between 250,000 and 500,000. For the non-parametric case, between 150,000 and 500,000, the mean variance was estimated to $8.3037 \cdot 10^{13}$ and the mean standard deviation was estimated to $4.9956 \cdot 10^6$. In the parametric case, between 250,000 and 500,000, the mean variance was estimated to $3.2739 \cdot 10^{13}$ and the mean standard deviation was estimated to $4.8887 \cdot 10^6$. Deciding on an actual large claims limit for fire claims would have to be a business decision. Using the absolute value loss function any limit between 250,000 and 500,000 would give approximately the same error. If choosing 250,000 as the limit around 85% of the claims would be standard reserved. If choosing 500 000 as the limit around 90% of the claims would be standard reserved. If the company can live with a slightly larger reservation error. An even higher limit can be chosen. Without enlarging the error a lot, one could put the limit at 3 000 000. Standard reserving would then be done on around 97.5% of the claims. This would save time for the claims handlers. However there are some negative sides. When putting the limit too high, some of the

assumptions made earlier might not hold. If the limit is 3 000 000, it could for example be very hard for a claims handler to decide whether a claim will land on $2,500,000$ or $4,000,000$ without detailed information about the claim. If the limit is low, large claims are more easily detected. Also the risk of one single claim upsetting the whole reserve error will be a lot lower if the limit is low.

In figure 5.4 and 5.7 we turn our attention to the other tested loss function, where we only measured negative reserving error. The behavior looks similar to the other loss function. The minimum error seems to be around the same interval as the absolute value loss function. The graphs probably looks similar because both types of reserving use a symmetrical method. The manual reserve uses a Student-t distribution, while the standard reserve uses the expected value.

In figures 5.8 and 5.9 we have plotted the absolute value loss function with errors coming only from standard reserves and then only from manual reserves. This gives us a way to analyze how the different components behave. We can see that the manual loss function has a quite negative slope in the beginning, when all or most of the claims will be booked as manual. As most claims are smaller than 50 000, the slope will level out quite quickly. Having a very low limit will certainly not be optimal. There would be a lot of work for the claim handlers and the error would be high. The standard reserve error also has a quite steep slope in the beginning. However, this graph is not as steep as in figure 5.9 and has a more steady growth. This makes the total error decrease in the beginning and then increase as the manual reserving error levels out.

5.2 Motor claims

The data used are from motor vehicle claims from 2000 to 2009. We have used 5000 (claims) as λ, when calculating the number of claims received each year. As before they consist of a matrix. The matrix has two columns. One column contains the original reserve of a single claim, and the other contains the actual payment of the same claim. The claims are considered to be closed and the payment is considered to have been paid out in a lump sum.

5.2.1 Deciding the distribution for claimhandlers

We use the same procedure as for the fire claims. We want to calculate α. Different values of α in equation 3.11 is plotted in a histogram until the graph looks approximately student-t-distributed. However, these data do not look student-t-distributed at all. It is too positively skewed. We therefore choose to use a generalized extreme value distribution instead. For these data $\alpha = 0.5$ made equation 3.11 approximately look like the generalized extreme value distribution. This can be seen in figure 5.10. With a maximum likelihood estimation of the parameters we find:

$$k = 0.155645$$

$$\sigma = 1961.26$$

$$\mu = 976.333 \tag{5.3}$$

Where μ is the location parameter, σ is the scale parameter and k is the shape parameter of the generalized extreme value distribution.

We can also see that the qq-plot in figure 5.11 looks ok. The real values could be approximately the same distribution as the fitted values.

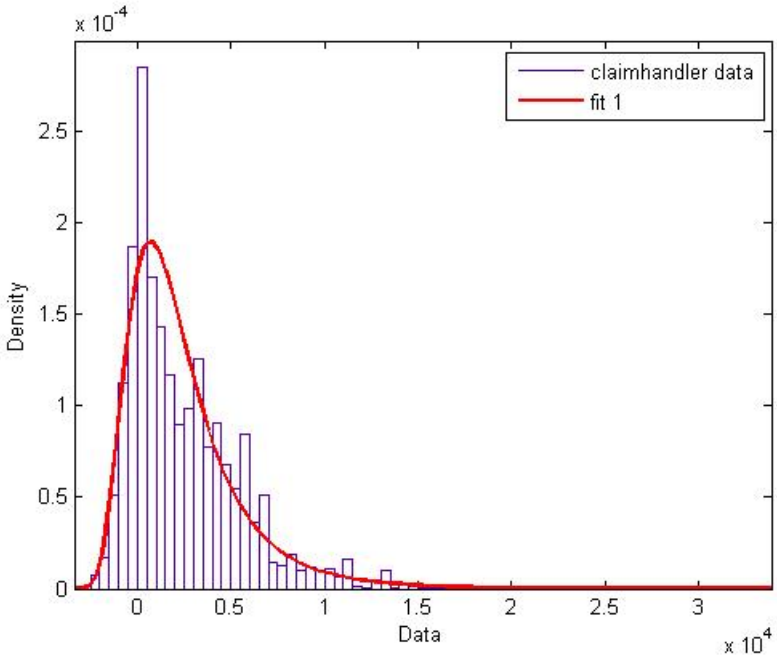

Figure 5.10: With a trial and error approach α was found to be roughly 0.5. This means that $\frac{s_i - r_i}{\sqrt{s_i}}$ is roughly generalized extreme value distributed.

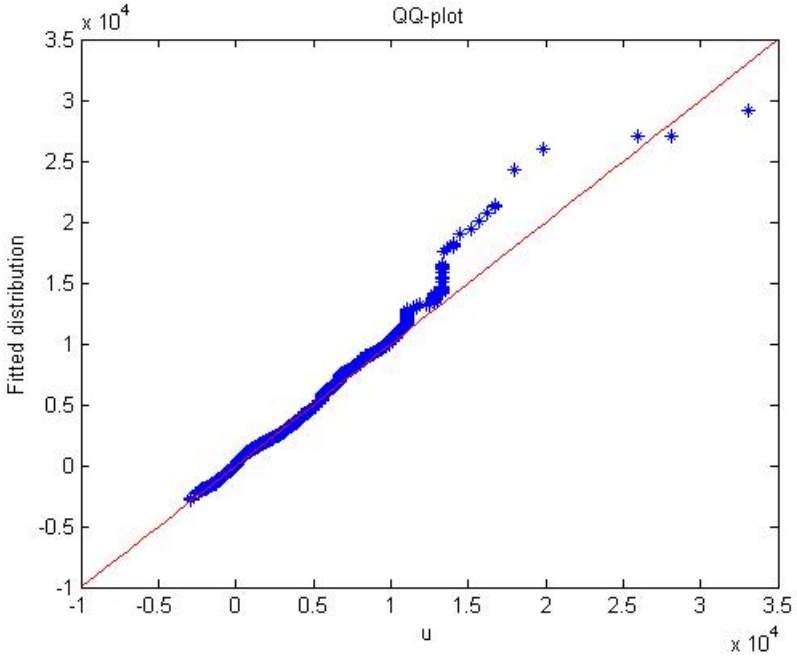

Figure 5.11: The qq-plot seems approximately linear. This means that the fitted distribution has roughly the same distribution as the real values coming from equation 3.11.

5.2.2 Non-parametric bootstrapping

We try to use non-parametric bootstrapping to measure the large claims limit. We create a loop that calculates the mean squared error for different u ranging from 0 to $3,000,000$. For every u a certain number of steps are made. We start with u=0.

First, the standard reserve is calculated by drawing values from the claim payments vector a large number of times. The ones below the large claims limit u is then sorted out. The mean from these values is then calculated as the standard reserve.

Second, we calculate how many samples, n , will be generated by sampling one value from $Po(\lambda)$. λ is calculated by taking $n_{tot} \cdot p$, where n_{tot} is the total number of insured and p is the probability of actually having a claim. Call the number of claims in the claims vector k. p is approximated to be k/n_{tot}. This means $\lambda = k$.

Third, we draw n random samples from the vector of claim payments.

For every sample s_i, if it is larger than the current u, give it a reserve by drawing a sample from $r_i^{manual} = s_i + \sqrt{s_i} \cdot t(\mu, \sigma, \nu)$. If the sample is smaller than u give it the standard reserve $r_i^{standard}$ calculated earlier.

Fourth, we go through the n samples and give every generated value a reserve. Calculate the loss function for every iteration by using r_i and s_i.As before we have chosen to use the loss functions of equation 3.4 and equation 3.6 with $k_2 = 0$. The latter loss function will only punish negative reservation errors.

Fifth, do the second, third and fourth step a large number of times and take an average of the result. This reduces the random factor involved.

We enlarge u and do all steps again. We do this procedure until u has reached $3,000,000$. Then we plot the loss function against its corresponding u. This is shown in figures 5.12 and 5.13.

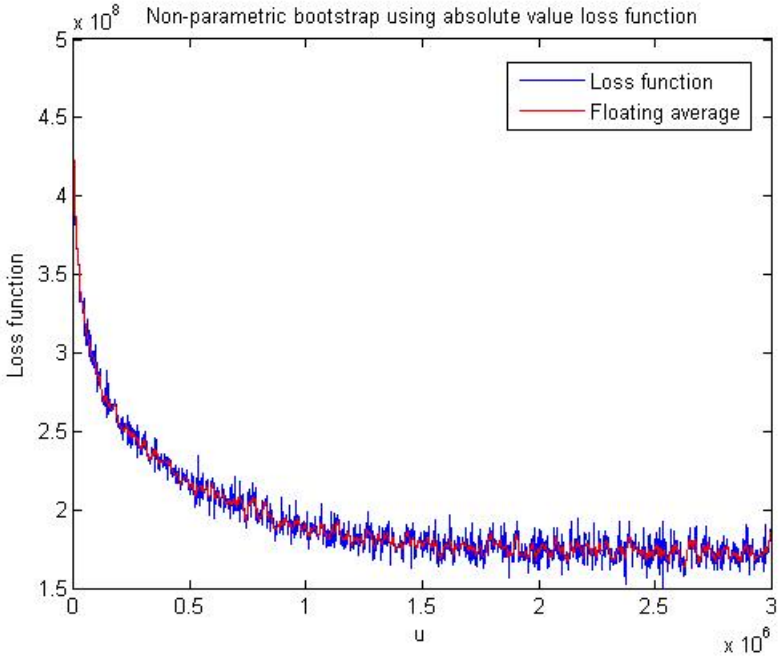

Figure 5.12: A non-parametric bootstrap of the large claims limit compared to the absolute value of the error. To make the graph a bit smoother a floating average has been added.

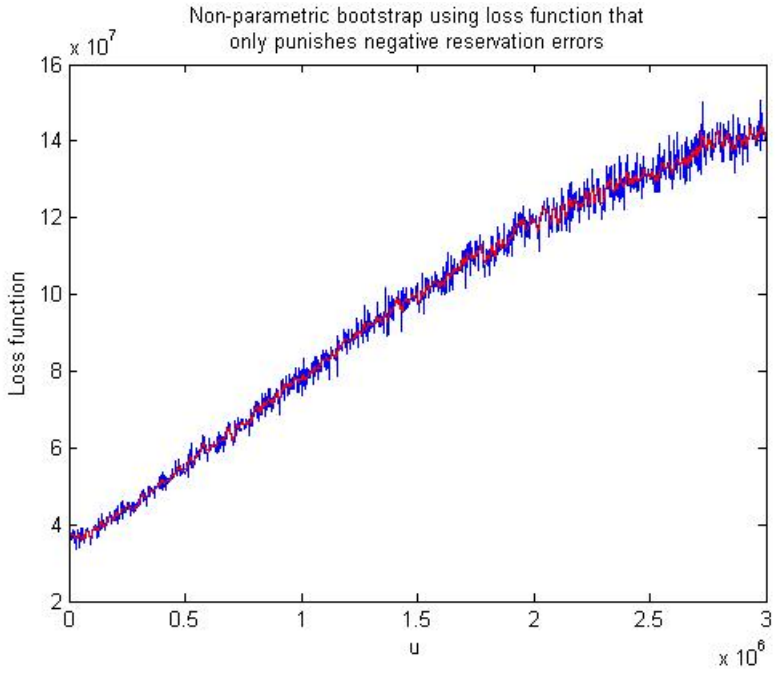

Figure 5.13: A non-parametric bootstrap of the large claims limit compared to the error generated by negative reservation errors. To make the graph a bit smoother a floating average has been added.

5.2.3 Parametric bootstrapping

We have to assume and test some distribution. The distribution should be fitting the motor insurance claim payments. The tail of motor insurance claims does not look as large as for fire insurance claims. Therefore we have assumed a lognormal distribution. We tried to fit the parameters to this distribution using maximum likelihood. The qq-plot of figure 5.14 looks fairly linear. The parameters for the lognormal distribution is estimated to be:

$$\mu = 9.1$$

$$\sigma = 2.2 \tag{5.4}$$

Next, we simulate the loss functions compared to the large claims limit by repeating the same steps as in the non-parametric bootstrapping. The difference will be that instead of drawing samples from the actual vector of samples we will generate samples and create a new vector from the fitted distribution and draw samples from that.

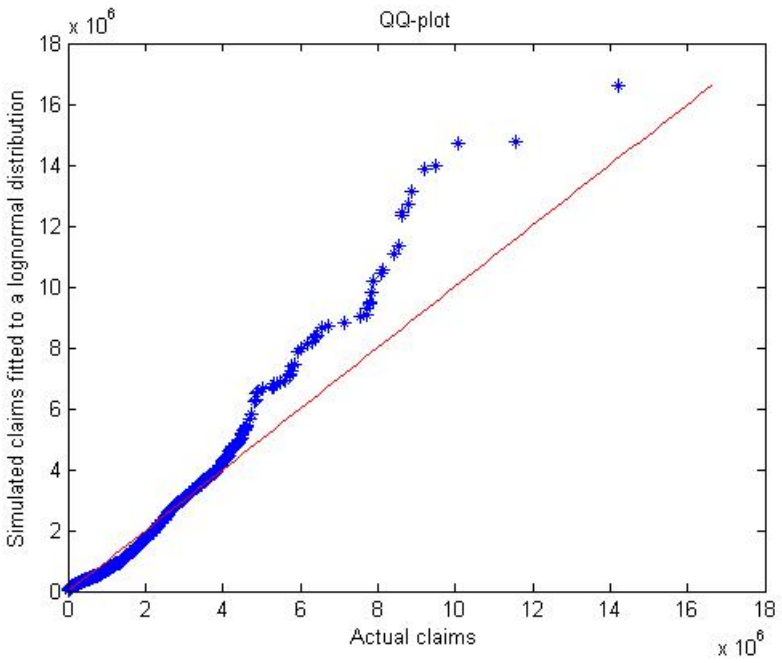

Figure 5.14: A qq-plot showing a fitted lognormal distribution against real samples from motor insurance claims. The plot looks linear. This means we can assume that the fitted distribution is roughly the same as the underlying distribution.

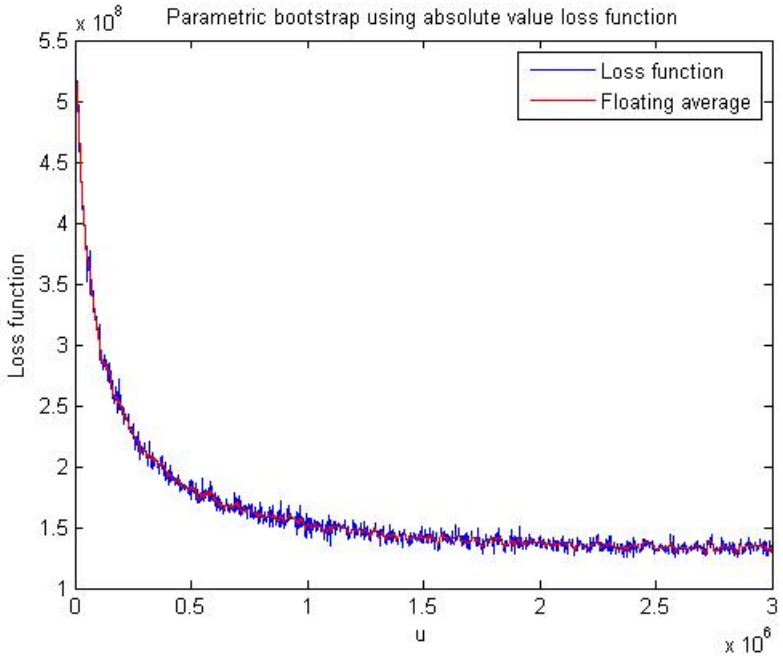

Figure 5.15: A parametric bootstrap of the large claims limit compared to the absoulte value of the error. To make the graph a bit smoother a floating average has been added.

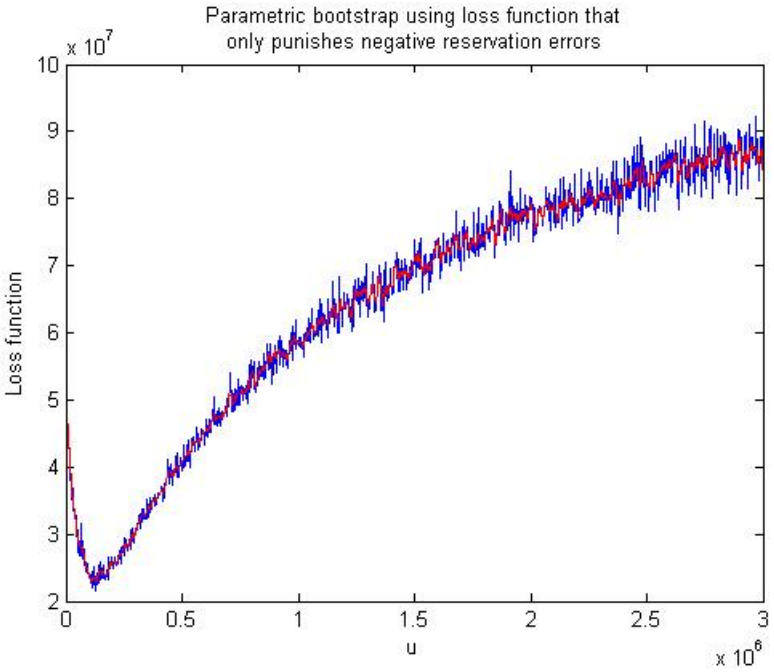

Figure 5.16: A parametric bootstrap of the large claims limit compared to the error generated by negative reservation errors. To make the graph a bit smoother a floating average has been added.

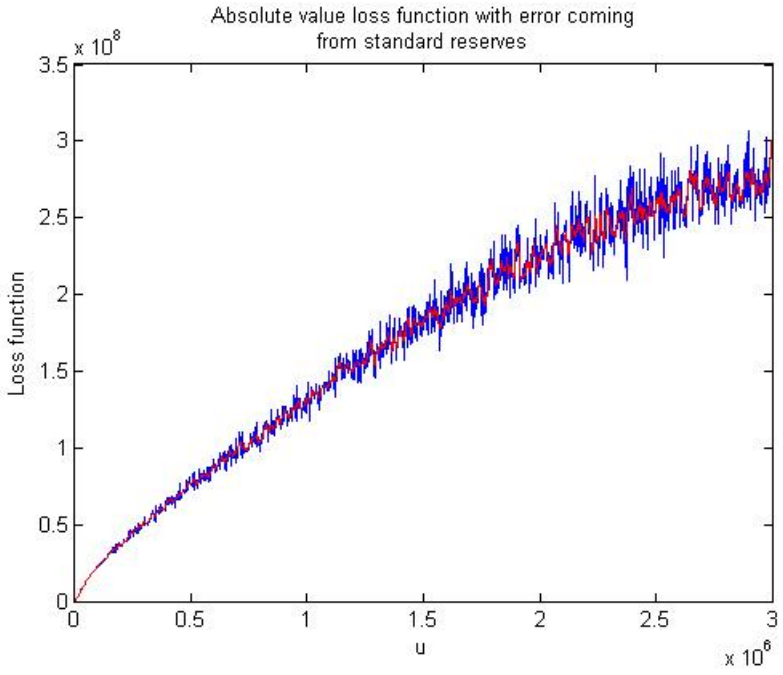

Figure 5.17: A non-parametric bootstrap containing only error coming from standard reserves. Absolute value loss function is used.

For further analysis we make two more plots of the the mean squared error against u. One where there is only error coming from manual reserve and one where there is only error coming from standard reserve.

5.2.4 Optimal large claims limit

If we look at figure 5.12 and 5.15, they look quite different from the fire claims. The figures both have a minimum of around $1,500,000$. For the non-parametric case, $1,500,000$, has a variance of about $3.5356 \cdot 10^{13}$ and the standard deviation was estimated to $5.7569 \cdot 10^6$. In the parametric case, $1,500,000$, had a variance of around $2.3465 \cdot 10^{13}$ and the standard deviation was estimated to $1.1269 \cdot 10^6$. What makes the motor claims special is that they have a very skewed claim handling distribution.
If we look at figures 5.17 and 5.18, we see that the manual reserving error is steeper than the standard in the beginning. However the slope of the manual reserve evens out to match the standard reserve around $1,500,000$.

This is also a good example telling us that different loss functions will produce different results. Figure 5.13 and 5.16 look totally different compared to figure 5.12 and 5.15. In the former we use the loss function generated by

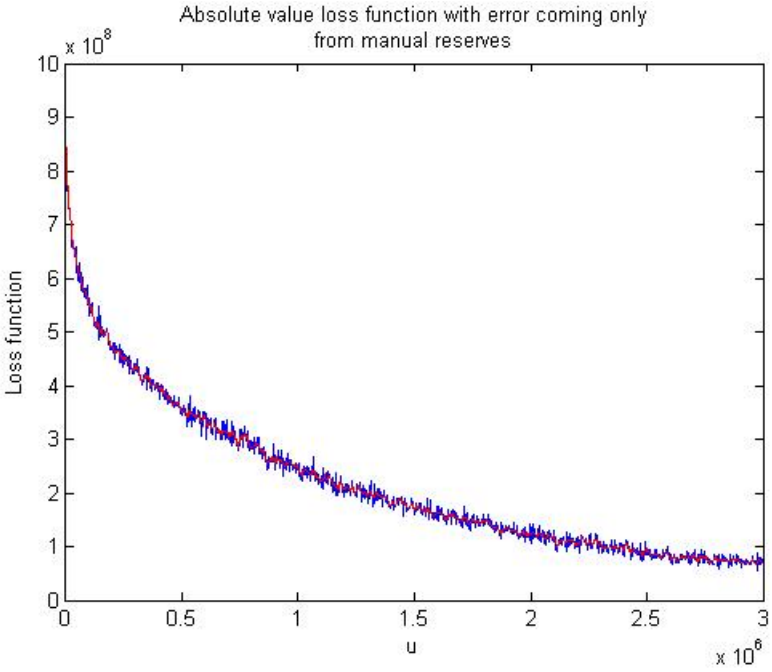

Figure 5.18: A non-parametric bootstrap containing only error coming from manual reserves. Absolute value loss function is used.

negative reserving errors, while in the latter we use the absolute value loss function. Because of the positive skewness of the claim handler distribution, one will have a much greater chance of having negative reserving errors if the large claims limit is high. Having a high limit means that a lot of claims will be standard reserved. Here, the probability of having a negative result will be as probable as having a positive result. While a manually reserved claim will have a much higher chance of having a positive result. Therefore there will be less error when a lot of claims are manually reserved.

5.3 Cargo claims

The data used are from cargo insurance claims from 1999 to 2010. We have used 4000 (claims) as λ, when calculating the number of claims received each year. The matrix of data has two columns. One column contains the original reserve of a single claim, and the other contains the actual payment of the same claim. The claims are considered to be closed and the payment is considered to have been paid out in a lump sum.

5.3.1 Deciding the distribution for claimhandlers

We want to calculate α. Different values of α in equation 3.11 is plotted in a histogram until the graph looks approximately student-t-distributed. For this data, $\alpha = 0.5$ made equation 3.11 approximately look like the student-t distribution. Altough slightly positively skewed. This can be seen in figure 5.19. With a maximum likelihood estimation of the parameters we find:

$$\mu = 60.3809$$

$$\sigma = 164.45$$

$$\nu = 2.9 \tag{5.5}$$

Where μ is the expected value, σ is the standard deviation and ν is the degrees of freedom of the student-t-distribution.

We can also see that the qq-plot in figure 5.20 looks ok. The real values could be approximately the same distribution as the fitted values.

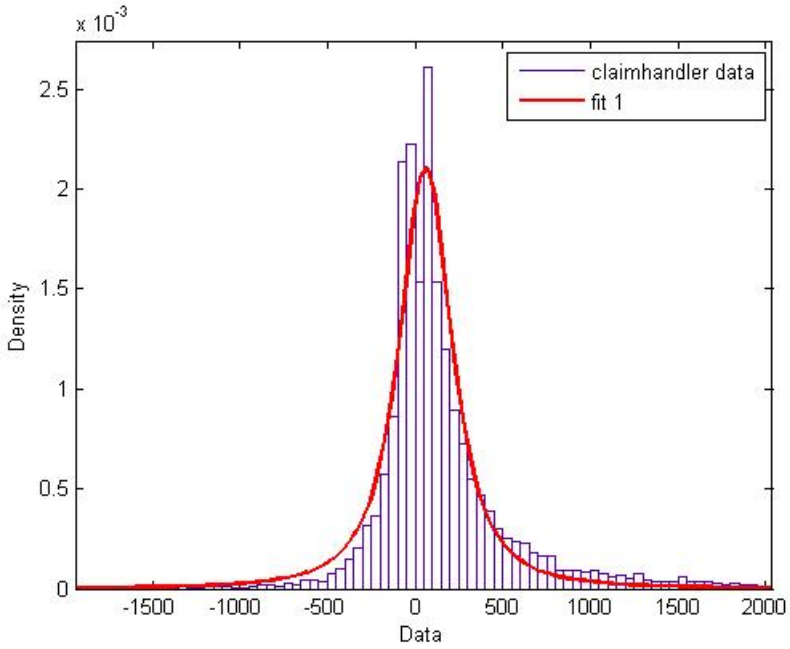

Figure 5.19: With a trial and error approach α was found to be roughly 0.5. This means that $\frac{s_i - r_i}{\sqrt{s_i}}$ is roughly student-t distributed.

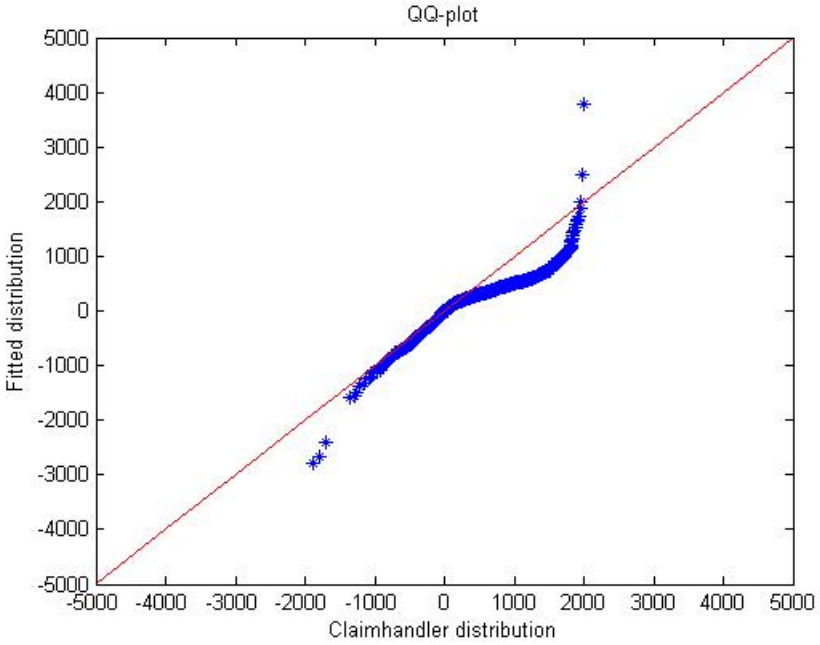

Figure 5.20: The qq-plot seems approximately linear. This means that the fitted distribution has roughly the same distribution as the real values coming from equation 3.11.

5.3.2 Non-parametric bootstrapping

We try to use non-parametric bootstrapping to measure the large claims limit. We create a loop that calculates the mean squared error for different u ranging from 0 to $3,000,000$. The cargo claims looks to be generally smaller than the other two types of claims. For every u a certain number of steps are made. We start with u=0.

First, the standard reserve is calculated by drawing values from the claim payments vector a large number of times. The ones below the large claims limit u is then sorted out. The mean from these values is then calculated as the standard reserve.

Second, we calculate how many samples, n, will be generated by sampling one value from $Po(\lambda)$. λ is calculated by taking $n_{tot} \cdot p$, where n_{tot} is the total number of insured and p is the probability of actually having a claim. Call the number of claims in the claims vector k. p is approximated to be k/n_{tot}. This means $\lambda = k$.

Third, we draw n random samples from the vector of claim payments. For every sample s_i, if it is larger than the current u, give it a reserve by drawing a sample from $r_i^{manual} = s_i + \sqrt{s_i} \cdot t(\mu, \sigma, \nu)$. If the sample is smaller than u give it the standard reserve $r_i^{standard}$ calculated earlier.

Fourth, we go through the n samples and give every generated value a reserve. Calculate the loss function for every iteration by using r_i and s_i.As before we have chosen to use the loss functions of equation 3.4 and equation 3.6 with $k_2 = 0$. The latter loss function will only punish negative reservation errors.

Fifth, do the second, third and fourth step a large number of times and take an average of the result. This reduces the random factor involved.

We enlarge u and do all steps again. We do this procedure until u has reached $3,000,000$. Then we plot the loss function against its corresponding u. This is shown in figures 5.21 and 5.22.

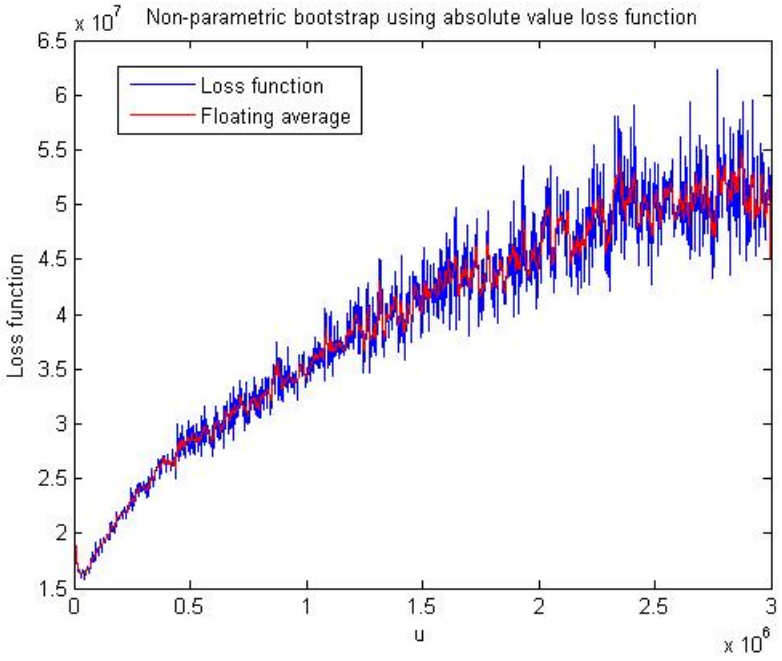

Figure 5.21: A non-parametric bootstrap of the large claims limit compared to the absolute value of the error. To make the graph a bit smoother a floating average has been added.

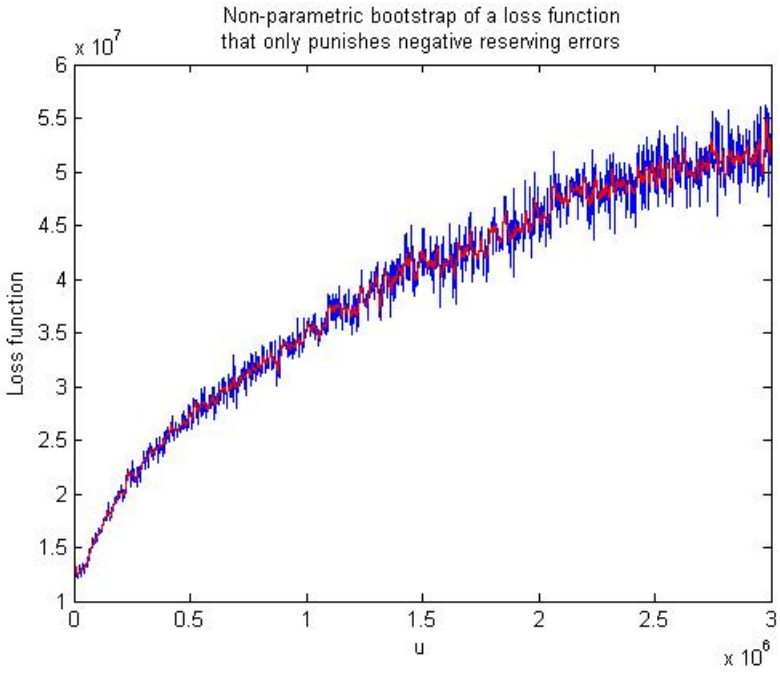

Figure 5.22: A non-parametric bootstrap of the large claims limit compared to the error generated by negative reservation errors. To make the graph a bit smoother a floating average has been added.

5.3.3 Parametric bootstrapping

We have to assume and test some distribution. The distribution should be fitting the cargo insurance claim payments. Like with the motor claims, the tail of cargo insurance claims does not look as large as for fire insurance claims. Therefore we have again assumed a lognormal distribution. We tried to fit the parameters to this distribution using maximum likelihood. The qq-plot of figure 5.23 looks fairly linear. In the tail, the figure does not look linear. In this case there was a trade of between a good fit close to zero and a good fit in the tail. Having the distribution fit being linear close to zero probably makes more sense. In real life, the extreme values will always be unpredictable. Smaller values will usually have a more predictable behavior. To catch this behavior a distribution that looks linear close to zero has been chosen. The parameters for the lognormal distribution is estimated to be:

$$\mu = 8.6$$

$$\sigma = 2.0 \tag{5.6}$$

Next, we simulate the loss functions compared to the large claims limit by repeating the same steps as in the non-parametric bootstrapping. The difference will be that instead of drawing samples from the actual vector of samples we will generate samples and create a new vector from the fitted distribution and draw samples from that.

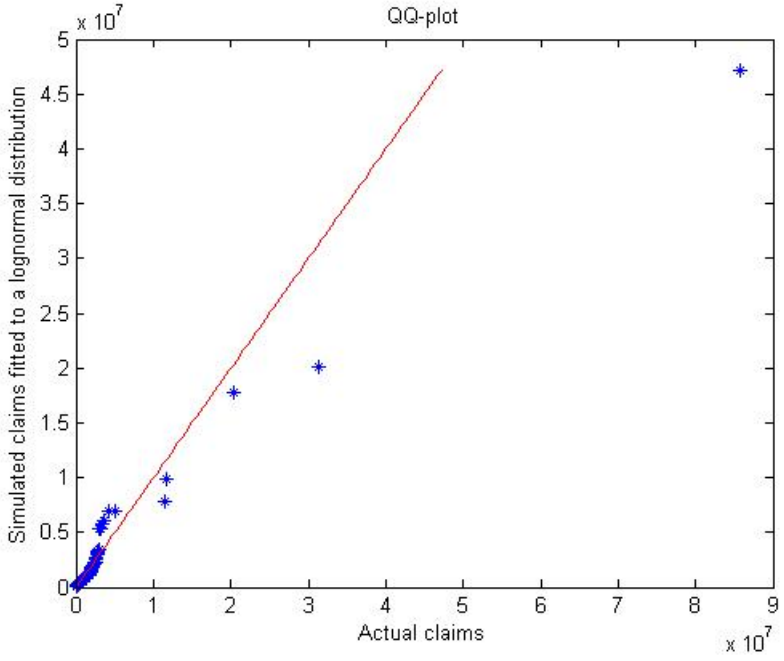

Figure 5.23: A qq-plot showing a fitted lognormal distribution against real samples from cargo insurance claims. Until 10^7 the plot looks approximately linear. The few observations in the tail does not look really linear. In this case I had to make a trade of between fitting the tail correctly and fitting the figure close to zero correctly. I have assumed that the fitted distribution is roughly the same as the underlying distribution.

For further analysis we make two more plots of the the mean squared error against u. One where there is only error coming from manual reserve and one where there is only error coming from standard reserve.

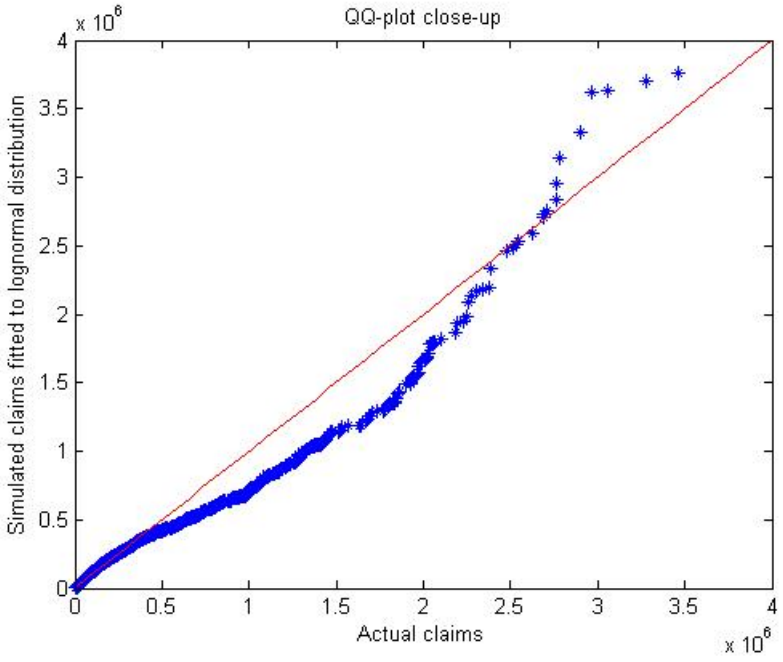

Figure 5.24: A zoom of figure 5.23.

5.3.4 Optimal large claims limit

Looking at figure 5.21 and 5.25, the minimum looks to be almost zero. In the parametric case one can more easily spot the minimum. It looks like there would be a minimum somewhere between $75,000$ and $125,000$. The non-parametric simulation an estimation of variance of about $1.9500 \cdot 10^{13}$ and a standard deviation estimate of around $4.4159 \cdot 10^6$. The parametric case had a variance of around $2.9121 * 10^{13}$ and the standard deviation was estimated to $5.3964 \cdot 10^6$. The reason for the appearance of the graphs is most likely the fact that 94% of cargo claims fall below $100,000$. Their are generally fewer "larger" claims than fire and motor. This will make the standard reserve value quite low. When claims larger than $100,000$ start being standard reserved, the error will grow quite rapidly. If we look at figures 5.28, we see that the manual reserving error drops rapidly in the beginning due to the sheer amount of small claims. However the slope levels out quickly and the standard reserve error of figure 5.27 will be dominating.

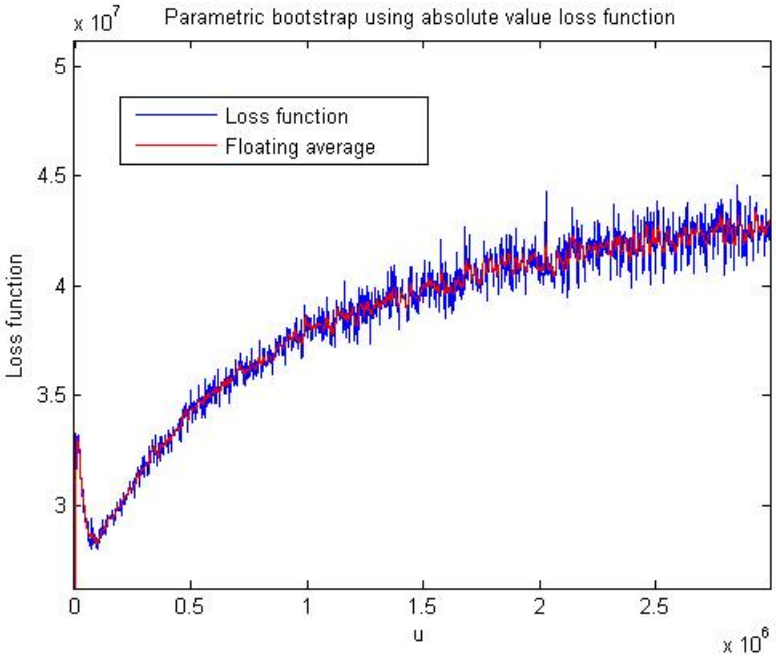

Figure 5.25: A parametric bootstrap of the large claims limit compared to the absoulte value of the error. To make the graph a bit smoother a floating average has been added.

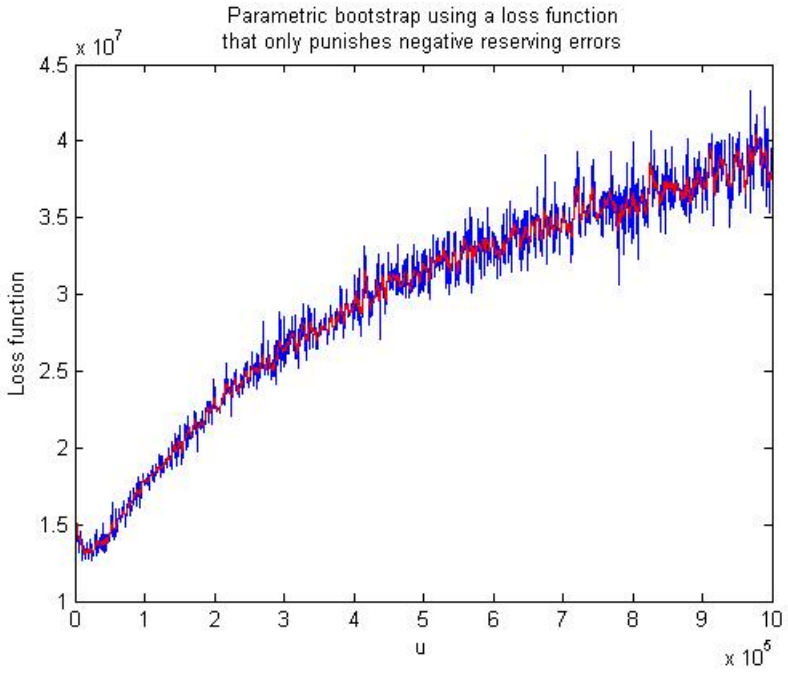

Figure 5.26: A parametric bootstrap of the large claims limit compared to the error generated by negative reservation errors. To make the graph a bit smoother a floating average has been added.

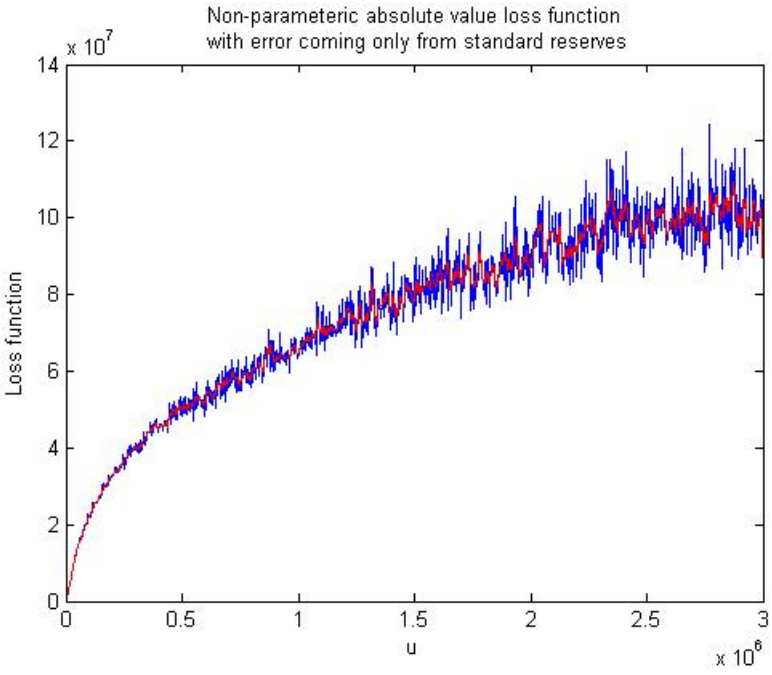

Figure 5.27: A non-parametric bootstrap containing only error coming from standard reserves. Absolute value loss function is used.

5.4 Mixed claims

Here we analyze if there is a point in differentiating among claim types. By making simulations and models around different claim types collectively, we will be able to compare it to a model where all types are treated individually. We have used the same data as for the fire, motor and cargo claims. In order to make a comparison, we have chosen $\lambda = 11,000$. Here $\lambda_{mixed} = \lambda_{fire} + \lambda_{motor} + \lambda_{cargo}$. This is supposed to simulate the total number of fire, motor and cargo claims received during a certain period. It is also supposed to simulate the distribution of different claim types. For example we have fewer fire claims than motor claims.

5.4.1 Deciding the distribution for claimhandlers

We use the same procedure as earlier. We want to calculate α. Different values of α in equation 3.11 is plotted in a histogram until the graph looks approximately student-t-distributed. However, the data is positively skewed. We therefore choose to use a generalized extreme value distribution, as with the motor claims. For this data $\alpha = 0.5$ made equation 3.11 approximately look like the generalized extreme value distribution. With a maximum like-

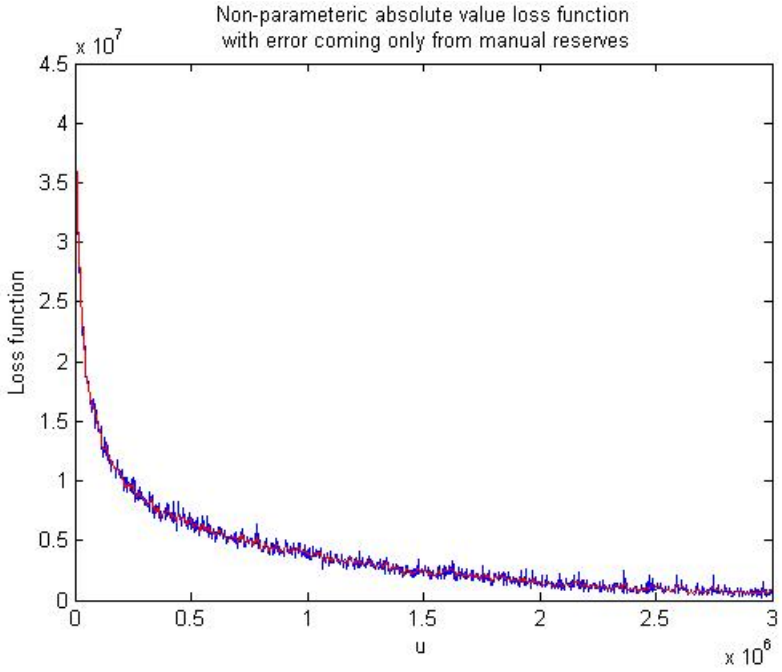

Figure 5.28: A non-parametric bootstrap containing only error coming from manual reserves. Absolute value loss function is used.

lihood estimation of the parameters we find:

$$k = 0.155645$$

$$\sigma = 1,961.26$$

$$\mu = 976.333 \tag{5.7}$$

Where μ is the location parameter, σ is the scale parameter and k is the shape parameter of the generalized extreme value distribution.

We can see that the qq-plot in figure 5.11 looks ok. The real values could be approximately the same distribution as the fitted values

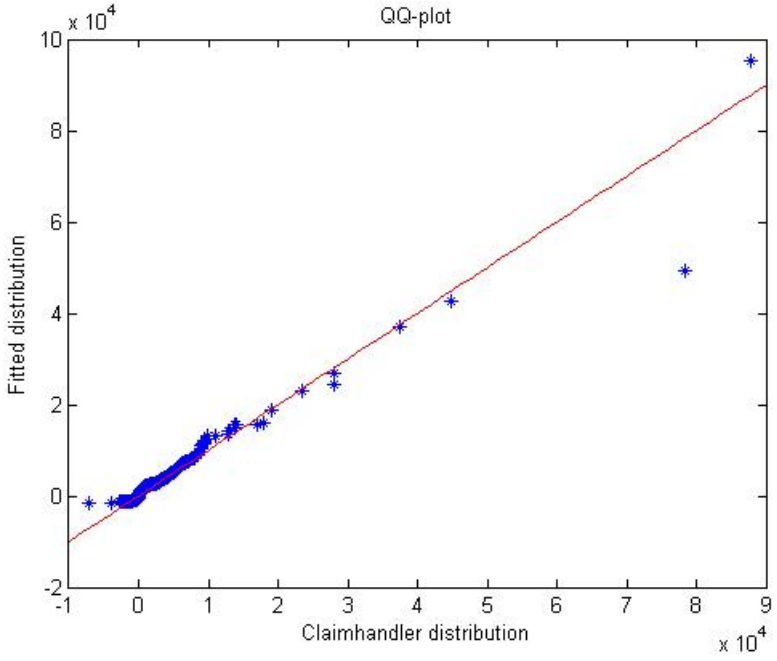

Figure 5.29: The qq-plot seems approximately linear. This means that the fitted distribution has roughly the same distribution as the real values coming from equation 3.11.

5.4.2 Non-parametric bootstrapping

We first try to use non-parametric bootstrapping to measure the large claims limit. We create a loop that calculates the mean squared error for different u ranging from 0 to $3,000,000$. For every u a certain number of steps are made. We start with u=0.

First, the standard reserve is calculated by drawing values from the claim payments vector a large number of times. The ones below the large claims limit u is then sorted out. The mean from these values is then calculated as the standard reserve.

Second, we calculate how many samples will be generated, n, by sampling one value from $Po(\lambda)$. λ is calculated by taking $n_{tot} \cdot p$, where n_{tot} is the total number of insured and p is the probability of actually having a claim. Call the number of claims in the claims vector k. p is approximated to be k/n_{tot}. This means $\lambda = k$.

Third, we draw n random samples from the vector of claim payments.

For every sample s_i, if it is larger than the current u, give it a reserve by drawing a sample from $r_i^{manual} = s_i + \sqrt{s_i} \cdot t(\mu, \sigma, \nu)$. If the sample is smaller than u give it the standard reserve $r_i^{standard}$ calculated earlier.

Fourth, we go through the n samples and give every generated value a reserve. Calculate the loss function for every iteration by using r_i and s_i. As before we have chosen to use the loss functions of equation 3.4 and equation 3.6 with $k_2 = 0$. The latter loss function will only punish negative reservation errors.

Fifth, do the second, third and fourth step a large number of times and take an average of the result. This reduces the random factor involved.

We enlarge u and do all steps again. We do this procedure until u has reached $3,000,000$. Then we plot the loss functions against its corresponding u. This is shown in figure 5.30 and 5.31.

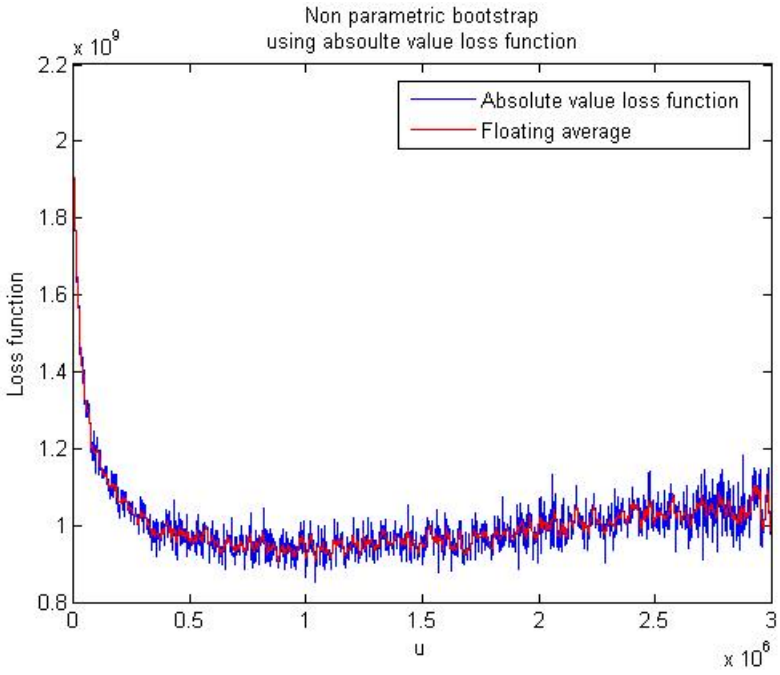

Figure 5.30: A non-parametric bootstrap of the large claims limit compared to the absolute value of the error. To make the graph a bit smoother a floating average has been added.

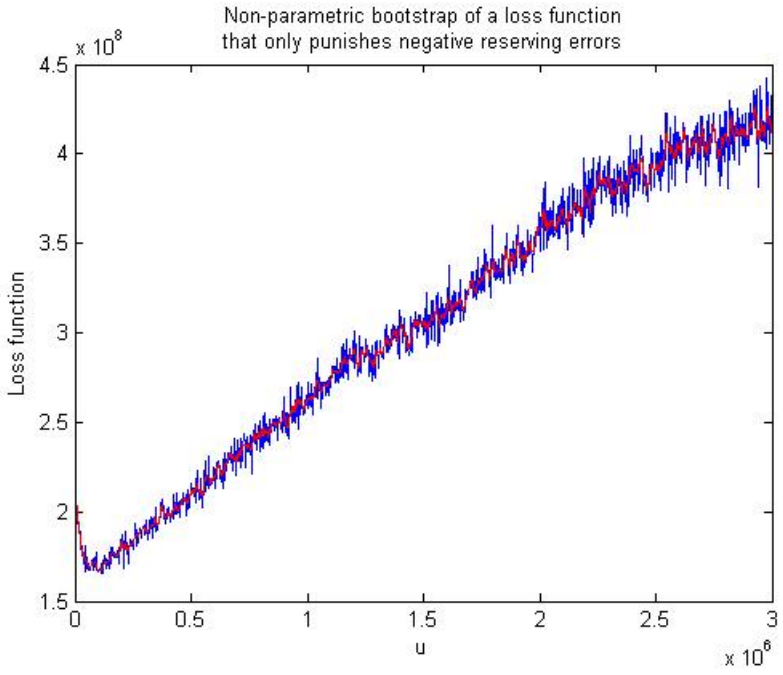

Figure 5.31: A non-parametric bootstrap of the large claims limit compared to the error generated by negative reservation errors. To make the graph a bit smoother a floating average has been added.

5.4.3 Parametric bootstrapping

The distribution should be fitting to all claim type at the same time. The best option seem to assume a lognormal distribution. The parameters to this distribution is estimated using maximum likelihood. The qq-plot of figure 5.32 looks fairly linear. The parameters for the lognormal distribution is estimated to be:

$$\mu = 9.774$$

$$\sigma = 2.257 \tag{5.8}$$

Next, we simulate the loss functions compared to the large claims limit by repeating the same steps as in the non-parametric bootstrapping. The difference will be that instead of drawing samples from the actual vector of samples we will generate samples and create a new vector from the fitted distribution and draw samples from that.

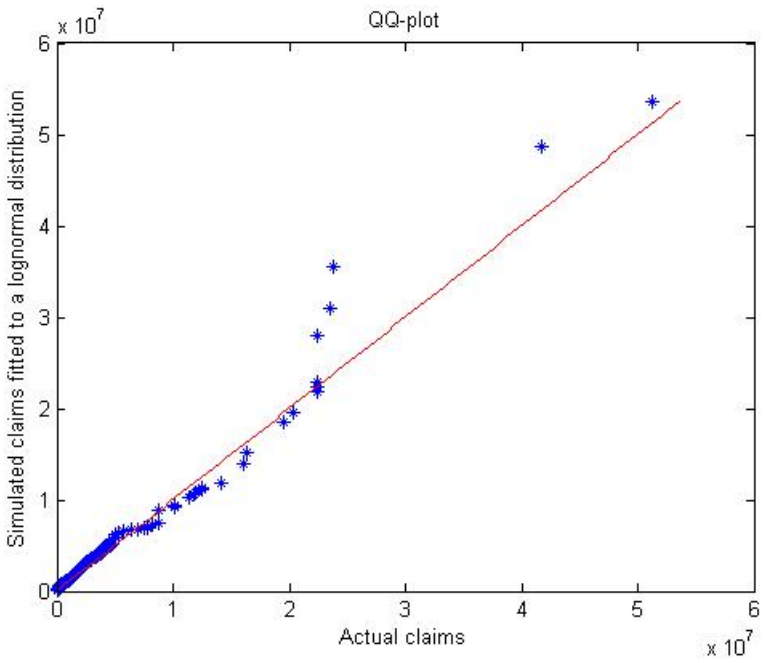

Figure 5.32: A qq-plot showing a fitted lognormal distribution against real samples from motor insurance claims. The plot looks linear. This means we can assume that the fitted distribution is roughly the same as the underlying distribution.

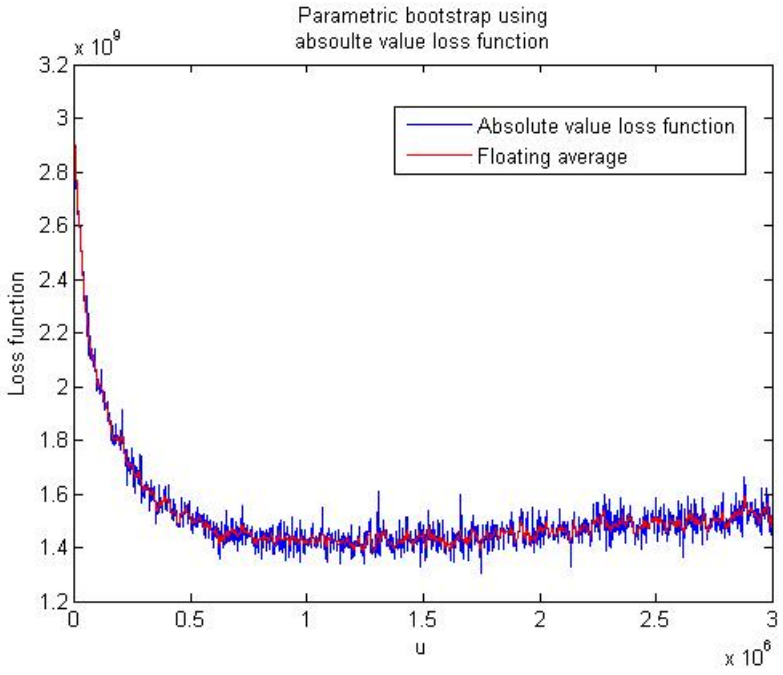

Figure 5.33: A parametric bootstrap of the large claims limit compared to the absoulte value of the error. To make the graph a bit smoother a floating average has been added.

For further analysis we make two more plots of the the mean squared error against u. One where there is only error coming from manual reserve and one where there is only error coming from standard reserve.

5.4.4 Comparison

To make a relevant analysis we need to compare the mixed claim simulation with the individual simulations. In figures 5.37 and 5.38 the upper graphs show the mixed claim simulation and the lower graphs show the individual estimates added together. This will simulate all claims received from the different claim types. It will make the graphs comparable. In both figures, a parametric bootstrap have been used.

In comparison the behavior of the mixed claims and individual claims look very similar. The large claims limit with minimal error looks to be the same in both graphs. However the scale of the error is quite different. In figure 5.37, the mixed version is around 5 times larger than the other one. In figure 5.38 the mixed version is about twice as large.

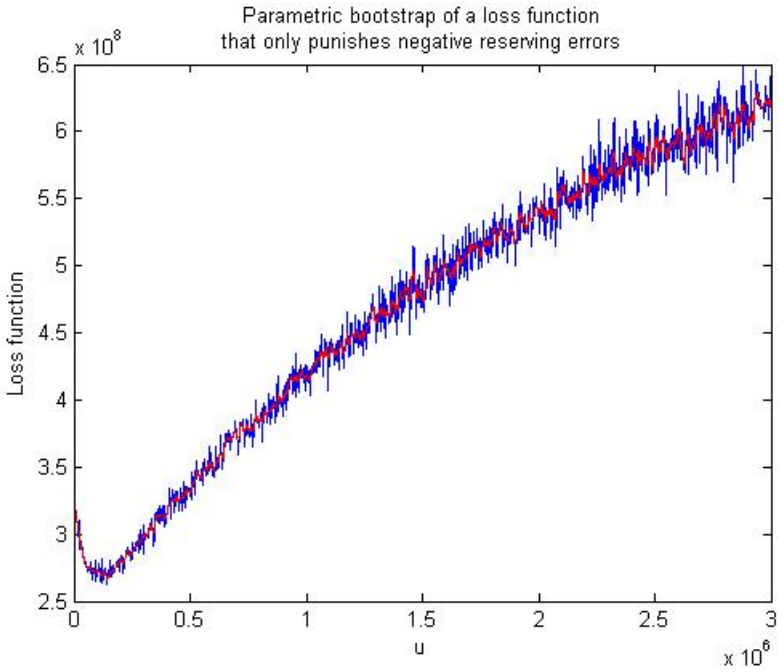

Figure 5.34: A parametric bootstrap of the large claims limit compared to the error generated by negative reservation errors. To make the graph a bit smoother a floating average has been added.

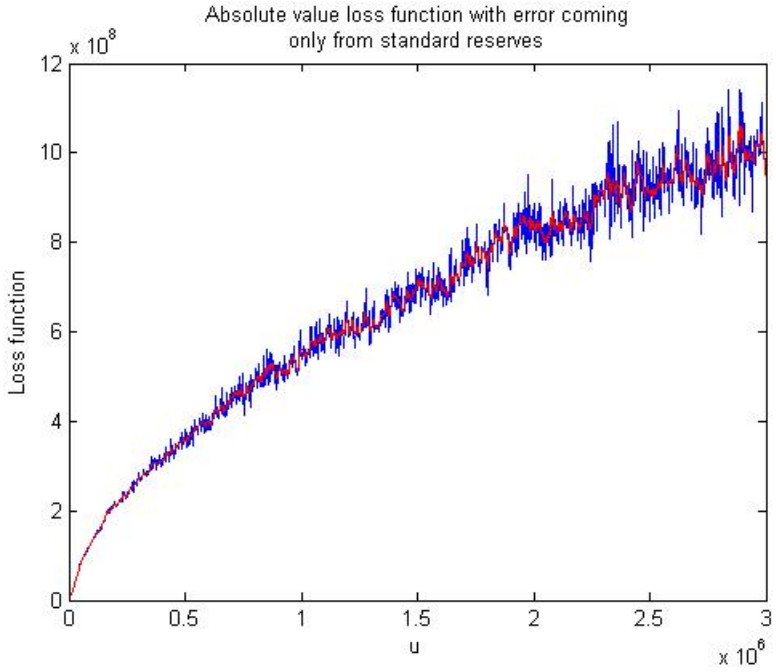

Figure 5.35: A non-parametric bootstrap containing only error coming from standard reserves. Absolute value loss function is used.

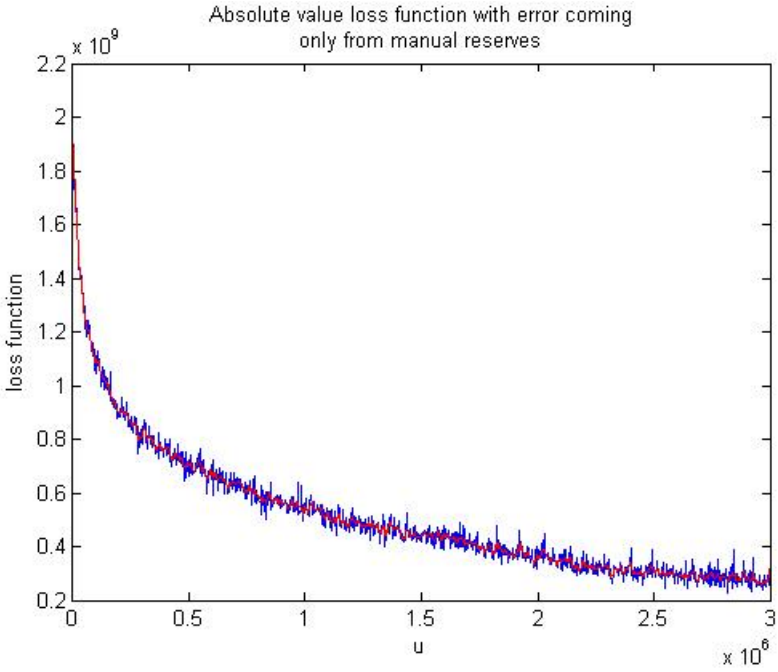

Figure 5.36: A non-parametric bootstrap containing only error coming from manual reserves. Absolute value loss function is used.

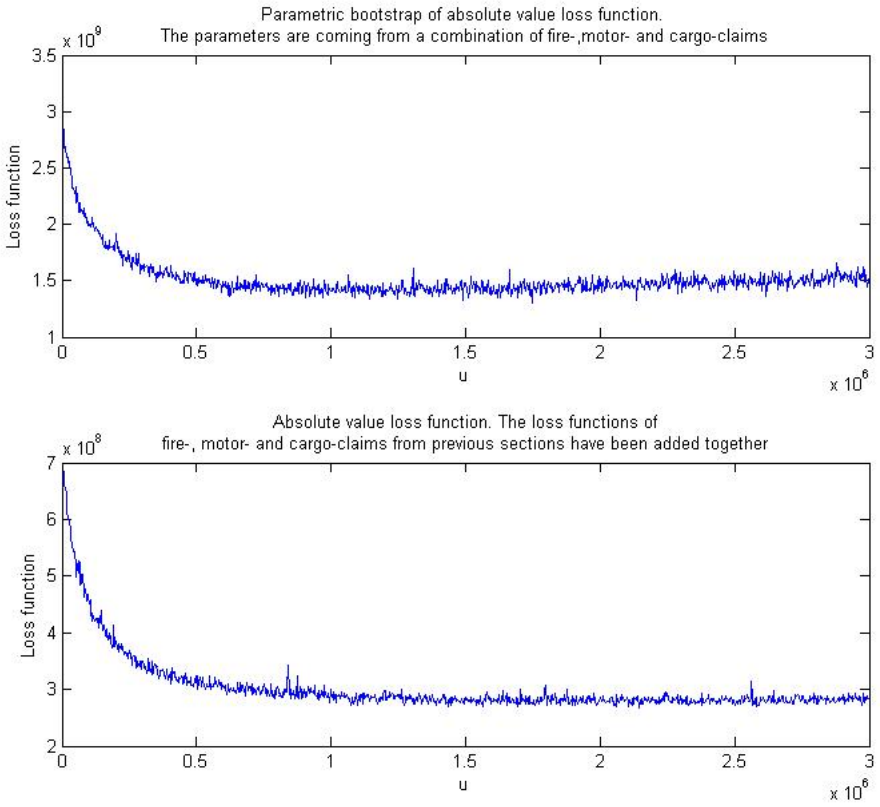

Figure 5.37: A comparison between the simulation of mixed claims compared to fire, motor and cargo claims modeled individually and added together. Absolute value loss function

The behavior of the figures also look a lot like the motor claim simulations. This is because their are a lot more motor claims (around 5000) than fire claims (around 2000). The cargo claims are too small to be dominating. This is probably one of the reasons for the larger error in the mixed claims simulation. This is logical since the model has to make a compromise when deciding both how the claim handlers will book their manual reserves and when estimating the actual distribution of the future payments. As we can see it will still catch the behavior that we see among the individual claims, but it will in turn have a larger error compared to individual estimations.

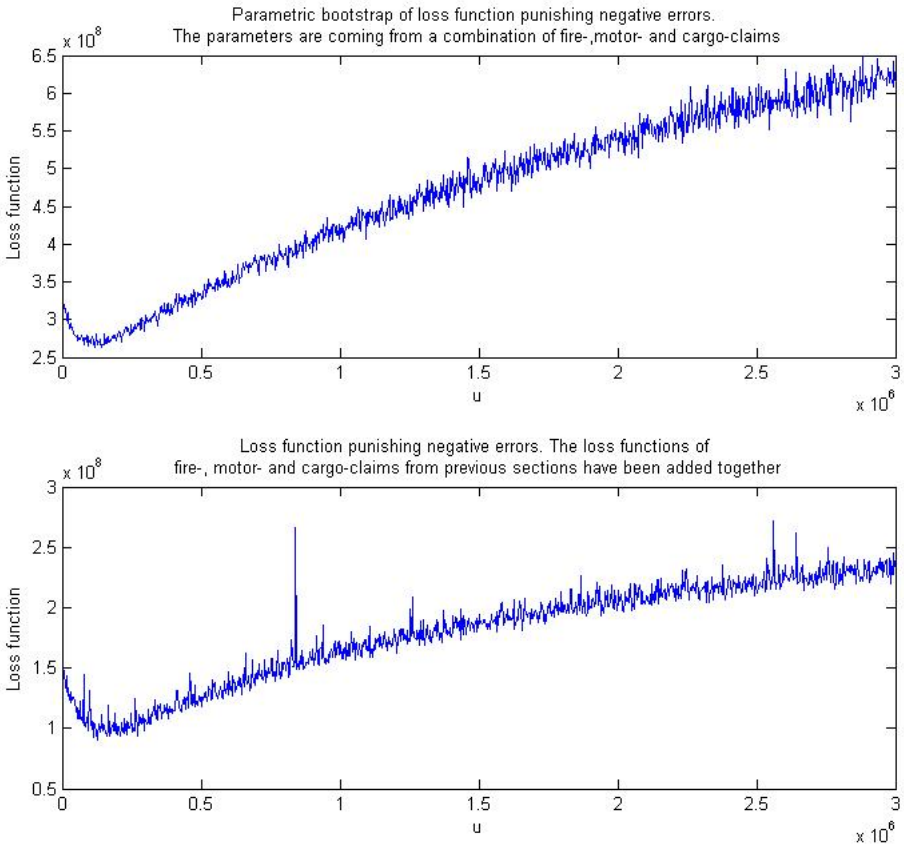

Figure 5.38: A comparison between the simulation of mixed claims compared to fire, motor and cargo claims modeled individually and added together. Loss function punishing only negative errors

Chapter 6

Conclusions

The aim of the thesis was to find a theory around the optimal claims limit and also to find a method of actually calculating this limit. The aim was also to make an analysis on categorization of claim types.

6.1 Theory

Assumptions regarding data and the human factor of the claim handlers have been made. What defines an optimal limit has been discussed. This can of course be done in many ways. Different definitions of loss functions have been presented and discussed.

In order to simulate the models, several problems had to be dealt with.

Modeling the actions of the claim handlers is risky. Here one is modeling human behavior. The model will be dependent on several aspects, for example the quality and speed of claims handling. There are probably many more aspects, but the model presented is perhaps good enough to predict some kind of trend.

When simulating, two different sampling methods have been described. The first type, non-parametric bootstrapping, is an easy and fast way to simulate complex distributions. Here we start with the actual historical vector and sample directly from it with replacement. The major benefit of using this kind of model is that it is easy and fast to implement. However no "new" values will be introduced and the largest value of historical data will be the largest possible value in any sampled distribution. The second type, parametric bootstrapping, is a bit harder to implement. Here we look at the data and from it, we assume a distribution. Maximum likelihood and qq-plotting are tools that can be used to calculate the parameters of the assumed distribution. Choosing a distribution for the claims data is

sometimes hard. Every type of claim will be somewhat different. To get a working model, some simplifications must always be done. In the thesis a few different typical distributions have been described.

6.2 Simulation

We have seen that different claim types can behave quite differently. For example we have seen that for motor claims, the claim handlers tend to overestimate the actual reserve. There must be a logical explanation to this. Perhaps expensive motor cases are very hard to estimate and the claim handlers might reserve a "safe" amount, just in case.

The behavior was also different when looking at the distribution of the payments. Fire claims had a very large tail, while cargo claims tend to have a very small one.

So how should one chose the optimal large claims limit?

Having individual claims limit calculations could reduce the error as we have seen in our comparison. As long as the error is not to large and the result is stable one can merge claim types into mixed models. Splitting up the types too much might however be bad. There is always a systematic error when making this kind of modeling and simulation. This systematic error will grow larger and dominate the total error as we do more and more claim type splits.

To reduce the workload of the claim handlers there is a point in having a high large claims limit. Here it is also a question regarding stability. If the error is sufficiently low and the result is somewhat stable, a higher limit than the "optimum" might be preferred. The cost of administrating claims by hand might outweigh the risk of loosing accuracy in estimations. There will always be a conflict regarding accuracy of the reserve versus cost of claim handling. In the end it really is a business decision.

Hopefully this thesis will be relevant regarding some basic theory concerning reserving and around the optimal large claims limit.

Bibliography

[1] Ananda M. M. A. Cooray, K. Modeling actuarial data with a composite lognormal-pareto model. *Scandinavian Actuarial Journal*, 5:321–334, 2005.

[2] Carlos A. R. Diniz Daiane Aparecida Zuanetti and Jose Galvao Leite. A lognormal model for insurance claims data. *REVSTAT - Statistical Journal*, 4:131–142, 2006.

[3] David C. M. Dickson. *Insurance risk and ruin.* Cambridge University Press, 2005.

[4] George S. Fishman. *Monte Carlo: concepts, algorithms, and applications.* Springer Verlag, 2000.

[5] A. Gut. *An intermediate course in Probability Theory.* Springer Verlag, 1995.

[6] Hult Lindskog. Kompendium till kursen riskvardering och riskhantering. print (2010), available at Matematikinstituionen KTH Stockholm Sweden.

[7] Paula A. Whitlock Malvin H. Kalos. *Monte Carlo methods.* Wiley-VCH, 2008.

[8] Thomas Mikosch Paul Embrechts, Claudia Kloppelberg. *Modelling extremal events for insurance and finance.* Springer Verlag, 2003.

[9] R. Sundberg. Kompendium i tillampad matematisk statistik. print (2010), available at Matematikinstituionen KTH Stockholm Sweden.

[10] Gregory Clive Taylor. *Loss reserving: an actuarial perspective.* Kluwer Academic Publishers, 2000.

[11] Chang Yen-Chang and Hung Wen-Liang. Linex loss functions with applications to determining the optimum process parameters. *Quality and Quantity*, 41:291–301, 2007.