UNIVERSITY OF TARTU

Faculty of Science and Technology

Institute of Mathematics and Statistics

Iuliia Kim

Modelling large claims in order to optimise the reinsurance program

Actuarial and Financial Engineering

Master's Thesis (30 ECTS)

Supervisors: Meelis Käärik, Ph.D.,

Hele-Liis Viirsalu, MSc

TARTU 2019

Modelling large claims in order to optimise the reinsurance program

Master's Thesis

Iuliia Kim

Abstract. Reinsurance is one of the cornerstones in the actuarial mathematics. When the company deals with large claims, the matter of choosing the optimal reinsurance strategy becomes especially important as it helps to reduce the risks of the insurance company. For defining the best reinsurance program, the insurer needs to know the behaviour of large claims in portfolio. Based on historical data of incurred claims, it is possible to estimate the ultimate values with chain ladder method. The most common claim distributions are used to fit to the ultimate amounts and the best distribution is chosen for generating claim severities. Based on generated number of claims and claim sizes, it's possible to simulate many scenarios for calculating the total loss without reinsurance and net losses after reinsurance. In case of excess of loss reinsurance, the reinsurer covers the part of claim which exceeds the retention level. Comparing net losses from all reinsurance programs, the insurance company can make a decision whether the reinsurance program is beneficial depending on the premium which the reinsurance company requests.

CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics

Keywords: actuarial mathematics, data analysis, non-life insurance, probability distributions, reinsurance, simulation

Suurte nõuete modelleerimine edasikindlustusprogrammi optimeerimiseks

Magistritöö

Iuliia Kim

Lühikokkuvõte. Edasikindlustus on kindlustusmatemaatika üks nurgakive. Kui ettevõte tegeleb suurte kahjunõuetega, muutub optimaalse edasikindlustusstrateegia valik eriti oluliseks, kuna see aitab vähendada kindlustusseltsi riske. Sobivaima edasikindlustusprogrammi valimiseks peab kindlustusandja kahjunõuete portfelli võimalikult hästi hinnata oskama. Kuna suurte nõuete käsitlemine võib kesta väga pikalt, tuleb kõigepealt hinnata nõuete arengut ajas, töös on selleks kasutatud ahel-redel meetodit. Ahel-redel meetodi abil saadud lõppsummadele rakendatakse levinumaid kahjujaotusi ja valitakse neist välja sobivaim. Samuti hinnatakse ajalooliste andmete põhjal kahjude sagedust. Edasikindlustusprogrammide võrdemiseks genereeritakse kahjude sagedused ja üksikute kahjude suurused ning võrreldakse kindustusandja kogukahju erinevate edasikindlustusprogrammide korral. Kui saadud tulemustele lisada edasikindlustuse pakkumised, saab kindlustusandja otsustada, milline edasikindlustusprogramm valida.

CERCS teaduseriala: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: andmeanalüüs, edasikindlustus, kahjukindlustus, kindlustusmatemaatika, simulatsioon, tõenäosusjaotused.

Contents

List of Figures and Tables

Figures

Tables

Introduction

Nowadays the insurance industry is of great importance as it is a form of economic remediation. Global insurance companies, especially those which are systemically important provide the stability of the whole financial system, mainly because they are large investors in financial markets. Insurance companies are safeguarding the stability of households and firms by insuring their risks.

As any financial institution, insurance company is concerned about its profitability and solvency. It analyses the inflows and outflows and which risks it takes when signs variety of policies. Obviously, each insurer has the variety of expenses to be covered and its task is to predict the future losses and to remain stable.

The analysis of claims is one of the cornerstones in the insurance activity, so the insurance company employ large numbers of analysts, including actuaries, to understand the claims behaviour.

Actuaries are not interested in the occurrence of the claims themselves but rather in the consequences of its random outcome. That means, they are concerned with the amount the company will have to pay rather than with the particular circumstances, which give rise to claim numbers. Actuaries should have an understanding of various models for the risk consisting of the aggregate amount of claims payable over a fixed period of time.

The other issue is that in case of large claims, the matter of reinsurance becomes quite significant. Strategic reinsurance programs are customized to provide more efficient risk protection, and to help insurers optimize their capital structures in order to improve capital returns and minimize capital costs. Increasingly, reinsurance is integrated into insurers' long-term strategy and growth plans [13].

The effect of reinsurance is the reduction in the mean and the variability of the amount paid out by the direct insurer on claims.

In particular reinsurance protects the direct insurer against having sole responsibility (or any responsibility) for the tails of the distributions of large claims [5].

Only by sharing some of their risk with reinsurers it is possible for direct insurers to offer cover against the key risks we face today and to keep prices at affordable levels.

There are many different forms and types of reinsurance contracts: they either cover entire insurance portfolios or just relate to single risks; they may involve a sharing of all premiums and losses or they may just cover losses exceeding a certain threshold. Whatever the differences between the various contracts, they all have the same ultimate goal: reinsurance contracts help provide capital relief, they smooth the volatility in an insurance company's earnings and protect their balance sheet [12].

Insurer intends to choose the best reinsurance strategy, which gives sufficient level of risk protection and requires the least amount of premium paid to reinsurer, so that the resulting outflows of the company would be minimal.

The problem of optimization of reinsurance program is more complex because reinsurers set their prices based on estimates and models. The more accurate estimates are, the better overview of the possible outcomes after reinsurance is presented. Therefore, the company is able to assess whether the reinsurance program is profitable or not, based on designed models. That is why the analysis of claims behaviour is very important.

The aim of the present thesis is to is to model large claims to estimate their impact to the direct insurer's losses before and after reinsurance. Our research will be multi-stage.

First chapter starts with the description of chain-ladder method, which is commonly used for estimating the claim reserves, its underlying idea, assumptions and practical implementation.

Chapter 2 is dedicated to principles of the collective risk model and the most common probability distributions used for analysis of claim severity. Also, the concept of limited loss is presented and formulas of limited moments are derived.

Chapter 3 provides an introduction to proportional and non-proportional types of reinsurance, their main features and describes in more detail how excess of loss (XL) reinsurance affects to claim distribution.

In the last chapter the practical research based on company's data will be described step-by-step. The theoretical background considered in previous chapters is implemented in a real numerical problem: we apply chain-ladder technique to get the ultimate estimates of the claim severity and claim frequency, then the ultimate values of claim severity are used to be fitted by different theoretical distributions and we choose that one which fits the best. Then, based on the obtained results, we simulate different scenarios for total loss and net losses after reinsurance depending on different retention levels and compare the characteristics based on simulated samples with theoretical values. Our final purpose is to evaluate several working XL reinsurance programs, define value at risk for different given probabilities and based on assumed amount of premium, to conclude whether reinsurance program is beneficial or not.

1. Description of chain-ladder method

The ideas presented in this chapter are based on [2] and [8].

Chain-ladder method is the most widely used actuarial technique for estimating loss reserves. There are many variations, but they have the same objective: to extract from the loss development triangle a pattern for the claims run-off that can be used to extrapolate the less mature years of account.

Claims data are given as a *run-off triangle* as follows:

	Development period j				
Year of origin i	1	2	3	...	n
1	C_{11}	C_{12}	C_{13}	...	C_{1n}
2	C_{21}	C_{22}	...		
3	C_{31}	...			
...	...				
n	C_{n1}				

Here C_{ij} denotes the claim amount (or claim number) occurred within a year of origin i and payable in development year j. Development period 1 means "current" or "running" period. The number of usable years of origin is defined by the history the insurance company has.

Chain ladder method is quite simple and is based on the assumption that the development factors extracted from the run-off triangle remain the same (or at least are similar) in the future.

The key point is to start with the incremental data, and then to get the cumulative numbers in second step, as the development patterns are extracted from the cumulative triangle.

Let's denote the cumulative claim amount (or claim number) the following way

$$CC_{ij} = \sum_{k=1}^{j} C_{ik}.$$

The chain ladder model assumes that the development between successive periods of development is the following:

$$CC_{i,j+1} \approx f_j CC_{ij}, \quad i = 1,2, \dots n, \quad j = 1,2, \dots n-1,$$

where

- CC_{ij} is the cumulative claim amount or claim number corresponding to year of origin i and year of development j,
- $CC_{i,j+1}$ is the cumulative claim amount or claim number corresponding to year of origin i and year of development j+1,
- f_j is the development factor between development periods j and j+1.

Equivalently, chain ladder model can be written as

$$CC_{ij} \approx S_i R_j, \quad i = 1,2, \dots n, \quad j = 1,2, \dots n,$$

where

- CC_{ij} is the cumulative claim amount (or claim number) corresponding to year of origin i and year of development j,
- S_i is the ultimate claim amount (or claim number) corresponding to year of origin i and is equal to CC_{in},
- R_j is the proportion factor of the ultimate value that arose by the end of development period j.

For estimating the development (age-to-age) factors we use all the available data between successive development periods:

$$\hat{f}_j = \frac{\sum_{i=1}^{n-j-1} CC_{i,j+1}}{\sum_{i=1}^{n-j-1} CC_{ij}}, \quad j = 0, \dots, n-1.$$

The cumulative development factors describe the development pattern between given development period j and the ultimate period n. Thus, the cumulative development factors are given by

$$\hat{F}_j = \prod_{k=j}^{n-1} \hat{f}_k, \quad j = 0, \dots, n-1.$$

Also, the inverses of cumulative factors represent the proportion of total claims emerged by the end of given development period. Thus, the estimates for inverses are given by

$$\widehat{R}_j = \frac{1}{\hat{F}_j}, \quad j = 0, \dots, n-1.$$

As the basic chain ladder does not take into account the inflation effect, then in case of significant influence of particular calendar years, the adjustment of initial data is required.

If the inflation-adjustment is essential, it is necessary to consider incremental figures while correcting with inflation rates. Then, in second step the cumulative data will include the inflated-adjusted increments.

2. Collective risk model. Limited loss.

Losses caused by occurrences of unexpected events are problems both for individuals and for society as a whole. The company is interested both in understanding the behaviour of the claim amounts and number of claims occurred within a fixed period of time.

First of all, it is necessary to specify key terms which are used in this chapter.

2.1. Basic terminology

Exposure is a unit of measure, which represents the extent of risk, or, rephrasing, it measures the volume of potential claims [2].

The exposure may represent different things depending the problem at hand. For example,

- in health insurance, the rate may be the occurrence of a specific disease per 1 000 people and the exposure is the number of people considered in the unit of 1 000.

- in auto insurance, the rate may be the number of accidents per year of a driver and the exposure is the length of the observed period for the driver in the unit of year.
- in credit risk modelling, the rate may be the number of default events per 1 000 firms and the exposure is the number of firms under consideration in the unit of 1 000 [1].

Claim frequency is the number of claims per unit of exposure.

Claim severity is the cost per claim. In practice actuaries often use the average severity:

$$Average\ claim\ severity = \frac{Total\ loss}{Number\ of\ claims}.$$

Number of claims is usually modeled using non-negative discrete probability distributions since the number of claims is discrete and not equal to zero. The most common choices for describing the behaviour of claim count are Poisson or negative binomial distribution.

Claim severity is known to be best modeled using non-zero continuous distributions, which are skewed to the right and have heavy tails [1].

Next, we will define a collective risk model and its main characteristics.

2.2. Aggregate claims and collective risk model

An insurance activity exists due to its ability to pool risks. By insuring many people, the individual risks are combined into an aggregate risk that is manageable and can be priced at a level that will attract customers [9]. When the loss occurs, aggregate claims are obtained by summing over all the policies in the portfolio.

For the *collective risk model,* we assume a random process that generates claims for a portfolio of policies. The process is characterised in terms of the portfolio as a whole rather than in terms of the individual policies comprising the portfolio [3].

The purposes of collective risk model are:

- to describe the distribution of aggregate claim amount with some known distribution;
- to take into account only those policies which actually caused claims (for reducing the volume of work) [8].

Thus, we define S as a random sum

$$S = \sum_{i=1}^{N} X_i,$$

where N is a random variable denoting the number of claims.

We also assume that:

- the claim severities $X_1, X_2, \ldots$ do not depend on the number of claims N.
- for any given n the claim severities $X_1, \ldots, X_n$ are i.i.d. random variables.

Following [5] and using the conditional expectation formula, we can define the expected value and the second moment of S as follows

$$ES = E\left(\sum_{i=1}^{N} X_i\right) = E\left(E\left(\sum_{i=1}^{N} X_i \,|N\right)\right) = E(N \cdot EX) = EN \cdot EX,$$

$$ES^2 = E\left(\left(\sum_{i=1}^{N} X_i\right)^2\right) = E\left(E\left(\left(\sum_{i=1}^{N} X_i\right)^2 \middle| N\right)\right) = E\left(E\left(\sum_{i=1}^{N} X_i^2 + \sum_{i\neq j} X_i X_j\right)\right)$$

$$= E(N \cdot EX^2 + N(N-1)(EX)^2) = EN(EX^2 - (EX)^2) + EN^2 \cdot (EX)^2$$

$$= EN \cdot DX + EN^2 \cdot (EX)^2.$$

Based on above, we also can derive the formula of the variance of S as

$$DS = ES^2 - (ES)^2 = EN \cdot DX + EN^2 \cdot (EX)^2 - (EN)^2(EX)^2$$

$$= EN \cdot DX + DN \cdot (EX)^2.$$

2.2. Distributions of claim severity and their properties

The definitions and properties given in this section are mostly taken from [1] and [7] unless otherwise stated.

2.2.1. *Gamma distribution*

The gamma distribution is commonly used in modeling claim severity.

Definition

A random variable X follows a gamma distribution if its density function is given by

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}\ , \qquad x > 0, \lambda > 0, \alpha > 0,$$

and cumulative distribution function is given by

$$F_X(x) = \ \gamma(\alpha, x\lambda),$$

where

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx = (\alpha-1)\Gamma(\alpha-1) \text{ is gamma function,}$$

$$\gamma(\alpha, t) = \frac{\int_0^t x^{\alpha-1} e^{-x} dx}{\Gamma(\alpha)} \text{ is incomplete gamma function,}$$

α is a shape parameter and λ is a rate parameter.

The gamma distribution can also be considered as generalization of exponential distribution. If parameter α is equal to 1, then we get exactly the density function of exponential distribution. If α is integer and greater than 1, we can interpret gamma distribution as sum of α independent exponentially distributed random variables [8].

Moments

The moments of the gamma distribution are defined as

$$\mathrm{E}(X^n) = \frac{(\alpha + n - 1)(\alpha + n - 2) \dots \alpha}{\lambda^n}.$$

Thus, the mean of the gamma distribution is defined as

$$\mathrm{E}(X) = \frac{\alpha}{\lambda},$$

and the variance is defined as

$$D(X) = \frac{(\alpha + 1)\alpha}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}.$$

Estimation of parameters

The parameters α and λ can be estimated using the maximum likelihood estimation method.

Let's assume we have an empirical sample of n observations $X_1, X_2, \dots, X_n$ which are independent and identically gamma distributed.

Using the likelihood function, we get:

$$L(x; \lambda; \alpha) = \frac{\lambda^{n\alpha}}{(\Gamma(\alpha))^n} \prod_{i=1}^{n} (x_i^{\alpha-1})\, e^{-\lambda \sum_{i=1}^{n} x_i}.$$

In practice it is more convenient to use the logarithm of the likelihood function:

$$l(x; \lambda, \alpha) = n\alpha \ln\lambda - n\ln\big(\Gamma(\alpha)\big) + (\alpha - 1) \sum_{i=1}^{n} \ln x_i - \lambda \sum_{i=1}^{n} x_i.$$

To find $\hat{\lambda}$, we compute partial derivative of the log-likelihood function with respect to λ and make it equal to zero:

$$\frac{\partial l}{\partial \lambda} = \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} x_i = 0,$$

from which we get the following estimate:

$$\hat{\lambda} = \frac{n\alpha}{\sum_{i=1}^{n} x_i} = \frac{\hat{\alpha}}{\bar{x}}.$$

To find $\hat{\alpha}$, we compute partial derivatives of the log-likelihood function with respect to α and make it equal to zero:

$$\frac{\partial l}{\partial \alpha} = n\alpha \ln\lambda - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n} \ln x_i = n\alpha \ln\frac{\alpha}{\bar{x}} - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n} \ln x_i = 0.$$

Since the last equation does not have closed form solution, it can be solved using Newton-Raphson method.

2.2.2. *Lognormal distribution*

The lognormal random variable has a lower bound of zero, is positively skewed and has a long right tail. A lognormal distribution is commonly used to describe distributions of financial assets such as stock prices. It is also used in fitting claim amounts for automobile as well as health insurance.

Definition

A positive random variable X is lognormally distributed if the logarithm of X is normally distributed. Hence X follows a lognormal (μ, σ^2) distribution if its density function is given by

$$f_X(x;\mu,\sigma^2) = \frac{(2\pi\sigma^2)^{-\frac{1}{2}}}{x} \exp\left\{-\frac{1}{2\sigma^2}(\ln(x)-\mu)^2\right\}$$

for $x > 0$, $-\infty < \mu < \infty$ and $\sigma > 0$,

where μ is a location parameter and σ is a shape parameter.

Products and quotients of lognormally distributed variables are themselves lognormally distributed, as well as X^a and aX, for $a \neq 0$ and X following a lognormal (μ, σ^2) distribution. However, the distribution of the sum of independent lognormally distributed variables, that appears in many practical problems is not lognormally distributed and does not present a recognizable probability density function [14].

Moments

The moments of the lognormal distribution can be calculated from the moment generating function of the normal distribution and are defined as

$$\mathrm{E}(X^n) = \exp\left(n\mu + \frac{1}{2}n^2\sigma^2\right).$$

Thus, the mean of the lognormal distribution is defined as

$$\mathrm{E}(X) = \exp\left(\mu + \frac{1}{2}\sigma^2\right),$$

and the variance is defined as

$$\mathrm{D}(X) = \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2).$$

Estimation of parameters

The parameters μ and σ can be estimated using the maximum likelihood estimation method.

Using the likelihood function, we get the following expression:

$$L(x;\mu,\sigma) = \frac{1}{\sigma^n(\sqrt{2\pi})^n \prod_{i=1}^n x_i} \cdot \prod_{i=1}^n e^{-\frac{(\ln x_i - \mu)}{2\sigma^2}} .$$

The log-likelihood function is defined as:

$$l(x;\mu,\sigma) = \ln L = -\left(n\ln\sigma + \frac{n}{2}\ln(2\pi) + \sum_{i=1}^n \ln x_i\right) - \frac{\sum_{i=1}^n (\ln x_i - \mu)^2}{2\sigma^2} .$$

To find $\hat{\mu}$ and $\hat{\sigma}$, we compute partial derivatives and determine where they are equal to zero:

$$\frac{dl}{d\mu} = \frac{(2\sum_{i=1}^{n} \ln x_i - 2n\mu)}{2\sigma^2} = 0,$$

$$\frac{dl}{d\sigma} = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(\ln x_i - \mu)^2}{\sigma^3} = \frac{-n\sigma^2 + \sum_{i=1}^{n}(\ln x_i - \mu)^2}{\sigma^3} = 0.$$

Thus, we get:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \ln x_i}{n} \quad ; \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(\ln x_i - \mu)^2}{n} \quad .$$

2.2.3. *Weibull distribution*

The Weibull distribution is widely used in reliability, life data analysis, weather forecasts and general insurance claims. Truncated data arise frequently in insurance studies. The Weibull distribution is particularly useful in modeling left-truncated claim severity distributions. Weibull was used to model excess of loss treaty over automobile insurance as well as earthquake inter-arrival times.

Definition

A random variable X follows a Weibull distribution if its density function is given by

$$f_X(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left\{-\left(\frac{x}{\lambda}\right)^k\right\} , \qquad x \geq 0, k > 0, \lambda > 0,$$

and cumulative distribution function is given by

$$F_X(x) = 1 - \exp\left\{-\left(\frac{x}{\lambda}\right)^k\right\} ,$$

where k is a shape parameter and λ is a scale parameter.

Moments

The moments of Weibull distribution are defined as

$$\mathrm{E}(X^n) = \lambda^n \Gamma(1 + \frac{n}{k}),$$

where $\Gamma(\cdot)$ is the gamma function $\Gamma(k) = \int_0^\infty \mathrm{e}^{-t} t^{k-1} dt$.

Thus, the mean and variance of Weibull distribution are defined as

$$\mathrm{E}(X) = \lambda\Gamma(1 + \frac{1}{k}),$$

$$\mathrm{D}(X) = \lambda^2\Gamma\left(1 + \frac{2}{k}\right) - \left[\lambda\Gamma\left(1 + \frac{1}{k}\right)\right]^2 .$$

Estimation of parameters

The parameters k and λ can be estimated using the maximum likelihood estimation method.

Using the likelihood function, we get the following expression:

$$L(x;k,\lambda)=\prod_{i=1}^{n}\frac{k}{\lambda}\left(\frac{x_i}{\lambda}\right)^{k-1}\exp\left\{-\left(\frac{x_i}{\lambda}\right)^{k}\right\}=\frac{k^n}{\lambda^{nk}}\prod_{i=1}^{n}x_i^{k-1}\exp\left\{-\sum_{i=1}^{n}\left(\frac{x_i}{\lambda}\right)^{k}\right\}.$$

The log-likelihood function is defined as

$$l(x;k,\lambda)=\ln L=n\ln(k)-kn\ln(\lambda)+(k-1)\sum_{i=1}^{n}\ln(x_i)-\sum_{i=1}^{n}\left(\frac{x_i}{\lambda}\right)^{k}.$$

The partial derivative with respect to the parameter $\hat{\lambda}$ is given by

$$\frac{\partial l}{\partial\lambda}=\;-\;\frac{kn}{\lambda}+\sum_{i=1}^{n}\frac{kx_i^k}{\lambda^{k+1}}=\;\frac{kn}{\lambda^{k+1}}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^k-\lambda^k\right).$$

Since $\frac{\partial l}{\partial\lambda}$ has to be equal to zero, then the estimation of the parameter λ is

$$\lambda=\left(\frac{1}{n}\sum_{i=1}^{n}x_i^k\right)^{\frac{1}{k}}.$$

Substituting λ into l, we get:

$$\begin{aligned}
l&=n\ln(k)-kn\ln\left(\frac{1}{n}\sum_{i=1}^{n}x_i^k\right)^{\frac{1}{k}}+(k-1)\sum_{i=1}^{n}\ln(x_i)-\;\sum_{i=1}^{n}\left(\frac{x_i}{\left(\frac{1}{n}\sum_{i=1}^{n}x_i^k\right)^{\frac{1}{k}}}\right)^{k}\\
&=n\ln k-n\ln\sum_{i=1}^{n}x_i^k+n\ln(n)+(k-1)\sum_{i=1}^{n}\ln(x_i)-n\frac{\sum_{i=1}^{n}x_i^k}{\sum_{i=1}^{n}x_i^k}\\
&=n\ln k-n\ln\sum_{i=1}^{n}x_i^k+n\ln(n)+(k-1)\sum_{i=1}^{n}\ln(x_i)-n.
\end{aligned}$$

The partial derivative from l after substitution with the respect to parameter k is given by

$$\frac{\partial l}{\partial k}=\frac{n}{k}-n\frac{\sum_{i=1}^{k}x_i^k\ln x_i}{\sum_{i=1}^{n}x_i^k}+\;\sum_{i=1}^{n}\ln x_i=n\left(\frac{1}{k}-\frac{\sum_{i=1}^{k}x_i^k\ln x_i}{\sum_{i=1}^{n}x_i^k}+\frac{1}{n}\sum_{i=1}^{n}\ln x_i\right).$$

Since $\frac{\partial l}{\partial k}$ has to be equal to zero, then the estimation of the parameter k can be found from the following equation

$$\frac{1}{k}-\frac{\sum_{i=1}^{k}x_i^k\ln x_i}{\sum_{i=1}^{n}x_i^k}+\frac{1}{n}\sum_{i=1}^{n}\ln x_i=0.$$

Since the last equation doesn't have a closed form solution it can be solved using Newton-Raphson method or any other numerical solution procedure to get the estimate of k.

2.2.4 *Pareto distribution*

The Pareto distribution has many economic and financial applications. It is a positively skewed and heavy-tailed distribution which makes it suitable for modelling income, high-risk insurance claims and severity of large casualty losses. For extreme insurance claims, the tail of the severity distribution (losses in excess of threshold) can be modelled using a Pareto distribution.

Definition

A random variable X follows a Pareto distribution if its density function is given by [8]

$$f_X(x) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}} \ , \qquad x \geq \beta$$

and cumulative distribution function is given by

$$1 - \left(\frac{\beta}{x}\right)^\alpha , \qquad x \geq \ \beta \ ,$$

where α is a shape parameter, β is a scale parameter.

Moments

The moments of Pareto distribution are defined as [8]

$$\mathrm{E}(X^n) = \begin{cases} \dfrac{\alpha\beta^n}{\alpha - n} \ , & \alpha > n, \\ \infty \ , & \alpha \leq n. \end{cases}$$

Thus, the mean of Pareto distribution is defined as

$$\mathrm{E}(X) = \frac{\alpha\beta}{\alpha - 1} \ , \quad \alpha > 1$$

and the variance is defined as

$$\mathrm{D}(X) = \ \frac{\alpha\beta^2}{\alpha - 2} - \frac{\alpha^2\beta^2}{(\alpha-1)^2} = \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)} \ , \quad \alpha > 2.$$

Estimation of parameters

The parameters α and β can be estimated using the maximum likelihood estimation method.

Using the likelihood function, we get the following expression

$$L(x;\alpha,\beta) = \frac{\alpha^n\beta^{n\alpha}}{\prod_{i=1}^n (x_i)^{\alpha+1}} \ .$$

The log-likelihood function is defined as

$$l(x;\alpha,\beta) = n\ln\alpha + n\alpha\ln\beta - (\alpha+1)\sum_{i=1}^n \ln x_i \ .$$

Since a higher β will always result in a higher likelihood, as $\ln\beta$ is monotonically increasing, we maximize the likelihood by setting $\hat{\beta}$ as high as possible. Since $\beta \le x_i$ for all i, we maximize the likelihood by setting $\hat{\beta} = \min_i x_i$, the smallest x_i in the sample.

For α, we set a partial derivative of l with respect to α equal to 0:

$$\frac{\partial l}{\partial \alpha} = \frac{n}{\alpha} + n\ln\beta - \sum_{i=1}^{n} \ln x_i = 0.$$

Therefore,

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \ln x_i - n\ln\hat{\beta}}.$$

2.3. Characteristics of limited loss

In many practical cases actuaries are interested not only in the claim size X, but in the claim size limited by some threshold level R.

In that case it is important to know such basic characteristics as mean and variance of the random variable $\tilde{X} = \min(X; R)$.

2.3.1. *Limited expected value function and limited k-th moment*

Definition

Let random variable X have a distribution function $F(x)$ and a density function $f(x)$ whose support is $0 < x < \infty$. Then the **limited expected value function** $E[X;x]$ of X is defined as [7]

$$E[X;x] = E(\min(X;x)) = \int_0^x y f(y)dy + x[1 - F(x)].$$

Similarly, the **limited *k*-th moment** $E[X;x]^k$ of X is defined as

$$E[X;x]^k = E(\min(X;x))^k = \int_0^x y^k f(y)dy + x^k[1 - F(x)].$$

2.3.2. *Derivation of limited k-th moment of lognormal distribution*

As an example, we will further derive the limited k-th moment of lognormal distribution.

Lemma 2.1 (Limited k-th moment of lognormal distribution)

If the positive random variable X has lognormal distribution with parameters μ and σ and its cumulative distribution function F(x) can be presented via the normal distribution function as $F(x) = \Phi(\frac{\ln x - \mu}{\sigma})$, then the limited k-th moment of X is defined as

$$E[X;x]^k = e^{k\mu + \frac{k^2\sigma^2}{2}} \Phi\left(\frac{\ln x - \mu - k\sigma^2}{\sigma}\right) + x^k \left(1 - \Phi\left(\frac{\ln x - \mu}{\sigma}\right)\right).$$

Proof

Based on the definition of limited k-th moment of random variable X, for lognormal distribution we will get the following equation:

$$E[X;x]^k = \int_0^x y^k \frac{1}{\sigma y\sqrt{2\pi}} e^{-\frac{(\ln y-\mu)^2}{2\sigma^2}} dy + x^k \left(1 - \Phi\left(\frac{\ln x - \mu}{\sigma}\right)\right).$$

Calculating the integral $I = \int_0^x y^{k-1} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\ln y-\mu)^2}{2\sigma^2}} dy$ by substitution z=ln$y$, we get that

$$I = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\ln x} e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2 + kz} dz,$$

where rewriting the power of e gives us the following expression

$$\begin{aligned}
-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2 + \frac{2kz\sigma^2}{2\sigma^2} &= \\
&= -\frac{(z-\mu)^2 - 2kz\sigma^2 + 2k\mu\sigma^2 + k^2\sigma^4 - 2k\mu\sigma^2 - k^2\sigma^4}{2\sigma^2} \\
&= -\frac{(z-\mu-k\sigma^2)^2}{2\sigma^2} + k\mu + \frac{k^2\sigma^2}{2},
\end{aligned}$$

and therefore

$$\begin{aligned}
I &= e^{k\mu + \frac{k^2\sigma^2}{2}} \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\ln x} e^{-\frac{(z-\mu-k\sigma^2)^2}{2\sigma^2}} dz \\
&= e^{k\mu + \frac{k^2\sigma^2}{2}} \Phi\left(\frac{\ln x - \mu - k\sigma^2}{\sigma}\right),
\end{aligned}$$

from which we can finally get the equation for limited k-th moment:

$$E[X;x]^k = e^{k\mu + \frac{k^2\sigma^2}{2}} \Phi\left(\frac{\ln x - \mu - k\sigma^2}{\sigma}\right) + x^k \left(1 - \Phi\left(\frac{\ln x - \mu}{\sigma}\right)\right). \quad \blacksquare$$

Following the derived formula of limited k-th moment of lognormal distribution, we can define the limited expected value as

$$E[X;x] = e^{\mu + \frac{\sigma^2}{2}} \Phi\left(\frac{\ln x - \mu - \sigma^2}{\sigma}\right) + x \left(1 - \Phi\left(\frac{\ln x - \mu}{\sigma}\right)\right),$$

and the limited second moment as

$$E[X;x]^2 = e^{2\mu + 2\sigma^2} \Phi\left(\frac{\ln x - \mu - 2\sigma^2}{\sigma}\right) + x^2 \left(1 - \Phi\left(\frac{\ln x - \mu}{\sigma}\right)\right).$$

Based on known values of limited expected value function and limited second moment, the variance of limited loss can be easily defined as

$$D[X;x] = E[X;x]^2 - (E[X;x])^2.$$

The problem of finding the mean and variance of limited loss becomes especially important when the insurance company chooses the optimal reinsurance program where the claim part retained by insurer depends on the specified level. This issue will be covered in more detail in Chapter 3.

3. Types of reinsurance. Effect of XL reinsurance on claim distribution

Buying insurance protects the policyholder against the effects of "large" losses. Similarly, the inclusion of a reinsurance arrangement often protects the direct insurer against the effects of "very large" claims. In particular, it protects the direct insurer against having sole responsibility (or any responsibility) for the tails of the distributions of large claims.

In other words, reinsurance "stabilizes" the direct insurer's payouts on claims [5].

By the compensation form, reinsurance is classified as:

- Proportional reinsurance;
- Non-proportional reinsurance.

3.1. Proportional reinsurance

Proportional reinsurance means that the direct insurer and reinsurer share insured sums in a clearly defined proportion as described within a contract. Premiums and claims are also split up according to the respective share of the risk (i.e. proportionally).

There are two distinct types of proportional reinsurance: *quota share* and *surplus* reinsurance.

Quota share reinsurance is a form of proportional reinsurance whereby the ceding company is indemnified for a fixed percent of loss on each risk covered by the treaty contract. All liability and premiums are shared from the first dollar. "Quota" or "definite" share relates to the fixed percentage as stated in the treaty [11].

Then, all claims in given portfolio are divided using proportion α as follows:

- $\tilde{X} = \alpha X -$ the part of individual claim retained by direct insurer;
- $\hat{X} = (1-\alpha)X -$ the part of individual claim ceded to reinsurer;
- $\tilde{S} = \alpha S -$ the part of aggregate claims retained by direct insurer;
- $\hat{S} = (1-\alpha)S -$ the part of aggregate claims ceded to reinsurer.

Similar holds for premiums

- $\tilde{P} = \alpha P -$ the part of premium retained by direct insurer;
- $\hat{P} = (1-\alpha)P -$ the part of premium ceded to reinsurer.

Under a *surplus* reinsurance, the proportion ceded depends on the size and type of risk. The ceding company has the right to decide how much it wants to retain on any one risk.

The parameters that define the amount that might be ceded are:

- R – the maximum amount that the direct insurer may retain;
- L – the maximum number of lines that may be ceded under the treaty.

One line is equal to the amount retained; the maximum amount that can be ceded is LR. It follows that the maximum exposure that the insurer can underwrite to be covered under the terms of the treaty is the retention R plus the amount ceded LR, i.e. $(1+L)R$ [2].

3.2. Non-proportional reinsurance

The traditional forms of nonproportional reinsurance cover are known as *excess of loss* and *stop loss*. Both provide cover once claims exceed a certain level and they usually have a limited insured amount.

In case of excess of loss (XL) reinsurance, the reinsurer has to cover the part of claims, which exceeds a certain excess level or retention.

XL reinsurance can be classified [8]:

- Working XL – retention applied to each individual claim;
- Aggregate XL – retention applied to aggregate claims;
- Cat XL – retention applied to losses occurred in a fixed period after a catastrophe.

Claims are divided by direct insurer and reinsurer as follows:

$$\tilde{X} = \min(X, R) \quad \text{and} \quad \hat{X} = \max(0, X - R),$$

where R is the retention of direct insurer.

Stop loss reinsurance is usually the final part of a reinsurance program to operate. It may be applied to aggregate claims arising from a specific class of business or on a whole-account basis. The upper and lower limits are expressed as proportions of the total net earned premium during the period, i.e. in terms of the *loss ratio* [2]:

$$\frac{\text{Net incurred claims}}{\text{Net earned premiums}}.$$

In case loss ratio exceeds a fixed proportion, exceeding part is covered by reinsurer.

3.3. Effect of XL reinsurance on claim distribution

The following subsection is based on [8].

In case of XL reinsurance, the distribution of the claim part retained by direct insurer: $\tilde{X} = \min(X, R)$ is given by

$$F_{\tilde{X}}(x) = P\left(\tilde{X} < x\right) = \begin{cases} F_X(x), & x < R, \\ 1, & x \geq R. \end{cases}$$

Then, the expected values of individual and aggregate claim amounts for direct insurer are defined as

$$E\tilde{X} = E(\min(X, R)) = E[X; R],$$

$$E\tilde{S} = E\tilde{N} \cdot E\tilde{X} = EN \cdot E[X; R],$$

and the variances are defined as

$$D\tilde{X} = D(\min(X, R)) = E[X; R]^2 - (E[X; R])^2,$$

$$D\tilde{S} = EN \cdot D\tilde{X} + \left(E\tilde{X}\right)^2 \cdot DN.$$

4. Modelling large claims and estimation of net losses after reinsurance

4.1. Description of the data and assumptions required for preliminary work

We work with a dataset of large claims that was provided by one of the leading Estonian insurance companies. The dataset contains information about 193 cases occurred within period 2007-2018 (all the figures are prepared on quarterly basis): accident period, reporting period, period of claim development and the amount of claim paid and reserve. The data are classified by two lines of business from Property & casualty insurance (but we don't have details what these lines of business are). In addition, we have information about quarterly earned exposure for each line of business and inflation rates within covered period.

We have to emphasise that we deal only with large claims.

The large claims are characterised by following conditions [10]:

- greater than certain amount of money (threshold);
- have adverse financial effect;
- the event that such large claim occurs is rare;
- reside at right tail of a distribution function

In our case we are modelling the claims that exceed the certain level. The threshold chosen by company was €150 000.

As the initial step of our analysis is estimating the ultimate values of claim severity and claim frequency with chain-ladder technique, we have to start with the data manipulation for the further numerical procedures.

The reason while we deal with claim severities and claim frequencies but not with claims incurred is that averaging claims should give more stable estimates, as in total triangle the periods with more claims will probably dominate the development pattern.

We first adjust the amounts of incurred claims taking into the account the quarterly inflation rates (each given rate is the average of rates of Lithuania, Latvia and Estonia for the corresponding period) so that we could treat them in money 2018.

We follow the assumption that one case is equal to one claim, independently on how many claims were paid for certain case.

The other question is whether lines of business (hereinafter – LoBs) should be considered separately or not. For that purpose we make a visual inspection of the claims development graph.

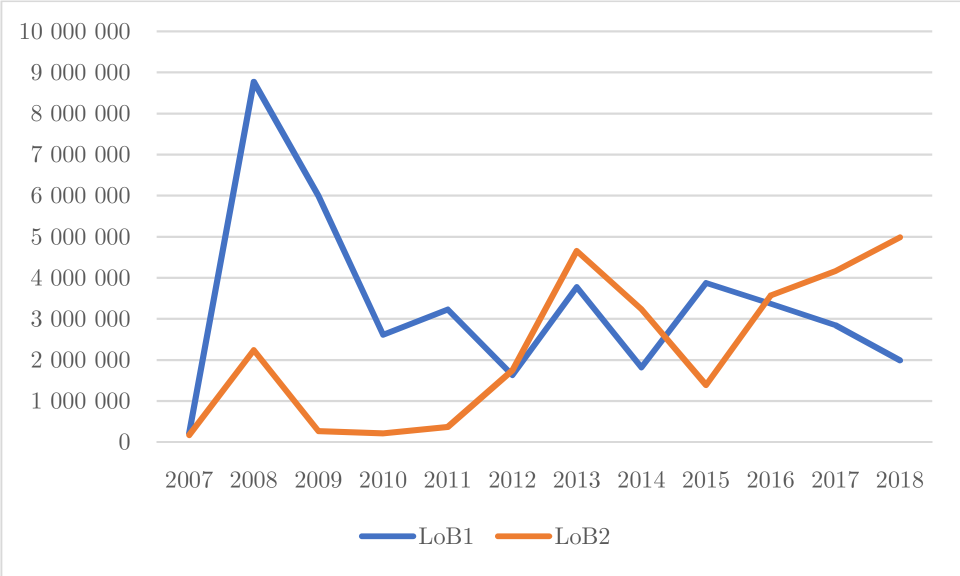

Figure 4.1: Development of aggregate claims

We may see from the graph that aggregate claims from both LoBs have different development patterns within time period that should be covered, so they are not homogeneous and should be modelled separately.

4.2 Application of chain ladder method

Starting with run-off triangles, we have to make assumptions as well.

Recalling the formulas from chapter 2:

- $\text{Claim frequency} = \dfrac{\text{Number of claims}}{\text{Exposure}},$
- $\text{Average claim severity} = \dfrac{\text{Claims paid} + \text{Reserve}}{\text{Number of claims}}.$

We need to choose the base period for each parameter to be modelled. For frequencies we base on the period when the claim *occurred* and for severities – the period when the claim was *reported*. So, the denominator equivalent to exposure (or number of claims) for corresponding period remains the same within a row of claim frequency (or severity) triangle.

The matter of choice is periodicity of chain ladder. Despite the fact that company's data are given on quarterly basis, we make a decision to choose annual basis as it should give smoother results.

We assume that there is no tail (i.e. claims reported in 2007 are not developed after 2018) in our triangle. Although in reality, there can be another situation but for our purposes and convenience, we decide to follow this assumption.

The run-off triangles of incremental frequencies for both LoBs are the following (values are scaled 10^6 times for convenience):

	Development period											
Accident period	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
2007	13,31	0	0	0	0	0	0	0	0	0	0	0
2008	102,05	0	0	0	0	0	0	0	0	0	0	
2009	62,18	0	0	0	0	0	0	0	0	0		
2010	45,30	0	0	0	0	0	0	0	0			
2011	43,11	0	0	0	0	0	0	0				
2012	45,14	0	0	0	0	0	0					
2013	46,19	0	0	0	0	0						
2014	47,84	0	0	0	0							
2015	73,26	0	6,66	0								
2016	63,15	5,74	0									
2017	70,78	0										
2018	62,71											

Table 4.1: Run-off triangle of incremental claim frequencies for LoB1

	Development period											
Accident period	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
2007	3,77	0	0	0	0	0	3,77	0	0	0	0	0
2008	20,15	3,36	0	0	0	3,36	0	0	0	0	0	
2009	3,49	0	0	0	6,98	0	0	0	0	0		
2010	3,37	0	0	6,75	0	0	0	0	0			
2011	6,30	0	0	0	0	0	0	0				
2012	20,08	2,87	0	0	0	0	0					
2013	14,69	2,94	0	2,94	0	0						
2014	22,92	8,59	2,86	0	0							
2015	2,28	2,28	0	0								
2016	19,49	2,17	0									
2017	20,10	4,02										
2018	17,36											

Table 4.2: Run-off triangle of incremental claim frequencies for LoB2

We may see that the claims related to LoB1 are more likely to be reported the same year the accident occurs, as the values of claim frequencies within each row mainly do not change. For the LoB2 some claims tend to be reported several years later than the accident occurs as the development of claim frequencies can be still observed after running period.

In the triangle diagonals represent calendar years and it is necessary to investigate specifically the calendar effects from the point of view of legal or other processes, in order to understand whether something unusual happened that year or not. This time the information was not available but it is worth keeping in mind that some specific factors may influence significantly on the claim development, especially if we deal with large claims, whose nature is irregular and not standard.

After applying the chain ladder method, we obtain the ultimate values based on development factors. For frequency the results (scaled 10^6 times) are as follows:

Accident year	**LoB1**	**LoB2**
2007	13,306	7,530
2008	102,055	26,865
2009	62,182	10,465
2010	45,298	10,122
2011	43,113	6,301
2012	45,136	22,948
2013	46,195	21,527
2014	47,836	37,221
2015	79,921	5,221
2016	68,893	26,752
2017	71,643	30,410
2018	64,070	26,095

Table 4.3: Ultimate values of claim frequency

Inspecting the figures, we may see that for the LoB1 the first and second value are outliers, i.e. they are significantly distant from other observations. For the LoB2 it is visible that the values within period 2007-2011 are quite unstable. This can be reasonably explained by the global financial crisis of 2008 which destabilized market conditions and economic position of many companies and consumers.

Our assumption is to avoid outliers caused by such specific circumstance as financial crisis and to deal with more persistent figures, so we clean the data and remove the first two values of claim frequency for LoB1 and the first 5 values for LoB2. The mean and variance of frequencies (scaled 10^6 times) are as follows:

	LoB1	LoB2
Mean	57.43	24.68
Variance	180.83	100.98

Table 4.4: Mean and variance of claim frequencies

As the actual data rows for ultimate frequencies are very small (10 and 7 observations for LoB1 and LoB2, respectively), fitting the theoretical distribution as well as some other methods (for example, bootstrapping) are not efficient. We are not able to find any appropriate characteristics for population based on very few observations.

So, for further steps we just leave the sample of frequencies as it is.

The run-off triangles of cumulative claim severities are following (values are given in thousands):

	Development period											
Reporting period	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
2007	315	285	285	285	285	285	285	285	285	285	285	285
2008	1 079	1 930	1 729	1 559	1 562	1 577	1 582	1 581	1 596	1 606	1 595	
2009	2 255	3 949	3 754	2 650	2 077	2 077	2 018	1 998	1 998	1 998		
2010	1 048	1 012	1 078	1 150	1 148	1 159	1 159	1 159	1 159			
2011	1 057	1 397	1 434	1 434	1 434	1 434	1 434	1 434				
2012	333	443	465	465	465	465	465					
2013	559	1 826	1 868	1 895	1 604	755						
2014	273	464	459	454	454							
2015	713	1 057	1 016	1 033								
2016	354	597	585									
2017	358	428										
2018	222											

Table 4.5: Run-off triangle of cumulative claim severities for LoB1

	Development period											
Reporting period	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
2007	0	6	27	41	44	44	163	159	160	167	196	218
2008	20	129	270	289	338	384	371	379	385	382	396	
2009	8	37	45	45	81	95	94	92	89	89		
2010	96	98	107	109	123	171	289	172	97			
2011	65	111	106	101	175	169	162	164				
2012	205	355	353	392	492	500	500					
2013	123	610	782	837	885	931						
2014	359	640	682	738	806							
2015	258	704	660	370								
2016	260	489	600									
2017	306	617										
2018	598											

Table 4.6: Run-off triangle of cumulative claim severities for LoB2

In case of claim severities, the main parameter needed is the cumulative development factor classified by reporting year (i.e. proportion required to reach from the last known claim severity value to the ultimate estimated value for claim reported in i-th year).

Report year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
LoB1	1,000	1,000	0,994	0,997	1,000	0,997	0,989	0,894	0,816	0,738	0,721	1,157
LoB2	1,000	1,112	1,201	1,207	1,101	0,984	1,142	1,225	1,413	1,361	1,556	3,471

Table 4.7: Cumulative development factors for claim severities

We may see that for claims reported in the last calendar period the values of cumulative factors are the highest (the difference in values for LoB2 is especially visible), which means that for new reported claims the proportion required to reach the ultimate value is greater than for claims reported earlier. It can be explained by the fact that often at the moment of claim reporting there is a tendency to overestimate the reserve amount but later when the situation is evaluated in more detail, the reserve is reduced, so the estimated ultimate claim and cumulative factor are reduced as well. In practice it is quite hard to get correct first estimates for ultimate claims straight away.

Knowing the cumulative factors for severities we may apply them to last known cumulative incurred individual claim amounts. Thus, we got the selection of individual ultimate claims for each LoB.

The next issue that we face is that some of the ultimate claims are less than our suggested threshold 150 000. As within the claim development process there could be several adjustments of claim reserve amount (based on case investigation details or other factors), the cumulative incurred amount changed quarterly and some claims which were estimated as exceeding threshold, at some point could be significantly adjusted, so the final incurred value can be less than threshold.

The threshold parameter is very important for our practical problem, so for next steps we leave only those ultimate individual claims which were higher than 150 000. Our sample sizes are 82 and 68 for LoB1 and LoB2, respectively. The mean and variance of severities from the original sample are as follows:

	LoB1	LoB2
Mean	424 806	653 476
Variance	$3.137 \cdot 10^{11}$	$6.407 \cdot 10^{11}$

Table 4.8: Mean and variance of the original sample of severities

4.3 Fitting distributions to the tail part of claim severity

After selecting the ultimate claims, we may fit the distribution to data. Our main focus is not on the individual claims themselves, but only on the part that exceeds the threshold. As was discussed in Chapter 2, our candidate models are gamma (and exponential as a particular case), lognormal, Weibull and Pareto.

We fit step-by-step all the models to exceeding part of severities from each LoB.

The parameters of the models are given in the following tables (gamma model did not converge, therefore is missing here):

	Exponential	**Lognormal**	**Weibull**	**Pareto**
LoB1	$\lambda = 3.639 \cdot 10^{-6}$	$\mu = 11.6584$, $\sigma = 1.3036$	$k = 0.7567$, $\lambda = 2.208 \cdot 10^5$	$\alpha = 2.1498$, $\beta = 3.103 \cdot 10^5$
LoB2	$\lambda = 1.986 \cdot 10^{-6}$	$\mu = 12.2248$, $\sigma = 1.4132$	$k = 0.7574$, $\lambda = 4.418 \cdot 10^5$	$\alpha = 1.8708$, $\beta = 4.864 \cdot 10^5$

Table 4.9: Parameters of candidate models for both LoBs

Let's compare the histogram of exceeding part with the density function plots of all candidate models.

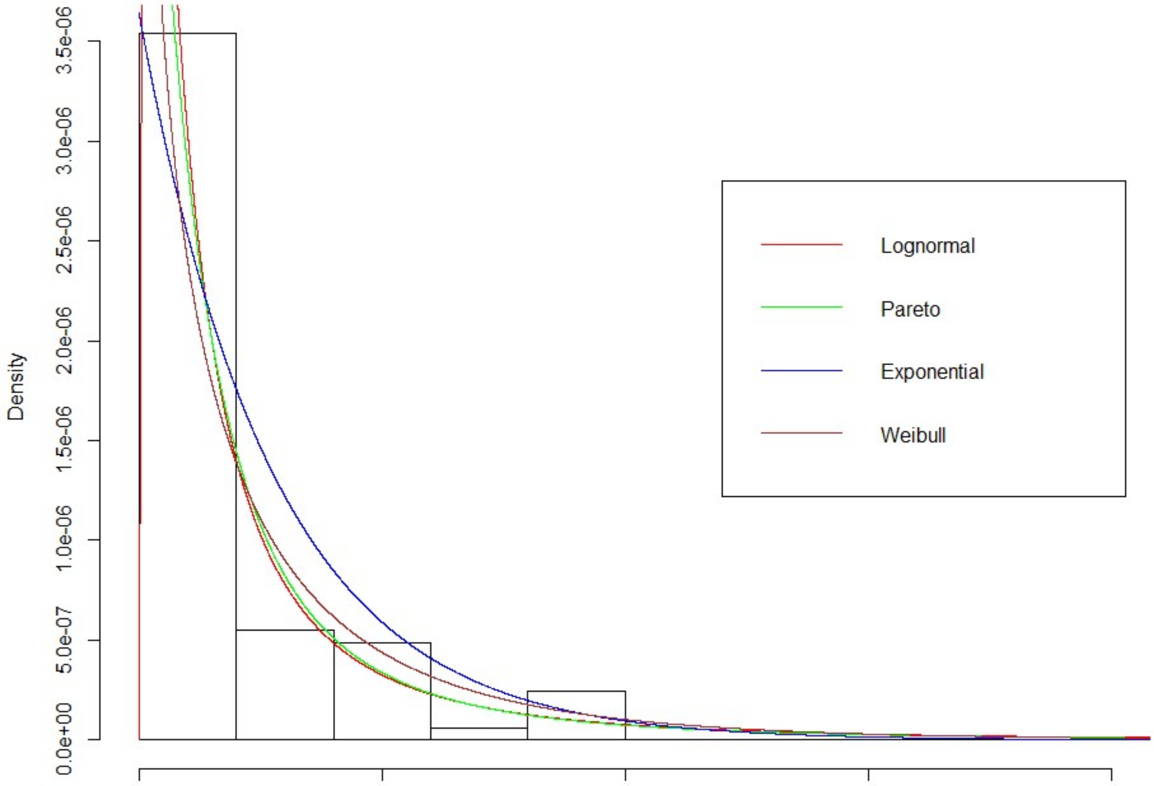

Figure 4.2: Histogram of severity tail part for LoB1 compared to density plots of fitted distributions

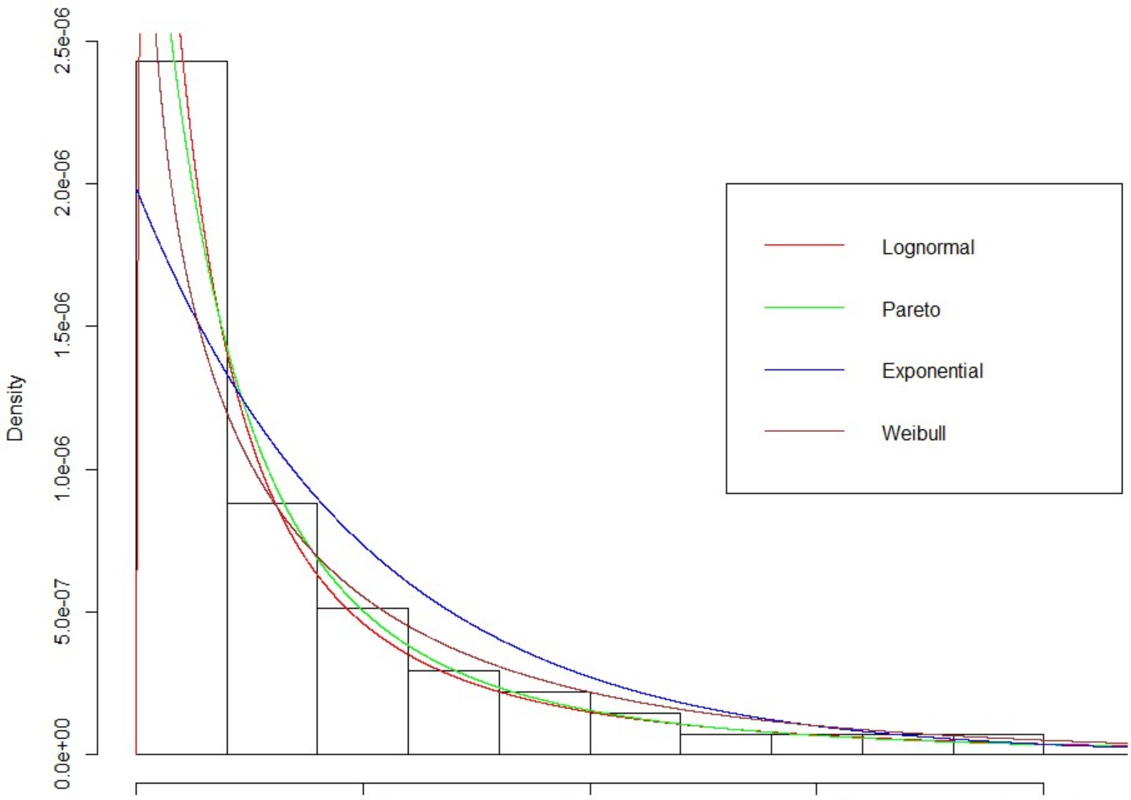

Figure 4.3: Histogram of severity tail part for LoB2 compared to density plots of fitted distributions

From the plots we see that both lognormal and Pareto distribution fit rather well to the data. To compare which model is the best we used the value of Akaike information criteria (AIC).

The AIC values are as follows (gamma model did not converge, therefore is missing here):

	Exponential	Lognormal	Weibull	Pareto
LoB1	2219.907	**2192.16**	2205.892	2192.547
LoB2	1923.584	**1906.581**	1913.381	1908.447

Table 4.10: AIC values of candidate models

It's visible that in both cases lognormal distribution gives the least AIC values, but Pareto values are very close so it's hard to say exactly which one is better.

To use lognormal model is quite conservative approach and it's efficient in statistical sense as the distribution is very close by its nature to normal, so it's easier to derive mathematically its characteristics based on known properties and make conclusions.

Kolmogorov-Smirnov test of the goodness-of-fit (GOF) for lognormal model shows us the following results:

LoB1	**LoB2**
One-sample Kolmogorov-Smirnov test	One-sample Kolmogorov-Smirnov test
D = 0.064819, p-value = 0.8588	D = 0.067803, p-value = 0.8923
alternative hypothesis: two-sided	alternative hypothesis: two-sided

Table 4.11: Kolmogorov-Smirnov GOF test results

We may see that in both cases p-values are higher than 0.05, which means that we can't reject the hypothesis that our data came from lognormal distribution with specified parameters.

Thus, we are able to conclude that lognormal distribution fits quite well to the tail part of severities and we may use it for our further steps.

4.4. Simulation of scenarios and comparison of different reinsurance programs

After fitting the distribution for individual claims, it is possible to model the total loss and amounts of net loss after reinsurance.

The reinsurance is based on excess-of-loss type. Company providing the data uses working XL strategy, i.e. retention is applied to each claim, not to aggregate amount.

By agreement with the company, we set six levels of retention: from 0.5 million to 3 million with step size = 0.5 million.

Our task is to simulate 10 000 scenarios for each reinsurance program and program without reinsurance and to find main characteristics of each program to suggest whether reinsurance is profitable.

For implementing this part, we have to generate the total loss before reinsurance. For each scenario we need random number of claims and claim severities generated from shifted distribution $X = 150\ 000 + Y,$ where Y has lognormal distribution

As claim number is the product of claim frequency and exposure, for generating we use our sample of frequencies and anticipated exposure.

The historical annual exposure values are as follows:

Year	**LoB1**	**LoB2**
2007	150 310	265 592
2008	156 778	297 784
2009	160 818	286 677
2010	176 607	296 391
2011	162 363	317 390
2012	155 086	348 612
2013	173 180	340 383
2014	146 332	349 091
2015	150 148	437 896
2016	174 184	461 833
2017	183 663	497 605
2018	191 353	518 383

Table 4.12: Historical annual exposure for both LoBs

As claim number has to be an integer, we should use reasonable values of anticipated exposure which correspond to historical trend and are convenient for rounding purposes as well.

We suggest, that anticipated exposure for LoB1 is equal to 200 000, and for LoB2 it is 500 000. Multiplying claim frequency values by corresponding exposure, we will get the number of claims sample.

Consolidating all that was mentioned above, for simulation we use two input parameters: number of claims as a sample (product of ultimate claim frequency and anticipated exposure) and claim severity coming from lognormal distribution. Their mean and variance are as follows:

	LoB1		**LoB2**	
	mean	**variance**	**mean**	**variance**
number of claims	11.486	7.233	12.34	25.246
claim severity	420 500	$3.27 \cdot 10^{11}$	703 110	$1.95 \cdot 10^{12}$

Table 4.13: Mean and variance of claim count and severity

To get various simulated results for number of claims, we generate 10 000 random outcomes from ultimate claim frequency sample, multiply them by anticipated exposure and round the result to integer value.

For each scenario we generate the set of claim severities depending on number of claims.

Then we apply for each scenario all possible reinsurance programs: if the individual claim size doesn't exceed the given retention level, the risk will be fully retained by direct insurer, otherwise the exceeding part of each individual claim should be ceded to reinsurer.

For each LoB we get the matrix with 10 000 net losses for each reinsurance program and program without reinsurance.

Visual inspection of density function plot for net losses after several reinsurance programs gives us the understanding that distribution of these losses is multimodal which means that net loss after reinsurance seems to be characterized by mixture of distributions.

This variability is mostly caused by values of claim frequency (especially for LoB2).

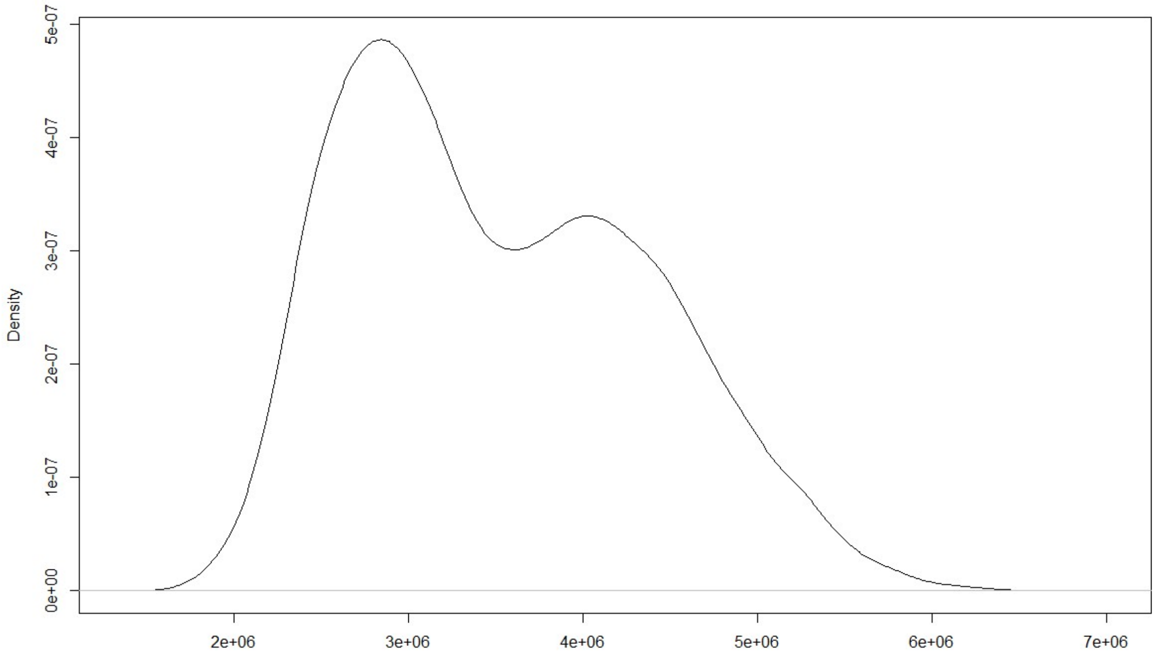

Figure 4.4: Density plot of the net loss after reinsurance with retention 0.5 million (LoB1)

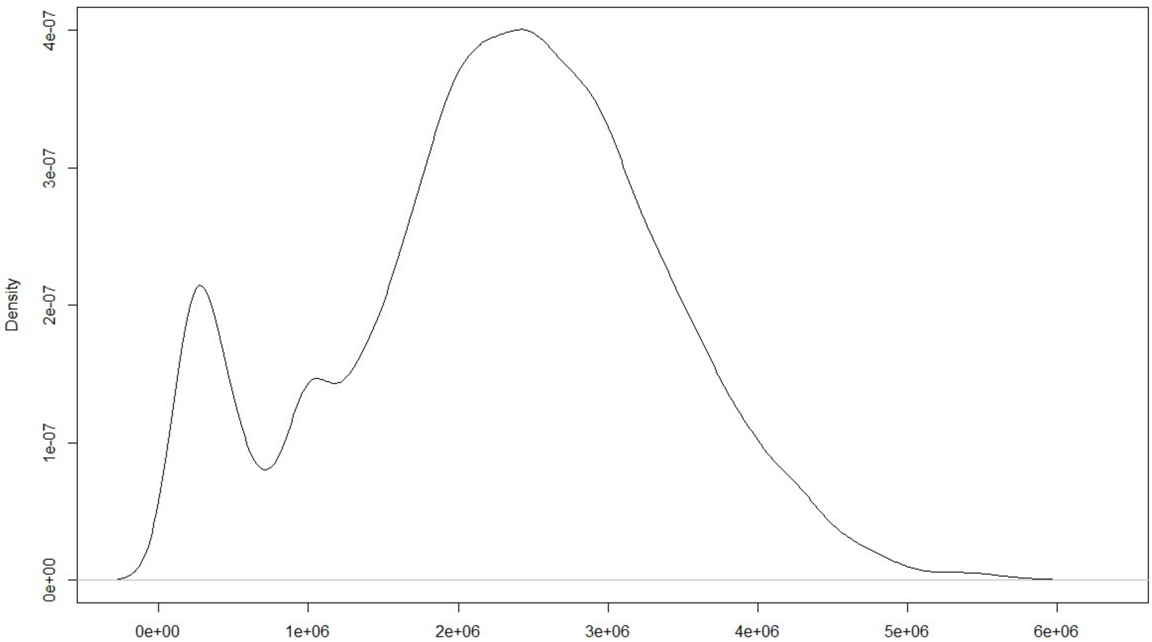

Figure 4.5: Density plot of the net loss after reinsurance with retention 1 million (LoB2)

To estimate the effect of reinsurance program, we can define the average net loss and value at risk (VaR) for different given probabilities.

Definition

The *value at risk* (*VaR)* of a non-negative random variable X *at given confidence level* $(1-\alpha)$ where $0<\alpha<1$ is defined as:

$$VaR_\alpha(X) \triangleq \inf\{x \geq 0\text{: } \mathrm{P}(X > x) \leq \alpha\}.$$

While dealing with continuous and strictly increasing distributions, the VaR can be simply defined as α-quantile of corresponding distribution, i.e.:

$$VaR_\alpha(X) = F^{-1}(\alpha).$$

In other words, the VaR_α represents the loss that with probability α will not be exceeded [6].

For finding the characteristics, we don't try to fit any specific distribution to the net loss, but calculate the necessary values directly from our generated sample as the number of observations is large enough. The values we are interested in are mean, median, $\mathrm{VaR}_{0.75}$ and $\mathrm{VaR}_{0.95}$.

Before analysing the figures and making conclusions, first we need to check whether simulation results are adequate or not. For that purpose, we may compare mean and variance of sample with mean and variance calculated analytically by formula. For the strategy without reinsurance, it is necessary to calculate the mean and variance of total loss, i.e. $E(S)$ and $D(S)$, and for each reinsurance program, the mean and variance of net loss of direct insurer, i.e. $E(\tilde{S})$ and $D(\tilde{S})$.

Based on the formulas from chapter 2 and 3,

$$ES = EN \cdot EX = EN \cdot E(150\ 000 + Y),$$

$$DS = EN \cdot DX + DN \cdot (EX)^2 = EN \cdot D(150\ 000 + Y) + DN \cdot (E(150\ 000 + Y))^2$$

$$E\tilde{S} = E\tilde{N} \cdot E\tilde{X} = EN \cdot E[150\ 000 + Y; R],$$

$$D\tilde{S} = EN \cdot D\tilde{X} + DN \cdot \left(E\tilde{X}\right)^2 = EN \cdot D[150\ 000 + Y; R] + DN \cdot (E[150\ 000 + Y; R])^2$$

where

- N is a number of claims,
- Y is the tail part of claim severity, which is lognormally distributed
- R is the given retention level.

The formulas of mean and variance of lognormal distribution were given in 2.2.2. and the formulas of limited k-th moment were derived in 2.3.2. We can apply them now in our calculations.

The following tables show the comparison results for theoretical and simulated means and variances:

Program	Theoretical mean	Simulated mean	Difference	Proportion of difference to theoretical mean
0.5mln	3 527 444	3 524 160	-3 284	-0.093%
1mln	4 169 751	4 165 741	-4 010	-0.096%
1.5mln	4 419 125	4 417 808	-1 317	-0.0298%
2mln	4 547 521	4 550 201	2 680	0.059%
2.5mln	4 623 571	4 630 440	6 869	0.149%
3mln	4 672 689	4 680 671	7 982	0.0017%
no reinsurance	4 829 767	4 828 111	-1 656	0.034%

Table 4.14: Comparison of means for LoB1

Program	Theoretical variance	Simulated variance	Difference	Proportion of difference to theoretical variance
0.5mln	$8.58 \cdot 10^{11}$	$7.5 \cdot 10^{11}$	$-1.08 \cdot 10^{11}$	-12.59%
1mln	$1.6 \cdot 10^{12}$	$1.44 \cdot 10^{12}$	$-1.66 \cdot 10^{11}$	-10.37%
1.5mln	$2.14 \cdot 10^{12}$	$1.95 \cdot 10^{12}$	$-1.895 \cdot 10^{11}$	-8.86%
2mln	$2.55 \cdot 10^{12}$	$2.35 \cdot 10^{12}$	$-1.95 \cdot 10^{11}$	-7.65%
2.5mln	$2.86 \cdot 10^{12}$	$2.67 \cdot 10^{12}$	$-1.89 \cdot 10^{11}$	-6.61%
3mln	$3.12 \cdot 10^{12}$	$2.92 \cdot 10^{12}$	$-1.93 \cdot 10^{11}$	-6.19%
no reinsurance	$5.04 \cdot 10^{12}$	$4.49 \cdot 10^{12}$	$-5.496 \cdot 10^{11}$	-10.9%

Table 4.15: Comparison of variances for LoB1

Program	Theoretical mean	Simulated mean	Difference	Proportion of difference to theoretical mean
0.5mln	4 400 333	4 420 648	20 315	0.46%
1mln	5 834 011	5 858 907	24 896	0.43%
1.5mln	6 566 220	6 587 984	21 764	0.33%
2mln	7 017 908	7 035 041	17 133	0.24%
2.5mln	7 325 045	7 338 298	13 251	0.18%
3mln	7 547 084	7 558 611	11 527	0.15%
no reinsurance	8 676 692	8 692 843	16 151	0.18%

Table 4.16: Comparison of means for LoB2

Program	Theoretical variance	Simulated variance	Difference	Proportion of difference to theoretical variance
0.5mln	$3.42 \cdot 10^{12}$	$2.91 \cdot 10^{12}$	$-5.07 \cdot 10^{11}$	-14.85%
1mln	$6.71 \cdot 10^{12}$	$5.87 \cdot 10^{12}$	$-8.48 \cdot 10^{11}$	-12.62%
1.5mln	$9.28 \cdot 10^{12}$	$8.23 \cdot 10^{12}$	$-1.05 \cdot 10^{12}$	-11.33%
2mln	$1.14 \cdot 10^{13}$	$1.02 \cdot 10^{13}$	$-1.18 \cdot 10^{12}$	-10.4%
2.5mln	$1.31 \cdot 10^{13}$	$1.19 \cdot 10^{13}$	$-1.26 \cdot 10^{12}$	-9.6%
3mln	$1.46 \cdot 10^{13}$	$1.33 \cdot 10^{13}$	$-1.31 \cdot 10^{12}$	-8.97%
no reinsurance	$3.65 \cdot 10^{13}$	$3.49 \cdot 10^{13}$	$-1.58 \cdot 10^{12}$	-4.32%

Table 4.17: Comparison of variances for LoB2

We notice that mean differences are not so big, especially taking into account that amounts of loss are measured in millions, so the difference doesn't exceed 0.46% of the loss size.

The variance differences, in contrast, are quite large. This can be explained by the big variability of initial data.

We could see that claim severity itself has large variance, which inevitably has an effect on the variance of total loss, and as a result, on the data generated from distribution and the differences in variance between simulated data and analytical results.

We are not able to increase the number of simulations in one run of a programming code, but we run it multiple times instead. As at the beginning we fixed the parameter set.seed(1), we should also try to simulate new outcomes several times to check whether results have systemic nature.

While repeating the procedure, the results for differences vary from time to time. For means, although for LoB1 the simulated results do not differ from theoretical so much and each new code run doesn't influence the proportion of differences significantly, but for LoB2 the results are more sensitive. The result with the biggest mean differences (for LoB2) that we managed to obtain is as follows

Program	Mean difference	Proportion of difference in the theoretical mean
0.5mln	60 506	1.375%
1mln	85 356	1.463%
1.5mln	103 942	1.583%
2mln	116 082	1.654%
2.5mln	126 470	1.727%
3mln	135 069	1.79%
no reinsurance	225 447	2.598%

Table 4.18: The biggest values of mean differences after 30 repetitions for LoB2

We may notice that the proportion of difference in theoretical mean which we got initially did not exceed 0.5% for LoB2. Based on new values, for reinsurance programs the proportion of difference in the theoretical mean is about 1.5% in average, and for the program without reinsurance, the proportion is equal to 2.598%. These figures still could be regarded as an acceptable result.

As for variances the differences are close to those values that we previously obtained and the proportion of difference in the theoretical variance range from 0 to 16%, and we can notice that mostly the simulated variance is less than the theoretical one. It means that we can't treat the results as exact, and for achieving higher accuracy we need to increase the computational power: either use bigger number of simulations (more than 10 million) or make more repetitions (>1000).

4.5 Analysis of results

We can finally comment the characteristics of net loss after reinsurance that we obtained and compare different programs.

The results from the following tables were generated based on the set.seed(1).

LoB1	**retention levels**						**no reinsurance**
	0.5 million	**1 million**	**1.5 million**	**2 million**	**2.5 million**	**3 million**	
mean	3 524 160	4 165 741	4 417 808	4 550 201	4 630 440	4 680 671	4 828 111
median	3 394 582	4 033 122	4 254 377	4 357 199	4 405 282	4 409 124	4 409 124
0.75 quantile	4 185 559	4 985 380	5 312 289	5 498 511	5 604 646	5 670 909	5 742 516
0.95 quantile	5 043 700	6 339 992	6 993 398	7 384 188	7 664 111	7 890 403	8 683 438

Table 4.19: Mean and quantiles of net losses for LoB1

LoB2	**retention levels**						**no reinsurance**
	0.5 million	**1 million**	**1.5 million**	**2 million**	**2.5 million**	**3 million**	
mean	4 420 648	5 858 907	6 587 984	7 035 041	7 338 298	7 558 611	8 692 843
median	4 581 932	6 026 908	6 711 357	7 103 120	7 352 429	7 514 382	7 775 801
0.75 quantile	5 504 445	7 446 910	8 472 546	9 104 815	9 543 061	9 870 179	10 903 386
0.95 quantile	7 012 475	9 588 041	11 165 120	12 255 120	13 110 017	13 759 361	18 533 627

Table 4.20: Mean and quantiles of net losses for LoB2

We may notice that for LoB1 mean loss without reinsurance is close to the $\mathrm{VaR}_{0.95}$ for the reinsurance with 0.5 million retention (4.82 million and 5.04 million, respectively). At the same time $\mathrm{VaR}_{0.95}$ without reinsurance is about 8.7 million. This situation demonstrates that if, for example, the mentioned reinsurance program costs 1 million, then it is beneficial for insurer to incur additional expenses but at the same time to decrease the level of risk.

Analogically for LoB2, we see that if reinsurance with 0.5 million retention level costs, for example, 1.5 million, then the sum of $\mathrm{VaR}_{0.95}$ and incurred costs for reinsurance is still lower than average loss without reinsurance (8.51 million and 8.69 million, respectively).

Thus, when insurance company knows the real reinsurance offers and knows VaRs and average loss for different programs, it may conclude whether the program is beneficial for it and which program would be the best.

As the variance of loss is very large (measured in 10^{12} or even in 10^{13}), which means that standard deviation is measured in millions, that indicates high level of uncertainty for reinsurance company as well, so it may request higher premium for reducing the risk and protecting its solvency.

Conclusion

The matter of finding the model which adequately describes the behaviour of large claims will be always the urgent task for insurance companies. Depending on how large the historical data is, the nature of data, how long the claims are being developed and which assumptions were made, the analysis process may vary. While simulating different scenarios, the company should know not only the average amount of net loss, but the different quantiles as well, to be able to predict the value at risk with given probability. Dealing with large claims requires more responsible and accurate approach, as the possible outcomes in case of incorrect assessment might be catastrophic.

Our modelling involved three stages, which required making reasonable assumptions, accurate and consistent calculations, as each intermediate result influenced the subsequent solution. We found the ultimate values of claim frequency and claim severity with chain ladder method, then fitted the distribution to the tail part of claim severity and chose lognormal model as the best. Then we simulated 10 000 scenarios to obtain the results for net loss before and after reinsurance with different retention levels. We compared the mean and variance of simulated samples with their theoretical values and found that the simulating results for means are accurate enough, but with variances the situation is more questionable, as the proportion of difference in the theoretical variance ranges from 0 to 16% (the simulated variance is mostly smaller than the theoretical). In that case, the higher number of simulations or repetitions is required to achieve more accurate results.

Comparing the means and values at risk for different programs, we could see that some reinsurance programs can be rated as advantageous since their values of $\mathrm{VaR}_{0.95}$ were less or almost equal to the mean of loss without reinsurance, so it is reasonable for direct insurer to pay the certain premium to reinsurance company but reduce the average level of risk. It is worth noting that we didn't have the information about real offers from reinsurers and which premium amount they might request, so we tried to make some adequate assumptions on our own.

As an improvement for a further research, it is possible also to include into the modelling procedure other distributions which should deal mainly with tails of data, such as generalized Pareto distribution (GPD) or some conditional distributions which are based on the assumption that the data is truncated from below. Also, it may be reasonable to assume that the development of claims reported in the first year of covered period still goes on, which means that we should include the extra development factor and multiply the ultimate values by it. The additional option is to estimate future inflation and adjust the ultimate values with it as well. For simulating different reinsurance programs, there can be applied the limitation of the upper bound of the reinsured sum, so that the reinsurer has to pay the amount over the retention R up to maximum level M.

We believe that this study may have practical application and will be useful for insurance company in decision-making process while choosing between different reinsurance programs.

Bibliography

[1] Actuarial Community. (2018) *Loss Data Analytics.* URL: http://www.researchgate.net/publication/327153234_Loss_Data_Analytics. Last visited: 12/05/2019.

[2] Booth P., Chadburn, R., Cooper, D., Haberman, S., James, D. (1999) *Modern Actuarial Theory and Practice.* Chapman & Hall / CRC.

[3] Bowers N.L., Gerber H.U., Hickman J.C., Jones D.A., Nesbitt C.J. (1997) *Actuarial Mathematics.* Society of Actuaries.

[4] Dutang C., Goulet V., Pigeon M. (2008) *actuar: An R Package for Actuarial Science.* Journal of Statistical Software, vol. 25, no. 7, 1-37. URL: http://www.jstatsoft.org/v25/i07. Last visited: 10/05/2019.

[5] Gray, R.J. & Pitts, S.M. (2012) *Risk Modelling in General Insurance. From Principles to Practice.* Cambridge University Press.

[6] Hardy M.R. (2006) *An Introduction to Risk Measures for Actuarial Applications.* Education and Examination Committee of the Society of Actuaries.

[7] Hogg, R. & Klugman, S. (1984) *Loss distributions.* Wiley, New York.

[8] Käärik M. (2018) *Non-Life Insurance Mathematics. Lecture notes.* University of Tartu. Institute of Mathematics and Statistics.

[9] Klugman, S.A., Panjer, H.H., Willmot, G.E. (1998) *Loss Models: From Data to Decisions.* Wiley, New York.

[10] Kravych Y., Mergel V. (2000) *Large Loss Distributions: probabilistic properties, EVT tools, maximum entropy characterization.* Proceedings of the 31st ASTIN Colloquium, Sardinia, Italy.

[11] Munich Re. (2010) *Reinsurance: A Basic Guide To Facultative And Treaty Reinsurance.* Munich Reinsurance America, Inc.

[12] Swiss Re. (2015) *The essential guide to reinsurance.* Swiss Reinsurance Company Ltd.

[13] Swiss Re. (2016) *Strategic reinsurance and insurance: the increasing trend of customised solutions.* Swiss Re Ltd, Economic Research & Consulting.

[14] Zuanetti D.A., Diniz C.A.R., Leite J.G. (2006) *A Lognormal Model for Insurance Claims Data.* Revstat – Statistical Journal, Vol. 4, №2, 131-142

Non-exclusive licence to reproduce thesis and make thesis public

I, Iuliia Kim,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1. reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, and

1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work from ***15/05/2024*** until the expiry of the term of copyright, graduation thesis "Modelling large claims in order to optimise the reinsurance program", supervised by Meelis Käärik and Hele-Liis Viirsalu,

2. I am aware of the fact that the author retains the rights specified in p. 1.

3. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Iuliia Kim

15/05/2019