Charles University in Prague
Faculty of Mathematics and Physics

MASTER THESIS

Bc. Kristýna Bílková

Granular loss models in reserving

Department of Probability and Mathematical Statistics

Supervisor of the master thesis:	RNDr. Michal Pešta, Ph.D.
Study programme:	Mathematics
Specialization:	Financial and Insurance Mathematics

Prague 2014

I would like to thank my supervisor RNDr. Michal Pešta, Ph.D. for his valuable advice, guidance and patience. Many thanks to RNDr. Petr Jedlička, Ph.D. for providing me with data for the practical part of this thesis. Last but not least, I would like to thank my family and friends for their support, especially my parents who have always been there for me.

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In .... on .... Author signature

Název práce: Granular loss models in reserving

Autor: Bc. Kristýna Bílková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Michal Pešta, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Většina metod pro odhad rezerv na pojistná plnění používá data agregovaná do vývojových trojúhelníků, díky čemuž velké množství informací, které má pojišťovna k dispozici, zůstává nevyužito. Tato práce předkládá přístup využívající granulární informace o jednotlivých škodách neseskupených do trojúhelníků. Je zde vytvořen statistický model vývoje škod, který může být dále využit pro odhady rezerv na pojistná plnění. Tento model se skládá z čítacího procesu, kterým se řídí doby nastání pojistných událostí, rozdělení doby mezi nastáním a nahlášením pojistné události a rozdělení výše škod. Je zde představeno několik vhodných rozdělení a metody pro odhad jejich parametrů. Teoretický aparát je aplikován na reálná data. Práce se dále zabývá srovnáním odhadu IBNR rezervy pomocí granulární metody a standartní metody Chain ladder. Toto srovnání je provedeno jak na reálných, tak i na uměle nasimulovaných datech. Pro data použitá v této práci se větší komplexnost a nároky na přesnost dat ukazují být ve prospěch lepší přesnosti odhadů a univerzálnosti.

Klíčová slova: Rezervy na pojistná plnění, granulární data, IBNR, Chain ladder

Title: Granular loss models in reserving

Author: Bc. Kristýna Bílková

Department: Department of Probability and Mathematical Statistics

Supervisor of the master thesis: RNDr. Michal Pešta, Ph.D., Department of Probability and Mathematical Statistics

Abstract: Claims reserving methods usually use data aggregated into development triangles, therefore a lot of information that insurance companies possess remains unused. This thesis shows a triangle-free approach using granular information from a claim by claim database. A statistical model for claims development which can further be used for estimation of reserves is built. The statistical model consists of a counting process that drives claims occurrence, distribution of reporting delay and distribution of claims severity. Several suitable distributions are presented, as well as methods for obtaining their parameters from data. Theoretical apparatus is used for real data. The thesis also pursues comparison of the IBNR reserve estimation using the triangle free approach and distribution free Chain ladder method for real data as well as for simulated data sets. For the data used in this thesis the complexity and data requirements of the triangle free approach are in favor of more preciseness and versatility.

Keywords: Claims reserving, granular data, IBNR, Chain ladder

Contents

Introduction and motivation

An insurance policy is a financial services contract according to which an insurer commits to providing a financial coverage against a random occurrence of specified events and an insured commits to pay deterministic payments called premiums to the insurer for this service. Occurrence time of such events and the amount the insurer would have to pay to the insured are random variables. The insurer's obligation to pay under defined conditions represents a claim and the insurer needs to make reserves for such claims sufficient enough to stay solvent. Also, according to Solvency II which shall come to effect in the year 2016 insurers should do the best estimate of their assets and liabilities. Determination of reserves amounts has a key impact on the financial position of an insurance company, therefore finding suitable methods for claims estimation is an important issue.

This thesis deals only with non-life insurance as types of claims, risk drivers, terms of contracts, etc. are different in life and non-life insurance, therefore different modeling approaches are suitable. Moreover, according to Solvency II if an insurance company provides both life and non-life insurance it shall manage those two branches separately.

According to the Insurance Act No. 277/2009 Coll. non-life insurance includes following lines of business:

- Accident and sickness insurance
- Motor/car insurance
- Marine and transport insurance
- Aviation insurance
- Insurance against fire and other property damage
- Liability insurance
- Credit insurance and guarantees
- Other insurance listed in the act (such as travel insurance, financial loss insurance, etc.)

Claims reserving

Figure 1 taken from Antonio and Plat (2013) illustrates an example of a general non-life insurance claim development. A claim occurs at time t_1 and is reported with some delay at time t_2. Time $(t_2 - t_1)$ is a random variable called reporting

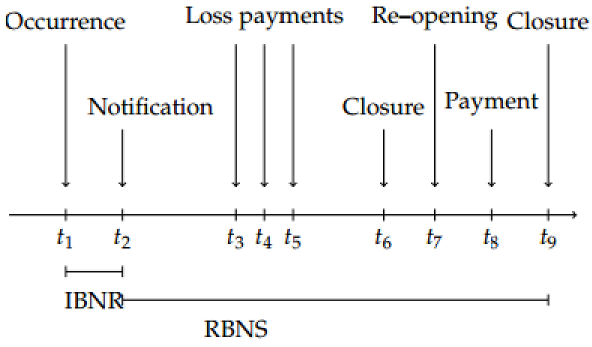

Figure 1: Development of a general insurance claim

delay. After some internal process is carried out, insurance company pays loss payments in times t_3, t_4, t_5. Sometime after closing the claim (in time t_6) some further facts might emerge and cause re-opening of the claim in time t_7 that could lead to revaluation of the claim and further payment in time t_8 and then final closure of the claim in time t_9. Time between reporting and settling the claim $(t_6 - t_2)$ or $(t_9 - t_2)$ is also a random variable called settlement delay.

Because of reporting and settlement delays claims cannot be settled immediately, therefore insurance companies need to hold reserves for covering these claims.

We distinguish between several types of claims reserves. RBNS stands for reported but not settled claims. These are the claims that had occurred and were reported before the present moment but their settlement would occur in the future, hence in Figure 1the present moment is somewhere between t_2 and t_6 or t_9. IBNR stands for incurred but not reported claims. These are claims that occured before the present moment but will be reported in the future, therefore the present moment is somewhere between t_1 and t_2. IBNER stands for incurred but not enough reserved/reported claims. It might happen that the initial reported or estimated claim severity might change during the time. UPR stands for unearned premium reserve. It is a reserve for claims that have not occurred yet but are covered by the policies currently in force. In such a case the present moment is before t_1.

There are several approaches to the claims reserving problem, such Chain Ladder model, Bayesian models, Over-Dispersed Poisson Model, Generalized linear models and others (they are described for example in Wüthrich and Merz (2008) and Endgland and Verrall (2002)). Models usually use aggregated data therefore a lot of information on claims development remains unused and this might lead to less appropriate results. Quoting Endgland and Verrall (2002, pp.507):

> ”However, it has to be borne in mind that traditional techniques were developed before the advent of desktop computers, using methods which could be evaluated using pencil and paper. With the continuing increase in computer power, it has to be questioned whether it would not be better to examine individual claims rather than use aggregated data. Databases of individual claims are routinely used for pricing purposes, so the provision of individual claims databases for reserving is feasible.”

Insurance companies collect information on all events illustrated in Figure 1. They have claim-by-claim databases including information on reported claims. For each reported claim there is an occurrence date, a reporting date and all the paid amounts with dates of payments if there are any. Insurance companies also have claim-by-claim databases with changes of the reserved amounts for each claim that has been reported.

This thesis presents a claims reserving approach that uses these granular data and is based mainly on the article Parodi (2014). We were provided a real data set on which we can illustrate the described apparatus.

Data description and main goal

We were provided data from the evidence of the Czech Insurer's Bureau Guarantee Fund for car insurance. The data consist of three data-sheets containing selections from the Czech Insurers' Bureau's database of twelve years of claims development from the beginning of the year 2000 to the spring 2013. Data was selected randomly from the complete database of the Czech Insurer's Bureau. The used currency is the Czech Koruna henceforward denoted as CZK.

The first data-sheet contains information on claims payments. Table 1 shows five rows of this data-sheet. It contains 47 301 records in total, each row represents one record and contains:

- Claim ID. There are 38 258 unique IDs, therefore if one claim is associated with more payments each of those payments is on a single row.
- Type of claim. It can be either a health claim or a material claim. There are 39 605 material claims payments and 7 696 health claims payments.
- Occurrence date. The first claim occurrence in this database is 1.1.2000 and the last is 29.4.2013.
- Reporting date. The first reporting is from 20.1.2000 and the last from 13.5.2013.
- Date of payment. It is a date when the payment is credited to the client's bank account. The first payment is from 28.7.2000 and the last from 21.5.2013.
- Amount of payment. It is denominated in CZK.

ID	Type	Occurrence date	Reporting date	Date of payment	Amount
3138	material	19.2.2000	8.3.2000	6.11.2000	142 284
3140	material	11.2.2000	9.3.2000	20.12.2000	10 287
3142	health	16.2.2000	9.3.2000	9.10.2003	48 170
3142	material	16.2.2000	9.3.2000	9.10.2003	74 626
3142	material	16.2.2000	9.3.2000	27.4.2001	53 123

Table 1: Five rows of claims payment database

The second data-sheet contains information on changes of reserves for individual claims. Figure 2 shows five rows of claims reserves development. We have no information on how exactly the reserves were determined. There are 206 418 records in total, each row represents one record of a change in reserve and contains:

- Claim ID. There are 58 272 unique IDs. Each claim from the first file can be found in this and more over there are reported claims for which no payments were realized before the end of the observed period.
- Type of claim. It can be either a health claim or a material claim. There are 173 561 material claims records and 32 857 health claims records.
- Occurrence date. The first occurrence time is 1.1.2000 and the last is 21.5.2013.
- Reporting date. The first reporting is from 20.1.2000 and the last from 23.5.2013.
- Date of change. This is a date when change of reserve for an individual claim was recognized. The first change of reserve happened 16.5.2000 and the last 23.5.2013.
- Amount. This an amount of change in reserve denominated in CZK. Negative value means decreasing reserve, positive value means increasing reserve.

ID	Type	Occurrence date	Reporting date	Date of change	Amount
2918	material	6.2.2000	16.2.2000	16.6.2000	-30 000
2918	health	6.2.2000	16.2.2000	17.5.2000	1 000
2918	health	6.2.2000	16.2.2000	16.6.2000	-1 000
2923	material	4.2.2000	16.2.2000	18.5.2000	68 832
2923	material	4.2.2000	16.2.2000	16.6.2000	-68 832

Table 2: Five rows of reserves database

This data also include annuity claims. The third file contains a list of IDs of claims with annuity. We use this file to identify claims associated with annuities and henceforward exclude them from our computations as such a type of claim has different properties and development and needs to be modeled in a different way.

The main goal of this thesis is to create a model of a claims development process for the purpose of reserving. We are going to use the information on each claim from the data-sheets described above. Based on this model we should be able to predict future development of the claims process and estimate reserves. We also want to show that the complexity of such an approach is in favor of more preciseness compared to the traditional triangle based technique.

Chapter 1

Triangle-free approach

Our aim is to find a model for claims count and claims severity. The course of the method we use is as follows. From the data available we estimate parameters of the model. Having done this we run simulations and then, using the results, obtain the final estimate of future development.

1.1 Claims count

In this section we show how to estimate number of claims that occurred before time t. We take claims occurrence as a counting process N_t, $t \geq 0$, where $N_0 = 0$ almost certainly, therefore the number of claims occurred until time t is $N_t = N_t - N_0$. In this thesis we focus on the homogeneous Poisson process and a general non-homogeneous counting process.

1.1.1 Homogeneous Poisson process

Claims occurrence N_t might be taken as a homogeneous Poisson process with intensity μ which is a counting process with independent increments over disjoint intervals of which an increment in an interval of length s follows Poisson distribution with mean μs, $\mu > 0$. Alternative definition taken from Prášková and Lachout (2001):

Definition 1 (Poisson process). *Poisson process with intensity μ, $\mu > 0$ is an integer-valued counting process $\{N_s, s \geq 0\}$ for which it is true that:*

1. *Numbers of events in disjoint intervals are independent.*
2. *For all $s \geq 0$ probability of exactly one event occurrence in an interval $(s, s+h]$ is $\mu h + o(h)$, $h > 0$.*
3. *For all $s \geq 0$ probability of at least two events occurrence in an interval $(s, s+h]$ is $o(h)$, $h > 0$,*

where $o(h)$ is a symbol for expressions for which it is true that

$$\lim_{h \to 0_+} \frac{o(h)}{h} = 0 \,.$$

It is also proven in Prášková and Lachout (2001) that:

1. Number of events that occurred in $(0, t]$ follows Poisson distribution with parameter μt: $\mathsf{P}(N_t = n) = \frac{(\mu t)^n}{n!} \exp\{-\mu t\}$.
2. Time between two successive events occurrence follows exponential distribution with parameter μ: $\mathsf{P}(\tau > x) = \exp\{-\mu x\}$, $\forall x > 0$.
3. Process is stationary in time.

Estimating intensity of the process

As the time between two successive events occurrence follows exponential distribution with parameter μ, μ is also the mean of this process and can be estimated as a sample mean. The problem with this process is that we only observe occurrence time for claims that were reported before the present moment t, but there might also be claims which occurred before t but will be reported later. If the observation period $[0, t]$ is long enough with many observations we can proceed the following way.

First we choose a time window $w \subseteq [0, t]$. Our aim is to estimate μ_w which is the mean of time between two successive claims occurrence in the time window w. We make an assumption that all claims that occurred in time window w were reported before t. Therefore we assume to have all the observations of occurrence times for this time window. Hence we estimate the value for w using the successive time period.

If claims occur in times $t_1 \leq \ldots \leq t_n$, $t_1, \ldots, t_n \in w$ where t_n is the last occurrence time before the end of w, the estimate of μ_w is

$$\hat{\mu}_w = \frac{1}{n-1} \sum_{i=2}^{n} (t_i - t_{i-1}) .$$

As we assume homogeneity of the claims occurrence process, what holds for time period w actually holds for any time thus $\mu_w = \mu$.

While choosing a time window w two contradictory aspects need to be borne in mind:

- The larger the time window w the more observations we have for the estimation of μ_w hence, arising from consistency of a sample mean as an estimate of mean, the estimate should be more accurate.
- On contrary if the time window w is too large or too close to the end of the observable period, we are not able to observe all claims that actually occur because of the reporting delay, therefore the assumption that all claims that had occurred in time window w were reported before t does not hold and the estimate $\hat{\mu}_w$ is biased.

As for fixed $t \geq 0$ random variable N_t follows Poisson distribution with parameter μt, its mean is equal to its variance which is μt. Sometimes data shows bigger variance than the mean and then another process should be chosen to model the data. In such a case we should probably choose another stochastic counting process, for example some of those described in Winkelmann (2008) such as a negative-binomial process. Another problem might be non-homogeneity of the process. In such a case we might be willing to use some non-homogeneous process.

1.1.2 Non-homogeneous counting process

We could take $\{N_t,\, t \geq 0\}$ as a non-homogeneous process with intensity at time t:

$$\mu_t = \int_0^t v(x)\, dx\,,$$

where $v(x)$ is the intensity of claims occurrence at time x.

We consider number of reported claims to be also driven by a non-homogeneous counting process R_t, where $R_0 = 0$ almost certainly, number of claims reported until time t is $R_t - R_0$ and the intensity of this process is

$$\rho_t = \int_0^t v(x) F(t-x)\, dx\,,$$

where $F(x)$ is the distribution function of the reporting delay

$$F(x) = \mathrm{P}(\text{reporting date} - \text{occurence date of a claim} \leq x)\,.$$

Increments over disjoint intervals are also independent. Unlike the number of claims incurred, values of the process R_t can be observed.

We can estimate mean of the process N_t the same way as in Parodi (2014):

$$\hat{\mu}_t = \frac{\int_0^t v(x)\, dx}{\int_0^t v(x) F(t-x)\, dx} R_t\,, \tag{1.1}$$

which is an unbiased estimate as can be seen:

$$E\hat{\mu}_t = \frac{\mu_t}{\rho_t}\rho_t = \mu_t\,.$$

An estimate of the expected number of claims that occurred until time $s < t$ where t is the longest observable period is

$$\hat{\mu}_s = \frac{\int_0^s v(x)\, dx}{\int_0^s v(x) F(t-x)\, dx} R_s\,.$$

For sake of simplicity we might assume that the intensity of claims occurrence does not change during a time period, therefore $v(x) = 1$. Hence

$$\hat{\mu}_t = \frac{t}{\int_0^t F(t-x)\, dx} R_t$$
$$s < t:\ \hat{\mu}_s = \frac{s}{\int_0^s F(t-x)\, dx} R_s\,.$$

If we take the time as a discrete variable we get

$$\begin{aligned}\hat{\mu}_t &= \frac{t}{\sum_{k=0}^{t} F(t-k)} R_t \\ s<t:\ \hat{\mu}_s &= \frac{s}{\sum_{k=0}^{s} F(t-k)} R_s\,.\end{aligned}$$

Obtaining the estimate of distribution of reporting delay is further described in Section 1.1.3. We need to realize that $\hat{\mu}_t$ is a point estimate of intensity of a non-homogeneous counting process at time t.

1.1.3 Distribution of reporting delay

Let D denote a random variable representing a reporting delay. We can assume that the distribution of D is known, e.g. exponential, or we can estimate it from the data. We just need to realize that the empirical cumulative distribution function estimated from the data would be biased towards smaller delays as we have observations until time t only therefore firstly, we are not able to observe delays longer than t and secondly, we cannot observe claims with a long reporting delay that have occurred recently.

What we actually can observe from data is

$$f_t(x) = \mathrm{P}(D = x \mid T_0 + D \le t)\,,$$

where T_0 is the occurrence time of a claim. From the Bayes' theorem that can be found for example in Anděl (2007) we get

$$f_t(x) = \mathsf{P}(D = x \mid T_0 + D \le t) = \frac{\mathrm{P}(T_0 + D \le t \mid D = x)\mathrm{P}(D = x)}{\mathrm{P}(D + T_0 \le t)}\,.$$

Assuming the intensity of a claim occurrence to be $v(x) = 1$, the distribution of the occurrence time of a claim during a period $[0, t]$ is uniform:

$$\mathsf{P}(T_0 \le t - x) = \frac{t-x}{t} = 1 - \frac{x}{t}$$

and by definition

$$\mathrm{P}(D = x) = f(x)\,.$$

Altogether we get

$$f_t(x) = \begin{cases} \dfrac{\left(1 - \frac{x}{t}\right) f(x)}{G(t)} & \text{for } x < t,\ , \\ 0 & \text{elsewhere.} \end{cases}$$

where $G(t) = \mathrm{P}(D + T_0 \le t)$ which is a function of t only and can be taken just as a normalizing factor. Hence

$$f(x) = \begin{cases} \dfrac{f_t(x)G(t)}{1 - \frac{x}{t}} & \text{for } x < t,\ , \\ \text{undefined} & \text{elsewhere.} \end{cases} \tag{1.2}$$

Example 1 (Estimating parameter of exponential delay distribution). Assume the delay distribution is exponential with unknown parameter λ and density

$$f(x) = \frac{1}{\lambda}\mathrm{exp}\left\{-\frac{x}{\lambda}\right\} .$$

We observe average delay distribution

$$\hat{\lambda}_{obs} = \frac{1}{n}\sum_{i=1}^{n}\left((reporting\, date\, of\, claim\, i) - (occurence\, date\, of\, claim\, i)\right) ,$$

where n is the total number of claims reported during the time interval $[0, t]$. If n is big enough

$$\begin{aligned}
\lambda_{obs} &= E(D \mid T_0 + D \leq t) = \int_0^t f_t(x)\, dx \\
&= \int_0^t x\frac{\left(1-\frac{x}{t}\right)\frac{1}{\lambda}\mathrm{exp}\left\{-\frac{t}{\lambda}\right\}}{1-\frac{\lambda}{t}\left(1-\mathrm{exp}\left\{-\frac{t}{\lambda}\right\}\right)}\, dx \\
&= \frac{1}{\lambda\left(1-\frac{\lambda}{t}\left(1-\mathrm{exp}\left\{-\frac{t}{\lambda}\right\}\right)\right)}\left(\int_0^t x\mathrm{exp}\left\{-\frac{x}{\lambda}\right\}\, dx - \int_0^t \frac{x^2}{t}\mathrm{exp}\left\{-\frac{x}{\lambda}\right\}\, dx\right) ,
\end{aligned}$$

where

$$\begin{aligned}
\int_0^t x\mathrm{exp}\left\{-\frac{x}{\lambda}\right\} dx &= \left[per\, partes : \begin{array}{cc} u = x & v = -\lambda\mathrm{exp}\{-\frac{x}{\lambda}\} \\ u' = 1 & v' = \mathrm{exp}\left\{-\frac{x}{\lambda}\right\} \end{array}\right] \\
&= \left[-\lambda x\mathrm{exp}\left\{-\frac{x}{\lambda}\right\}\right]_0^t + \lambda\int_0^t \mathrm{exp}\left\{-\frac{x}{\lambda}\right\} dx \\
&= t\lambda\mathrm{exp}\left\{-\frac{t}{\lambda}\right\} + \lambda\left[-\lambda\mathrm{exp}\left\{-\frac{x}{\lambda}\right\}\right]_0^t \\
&= -t\lambda\mathrm{exp}\left\{-\frac{t}{\lambda}\right\} - \lambda^2\mathrm{exp}\{-\frac{t}{\lambda}\} + \lambda^2 \\
&= \lambda\left(\lambda - \mathrm{exp}\left\{-\frac{t}{\lambda}\right\}(\lambda + t)\right)
\end{aligned}$$

and

$$\begin{aligned}
\int_0^t \frac{x^2}{t}\mathrm{exp}\left\{-\frac{x}{\lambda}\right\} dx &= \left[per\, partes : \begin{array}{cc} u = x^2 & v = -\lambda\mathrm{exp}\left\{-\frac{x}{\lambda}\right\} \\ u' = \frac{2x}{t} & v' = \mathrm{exp}\left\{-\frac{x}{\lambda}\right\} \end{array}\right] \\
&= \left[-\lambda\frac{x^2}{t}\mathrm{exp}\left\{-\frac{x}{\lambda}\right\}\right]_0^t + \frac{2\lambda}{t}\int_0^t x\mathrm{exp}\left\{-\frac{x}{\lambda}\right\} dx \\
&= -\lambda t\mathrm{exp}\left\{-\frac{t}{\lambda}\right\} + \frac{2\lambda}{t}\lambda\left(\lambda - \mathrm{exp}\left\{-\frac{t}{\lambda}\right\}(\lambda + t)\right) \\
&= \frac{2\lambda^3 - \mathrm{exp}\left\{-\frac{t}{\lambda}\right\}\lambda\left(t^2 + 2t\lambda 2\lambda^2\right)}{t} ,
\end{aligned}$$

therefore alltogether

$$
\begin{aligned}
\lambda_{obs} &= \frac{1}{\lambda\left(1-\frac{\lambda}{t}\left(1-\exp\left\{-\frac{t}{\lambda}\right\}\right)\right)}\left(\lambda\left(\lambda-\exp\left\{-\frac{t}{\lambda}\right\}(\lambda+t)\right)\right. \\
&- \left.\left(\frac{2\lambda^3-\exp\left\{-\frac{t}{\lambda}\right\}\lambda\left(t^2+2t\lambda+2\lambda^2\right)}{t}\right)\right) \\
&= \frac{1}{t\lambda\left(1-\frac{\lambda}{t}\left(1-\exp\left\{-\frac{t}{\lambda}\right\}\right)\right)}\left(\lambda^2 t-\lambda^2 t\exp\left\{-\frac{t}{\lambda}\right\}-\lambda t^2\exp\left\{-\frac{t}{\lambda}\right\}\right. \\
&- \left.2\lambda^3+\lambda t^2\exp\left\{-\frac{t}{\lambda}\right\}+2t\lambda^2\exp\left\{-\frac{t}{\lambda}\right\}+2\lambda^3\exp\left\{-\frac{t}{\lambda}\right\}\right) \\
&= \frac{\lambda}{\left(1-\frac{\lambda}{t}\left(1-\exp\left\{-\frac{t}{\lambda}\right\}\right)\right)}\left(1-\exp\left\{-\frac{t}{\lambda}\right\}-\frac{t}{\lambda}\exp\left\{-\frac{t}{\lambda}\right\}-2\frac{\lambda}{t}\right. \\
&+ \left.\frac{t}{\lambda}\exp\left\{-\frac{t}{\lambda}\right\}+2\exp\left\{-\frac{t}{\lambda}\right\}+2\frac{\lambda}{t}\exp\left\{-\frac{t}{\lambda}\right\}\right) \\
&= \frac{\lambda}{\left(1-\frac{\lambda}{t}\left(1-\exp\left\{-\frac{t}{\lambda}\right\}\right)\right)}\left(1+\exp\left\{-\frac{t}{\lambda}\right\}-2\frac{\lambda}{t}\left(1-\exp\left\{-\frac{t}{\lambda}\right\}\right)\right) \\
&= \lambda\left(1+\frac{\exp\left\{-\frac{t}{\lambda}\right\}-\frac{\lambda}{t}\left(1-\exp\left\{-\frac{t}{\lambda}\right\}\right)}{\left(1-\frac{\lambda}{t}\left(1-\exp\left\{-\frac{t}{\lambda}\right\}\right)\right)}\right), \qquad (1.3)
\end{aligned}
$$

from which λ can be derived numerically.

Tail of the distribution of reporting delay

As we observe data only until time t the density 1.2 is defined only for values smaller than t. Here we describe how to estimate the distribution for tail values (bigger than t).

If we assume that the delay distribution is exponential with parameter λ, then

$$f(x) = \frac{1}{\lambda}\exp\left\{-\frac{x}{\lambda}\right\} \ \forall x \geq 0 \tag{1.4}$$

which also includes values bigger than t.

In case of general density $f(x)$ from 1.2 we can approximate the tail distribution by exponential distribution with parameter λ obtained as in example 1 and we get

$$f(x) = \begin{cases} \dfrac{f_t(x)G(t)}{1-\frac{x}{t}} & \text{for } x < t, \, , \\ \frac{1}{\lambda}\exp\left\{-\frac{x}{\lambda}\right\} & \text{for } x > t\,. \end{cases} \tag{1.5}$$

1.2 Claims severity

The severity of claims is influenced by several things. Very important might be the accident year from at least two points of view. The first is that the business environment or technology might change over time so that some risks are higher or lower in different years. The second is the value change due to the inflation effect - generally losses occurring now would have different costs than losses that occurred several years ago. Another influential factor might be the reporting delay. Losses that had occurred several years ago and were reported early after occurrence may have different distribution than losses that occurred in the same year but are reported today.

In the model we do not take into account the business environment changes nor the technology changes as these are difficult to measure and quantify. If we observe these effects in some lines of the business we can do the whole analysis for these lines separately.

To take the inflation and delay effects into account we actually search for the distribution function

$$F_X^{t_0,d}(x) = \mathrm{P}(X \leq x, \mathrm{X \text{ occured at }} t_0, \mathrm{X \text{ reported with delay }} d)\,.$$

For the sake of simplicity we assume that this distribution is a scaled version of the kernel distribution $F_X(x)$:

$$F_X^{t_0,d}(x) = F_X\left(\frac{(1+r)^{(t-t_0)}x}{(1+s)^d}\right),$$

where t is the current date, r it the rate of claims inflation and s is the rate that captures the effect of the reporting delay.

1.2.1 Adjusting for IBNER

In data collection there might also be some claims that are not fully paid yet, therefore their value might change. This type of data uncertainty is called IBNER, which stands for incurred but not enough reserved/reported. If we find a systemic underestimation or overestimation we would like to obtain factors to adjust the claims.

For estimating the IBNER factors we might use general linear modeling such as described in Branda (2013). In such a case we assume that IBNER depends on several things which we can be included in the model, such as:

- Development year - the later the occurrence of the claim the more it might deviate from the incurred estimate.
- Size of the claim - the larger the claim the more uncertainty about its development there might be.
- Outstanding ratio - the larger the outstanding amount the more conjecture about the final value there might be as some of the fully outstanding claims might show up to be zero or on the other hand much bigger than were initially expected (for example in health insurance where some of the consequences might occur later than at the reporting date).
- Type of claim - different types of claims would probably behave differently.
- etc.

These assumptions are just theoretical and can not be justified hence only empirical evidence can show us the real influence of these factors to the claims behavior. The realization of IBNER factor for claim i between two successive years s and $s+1$ is a ratio between the incurred amounts of those two years.

To make this subsection more clear in writing we denote IBNER for claim i as a random variable Y_i and the values of corresponding observable parameters such as development year, size of claim, etc. as components of a vector $\mathbf{x_i}$. We suppose that $\mathbf{x_i}$ are vectors of observable parameters and Y_i are independent random variables with distribution from exponential family with density that can be written in form

$$f(y,\theta,\varphi) = exp\left\{\frac{y\theta - b(\theta)}{\varphi} + c(y,\varphi)\right\}\ , \tag{1.6}$$

where $y \in \mathbb{R}$, $\theta \in \mathbb{R}$ and $\varphi \in (0,\infty)$. Among distributions from the exponential family belong

- Normal, gamma, inverse Gaussian, Poisson, alternative
- Chi-square, exponential, binomial, geometric, multinomial, beta
- with known parameter: Weibull, negative-binomial, Pareto.

We suppose that IBNER has gamma distribution $Y \sim \Gamma(a,p)$ of which the density can be written in form 1.6:

$$\begin{aligned}
f(y,a,p) &= \frac{a^p}{\Gamma(p)} y^{p-1} \exp\{-ay\} \\
&= \exp\left\{-ay + \ln\left(y^{p-1}\right) + \ln\left(\frac{a^p}{\Gamma(p)}\right)\right\} \\
&= \exp\{-ay + (p-1)\ln(y) + p\ln(a) - \ln(\Gamma(p))\} \\
&= \exp\left\{\frac{-y\cdot\frac{a}{p} + \ln\left(\frac{a}{p}\right)}{\frac{a}{p}} + p\cdot\ln(p) - \ln(\Gamma(p)) + (p-1)\ln(y)\right\} \\
&= \exp\left\{\frac{y\theta + \ln(-\theta)}{\varphi} + \frac{1}{\varphi}\ln\left(\frac{1}{\varphi}\right) - \ln\left(\Gamma\left(\frac{1}{\varphi}\right)\right) + \left(\frac{1}{\varphi} - 1\right)\ln(y)\right\}
\end{aligned}$$

where $\theta = \frac{p}{a}$, $\varphi = \frac{1}{p}$ is a dispersion parameter, $b(\theta) = -\ln\left(\frac{a}{p}\right) = -\ln(-\theta)$ and $p\ln(p) - \ln\left(\Gamma(p)\right) + (p-1)\ln(y) = \frac{1}{\varphi}\ln\left(\frac{1}{\varphi}\right) - \ln\left(\Gamma\left(\frac{1}{\varphi}\right)\right) + \left(\frac{1}{\varphi} - 1\right)\ln(y) = c(y,\varphi)$.

We can also use another parameterization where $E(Y) = \mu = \frac{p}{a} = -\frac{1}{\theta}$ and $Var(Y) = \frac{p}{a^2} = \frac{1}{\nu}\mu^2$, hence $\varphi = \frac{1}{\nu}$ and the density can be written as

$$f(y,\mu,\nu) = \frac{1}{\Gamma(\nu)y}\left(\frac{\nu y}{\mu}\right)^{\nu} \exp\left\{-\frac{y\nu}{\mu}\right\}\ .$$

Variance function $V(\mu)$ shows relationship between the mean and variance of the random variable and unambiguously identifies concrete distribution from the exponential family. For $V(\mu)$ it holds that $Var(Y) = \varphi V(\mu)$. For gamma distribution $Var(Y) = \frac{1}{\nu}\mu^2$ hence the variance function is $V(\mu) = \mu^2$.

Another component of a general linear model is a linear predictor η_i which is a linear combination of observed parameters $\mathbf{x_i}$ and coefficients $\boldsymbol{\beta}_i$ that we would like to estimate:

$$\eta_i = \mathbf{x_i}^T \boldsymbol{\beta} = \sum_{j=1}^{m} x_{i,j}\beta_j \, .$$

As the last component of the model we need to determine a link function g. Link function needs to be strictly monotonous and twice differentiable and links mean of the dependent variable Y_i and a linear predictor η_i:

$$E(Y_i) = \mu_i = g^{-1}(\eta_i)$$

hence

$$g(\mu_i) = \eta_i \, .$$

Based on the nature of the data we might take natural logarithm as the appropriate link function which would lead for example to a model

$$\begin{aligned} E(IBNER_i) &= \exp\{\beta_1(\text{development year})_i + \beta_2(\text{size of claim})_i \\ &+ \beta_3(\text{outstanding ratio})_i + \beta_4(\text{type of claim})_i\} \end{aligned}$$

Having all of this we can estimate values of parameters $\boldsymbol{\beta}_i$ by maximum likelihood method and finally obtain estimations of IBNER factors

$$\widehat{IBNER}_i = \exp\left\{\mathbf{x_i}^T \hat{\boldsymbol{\beta}}\right\} \, .$$

The logarithmic likelihood function generally uses the form of density 1.6 and relationships between parameters $\eta_i = \mathbf{x_i}^T\boldsymbol{\beta}$, $\eta_i = g(\mu_i)$, $\mu_i = b'(\theta_i)$ and $V(\mu_i) = b''(\theta_i)$:

$$\begin{aligned} l(\mathbf{y}, \boldsymbol{\beta}, \varphi) &= \sum_{i=1}^{n} \ln\left(f(y_i, \theta_i, \varphi)\right) \\ &= \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\varphi} + c(y_i, \varphi) \, , \end{aligned}$$

therefore the partial derivation with respect to β_j is

$$\begin{aligned} \frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial f}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{i,j}}{g'(\mu_i)\varphi V(\mu_i)} \end{aligned}$$

as it holds:

$$\begin{aligned} \frac{\partial f}{\partial \theta_i} &= \frac{y_i - b'(\theta_i)}{\varphi} = \frac{y_i - \mu_i}{\varphi} \\ \frac{\partial \theta_i}{\partial \mu_i} &= \frac{1}{b''(\theta_i)} = \frac{1}{V(\mu_i)} \\ \frac{\partial \mu_i}{\partial \eta_i} &= \frac{1}{g'(\mu_i)} \\ \frac{\partial \eta_i}{\partial \beta_j} &= x_{i,j} \, . \end{aligned}$$

Therefore in our case if $g(\mu_i) = \ln(\mu_i)$, $V(\mu_i) = \mu_i^2$ we get

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{m} \frac{\left(y_i - \exp\left\{\mathbf{x_i}^T \boldsymbol{\beta}\right\}\right) x_{i,j}}{\varphi \exp\left\{\mathbf{x_i}^T \boldsymbol{\beta}\right\}}$$

and for finding the maximum likelihood estimates we use numerical methods.

1.2.2 Estimating kernel severity distribution

After obtaining IBNER factors for individual years we adjust outstanding claims, revalue past claim into current terms by using appropriate claims inflation rate and estimate the kernel severity distribution. The severity distribution should be non-negative valued and likely with heavy tails. We show the maximum likelihood estimates of parameters of log-normal and gamma distributions which are both two parametric hence provide one degree of freedom for scale and one for scope but are not that number of data intensive as distributions with more degrees of freedom (such as generalized Pareto distribution).

Log-normal distribution

The density of log-normal distribution is

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right\} \; for\, x \geq 0. \tag{1.7}$$

If we assume the independence of the claims, the likelihood function for claims $1, \ldots, n$ is

$$\mathrm{L}\left(\mu, \sigma, x_1, \ldots, x_n\right) = \prod_{i=1}^{n} \left(\left(\frac{1}{x_i}\right) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\ln(x_i) - \mu)^2}{2\sigma^2}\right\}\right).$$

Using logarithms on both sides we get the logarithmic likelihood function

$$\mathrm{l}\left(\mu, \sigma, x_1, \ldots, x_n\right) = \sum_{i=1}^{n} \left(-\ln(x_i) - \frac{1}{2}\ln(2\pi\sigma^2) - \frac{(\ln(x_i) - \mu)^2}{2\sigma^2}\right).$$

As we are searching for the maximum of the likelihood function with respect to parameters μ, σ we put the derivations with respect to them equal to zero:

$$\begin{aligned}
\frac{\partial l(\mu, \sigma, x_1, \ldots, x_n)}{\partial \mu} &= \sum_{i=1}^{n} \left(-\frac{2}{2\sigma^2} 2(\ln(x_i) - \mu)\right) = 0 \\
\sum_{i=1}^{n} \frac{\ln(x_i)}{\sigma^2} &= \sum_{i=1}^{n} \frac{\mu}{\sigma^2}
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial l(\mu, \sigma, x_1, \ldots, x_n)}{\partial \sigma} &= \sum_{i=1}^{n} \left(-\frac{1}{2} \cdot \frac{1}{2\pi\sigma^2} \cdot 2 \cdot 2\pi\sigma - \frac{(\ln(x_i) - \mu)^2}{2} \cdot (-2) \cdot \sigma^{-3}\right) \\
&= \sum_{i=1}^{n} \left(-\frac{1}{\sigma} + \frac{(\ln(x_i) - \mu)^2}{\sigma^3}\right) = 0 \\
\sum_{i=1}^{n} \frac{1}{\sigma} &= \sum_{i=1}^{n} \frac{(\ln(x_i) - \mu)^2}{\sigma^3}
\end{aligned}$$

hence the maximum-likelihood estimates of the parameters would be

$$\begin{aligned}\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n}\ln(x_i)\\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(\ln(x_i)-\hat{\mu}\right)^2 .\end{aligned}$$

Gamma distribution

The density of gamma distribution is

$$f(x)=\frac{a^p}{\Gamma(p)}x^{p-1}\exp\{-ax\}\ x\in(0,\infty)\,. \tag{1.8}$$

The likelihood function for n observations is

$$L(a,p,x_1,\ldots,x_n)=\prod_{i=1}^{n}\frac{a^p}{\Gamma(p)}x_i^{p-1}\exp\{-ax_i\}$$

and the logarithmic likelihood function

$$l(a,p,x_1,\ldots,x_n)=\sum_{i=1}^{n}\left(p\ln(a)-ln\left(\Gamma(p)\right)+(p-1)\ln x_i-ax_i\right)\,.$$

To obtain the estimates we set the partial derivations with respect to a and p equal to zero

$$\begin{aligned}\frac{\partial l(a,p,x_1,\ldots,x_n)}{\partial a} &= \sum_{i=1}^{n}\left(\frac{p}{a}-x_i\right)=0\\ \frac{p}{a} &= \frac{1}{n}\sum_{i=1}^{n}x_i && (1.9)\\ \frac{\partial l(a,p,x_1,\ldots,x_n)}{\partial p} &= \sum_{i=1}^{n}\left(\ln(a)-\frac{\Gamma\prime(p)}{\Gamma(p)}+\ln(x_i)\right)=0 && (1.10)\end{aligned}$$

and from 1.9 and 1.10 obtain estimates of a and p numerically.

To choose which distribution fits better the data we can do a graphical analysis using histograms or Q-Q plots or we might run some of the goodness-of-fit tests, for example the Kolmogorov-Smironov or Pearson Chi Square tests which can be found for example in Anděl (1998).

1.3 Aggregate model for outstanding liabilities

The method is generally as follows. Firstly we decide which model we think would be the most appropriate to use according to data (for example homogeneous Poisson process for claim count, exponential distribution for reporting delay, log-normal distribution for claims severity, ...). Based on this we estimate parameters of such a model. Then we make simulations and obtain statistical results from which probably the most important for reserves estimation might be mean and empirical distribution function.

We show methods separately for IBNR reserves, RBNS reserves, UPR and then method for overall outstanding liability estimation.

1.3.1 IBNR

IBNR stands for incurred but not reported claims which means that the occurrence date was prior to the present moment t and the reporting will come after t.

Inputs for the simulations are:

- Parameters of the claims counting process (μ for homogeneous process or μ_t for non-homogeneous,...)
- Kernel severity distribution $F_X(x)$ and rate of claims inflation r and reporting delay effect rate s if taken into account.

The course of the algorithm is:

1. Set number of scenarios to be n.
2. For each scenario $j = 1, \ldots, n$ simulate claims occurrence from the chosen claim count process. Therefore we get claim occurrence for each simulated claim and the total number of claims occurred in j-th scenario is n_j.
   (a) For each claim $i = 1, \ldots, n_j$ simulate reporting delay from the reporting delay distribution. The number of claims that were reported before the present moment t is
   $$\tilde{n}_j = \sum_{i=1}^{n_j} \mathbf{I}[(\text{occurence} \quad \text{time})_i \leq t \,\&\, (\text{reporting} \quad \text{time})_i > t] \, .$$
   We can rearrange the order of the claims and from now on we suppose that the first $\tilde{n}_j$ claims are those claims that incurred upon time t but have not been reported yet.
   (b) For each claim $i = 1, \ldots, \tilde{n}_j$ simulate the claim severity $x_i^{(j)}$ from the kernel severity distribution and then adjust the severity $\tilde{x}_i^{(j)}$. If we take claims inflation rate r into account it would be
   $$\tilde{x}_i^{(j)} = (1+r)^{(t_i^{(j)}-t)} x_i^{(j)} \, ,$$
   where $t_i^{(j)}$ is the occurrence time of this claim.
3. For each scenario $j = 1, \ldots, n$ we count the total loss $S_j = \sum_{i=1}^{\tilde{n}_j} \tilde{x}_i^{(j)}$.
4. We put total losses in ascending order $S_{(1)} \leq S_{(2)} \leq \ldots S_{(n)}$ and obtain the empirical distribution function
   $$\hat{F}_S(s) = \frac{max \left\{ j \mid S_{(j)} \leq s \right\}}{n}$$
   and their empirical mean $\bar{S} = \frac{1}{n} \sum_{j=1}^{n} S_j$ and empirical variance $\hat{\sigma}^2(S) = \frac{1}{n-1} \sum_{i=1}^{n} \left(S_j - \bar{S} \right)^2$.

1.3.2 RBNS

RBNS stands for reported but not settled claims. Those are the claims that occurred and were reported before the present moment t but have not been settled yet. Crucial is the IBNER model which we described in Section 1.2.1. Once more IBNER reserve is a reserve for incurred but not enough reported/reserved. It it that part of a reserve that should be put aside to deal with the adverse development of RBNS claims.

The main idea is that the ultimate RBNS claims liability is a sum of all outstanding claims adjusted for IBNER. To do this we need to estimate the setllement year for each outstanding claim.

The input for this algorithm is:

- All open losses, number of those losses is m
- Setllement delay distribution
- IBNER model.

The course of the algorithm is:

1. Set number of scenarios to be n.
2. For each scenario $j = 1, \dots, n$ do:
   (a) For each open loss x_i, $i = 1, \dots, m$ which is in development year d_i:
      i. Sample a value of settlement year $s_{(j,i)}$.
      ii. Sample IBNER factors between years $d_i, \dots, s_{(j,i)}$:
      $$\mathrm{IBNER}_{d_i,d_i+1}, \mathrm{IBNER}_{d_i+j,d_i+2}, \dots, \mathrm{IBNER}_{s_{(j,i)}-1,s_{(j,i)}} . \tag{1.11}$$
      iii. Calculate the cumulative IBNER factor for loss x_i
      $$\mathrm{IBNER}_i^j = \mathrm{IBNER}_{d_i,d_i+1} \cdot \mathrm{IBNER}_{d_i+j,d_i+2} \cdot \dots \cdot \mathrm{IBNER}_{s_{(j,i)}-1,s_{(j,i)}} .$$
      iv. Calculate the claim value adjusted for IBNER $x_i^j = \mathrm{IBNER}_i^j \cdot x_i$
   (b) Calculate the ultimate RBNS claims for simulation j:
   $$S^j = \sum_{i=1}^{m} x_i^j$$
3. Sort all the values S^j, $j = 1, \dots, n$ in ascending order $S_{(1)} \leq S_{(2)} \leq \dots \leq S_{(n)}$ and obtain the empirical distribution function
$$\hat{G}_S(x) = \frac{max\left\{j \mid S_{(j)} \leq x\right\}}{n} .$$
and their empirical mean $\bar{S} = \frac{1}{n} \sum_{j=1}^{n} S_j$ and empirical variance $\hat{\sigma}^2(S) = \frac{1}{n-1} \sum_{i=1}^{n} \left(S_j - \bar{S}\right)^2$.

1.3.3 UPR

UPR stands for unearned premium reserve which is a reserve for claims that have not occurred yet but are covered by the policies currently in force. We proceed almost the same way as when estimating IBNR. The kernel severity distribution would be the same and the frequency would be driven by the same counting process.

Therefore the inputs are the same as in Section 1.3.1 and the course of the algorithm is:

1. Set number of scenarios to be n.
2. For each scenario $j = 1, \ldots, n$ simulate a claims occurrence from the chosen claim count process. Therefore we get a claim occurrence for each simulated claim and the total number of claims occurred in j-th scenario is n_j.
   (a) For each claim $i = 1, \ldots, n_j$ simulate reporting delay from the reporting delay distribution. The number of claims that occurred after the present moment t is $\tilde{n}_j = \sum_{i=1}^{n_j} \mathbf{I}[occurrence\, time_i \geq t]$. We can rearrange the order of the claims and from now on we suppose that the first $\tilde{n}_j$ claims are those claims that occurred after time t. Here we see, that this is the only point where UPR and IBNR modeling differs and it is clear that a claim that occurred after t could not be reported before t therefore we can model UPR and IBNR together and the condition would be $[(\text{reporting} \quad \text{time})_i \geq t]$.
   (b) For each claim $i = 1, \ldots, \tilde{n}_j$ simulate the claim severity $x_i^{(j)}$ from the kernel severity distribution and then adjust the severity $\tilde{x}_i^{(j)}$. If we take inflation rate r into account it would be $\tilde{x}_i^{(j)} = (1+r)^{(t_i^{(j)}-t)} x_i^{(j)}$, where $t_i^{(j)}$ is the occurrence time of claim $x_i^{(j)}$.
3. For each scenario $j = 1, \ldots, n$ we count the total loss $S_j = \sum_{i=1}^{\tilde{n}_j} \tilde{x}_i^{(j)}$.
4. We put total losses in ascending order $S_{(1)} \leq S_{(2)} \leq \ldots S_{(n)}$ and obtain the empirical distribution function

$$\hat{F}_S(s) = \frac{max\left\{j \mid S_{(j)} \leq s\right\}}{n}$$

   and their empirical mean $\bar{S} = \frac{1}{n}\sum_{j=1}^{n} S_j$ and empirical variance $\hat{\sigma}^2(S) = \frac{1}{n-1}\sum_{i=1}^{n}\left(S_j - \bar{S}\right)^2$.

1.3.4 Overall outstanding liability

The overall outstanding liability for a considered period consists of claims that incurred before the present moment t and have not been settled (that also includes claims that have not been reported yet) and claims that will occur in the future but are covered by the policies currently in force. From this we need to subtract the amount that we have already paid for RBNS claims. Therefore we can write:

$$\text{Outstanding liabilities} = \text{IBNR} + (\text{RBNS} - \text{paid amount}) + \text{UPR}\,.$$

Paid amount is a deterministic value which will be subtracted in the end from the estimated reserve for outstanding liabilities. In Sections 1.3.1, 1.3.2, 1.3.3 we showed how to find the empirical distributions of IBNR, RBNS and UPR. If we assume these three random variables to be independent then the distribution of the outstanding liabilities would be a convolution of those three distributions. To obtain an estimate of the distribution function we do not actually need to derive the convolution as for three variables it might be quite complicated. Instead we can again use simulations similar way as above.

The input is:

- IBNR distribution
- RBNS distribution
- UPR distribution .

The algorithm proceeds:

1. Set number of simulations to be n.
2. For each scenario $j = 1, \dots, n$:
   (a) Sample IBNR amount from IBNR distribution.
   (b) Sample RBNS amount from RBNS distribution.
   (c) Sample UPR amount from UPR distribution.
   (d) Count the total liability for scenario j: $S_j = IBNR + RBNS + UPR$.
3. Sort all S_j, $j = 1, \dots, n$ in ascending order $S_{(1)} \leq S_{(2)} \leq \dots \leq S_{(n)}$ and obtain the empirical distribution function of overall outstanding liabilities:

$$\hat{F}_S(x) = \frac{max\left\{j \mid S_{(j) \leq x}\right\}}{n}$$

   and the empirical mean $\bar{S} = \frac{1}{n}\sum_{j=1}^{n} S_j$ and empirical variance $\hat{\sigma}^2(S) = \frac{1}{n-1}\sum_{i=1}^{n}\left(S_j - \bar{S}\right)^2$.

When estimating the reserve for overall outstanding claims we shall not forget to subtract the already paid amount after this algorithm proceeds.

Chapter 2

Chain ladder

In this chapter we describe a more traditional approach to claims estimation using Mack's distribution-free Chain ladder model as described for example in Wüthrich and Merz (2008). The method is based on aggregating data on the claims development into development triangles.

A development triangle is a specific organization of data on claims development. Reported claims are organized according to occurrence year and development year. Triangles might be incremental or cumulative where each item of the cumulative triangle contains also a sum of the previous values on the same row in the corresponding incremental triangle.

Table 2.1 shows a cumulative triangle with values known until the end of the year t.

Value $C_{j,s}$, where $j + s \leq t$ is sum of loss payments from years $j, \ldots, j+s$ related to claims that occurred in year j.

Our aim is to estimate values $C_{j,s}$, $j + s > t$ and therefore extend the triangle to a rectangle respectively a square as we suppose that the claim development is terminated after t years.

Occurence	Development year						
year	0	1	$\ldots$	s	$\ldots$	t-2	t-1
1	$C_{1,0}$	$C_{1,1}$	$\ldots$	$C_{1,s}$	$\ldots$	$C_{1,t-2}$	$C_{1,t-1}$
2	$C_{2,0}$	$C_{2,1}$	$\ldots$	$C_{2,s}$	$\ldots$	$C_{2,t-2}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		
j	$C_{j,0}$	$C_{j,1}$	$\ldots$				
$\vdots$	$\vdots$	$\vdots$					
t-1	$C_{t-1,0}$	$C_{t-1,1}$					
t	$C_{t,0}$						

Table 2.1: Cumulative development triangle

The assumptions of the model are:

1. $E\left(C_{j,s+1} \mid C_{j,0}, \ldots, C_{j,s}\right) = f_s \cdot C_{j,s}$ for $j = 1, \ldots, t$,
   where f_s is the development factor we would like to estimate by Chain

Ladder Method.

2. The cumulative claims with different occurrence year are mutually independent: $\{C_{j,0}, \ldots, C_{j,t-1}\} \perp \{C_{k,0}, \ldots, C_{k,t-1}\}, \forall j \neq k$.

3. $\text{Var}\left(C_{j,s+1} \mid C_{j,0}, \ldots, C_{j,s}\right) = \sigma_s^2 \cdot C_{j,s}\, j = 1, 2, \ldots, t, \;\; s = 0, \ldots, t-1$
   $\text{Var}\left(\frac{C_{j,s+1}}{C_{j,s}} \mid C_{j,0}, \ldots, C_{j,s}\right) = \frac{\sigma_s^2}{C_{j,s}}$.

The Chain-Ladder estimation of the development factor is

$$\hat{f}_s = \frac{\sum_{j=1}^{t-s-1} C_{j,s+1}}{\sum_{j=1}^{t-s-1} C_{j,s}} \qquad\qquad s = 0, \ldots, t-2\,.$$

Under assumptions 1 and 2 it is true that

$$E\left(C_{j,t-1} \mid \triangle\right) = C_{j,t-j} \cdot f_{t-j} \cdot \ldots \cdot f_{t-2}$$

and

$$E\left(\hat{f}_{t-j} \cdot \ldots \cdot \hat{f}_{t-2}\right) = f_{t-j} \cdot \ldots \cdot f_{t-2} \qquad\qquad \text{for}\, j = 3, \ldots, t\,,$$

where $\triangle = \{C_{j,s},\; j = 1, \ldots, t,\; s = 0, \ldots, t-j\}$ is the known development triangle

The estimate of all claims that occurred in year j is $\hat{C}_{j,\infty}$ and if we assume termination of the development after t years it is

$$\hat{C}_{j,\infty} = \hat{C}_{j,t-1} = C_{j,t-j} \cdot \hat{f}_{t-j} \cdot \hat{f}_{t-j+1} \cdot \ldots \cdot \hat{f}_{t-2}\,.$$

Claims that occurred in year j but are still outstanding are $R_j = X_{j,\infty} - X_{j,t-j}$ and their estimate is

$$\hat{R}_j = \hat{C}_{j,\infty} - C_{j,t-j} = C_{j,t-j} \cdot \left(\hat{f}_{t-j} \cdot \ldots \cdot \hat{f}_{t-2} - 1\right)\,.$$

The consequence of the relationships above is the unbiasedness of the estimates

$$\begin{aligned} E\left(\hat{C}_{j,\infty} \mid \triangle\right) &= C_{j,t-j} \cdot f_{t-j} \cdot \ldots \cdot f_{t-2} = E\left(C_{j,\infty} \mid \triangle\right) \\ E\left(\hat{R}_j \mid \triangle\right) &= E\left(\hat{C}_{j,\inf} - C_{j,t-j} \mid \triangle\right) = C_{j,t-j}(f_{t-j} \cdot \ldots \cdot f_{t-2} - 1) = E\left(R_j \mid \triangle\right)\,. \end{aligned}$$

The unbiased estimation of variance of the development factor $\hat{f}_s$ is

$$\frac{1}{t-s-1} \sum_{j=1}^{t-s-1} C_{j,s} \left(\frac{C_{j,s+1}}{C_{j,s}} - \hat{f}_s\right)^2\,.$$

Compared to the method described in the second chapter this triangle approach is more straightforward and easy to use, but during the process of aggregating data into the development triangles a lot of information remains unused.

Chapter 3

Practical application

In this chapter we demonstrate the triangle free approach on the real data which we have described in the introduction. Firstly, we have a look at the data and exclude incorrect values and claims connected with annuities. Having this final set we show some basic characteristics. Then we apply the method as described in Chapter 1. We build a model describing the claims development and using this we estimate an IBNR reserve.

3.1 Data

We have already described the data available in the introduction. As the data is real it shows some imperfections, therefore we need to exclude some values.

First of all, we exclude all claims associated with annuities. In our data this is only the case for some health claims, we are not going to exclude any material claims due to this criterion.

There are also some mistakes in the data. We excluded 7 claims out of material payments, 13 claims out of material reserves and 2 claims out of health reserves due to the reporting day being before the occurrence day which is inconsistent with principles of insurance that the insured events must be random.

From now on we take this adjusted data set as the initial one. As we do not know how the reserves were determined, we are going to use the database of reserve changes only for purpose of determining whether a claim in the payments database is settled or not. We consider claims with the total reserve being zero to be settled. As the data is real there might be some inaccuracies, for example sometimes the total reserve for a claim remains outstanding just a little bit (e.g. 5 CZK) or negative (e.g. -5 CZK) so we take all the claims for which the total reserve is less than 10 CZK as settled (the smallest total reserve bigger than 10 CZK is 118 CZK). Table 3.1 summarizes some basic characteristics of the data from payments database.

It can be seen that material claims and health claims behave differently. Health claims usually have a much longer reporting delay. That is because of the nature of these claims, as some of the health consequences of the accident might show up later such as head injures or internal bleeding. Material damages can usually be quantified more quickly. This is why we model health and material claims separately.

	material claims	health claims	all claims
number of records	39 595	4 021	43 616
number of unique claims	35 789	2 974	38 763
number of settled claims	34 751	2 578	37 329
percentage of settled claims	98.1 %	86.7 %	96.3%
observed average reporting delay	75.8	201.9	85.5
average settlement delay for settled claims	232.5	484.8	249.9

Table 3.1: Basic characteristics of data

3.2 Claims count

Our aim is to create a model for claims occurrence. We are dealing with car insurance. The number of car accidents is not subordinate to natural disasters of things that change during time. Assuming homogeneity over time therefore seems reasonable. Hence we assume that claims occurrence is driven by a homogeneous Poisson process.

In Section 1.1.1 we described methods for estimating parameters of a homogeneous Poisson process using a time window. For this estimation we use only data from the file containing payments as we do not have information on how reserves were determined and some of the reserved claims might end up being zero therefore we only want to take claims associated with some payments.

The estimate is based on the property of homogeneous Poisson process with intensity μ that the time between two successive events follows exponential distribution with mean $\frac{1}{\mu}$.

For a fixed time window w we estimate the mean time between two successive events as a sample mean of times between occurrences of claims that happened in the time window w and we assume that all of them were reported in the observed period.

3.2.1 Material claims

We have already discussed the issue of choosing the length of a time window. Maximum reporting delay in our data-set is 1512 days which is approximately four years, therefore any time window ending before the end of the year 2008 should not yield much biased estimates. The mean of reporting delay is 76 days therefore using even more recent time window should not pose a problem. Table 3.2 shows number of material claims in different years.

As can be seen, in year 2000, 2001 and 2002 the number of claims is quite low and since than it increases and is quite similar up to recent years where lower number is a result of reporting delay. Our data is not a full database but a random sample therefore there might be some distortions. Choosing the estimate is subjective to the actuary what he or she considers the most suitable based also on his or her experience. We find it reasonable to use time window 1.1.2003-31.12.2008 for which the estimate is $\hat{\mu} = 0.1058$.

year	number of claims	year	number of claims
2000	956	2007	3 499
2001	1 501	2008	3 696
2002	2 624	2009	2 895
2003	3 010	2010	2 788
2004	3 289	2011	2 438
2005	3 723	2012	1 764
2006	3 496	2013	110

Table 3.2: Number of material claims in different years

3.2.2 Health claims

For health claims we proceed the same way. Maximum reporting delay is 1 703 days which is less than five years therefore using time windows shorter ending before the end of the year 2007 should not yield biased estimates. However, the mean reporting delay is 202 days so even more recent time windows should not produce much biased estimates.

Table 3.3 shows number of health claims in different years. Again, we observe a low number of claims in the first two years so we exclude them. The low number of claims in recent years is again caused by the reporting delay. We can see that for the years 2002 to 2008 the number of claims is quite similar, therefore we use this time window for which the estimate is $\hat{\mu} = 1.25$. Note that the frequency of health claims occurrence is much lower compared to the frequency of material claims.

year	number of claims	year	number of claims
2000	128	2007	289
2001	162	2008	268
2002	273	2009	198
2003	314	2010	180
2004	323	2011	167
2005	285	2012	86
2006	300	2013	1

Table 3.3: Number of health claims in different years

3.3 Reporting delay

Again, we estimate reporting delay only from the file containing payments for the same reason, which is that we have no information on how the reserves are determined. Even though there are less records in the payment file it is still a representative sample and we shall obtain reliable results.

We assume the reporting delay to be exponential with the parameter λ and density $f(x) = \frac{1}{\lambda}\exp\left\{-\frac{x}{\lambda}\right\}$ and from the data we want to obtain the estimation of

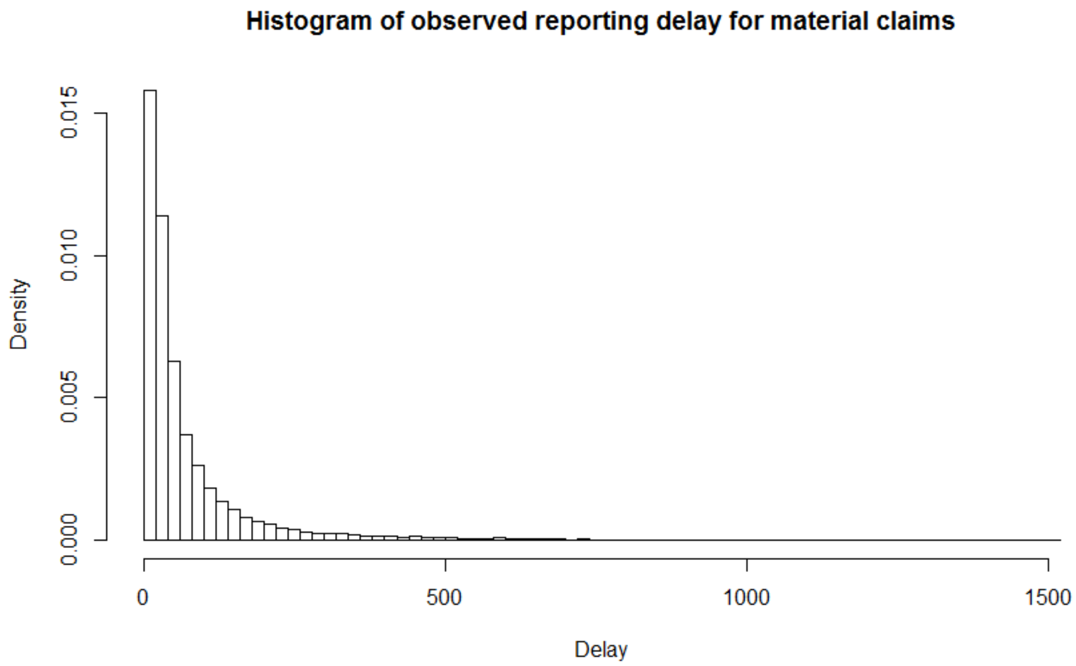

Figure 3.1: Histogram of the reporting delay for material claims

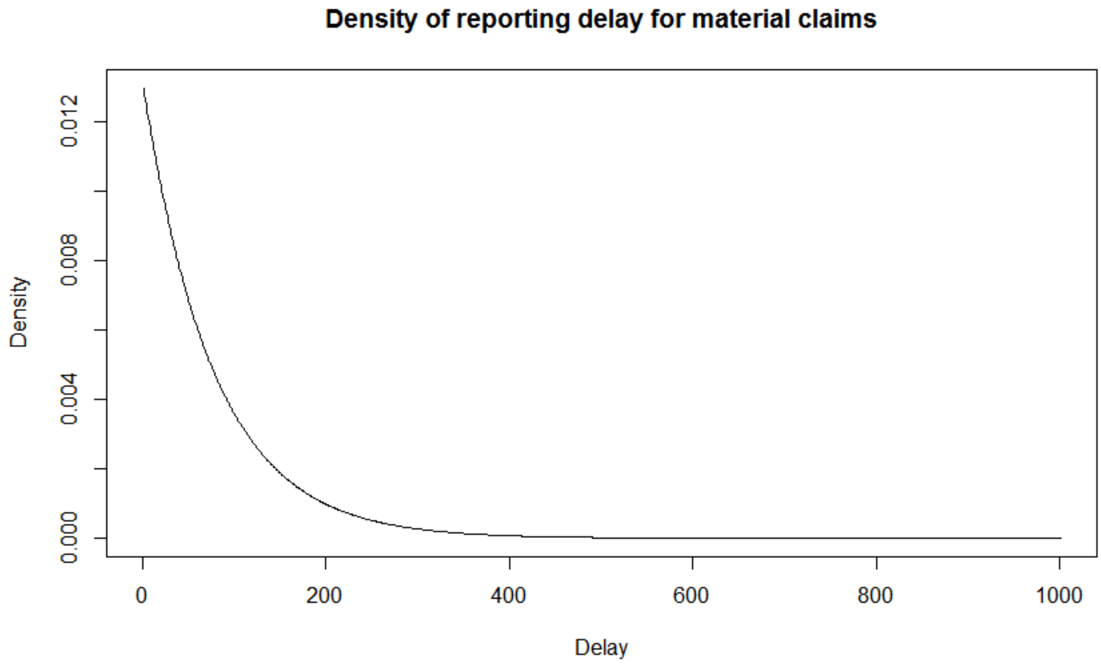

Figure 3.2: Density of reporting delay for material claims

the parameter of this exponential distribution $\hat{\lambda}$. In Section 1.1.3 we have already discussed the issue that we cannot observe all the values due to the limited time window. Ignoring this fact would lead to bias towards shorter delays. Therefore we are going to use Equation 1.3 to get an estimate of $\hat{\lambda}$.

Figure 3.1 shows a histogram of the observed reporting delay for material

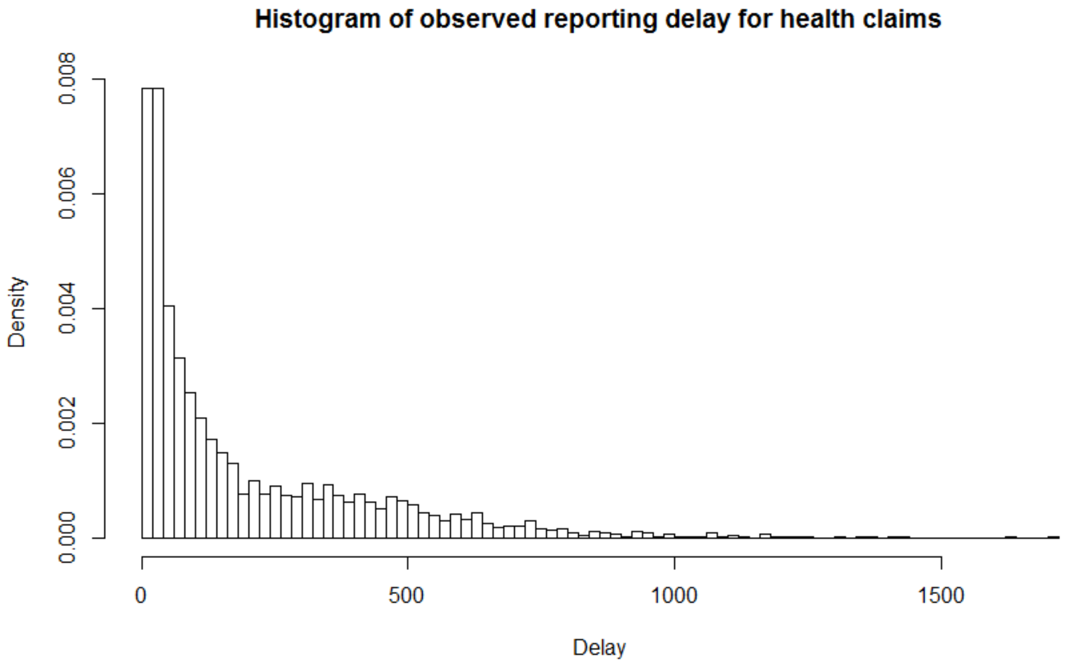

Figure 3.3: Histogram of reporting delay for material claims

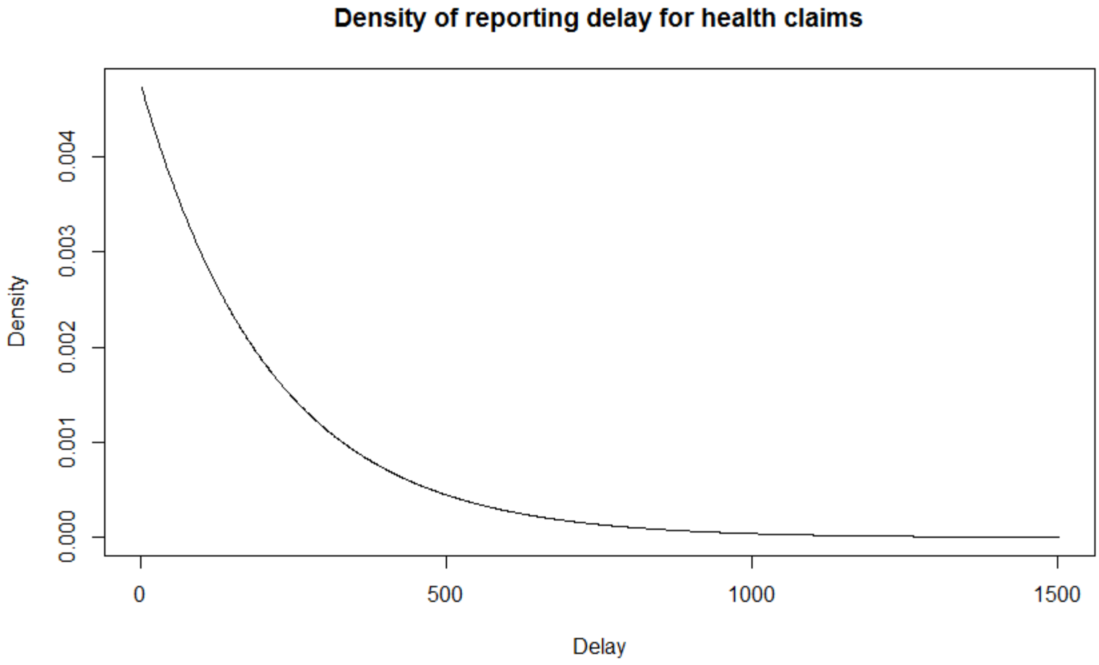

Figure 3.4: Density of reporting delay for health claims

claims and Figure 3.3 for health claims. As can be seen, they do not seem to be inconsistent with exponential distribution. The observed mean of reporting delay for material claims is 75.78261 and for health claims it is 201.9284. However, these estimates are biased towards shorter delay therefore we use them as $\hat{\lambda}_{obs}$ in Equation 1.3 and obtain estimates of λ numerically. For material claims we get

$\hat{\lambda} = 77.0174$ and for health claims we get $\hat{\lambda} = 211.552$. They are both greater then the initial estimates λ_{obs} which complies with the theory.

Figures 3.2 and 3.4 show the density of exponential distributions with the estimated parameters for material and health claims.

Henceforward we will assume the distribution of reporting delay to be exponential with parameter $\hat{\lambda} = 77.0174$ for material claims and $\hat{\lambda} = 211.552$ for health claims.

3.4 Claims severity

For estimating severity distribution we the information on settled claims only. We do not use payments for unsettled claims as that might cause bias towards smaller amounts because we do not know the final total severity. Another approach might be to use the database of claims reserves and adjust that for IBNER but as we have already discussed, we do not know how the reserves were determined and therefore we are not using this approach and stay just with those settled ones. From all material claims 98,1% are settled and for health claims it is 86,7% hence we shall still have enough data to obtain reliable results.

The average number of payments for settled claims is 1,007 for material claims and 1,3281 for health claims. As a claim severity we are going to take the sum of all payments associated with the claim. Mostly it is going to be just one payment especially for material claims.

3.4.1 Claims inflation

Claims inflation is an important fact that shall not be forgotten to be taken into account. Claims which occurred several years ago might have lower costs than claims which occurred recently. In case of a car insurance the price of repairs and spare parts, as well as the price of medical care and wages to be compensated, are changing and usually increasing.

As we do not have the information on the claims inflation we are going to use the historical average annual inflation rate reported by the Czech statistical office as an approximation. This inflation rate is the increment of average annual consumer price index and represents a percentage change in the average price level in the last 12 months compared to the previous 12 months. Table 3.4 shows the inflation rate in percentage for each year from 2012 to 2013.

Year	00	01	02	03	04	05	06	07	08	09	10	11	12	13
Rate	3.9	4.7	1.8	0.1	2.8	1.9	2.5	2.8	6.3	1.0	1.5	1.9	3.3	1.4

Table 3.4: Historical annual inflation rate reported by the Czech statistical office

To apply the inflation effect we use the principles of compound interest, as described in Cipra (2006) using the $\frac{\text{ACT}}{365}$ convention. We revalue each claim in the following way. If claim severity is x occurred in time t_0, let y_0 denote the year of t_0 and i_k the inflation rate corresponding to the year k. Revaluing to current time t_1 which is in the year y_1 is

$$\tilde{x} = x \cdot (1 + i_{y_0})^{(\text{end of year}\, y_0 - t)} \, (1 + i_{y_0+1}) \cdot \ldots \cdot (1 + i_{y_1})^{\frac{(t_1 - \text{beginning of year } y_1)}{365}} \tag{3.1}$$

for claims that occurred before the current year y_1 and

$$\tilde{x} = x \cdot (1 + i_{y_1})^{\frac{(t_1 - t_0)}{365}} \ , \tag{3.2}$$

for claims that occurred during the current year y_1. In our data the current date t_1 is 21.5.2013 which is the day of the last payment.

3.4.2 Material claims

There are 34 751 settled material claims observations. The maximal amount of settled material claims severity is 4 261 339 CZK, the minimal is 18 CZK and the average material claims severity is 42 866.78 CZK. After adjusting for the inflation effect the minimum is 19.8 CZK, maximum 4 951 945 CZK and the average claims adjusted severity is 50 712.8 CZK. Even from this overview we can see that the inflation effect is significant.

Using maximum likelihood method implemented in the statistical software R we estimate the parameters of log-normal and gamma distributions with densities expressed by Equations 1.7 and 1.8. For log-normal distribution the estimated parameters are $\hat{\mu} = 10.3013$ and $\hat{\sigma} = 1.0101$ and for gamma distribution the estimated parameters are $\hat{a} = 0.01$ and $\hat{p} = 298.2351$. Figure 3.5 shows a histogram of the adjusted material claims severity together with the estimated densities of log-normal and gamma distribution. Figure 3.6 shows Quantile-Quantile plots for the estimated log-normal and gamma distributions, scale of the axes is logarithmic for better visualization. It is demonstrated that the log-normal distribution seems to fit the data better, therefore we use the log-normal distribution with parameters $\hat{\mu} = 10.3013$ and $\hat{\sigma} = 1.0101$ to model material claims kernel severity.

3.4.3 Health claims

For settled health claims there are 2 578 observations. The maximal amount is 2 318 863 CZK, the minimal 85 CZK and the average 46 664.61 CZK. After adjusting for the inflation effect the maximal amount is 2 967 376 CZK, the minimal 101.7 CZK and the average claim severity 56 336.3 CZK.

We proceed the same way as with material claims and obtain parameter estimates of log-normal distribution $\hat{\mu} = 9.5470$ and $\hat{\sigma} = 1.5978$ and gamma distribution $\hat{a} = 0.01$ and $\hat{p} = 1.4052$. According to the histogram with estimated densities shown in Figure 3.7 and Quantile-Quantile plots with logarithmic scale of axes shown in Figure 3.8, it seems that log-normal distribution fits the data better, therefore we use log-normal distribution with parameters $\hat{\mu} = 9.5470$ and $\hat{\sigma} = 1.5978$ to model material claims kernel severity.

3.5 Overall model

Based on our data we have developed a model of the whole insurance events process except for the reserving development and settlement process as we do not have enough information on these.

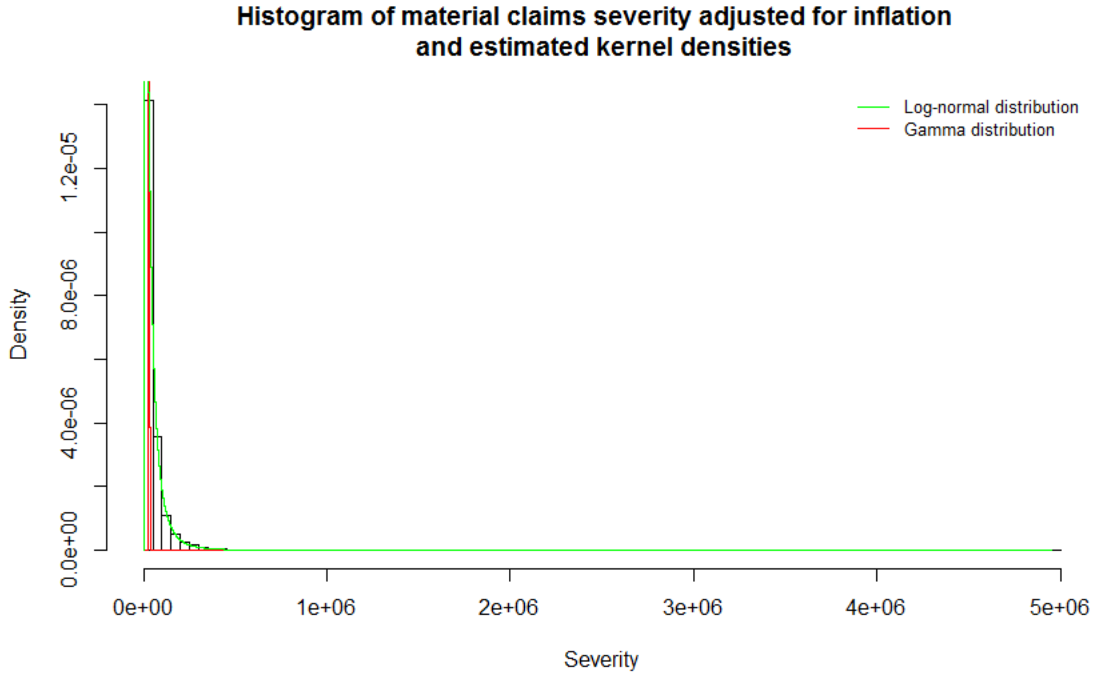

Figure 3.5: Histogram of material claims severity adjusted for the inflation effect and estimated densities

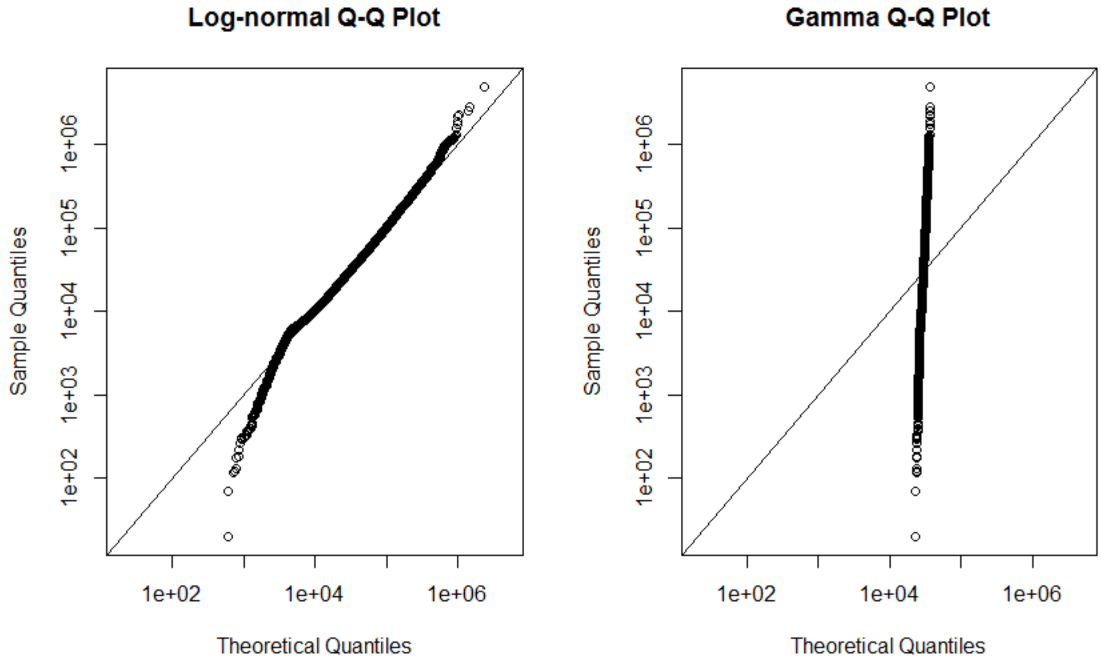

Figure 3.6: Quantile-Quantile plots of material claims severity

To sum the models up, claim count is driven by homogeneous Poisson process with intensity $\hat{\mu} = 0.105827$ for material claims and $\hat{\mu} = 1.25$ for health claims. Reporting delay follows exponential distribution with mean $\hat{\lambda} = 77.0174$ for material claims and $\hat{\lambda} = 211.552$ for health claims. Severity follows log-normal

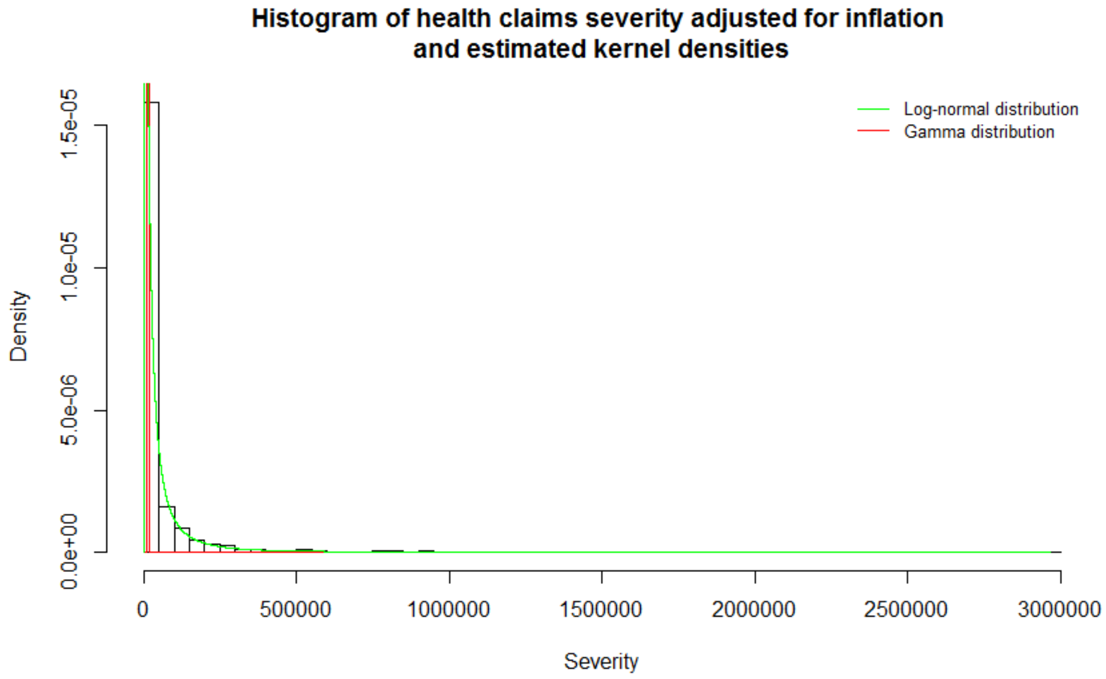

Figure 3.7: Histogram of health claims severity adjusted for inflation and estimated densities

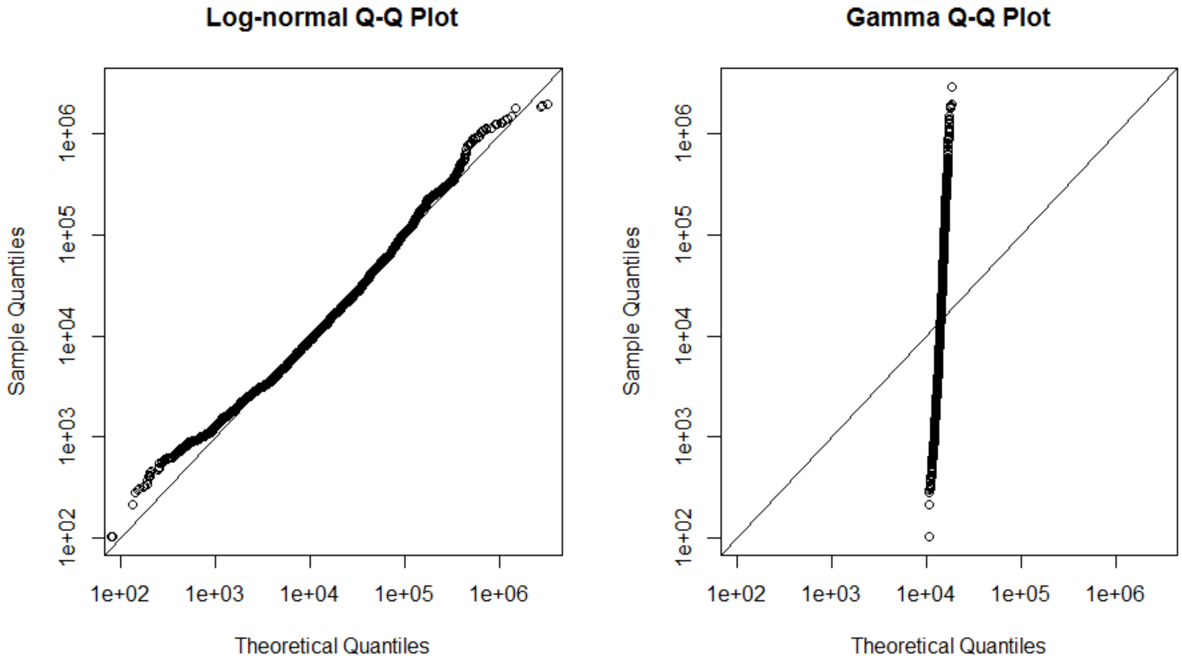

Figure 3.8: Histogram of health claims severity adjusted for inflation and estimated densities

distribution with parameters $\hat{\mu} = 10.3013$ and $\hat{\sigma} = 1.0101$ for material claims and $\hat{\mu} = 9.5470$ and $\hat{\sigma} = 1.5978$ for health claims.

We use these models for producing simulations and predicting the future development.

3.6 Estimating IBNR reserve

One application of the model might be estimating a reserve for incurred but not reported claims. We have described this method in Section 1.3.1. Basically we use the model which we developed just above and make 1000 simulations of the claims development in software R. The source code for both material and health claims can be found in the Appendix to this thesis.

The observed period is from the 1.1.2000 to 21.5.2013 which is 4 889 days both for material and health claims. Once more we shall emphasize that the database we are working with is not complete but just a selection, therefore our results are not those actual reserves that the Czech Insurer's Bureau shall hold.

According to this method we obtained estimate of the incurred but not reported claims for material claims 35 581 077 CZK health claims 8 223 996 CZK. The empirical cumulative distribution function of IBNR material claims is shown in Figure 3.9 and for health claims in Figure 3.10 and quantiles are summarized in Table 3.5.

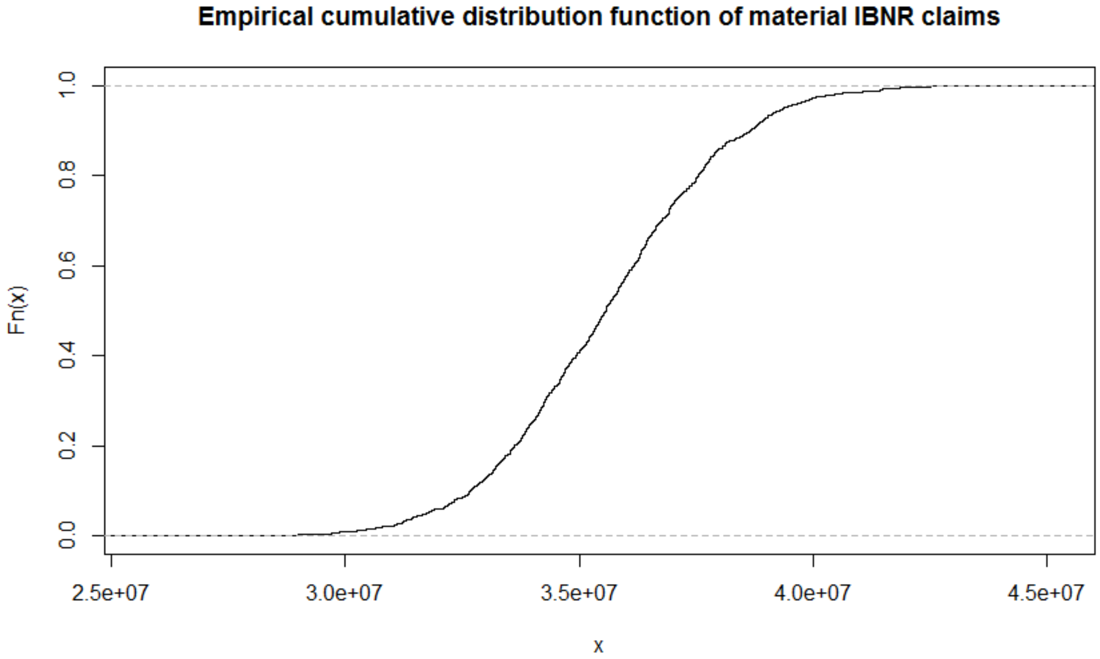

Figure 3.9: Empirical cumulative distribution function of IBNR material claims

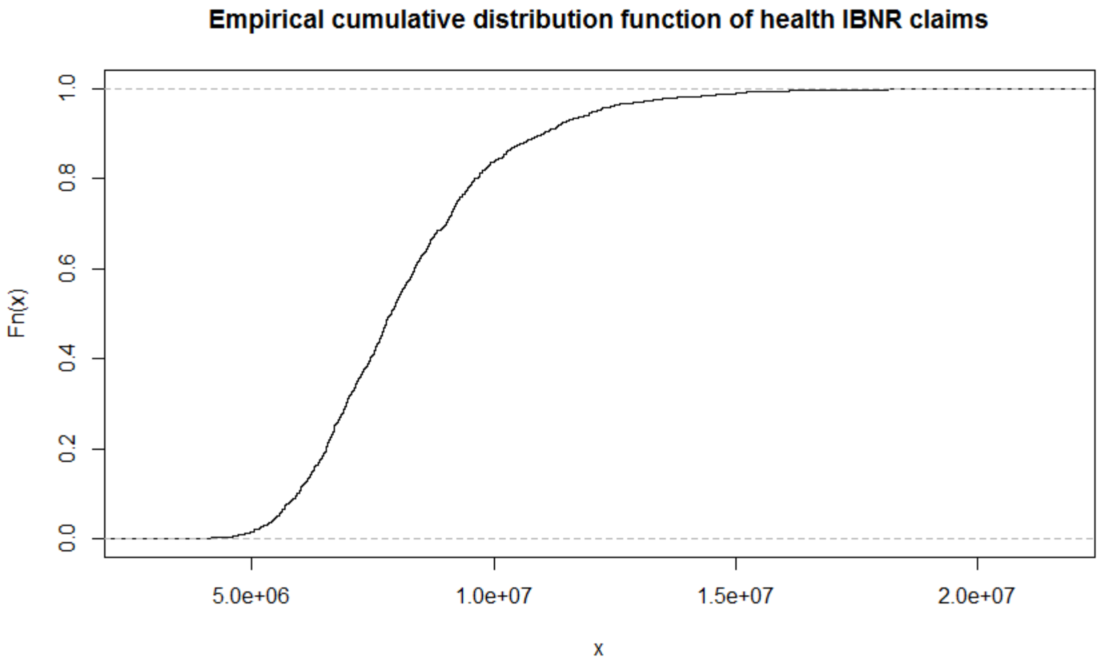

Figure 3.10: Empirical cumulative distribution function of IBNR health claims

quantile	material claims	health claims
0%	27 000 949	4 025 776
25%	33 993 155	6 709 962
50%	35 575 703	7 871 670
75%	37 118 707	9 258 030
100%	43 893 812	20 355 547

Table 3.5: Quantiles of the estimated IBNR

Chapter 4

Comparison of triangle-free approach and Chain ladder

In Chapter 3 we have showed how to use the triangle free approach as described in Chapter 1 on real data. Compared to the Chain ladder as described in Chapter 2 it is more complicated and computational demanding. In this chapter we aim to show that this complexity is in favor of more preciseness.

4.1 Real data

We will retain the same data which is described in the introduction. However, we use only data known as at the end of the year 2006 for estimating the IBNR reserve using method from Chapter 1 and Chain ladder.

Table 4.1 shows the sums of payments for claims that occurred in years 2000-2006. Out of these claims there are only 5 claims which are not settled until the spring of 2013, therefore we might assume that the development of claims is less than six years and we can take the sum of payments from the years 2007-2013 for claims that occurred before the end of the year 2006 and were reported after the year 2006 as the value of IBNR claims as at the end of 2006. Hence for material claims the IBNR reserve should be 30 936 696 CZK and for health claims 8 042 851 CZK.

	Material claims	Health claims
total paid until end of 2006	659 117 739	55 706 120
reported before end of 2006 paid after 2006	111 013 377	13 637 437
reported after 2006	30 936 696	8 042 851
total	**801 067 812**	**77 386 408**

Table 4.1: Overview of payments for claims that occurred in years 2000-2006

4.1.1 Triangle free approach

We proceed by replicating the methods as in Chapter 3 just using he database as at the end of the year 2006, which means that we use the database of claims payments from 1.1.2000 to the end of the year 2006. The last payment was

year	2000	2001	2002	2003	2004	2005	2006
number of claims	956	1 499	2 619	2 991	3 232	3 316	1 222

Table 4.2: Number of material claims in different years

year	2000	2001	2002	2003	2004	2005	2006
number of claims	128	162	271	301	268	130	37

Table 4.3: Number of health claims in different years

29.12.2006 both for material and health claims therefore the observed period is 2 554 days.

Claim count

We assume that claims occurrence is driven by a homogeneous Poisson process. To estimate the intensity we again use the trick with the time window. Table 4.2 shows the number of material claims in different years. The number of claims in the year 2000 is much lower then in other years therefore we exclude this year. The low number of claims in the year 2006 is caused by reporting delay. However, the number of claims in the other years is quite similar, therefore we are going to use this time window 1.1.2001-31.12.2005 and obtain estimate $\hat{\mu} = 0.1337$ for material claims.

In the full database there are 18 599 material claims that occurred in the years 2000 to 2006. We estimated that the claims occurrence is driven by the homogeneous Poisson process with intensity $\hat{\mu} = 0.1337$. We run 1000 simulation of the claim count process and from these simulations obtain an estimate of number of claims in the considered period 19 044.96 which is quite similar to the number from the database. At least for this data set the triangle free approach seems to be quite precise in estimating the material claims count in our database.

Table 4.3 shows the number of health claims in different years. Again, we exclude year 2000 and from then on assume homogeneity. The low number of claims in 2005 and 2006 is caused by a reporting delay. Note that the reporting delay is longer for health claims then for material claims hence the claim count is lower also in the year 2005 compared to the material claims. Therefore we are going to use the time window 1.1.2001-31.12.2004 and obtain the estimate $\hat{\mu} = 1.458$.

There are 1 758 health claims with the occurrence year from 2000 to 2006. After running 1000 simulations of homogeneous Poisson process with the estimated intensity $\hat{\mu} = 1.458$ we obtain an estimate of the claim count of 1 751.11 which is quite similar to the number of claims from our database. This result prompts precision of our triangle free approach in estimating claim count also for health claims in our database.

Reporting delay

We again assume reporting delay to follow the exponential distribution which does not seem to be inconsistent with histograms 4.1 and 4.3 of the observed reporting delay which is however biased towards smaller values. The average observed reporting delay is $\lambda_{obs} = 90.39406$ for material claims and $\lambda_{obs} = 193.0069$ for health claims. Using Equation 1.3 we obtain estimate $\hat{\lambda} = 93.9847$ for material claims and $\hat{\lambda} = 193.0069$ for health claims.

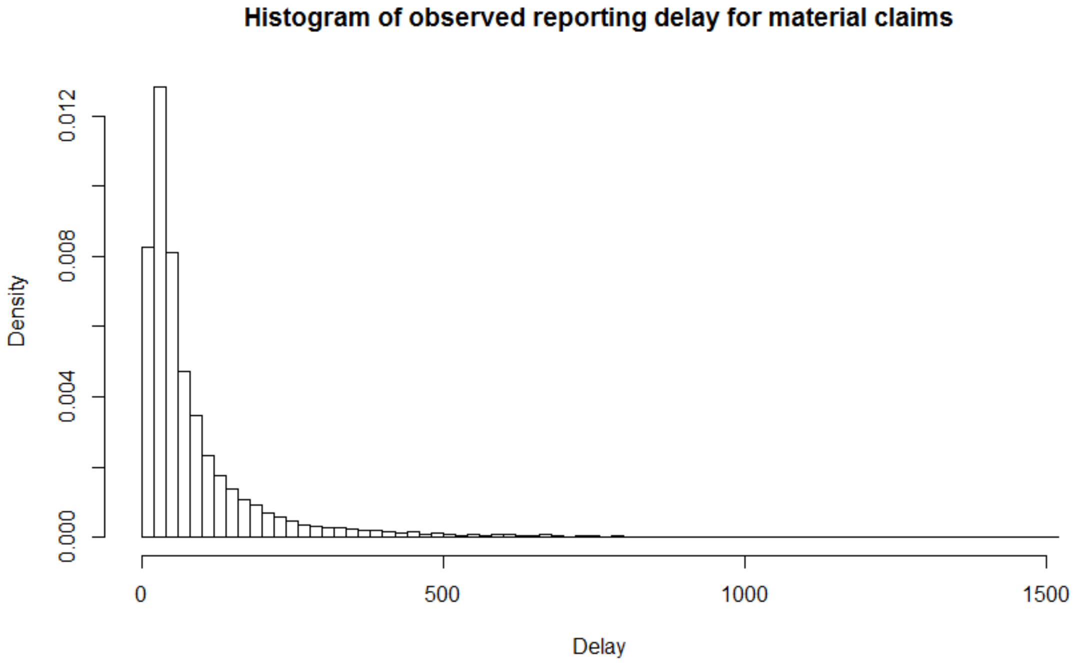

Figure 4.1: Histogram of the observed reporting delay for health claims

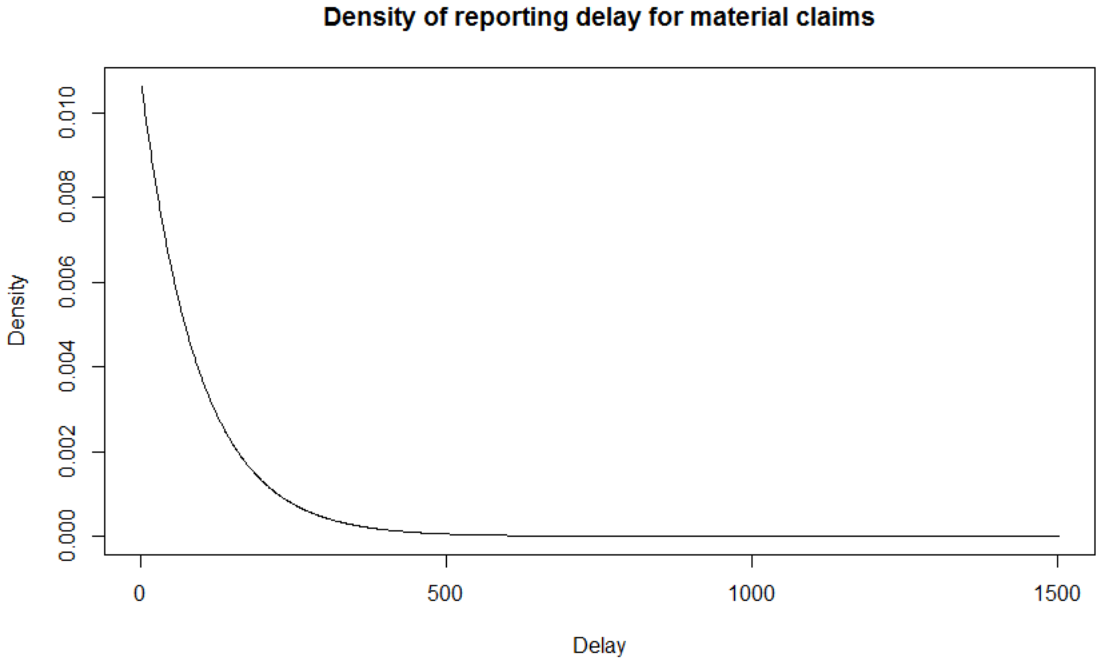

Figure 4.2: Estimated density of reporting delay of material claims

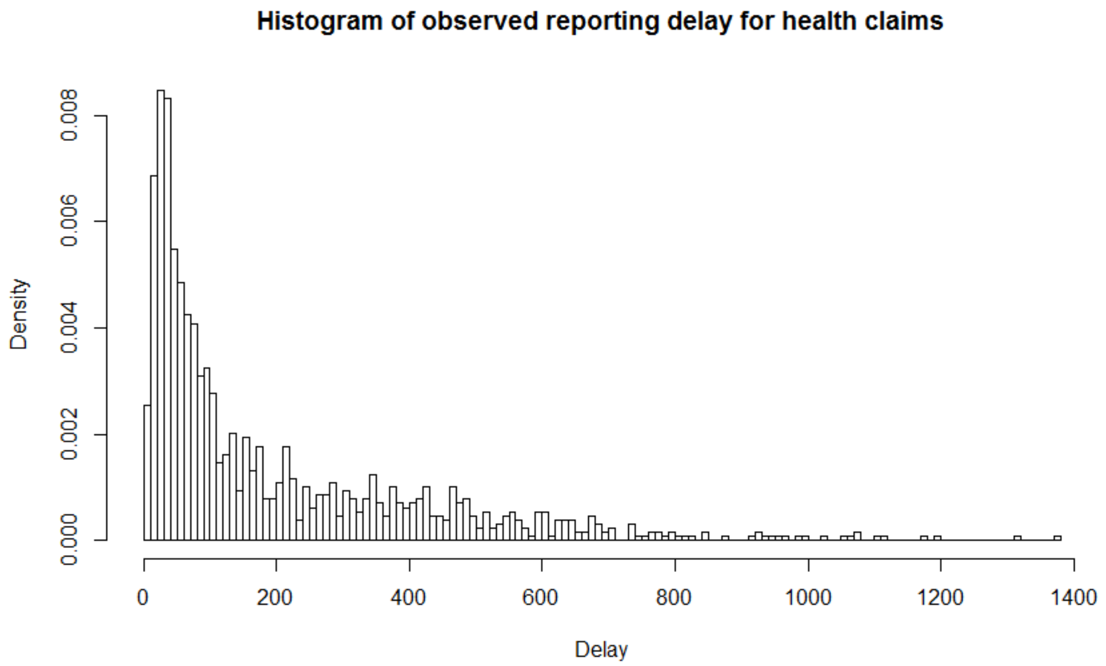

Figure 4.3: Histogram of the observed reporting delay for health claims

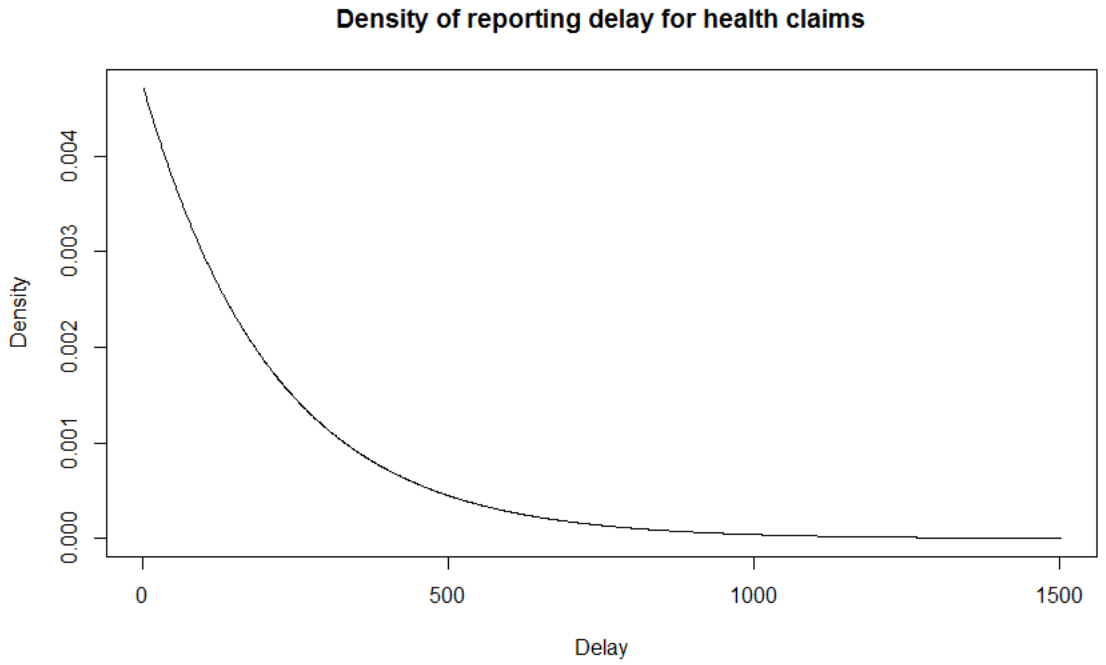

Figure 4.4: Estimated density of reporting delay of health claims

Figures 4.2 and 4.4 show the estimated density of reporting distribution for material and health claims.

The average reporting delay of claims that occurred in years 2000 to 2006 from the full database is 92.617 days for material claims and 215.546 days for health claims. Both estimates for material and health claims are pretty close to the numbers obtained from the database therefore our model seems to be quite precise in estimating reporting delay for our data set.

4.1.2 Claims severity

To estimate kernel severity distribution we need to take into account the inflation and revalue claims using the same inflation rates as in Chapter 3. After such an adjustment the minimum value is 142 CZK for material claims and 178 CZK for health claims. The average amount is 42 071.25 CZK for material claims and 42 636 CZK for health claims. Maximal amount is 2 340 250 CZK for material claims and 1 693 830 CZK for health claims.

Using maximum likelihood method we estimate parameters of log-normal and gamma distribution for both material and health claims. Values of the estimated parameters for material claims are for log-normal distribution $\hat{mu} = 10.1593$ and $\hat{\sigma} = 0.9265$ and for gamma distribution $\hat{a} = 0.001$ and $\hat{p} = 258.81550$. For health claims the estimates of parameters of log-normal distribution are $\hat{\mu} = 9.2436$ and $\hat{\sigma} = 1.5714$ and of gamma distribution $\hat{a} = 0.01$ and $\hat{p} = 103.855$. According to Figures 4.5 and 4.7 which show histograms and estimated densities and Figures 4.6 and 4.8 which show Quantile-quantile plots, for both material and health claims log-normal distribution seems to fit the data better. Therefore we use log-normal distribution with the estimated parameters to model material and health claims kernel severity.

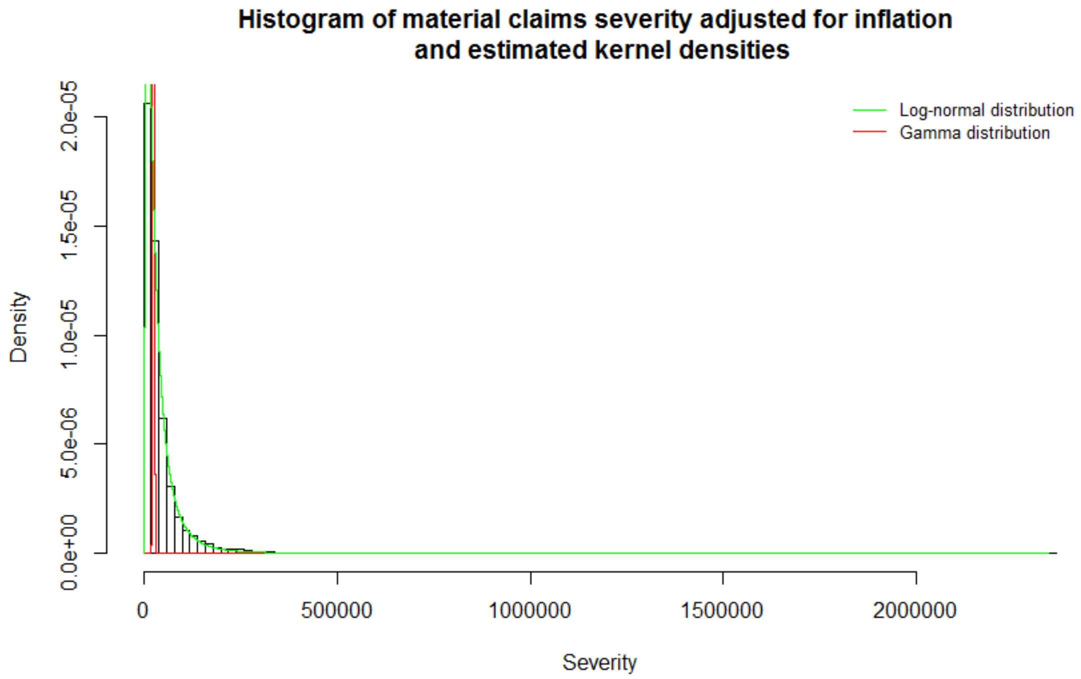

Figure 4.5: Histogram of material claims severity adjusted for inflation and estimated densities

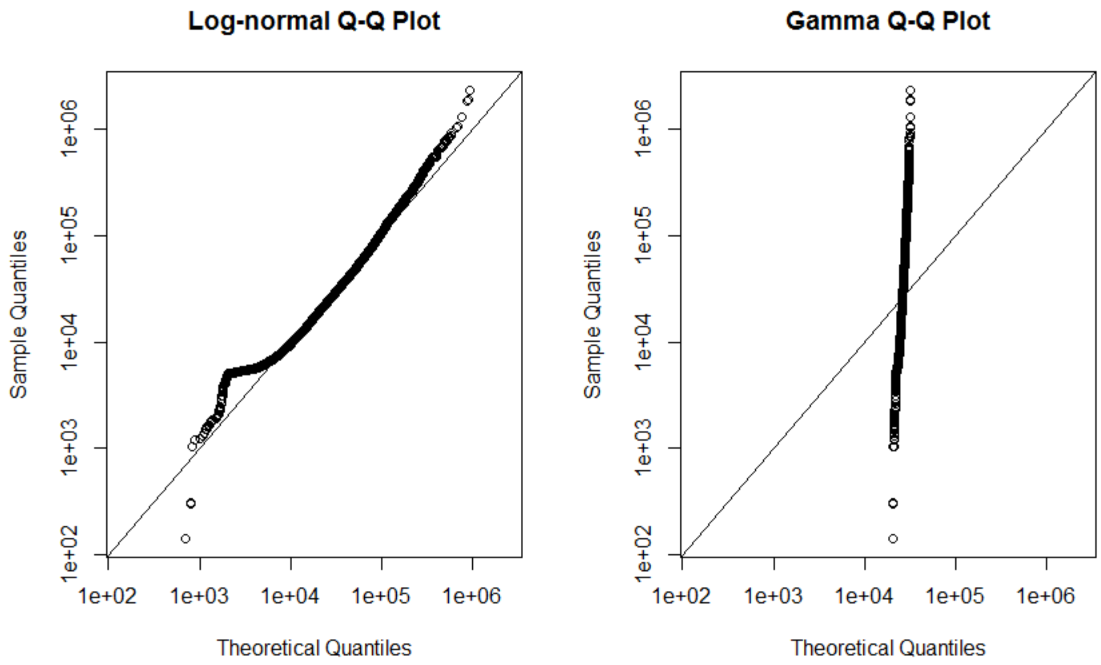

Figure 4.6: Quantile-Quantile plots of material claims severity

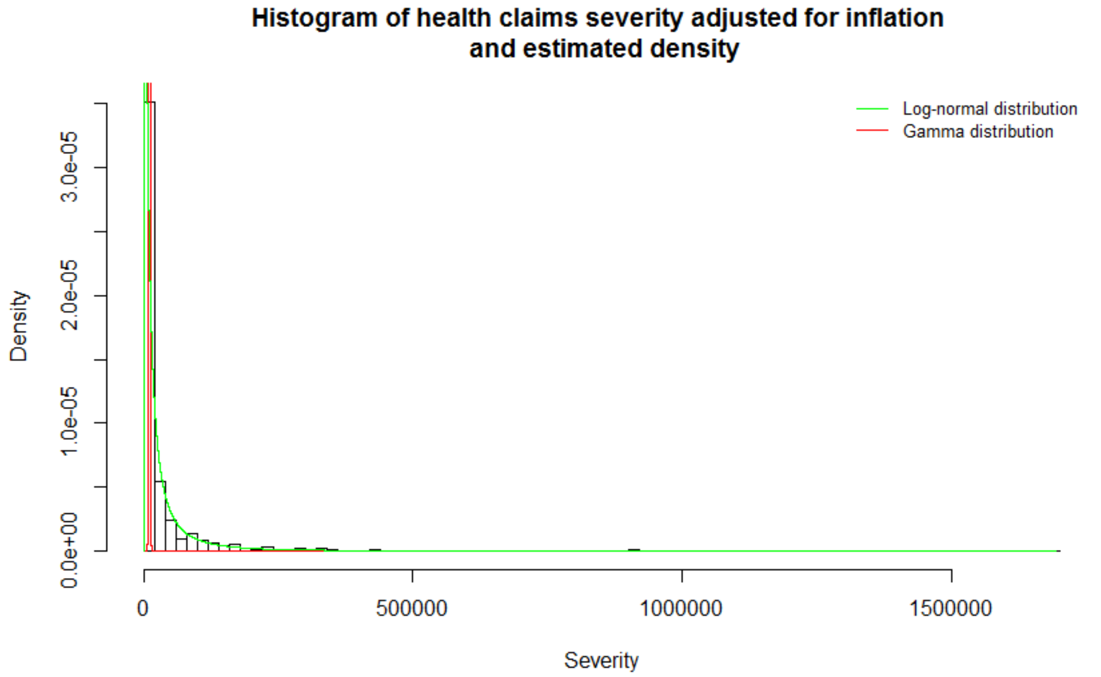

Figure 4.7: Histogram of health claims severity adjusted for inflation and estimated densities

IBNR reserve

Using the information above we have models for both material and health claims behavior and run simulations to obtain the estimate of the IBNR reserve such

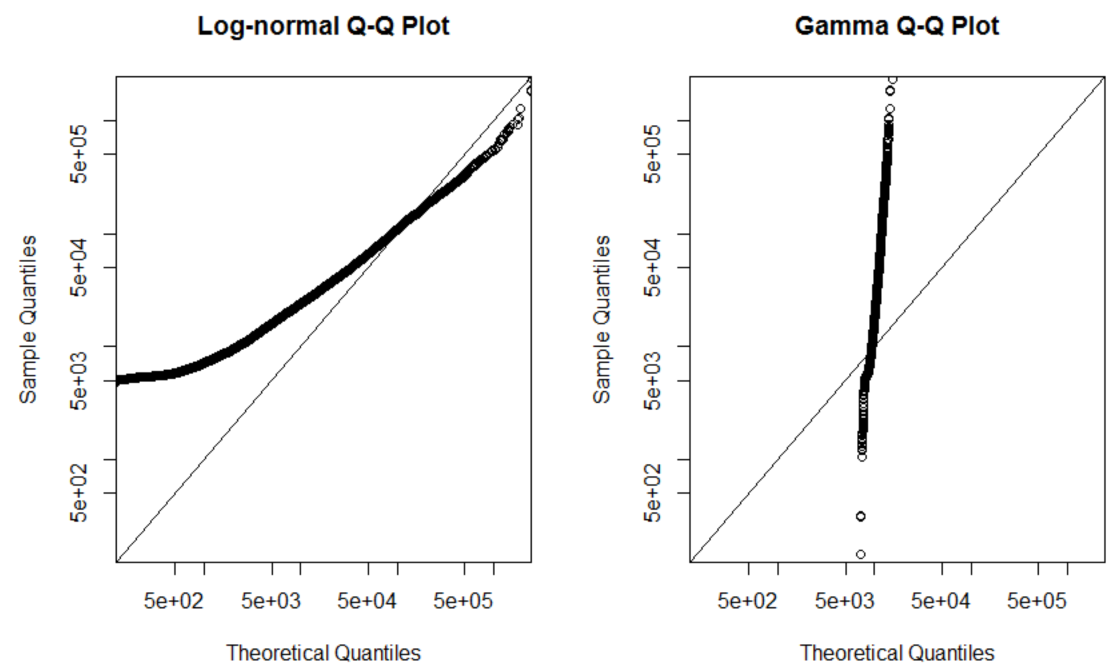

Figure 4.8: Quantile-Quantile plots of material claims severity

as described in Chapter 1. We obtained the estimate of 27 575 057.08 CZK for material claims and of 5 103 851.81 CZK for health claims.

	Material claims			Health claims		
	Database	Est.	Diff.	Database	Est.	Diff.
number of claims	18 599	19 044	2.39%	1 785	1 751	-1.90%
reporting delay	92.617	93.9847	1.48%	215.546	212.242	-1.533%
IBNR re-serve	30 936 696	27 575 057	10.8%	8 042 851	5 103 852	-36.54%

Table 4.4: Comparison of results of model and full database of claims

Table 4.1.2 summarizes all the estimates of our models and compares them with the corresponding values from the database. As can be seen, our estimate of the number of claims and reporting delay is quite close to the number from the database. However, the IBNR estimate differs from the database by 10.8 % for material claims and by 36.54% for health claims. We take a closer look for the health claims, as the difference is even greater there and the cause of the problem might be similar for both health and material claims.

As the estimates of the number of claims and reporting delay are quite close to the values from the database, the cause of such a difference of the IBNR health claims estimate is probably the estimate of the claims severity, which is estimated to be too low. We used the values of the inflation based on the change of the consumer price index in the Czech republic to approximate the health claims inflation. However, in the Czech republic the health claims inflation is higher than the price inflation. For example, according to Jedlička (2011), in the year 2011 the health claims inflation was 13.2% but the price inflation was 1.9%. Unfortunately we do not have the information on the claims inflation in other years. We used different values for claims inflation and estimated the IBNR reserve. The model was exactly the same except for the claims severity modeling for which we again used log-normal kernel distribution and estimated new parameters. For the sake of simplicity we used one rate for all years. Table 4.5 shows that the IBNR estimate increases with the increasing inflation rate. Therefore, we assume that in the years 2000 to 2006 the claims inflation was much higher than the price inflation and using the appropriate claims inflation should increase the precision of the estimate.

The same argumentation could be used for the material claims. The deviation from the database is a little bit smaller than the deviation of health claims. This suggests that the material claims inflation was much higher than the price inflation yet smaller than the health claims inflation during the observed years.

inflation rate	$\hat{\mu}$	$\hat{\sigma}$	IBNR
5%	9.3580	1.5743	5 950 122
8%	9.4736	1.5770	6 564 330
10%	9.5488	1.5794	7 007 448
13.2%	9.6664	1.5840	7 777 699

Table 4.5: Estimates of log-normal kernel severity distribution and IBNR health claims reserve using different inflation rates

4.1.3 Chain Ladder

Using the described triangle free method is pretty demanding and complex. In this section we will estimate the IBNR reserve for both material and health claims using the Chain ladder method such as described in Chapter 2.

Firstly, we organize the data into the incremental development triangles. An item $X_{i,j}$ of the incremental triangle is a sum of payments for claims that occurred in year i and were paid in year $j+i$, $i = 2000, \dots, 2006$ and $j = 0, \dots, 6$ and $i+j \leq 2006$. Having prepared these incremental triangles, we prepare the cumulative triangles. An item of the cumulative triangle $C_{i,j}$ includes also a sum of the previous items on the same row of the corresponding incremental triangle $C_{i,j} = \sum_{k=0}^{j} X_{i,k}$.

Using these cumulative development triangles we obtain the estimates of the total claims reserve of 364 756 411.37 CZK for material claims and of 194 807 613.54 CZK for health claims. These estimates also include reserves for reported but not settled claims which we need to subtract in order to obtain the IBNR reserves. The sum of the payments from the years 2007-2013 for claims that occurred and were reported during the years 200-2006 is 111 013 377 CZK for material claims and 13 637 437 CZK for health claims. Therefore the IBNR estimate obtained using the Chain ladder is 253 743 034.37 CZK for material claim and 181 170 16.5

Development	0	1	2	3	4	5
year			Individual factors			
2001	6.2381	1.4170	1.1729	1.0408	1.0092	1.0055
2002	5.7784	1.5888	1.1261	1.0417	1.0111	
2003	7.2596	1.3926	1.0950	1.0469		
2004	4.7220	1.3758	1.0704			
2005	3.8158	1.1312				
2006	3.2851					
			Chain ladder estimated factors			
	4.2099	1.3110	1.0994	1.0442	1.0104	1.0055

Table 4.6: Chain ladder estimated development factors and individual factors of material claims

Development	0	1	2	3	4	5
year			Individual factors			
2001	83.0062	3.4811	2.9168	1.3404	1.1088	1.0070
2002	12.1277	8.4252	2.4880	1.1796	1.0114	
2003	268.8179	4.5339	1.7557	1.0828		
2004	24.7397	5.7585	1.2503			
2005	14.5154	2.2551				
2006	27.0783					
			Chain ladder estimated factors			
	24.3136	4.0077	1.6742	1.1398	1.0409	1.0070

Table 4.7: Chain ladder estimated development factors and individual factors of health claims

CZK for health claims. The corresponding values from the full database are 30 963 696 CZK for material claims and 8 042 851 CZK for health claims.

A large margin of difference suggests that the data violate some of the assumptions 1, 2 or 3 of the Chain ladder method. We compare the estimates of the development factors $\hat{f}_s = \frac{\sum_{j=1}^{t-s-1} C_{j,s+1}}{\sum_{j=1}^{t-s-1} C_{j,s}}$ with the individual estimated factors that also depend on the occurrence year $\hat{F}_{r,s} = \frac{C_{r,s+1}}{C_{r,s}}$. Values obtained from the data can be found in Table 4.6 for material claims and Table 4.7 for health claims. If the assumption 1 holds, the values for one development year (which can be found in the columns of the tables) would be similar and approximately equal to the corresponding Chain ladder estimated factor $\hat{F}_{r,s} \cong \hat{f}_s$, $\forall r$. However, as can be seen the values of the individual factors corresponding to the same development stage vary and significantly differ from the development factors $\hat{f}_s$.

As at least one of the assumptions of the Chain ladder is strongly violated, therefore the method cannot be used neither for material nor for health claims from the database we are working with.

We should once more stress out that the data set we are working with is just a random sample of a database. We do not have the information on how exactly the data set was sampled, therefore even though we cannot use the Chain ladder method it does not mean that using the Chain ladder for the whole database of Czech Insurer's Bureau would be inappropriate.

4.2 Simulation study

In this section we will generate artificial data sets for which we would know the true distribution, and therefore will be able to compare the results obtained by using the triangle free approach with the results obtained by using the Chain ladder method for estimating the IBNR reserve. Generating the data and most of the estimating is run in software R and the source code can be found in the appendix to this thesis.

To confirm generally more preciseness of the triangle free approach than the Chain ladder, running many simulations with different setups would be needed. For purpose of this thesis we will only run 10 simulations with one setup to show the procedure and leave the more refined simulation study for the follow up of this thesis.

4.2.1 Generating artificial data sets

For simplicity we assume no claims inflation and negligible IBNER. We are also interested only in estimating the IBNR reserve, therefore we only need to generate claims occurrence, severity and reporting delay. Choosing the parameters is just incidental and has no deeper meaning as the purpose of this section is mainly to show a guidance for further simulation studies and demonstrate that the triangle free approach can be used even for different values than those from the real data set above. The chosen distributions are following:

- Claims occurrence is driven by a homogeneous Poisson process with rate $\mu = 30$ days, which means that the time between two successive events is approximately one month.
- Reporting delay follows exponential distribution with mean $\lambda = 730$ days, which means that the average reporting delay is two years.
- Claims severity follows log-normal distribution with density 1.7 and parameters $\mu = 12$ and $\sigma = 1.5$.

We consider a period of 10 years, each year beginning on the 1.1. and ending 31.12. which is 3650 days (we do not consider leap-years). We run 10 simulations to obtain 10 data sets. The algorithm for each simulation is following:

1. Simulate claims occurrence times using the property 2 of a homogeneous Poisson process that the time between two successive events follows the exponential distribution.
2. For each claim simulate a severity from the chosen log-normal distribution.
3. For each claim simulate a reporting delay from the chosen exponential distribution.

Sim.	Poisson process		Log-normal distribution	severity	Reporting delay	
	time window	$\hat{\mu}$	$\hat{\mu}$	$\hat{\sigma}$	λ_{obs}	$\hat{\lambda}$
1	1-5	30.6271	11.9288	1.4764	556.6703	732.967
2	1-5	29.88525	12.0972	1.4677	552.0947	723.9360
3	1-7	32.16456	12.3782	1.6824	524.2500	671.0360
4	1-7	30.0714	12.0250	1.6281	504.7411	635.907
5	1-7	31.4321	11.9596	1.2577	597.3265	817.9410
6	1-5	29.0806	12.0763	1.4543	514.8018	653.8360
7	1-7	33.9200	11.8562	1.5855	640.9887	920.2570
8	1-9	30.3333	11.9174	1.5842	509.6545	644.6150
9	1-6	28.7105	11.6955	1.3715	514.7843	653.8040
10	1-7	32.6667	12.0040	1.5305	604.7938	834.5600

Table 4.8: Results of estimates for simulations

4. Export this whole data set for the purpose of comparing estimates with the true values.

5. Export only claims that were reported before the end of the observable period which is here 3650 days and use this as the initial data set for estimating the IBNR reserve and the total loss.

4.2.2 Results

For each data set we use the triangle free approach exactly the same way as we did for the real data set using the appropriate distributions. The estimated parameters for each simulation can be found in Table 4.8.

We estimated the IBNR reserve for the period of ten years using the triangle free approach and the Chain ladder method. Results are compared in Table 4.9. We compare the absolute values of the differences between the estimates and corresponding true values. We can see that the triangle free approach is more accurate in estimating the IBNR for all ten simulations.

We can also compare the precision of the triangle free approach and the Chain ladder using the standard error which is the square root of a sum of squares of the difference between an estimate and corresponding true value, summed over all simulations, which gives 18 302 394.13 for the triangle free approach and 68 867 129.65 for the Chain ladder. We can see that the triangle free approach has a smaller standard error, which is in favor with a hypothesis that the triangle free approach generally provides more accurate estimates, as it uses more granular information compared to the Chain ladder. However, more different simulations would be needed to confirm that.

Sim.	True value	Triangle free	Distance	Chain ladder	Difference
1	12 555 833	10 516 594	2 039 239	8 957 298	3 598 535
2	6 948 981	12 649 829	5 700 848	20 074 781	13 125 780
3	8 198 330	19 487 810	11 289 480	70 635 697	62 437 367
4	5 422 638	12 460 439	7 037 801	13 960 093	8 537 455
5	9 911 635	8 934 232	977 403	19 899 295	9 987 660
6	9 085 227	11 492 898	2 407 671	3 960 207	5 125 020
7	13 457 810	13 044 506	413 303	1 867 713	11 590 096
8	18 302 493	10 977 957	7 324 537	6 072 373	12 230 121
9	14 751 339	7 194 231	7 557 108	3 233 273	11 518 066
10	14 873 370	12 961 637	1 911 732	8 274 787	6 598 583

Table 4.9: Comparison of IBNR estimates and corresponding true values

Conclusion

Claims reserving has been a challenging topic for actuaries for many years. With stricter regulations the willingness to find the best estimate of future liabilities is even more increasing. Most of the traditional techniques work with data aggregated into development triangles. This thesis presented an approach to claims reserving that uses a more granular information on individual claims. This approach was illustrated on real data from the Czech Insurers' Bureau. It was also compared with the basic Chain ladder method.

The triangle free approach is based on creating a model for claims occurrence and claims severity and then putting them together using simulations. The first chapter describes the components of this approach and illustrates how these can be used to obtain the claims reserves estimates.

For the claims occurrence a homogeneous Poisson process and a general non-homogeneous process are presented as well as ways to obtain point estimates of their intensities. Dealing with the bias of the observed distribution of a reporting delay is described generally and then shown on an example of the exponential distribution. Severity distribution is assumed to be a scaled version of a kernel severity distribution which is presented to be log-normal or gamma in this thesis. Claims severity values used for estimating the parameters of the kernel severity distribution shall be adjusted for IBNER. For such an adjustment a method using general linear models is described. Finally, the algorithms for obtaining the estimates of the claims reserves are described.

The second chapter gives an overview of the distribution free Chain ladder method. The assumptions of the method are described as well as the estimates of the ultimate loss. We use this knowledge in the practical part for comparison with the triangle free approach.

In the third chapter the triangle free approach is illustrated on real data. Two separate models, one for material claims and the other for health claims, are created. All the steps of the parameters estimations are shown in detail. These models are then used for estimation of the IBNR reserve.

The fourth chapter compares the triangle free approach and the basic Chain ladder method. Firstly, a part of the real database known as at the end of the year 2006 is used for the IBNR estimating using the triangle free approach and the Chain ladder. The estimates are then compared with the rest of the database.

Triangle free approach gave estimates of the same order as the values from the database, however, they were underestimated. It was shown that the underestimation was caused by using low claims inflation rates. The true values of the claims inflation were not available during creating this thesis, therefore the price inflation was used instead. However, the price inflation is generally lower then the claims inflation in the Czech republic. It was shown that the estimated claim count as well as the estimated reporting delay were close to the values from the full database and with increasing claims inflation used, the reserve estimate increased as well.

The estimates obtained using the Chain ladder were of much higher order than the values obtained from the full database. It was shown that the data do not meet the assumption of the approximate proportionality of the cumulative triangle columns, therefore the Chain ladder method could not be used.

Finally, ten artificial data sets were generated. For all of them the triangle free estimate was closer the the true values than the Chain ladder estimate.

Compared to the Chain ladder, the triangle free approach is more complex and requires more granular and accurate data. However, that should not be a problem for insurance companies that possess databases containing all the information needed. For the real data we worked with, as well as for the artificial data sets, we showed not only that triangle free approach gives more precise results compared to the Chain ladder, but also that it gives accurate estimates even when the Chain ladder method cannot be used.

To be able to state that all the advantages of the triangle free approach presented in this thesis, further analyses and especially more simulation studies using different parameters and distributions would be needed. It would also be interesting to compare the triangle free approach to other more sophisticated triangular methods, not only the basic Chain ladder.

Bibliography

Anděl, J. *Statistické metody.* Second Edition. Matfyzpress, Praha, 1998. ISBN 80-85863-27-8.

Anděl, J. *Základy matematické statistiky.* Second Edition. Matfyzpress, Praha, 2007. ISBN 80-7378-001-1.

Antonio, K. and Plat, R. Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal*, 2014(7):649–669, May 2013. doi: 10.1080/03461238.2012.755938. Electronic copy available at: `http://www.tandfonline.com/doi/abs/10.1080/03461238.2012.755938`.

Branda, M. Zobecněné lineární modely v pojišťovnictví. 2013. Electronic copy available at: `http://www.nfvp.cz/res/data/000170.pdf` [Last accessed: 22 July 2014].

Cipra, T. *Finanční a pojistné vzorce.* Grada Publishing, a.s., Praha, 2006. ISBN 80-247-1633-X.

Endgland, P. and Verrall, R. Stochastic Claims Reserving in General Insurance. *British Actuarial Journal*, 8(03):443–518, 2002. doi: 10.1017/S1357321700003809. Electronic copy available at: `http://www.journals.cambridge.org/abstract_S1357321700003809`.

Insurance Act No. 277/2009 Coll. Zákon o pojišťovnictví 277/2009 Sb. Electronic copy available at: `www.mvcr.cz/soubor/sb085-09-pdf.aspx` [Last accessed: 22 June 2014].

Jedlička, P. Kvantifikace a statistiky škod na zdraví POV se zaměřením na typové znaky vážných škod na zdraví, December 2011. Electronic copy available at: `http://www.actuaria.cz/sdeleni.asp?ID=496` [Last accessed: 13 July 2014].

Parodi, P. Triangle-free reserving. *British Actuarial Journal*, 19(1):168–218, 2014. doi: 10.1017/S1357321713000093. Electronic copy available at: `http://www.journals.cambridge.org/abstract_S1357321713000093` [Last accessed: 23 July 2014].

Prášková, Z. and Lachout, P. *Základy náhodných procesů.* Nakladatelství Karolinum, Praha, 2001. ISBN 80-7184-688-0.

Solvency II. Directive 2009/138/EC of the European Parliament and of the Council of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance (Text with EEA relevance). Electronic copy available at: `http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32009L0138` [Last accessed: 22 June 2014].

Winkelmann, R. *Econometric Analysis of Count Data.* Fifth Edition. Springer, Verlag, Berlin, Heidelberg, 2008. ISBN 978-3-540-77648-2.

Wüthrich, M. and Merz, M. *Stochastic claims reserving methods in insurance.* Wiley, Chichester, England, 2008. ISBN 978-0-470-72346-3.

List of Figures

List of Tables

Appendix

This appendix contains the source code of the estimates and the simulations from Chapters 3 and 4 written in software R.

Firstly, we need to read the data sets and estimate the distribution of the reporting delay and severity distribution. Preparing the data sets into appropriate form as well as estimating the intensities of the Poisson processes was done in Microsoft Office Excel. Computing the parameter of reporting delay was done in Wolfram Mathematica. Due to the sensitivity of the information included, the data sets cannot be provided.

```
(* Source code in Wolfram Mathematica for computing parameter
of reporting delay *)
t = 3650
lobs = 604.79381443299
Solve[lobs == l (1 + (E^(-t/l) - l/t (1 - E^(-t/l)))/(1 - l/t
(1 - E^(-t/l)))), l, Reals]
```

Rest of the source code is written in R.

```
###### Load packages ####
require("XLConnect")
require("MASS")

###reporting delay###
######################
material=readWorksheetFromFile("C:/Users/Kika/Cloud_
   kika/diplomka/data/rep_delay.xlsx",sheet="vec")
health=readWorksheetFromFile("C:/Users/Kika/Cloud_kika/
   diplomka/data/rep_delay.xlsx",sheet="zdravi")
#histograms of observed reporting delay
#material
hist(material$delay,breaks=100,prob=TRUE, main="
   Histogram of observed reporting delay for material
   claims", xlab="Delay",ylab="Density")
plot(dexp(0:1500,rate=1/77.0174),type="l",main="Density
    of reporting delay for material claims", xlab="
   Delay",ylab="Density")
#health
hist(health$delay,breaks=100,prob=TRUE, main="Histogram
    of observed reporting delay for health claims",xlab
   ="Delay",ylab="Density")
```

```
plot(dexp(0:1500,rate=1/211.552),type="l",main="Density
    of reporting delay for health claims", xlab="Delay"
   ,ylab="Density")

###severity distribution###
###########################
##material
materials=readWorksheetFromFile("C:/Users/Kika/Cloud_
   kika/diplomka/data/severity.xlsx",sheet="vec")
min(materials$adjusted)
max(materials$adjusted)
mean(materials$adjusted)
fitdistr(materials$adjusted,"log-normal")
fitdistr(materials$adjusted,"gamma",lower=c(0.01,0.01))

#histograms
hist(materials$adjusted,breaks=50,prob=TRUE, main="
   Histogram of material claims severity adjusted for
   inflation", xlab="Severity", ylab="Density")
hist(log(materials$adjusted),breaks=100,prob=TRUE, main
   ="Histogram of logarithm of material claims severity
    adjusted for inflation", xlab="Severity", ylab="
   Density")
hist(materials$adjusted,breaks=100,prob=TRUE, main="
   Histogram of material claims severity adjusted for
   inflation  \n and estimated kernel densities", xlab=
   "Severity", ylab="Density")
lines(dgamma(0:4951945,shape=2.982351e+02,rate=1.000000
   e-02),col="red")
lines(dlnorm(0:4951945,meanlog=10.301374872,sdlog
   =1.0010121441),col="green")
legend("topright",legend=c("Log-normal distribution","
   Gamma distribution"),lwd=c(1,1), col = c("green", "
   red"), merge=TRUE,inset=.01,cex=.8,adj=0,bty="n")

#Q-Q plots
attach(mtcars)
par(mfrow=c(1,2))
qqplot(rlnorm(length(materials$adjusted),meanlog
   =10.301374872,sdlog=1.0010121441), materials$
   adjusted, log="xy",xlim=c(19.80417,4951945), ylim=c
   (19.80417,4951945),xlab="Theoretical Quantiles",ylab
   ="Sample Quantiles", main="Log-normal Q-Q Plot")
abline(0,1)
qqplot(rgamma(length(materials$adjusted),shape=2.982351
   e+02,rate=1.000000e-02), materials$adjusted, log="xy
   ",xlim=c(19.80417,4951945), ylim=c(19.80417,4951945)
   , xlab="Theoretical Quantiles", ylab="Sample
   Quantiles", main="Gamma Q-Q Plot")
abline(0,1)

##health
healths=readWorksheetFromFile("C:/Users/Kika/Cloud_kika
   /diplomka/data/severity.xlsx",sheet="zdravi")
min(healths$adjusted)
max(healths$adjusted)
mean(healths$adjusted)
fitdistr(healths$adjusted,"log-normal")
fitdistr(healths$adjusted,"gamma",lower=c(0.01,0.01))

#histograms
hist(healths$adjusted,breaks=100,prob=TRUE, main="
   Histogram of health claims severity adjusted for
   inflation", xlab="Severity", ylab="Density")
hist(log(healths$adjusted),breaks=100,prob=TRUE, main="
   Histogram of logarithm of health claims severity
   adjusted for inflation", xlab="Severity", ylab="
   Density")
hist(healths$adjusted,breaks=100,prob=TRUE, main="
   Histogram of health claims severity adjusted for
   inflation \nand estimated kernel densities", xlab="
   Severity", ylab="Density")
lines(dgamma(0:2967376,shape=1.405290e+02,rate=1.000000
   e-02),col="red")
lines(dlnorm(0:2967376,meanlog=9.54702199,sdlog
   =1.59786844),col="green")
legend("topright",legend=c("Log-normal distribution","
   Gamma distribution"),lwd=c(1,1), col = c("green", "
   red"), merge=TRUE,inset=.01,cex=.8,adj=0,bty="n")

#Q-Q plots
attach(mtcars)
par(mfrow=c(1,2))
qqplot(rlnorm(length(healths$adjusted),meanlog
   =9.54702199,sdlog=1.59786844), healths$adjusted, log
   ="xy",xlim=c(101.7938,2967376), ylim=c
   (101.7938,2967376),xlab="Theoretical Quantiles",ylab
   ="Sample Quantiles", main="Log-normal Q-Q Plot")
abline(0,1)
qqplot(rgamma(length(healths$adjusted),shape=1.405290e
   +02,rate=1.000000e-02), healths$adjusted, log="xy",
   xlim=c(101.7938,2967376), ylim=c(101.7938,2967376),
   xlab="Theoretical Quantiles", ylab="Sample Quantiles
   ", main="Gamma Q-Q Plot")
abline(0,1)
```

```
###simulations###
#################
#interest
factor<-function(ot){
  #inflation rates and corresponing dates
  ir<-c
     (1.039,1.047,1.018,1.01,1.028,1.019,1.025,1.028,1.063,1.01,1.

  due<-c
     (365,730,1095,1460,1826,2191,2556,2921,3287,3652,4017,4382,47

  #which year is ot
  if(ot>due[[13]]){y<-14}else{
    if(ot>due[[12]]){y<-13}else{
      if(ot>due[[11]]){y<-12}else{
        if(ot>due[[10]]){y<-11}else{
          if(ot>due[[9]]){y<-10}else{
            if(ot>due[[8]]){y<-9}else{
              if(ot>due[[7]]){y<-8}else{
                if(ot>due[[6]]){y<-7}else{
                  if(ot>due[[5]]){y<-6}else{
                    if(ot>due[[4]]){y<-5}else{
                      if(ot>due[[3]]){y<-4}else{
                        if(ot>due[[2]]){y<-3}else{
                          if(ot>due[[1]]){y<-2}else{
                            y<-1
                          }}}}}}}}}}}}}
  if(y==14){interest<-1/(ir[[14]]^((due[[14]]-ot)/365))
     }else{ #occurrence time is in 2013
    interest<-1/(ir[[14]]^(140/365)) #year 2013: 140
       days between 21.5.2013 and 1.1.2013
    for(j in (y+1):13){
      interest<-interest*1/(ir[[j]]) #other years
    }
    interest<-interest*1/(ir[[y]]^((due[[y]]-ot)/365))
       #occurrence year
  }
  return(interest)
}

##occurence time follow poisson process##
poisson<-function(lambda,end){
  sim<-rexp(1,rate=lambda)
  #sim<-NULL
  repeat{
  diff<-rexp(1,rate=1/lambda)
  time<-sim[length(sim)]+diff
  sim<-c(sim,time)
  if(sim[[length(sim)]]>end) break
   }
  sim<-sim[1:(length(sim)-1)]
  sim<-lapply(sim,round)
  return(sim)
  }

###severity for time log-normal distr###
severity<-function(dat,a,b){
  sev<-NULL
  for(i in 1:length(dat)) {
    sev1<-rlnorm(1,meanlog=a,sdlog=b)
    sev1<-sev1*factor(dat[[i]])
    sev<-c(sev,sev1)
  }
  return(sev)
}

###reporting delay exponential ###
###returns dates of reporting###
reporting<-function(dat,lambda){
  del<-NULL
  for(i in 1:length(dat)){
    delay<-rexp(1,rate=lambda)
    time<-dat[[i]]+delay
    del<-c(del,time)
  }
  del<-lapply(del,round)
  return(del)
}

###IBNR material claims all###
#Defining simulation

IBNR<-function(nsim,mu,lambda,a,b,end){
                                          #nsim is the
   number of simulations
  loss<-seq(0,0,length.out=nsim)
  for (i in 1:nsim){
    datesibnr<-poisson(mu,end)
                                 #generating claims
       occurrences
    severities<-severity(datesibnr,a,b)   #generating
       severity of each claim
    rep<-reporting(datesibnr,1/lambda)
                             #generating reporting
       delay
    for(j in 1:length(datesibnr)){
      if(rep[[j]]>end){loss[[i]]<-loss[[i]]+severities
         [[j]]} #taking only IBNR
    }
  }
  return(loss)
}

#Running simulations
set.seed(1) #to be able to replicate simulations
ibnrmat<-IBNR(1000,0.1058,77.0174,10.301374872,4899)
ibnrmaterial<-mean(ibnrmat)
plot.ecdf(ibnmat,main="Empirical cumulative
   distribution function of material IBNR claims")
quantile(ibnmat)

set.seed(1)
ibnrh<-IBNR
   (1000,1.25,211.552,9.54702199,1.59786844,4899)
ibnrhealth<-mean(ibnrh)
plot.ecdf(ibnrh,main="Empirical cumulative distribution
    function of health IBNR claims")
```

The procedure for claims from the year 2000 to 2006 is very similar.

```
###### Load packages ####
require("XLConnect")
require("MASS")

###reporting delay###
#####################
material=readWorksheetFromFile("C:/Users/Kika/Cloud_
   kika/diplomka/data/delay_2006.xlsx",sheet="vec")
health=readWorksheetFromFile("C:/Users/Kika/Cloud_kika/
   diplomka/data/delay_2006.xlsx",sheet="zdravi")
#histograms of observed reporting delay
#material
hist(material$delay,breaks=100,prob=TRUE, main="
   Histogram of observed reporting delay for material
   claims", xlab="Delay",ylab="Density")
plot(dexp(0:1500,rate=1/93.9847),type="l",main="Density
    of reporting delay for material claims", xlab="
   Delay",ylab="Density")
#health
hist(health$delay,breaks=100,prob=TRUE, main="Histogram
    of observed reporting delay for health claims",xlab
   ="Delay",ylab="Density")
plot(dexp(0:1500,rate=1/212.242),type="l",main="Density
    of reporting delay for health claims", xlab="Delay"
   ,ylab="Density")
```

```
###severity distribution###
###########################
#material
materials=readWorksheetFromFile("C:/Users/Kika/Cloud_
   kika/diplomka/data/severity_2006.xlsx",sheet="vec")
min(materials$adjusted)
max(materials$adjusted)
mean(materials$adjusted)
fitdistr(materials$adjusted,"log-normal")
fitdistr(materials$adjusted,"gamma",lower=c(0.01,0.01))

#histograms
hist(materials$adjusted,breaks=100,prob=TRUE, main="
   Histogram of material claims severity adjusted for
   inflation", xlab="Severity", ylab="Density")
hist(log(materials$adjusted),breaks=100,prob=TRUE, main
   ="Histogram of logarithm of material claims severity
    adjusted for inflation", xlab="Severity", ylab="
   Density")
hist(materials$adjusted,breaks=100,prob=TRUE, main="
   Histogram of material claims severity adjusted for
   inflation \n and estimated kernel densities", xlab="
   Severity", ylab="Density")
lines(dlnorm(0:2340250,meanlog=10.159352526,sdlog
   =0.926595463),col="green")
lines(dgamma(0:2340250,shape=2.588155e+02,rate=1.000000
   e-02),col="red")
legend("topright",legend=c("Log-normal distribution","
   Gamma distribution"),lwd=c(1,1), col = c("green", "
   red"), merge=TRUE,inset=.01,cex=.8,adj=0,bty="n")

#Q-Q plots
attach(mtcars)
par(mfrow=c(1,2))
qqplot(rlnorm(length(materials$adjusted),meanlog
   =10.159352526,sdlog=0.926595463), materials$adjusted
   , log="xy",xlim=c(142.0117,2340250), ylim=c
   (142.0117,2340250),xlab="Theoretical Quantiles",ylab
   ="Sample Quantiles", main="Log-normal Q-Q Plot")
abline(0,1)
qqplot(rgamma(length(materials$adjusted),shape=2.588155
   e+02,rate=1.000000e-02), materials$adjusted, log="xy
   ",xlim=c(142.0117,2340250), ylim=c(142.0117,2340250)
   , xlab="Theoretical Quantiles", ylab="Sample
   Quantiles", main="Gamma Q-Q Plot")
abline(0,1)

#health
```

```
healths=readWorksheetFromFile("C:/Users/Kika/Cloud_kika
   /diplomka/data/severity_2006.xlsx",sheet="zdravi")
fitdistr(healths$adjusted,"log-normal")
fitdistr(healths$adjusted,"gamma",lower=c(0.01,0.01))
min(healths$adjusted)
max(healths$adjusted)
mean(healths$adjusted)

#histograms
hist(healths$adjusted,breaks=100,prob=TRUE, main="
   Histogram of health claims severity adjusted for
   inflation", xlab="Severity", ylab="Density")
hist(log(healths$adjusted),breaks=100,prob=TRUE, main="
   Histogram of logarithm of health claims severity
   adjusted for inflation", xlab="Severity", ylab="
   Density")
hist(healths$adjusted,breaks=100,prob=TRUE, main="
   Histogram of health claims severity adjusted for
   inflation \n and estimated density", xlab="Severity"
   , ylab="Density")
lines(dlnorm(0:1693830,meanlog=9.24363847,sdlog
   =1.57143331),col="green")
lines(dgamma(0:1693830,shape=1.038855e+02,rate=1.000000
   e-02),col="red")
legend("topright",legend=c("Log-normal distribution","
   Gamma distribution"),lwd=c(1,1), col = c("green", "
   red"), merge=TRUE,inset=.01,cex=.8,adj=0,bty="n")

#Q-Q plots
attach(mtcars)
par(mfrow=c(1,2))
qqplot(rlnorm(length(materials$adjusted),meanlog
   =9.24363847,sdlog=1.57143331), materials$adjusted,
   log="xy",xlim=c(178.2757,1693830), ylim=c
   (178.2757,1693830),xlab="Theoretical Quantiles",ylab
   ="Sample Quantiles", main="Log-normal Q-Q Plot")
abline(0,1)
qqplot(rgamma(length(materials$adjusted),shape=1.038855
   e+02,rate=1.000000e-02), materials$adjusted, log="xy
   ",xlim=c(178.2757,1693830), ylim=c(178.2757,1693830)
   , xlab="Theoretical Quantiles", ylab="Sample
   Quantiles", main="Gamma Q-Q Plot")
abline(0,1)

###simulations###
#inflation adjustment
factor<-function(ot){
  #inflation rates and corresponing dates
  ir<-c(1.039,1.047,1.018,1.001,1.028,1.019,1.025)
  due<-c(365,730,1095,1460,1826,2191,2554)
  #which year is ot
  if(ot>due[[6]]){y<-7}else{
    if(ot>due[[5]]){y<-6}else{
      if(ot>due[[4]]){y<-5}else{
        if(ot>due[[3]]){y<-4}else{
          if(ot>due[[2]]){y<-3}else{
            if(ot>due[[1]]){y<-2}else{
              y<-1
   }}}}}}
  if(y==7){interest<-1/(ir[[7]]^((due[[7]]-ot)/365))}
     else{ #occurrence time is in 2006
    interest<-1/(ir[[7]]^(362/365)) #year 2006: 362
       days between 29.12.2006 and 1.1.2006
    for(j in (y+1):6){
      interest<-interest*1/(ir[[j]]) #other years
    }
    interest<-interest*1/(ir[[y]]^((due[[y]]-ot)/365))
       #occurrence year
  }
  return(interest)
}

#occurence time follow poisson process
poisson<-function(lambda,end){
  sim<-rexp(1,rate=lambda)
  repeat{
  diff<-rexp(1,rate=1/lambda)
  time<-sim[length(sim)]+diff
  sim<-c(sim,time)
  if(sim[[length(sim)]]>end) break
   }
  sim<-sim[1:(length(sim)-1)]
  sim<-lapply(sim,round)
  return(sim)
  }

#material claim count simulation
set.seed(1)
ccm<-NULL
for(i in 1:1000){
  claimcount<-poisson(0.133704,2554)
  ccm<-c(ccm,length(claimcount))
}
mean(ccm)

#health claim count simulation
```

```
set.seed(1)
cch<-NULL
for(i in 1:1000){
  claimcount<-poisson(1.458,2554)
  cch<-c(cch,length(claimcount))
}
mean(cch)

#severity for given time- log-normal distribution
severity<-function(dat,a,b){
  sev<-NULL
  for(i in 1:length(dat)) {
    sev1<-rlnorm(1,meanlog=a,sdlog=b)
    sev1<-sev1*factor(dat[[i]])
    sev<-c(sev,sev1)
  }
  return(sev)
}

###reporting delay - exponential distribution ###
#returns dates of reporting
reporting<-function(dat,lambda){
  del<-NULL
  for(i in 1:length(dat)){
    delay<-rexp(1,rate=lambda)
    time<-dat[[i]]+delay
    del<-c(del,time)
  }
  del<-lapply(del,round)
  return(del)
}

###IBNR material claims all###
#Defining simulation

IBNR<-function(nsim,mu,lambda,a,b,end){ #nsim is the
   number of simulations
  loss<-seq(0,0,length.out=nsim)
  for (i in 1:nsim){
    datesibnr<-poisson(mu,end) #generating claims
       occurrences
    severities<-severity(datesibnr,a,b) #generating
       severity of each claim
    rep<-reporting(datesibnr,1/lambda)
                             #generating reporting
       delay
    for(j in 1:length(datesibnr)){
      if(rep[[j]]>end){loss[[i]]<-loss[[i]]+severities
```

```
         [[j]]} #taking only IBNR
    }
  }
  return(loss)
}

#Running simulations
set.seed(1) #to be able to replicate simulations
ibnrm<-IBNR
   (1000,0.133704,93.9847,10.159352526,0.926595463,2554)

ibnrmaterial<-mean(ibnrm)

ibnrh<-IBNR
   (1000,1.458,212.242,9.2436387,1.57143331,2554)
ibnrhealth<-mean(ibnrh)
```

In Microsoft Office Excel we also prepared the development triangles and then again used software R for the Chain ladder estimations.

```
###Chain ladder###
require(ChainLadder)
require("XLConnect")

material=readWorksheetFromFile("C:/Users/Kika/Cloud_
   kika/diplomka/data/chl_2006.xlsx",sheet="vec11")
M1 <- MackChainLadder(material, est.sigma="Mack")
M1

health=readWorksheetFromFile("C:/Users/Kika/Cloud_kika/
   diplomka/data/chl_2006.xlsx",sheet="zdravi1")
M2 <- MackChainLadder(health, est.sigma="Mack")
M2
```

Generating the artificial data sets was all done the same way only using seeds 1, 5, 10, 50, 100, 200, 500, 1000, 2000, 5000.

```
#occurence time follow poisson process
poisson<-function(lambda,end){
  sim<-rexp(1,rate=lambda)
  repeat{
    diff<-rexp(1,rate=1/lambda)
    time<-sim[length(sim)]+diff
    sim<-c(sim,time)
    if(sim[[length(sim)]]>end) break
  }
  sim<-sim[1:(length(sim)-1)]
  sim<-lapply(sim,round)
  return(sim)
}
```

```
#severity follows lognormal distribution
severity<-function(dat,a,b){
  sev<-NULL
  for(i in 1:length(dat)) {
    sev1<-rlnorm(1,meanlog=a,sdlog=b)
    sev<-c(sev,sev1)
  }
  return(sev)
}

###reporting delay - exponential distribution ###
#returns dates of reporting
reporting<-function(dat,lambda){
  del<-NULL
  for(i in 1:length(dat)){
    delay<-rexp(1,rate=lambda)
    time<-dat[[i]]+delay
    del<-c(del,time)
  }
  del<-lapply(del,round)
  return(del)
}

###Simulations
set.seed(1) #here we use different seeds for different
   simulations
datum<-poisson(30,3650)
sev<-severity(datum,12,1.5)
rep<-reporting(datum,1/730)
#full database
mydata<-cbind(datum,sev,rep)
#database of claims reported before the end of 10th
   year
mydata2<-subset(mydata,rep<3651)
##export full database to csv
write.csv(mydata,file="mydata1.csv")
write.csv(mydata2,file="2mydata1.csv")
```

Further we proceeded similarly as with the real data set. We did not take into account the claims inflation. Here we show the source code just for one simulation, as it is then used for the others just with the appropriate values of the parameters. Deciding whether log-normal or gamma distribution is more suitable for claims severity was done using Kolmogorov-Smirnov test as described in (Anděl, 1998). For all simulations the results of this test were correctly in favor of using log-normal distribution which is the distribution also used for generating the data.

```
###triangle free approach for simulated data
require("MASS")
```

```
#occurence time follow poisson process
poisson<-function(lambda,end){
  sim<-rexp(1,rate=lambda)
  repeat{
    diff<-rexp(1,rate=1/lambda)
    time<-sim[length(sim)]+diff
    sim<-c(sim,time)
    if(sim[[length(sim)]]>end) break
  }
  sim<-sim[1:(length(sim)-1)]
  sim<-lapply(sim,round)
  return(sim)
}

#severity for given time- log-normal distribution
severity<-function(dat,a,b){
  sev<-NULL
  for(i in 1:length(dat)) {
    sev1<-rlnorm(1,meanlog=a,sdlog=b)
    sev<-c(sev,sev1)
  }
  return(sev)
}

#reporting delay - exponential distribution
#returns dates of reporting
reporting<-function(dat,lambda){
  del<-NULL
  for(i in 1:length(dat)){
    delay<-rexp(1,rate=lambda)
    time<-dat[[i]]+delay
    del<-c(del,time)
  }
  del<-lapply(del,round)
  return(del)
}

IBNR<-function(nsim,mu,lambda,mlog,slog,end){
  loss<-seq(0,0,length.out=nsim)
  for (i in 1:nsim){
    datesibnr<-poisson(mu,end)
                                    #generating claims
       occurrences
    severities<-severity(datesibnr,mlog,slog)    #
       generating severity of each claim
    rep<-reporting(datesibnr,1/lambda)
                                #generating reporting
       delay
    for(j in 1:length(datesibnr)){
      if(rep[[j]]>end){loss[[i]]<-loss[[i]]+severities
         [[j]]} #taking only IBNR
    }
  }
  return(loss)
}

totalloss<-function(nsim,mu,lambda,mlog,slog,end){
  loss<-seq(0,0,length.out=nsim)
  for (i in 1:nsim){
    datesibnr<-poisson(mu,end) #generating claims
       occurrences
    severities<-severity(datesibnr,mlog,slog) #
       generating severity of each claim
    rep<-reporting(datesibnr,1/lambda) #generating
       reporting delay
    for(j in 1:length(datesibnr)){
      loss[[i]]<-loss[[i]]+severities[[j]]
    }
  }
  return(loss)
}

#estimating parameters
data<-read.csv("C:/Users/Kika/Cloud_kika/diplomka/data/
   simulace/2mydata1.csv")
repdelay<-data$rep-data$datum
lambda<-mean(repdelay)

#severity distribution
fitdistr(data$sev,"log-normal")
ks.test(data$sev,"plnorm", 11.9288955, 1.4764946)
fitdistr(data$sev,"gamma",lower=c(0.01,0.01))
ks.test(data$sev,"pgamma", 1.51634e+03,1.000000e-02,
   exact=FALSE)

#estimate total loss and ibnr
set.seed(1)
ibnr<-IBNR
   (1000,30.62712,732.967,11.9288955,1.4764946,3650)
ibnrmean<-mean(ibnr)

set.seed(1)
tl<-totalloss
   (1000,30.62712,732.967,11.9288955,1.4764946,3650)
tlmean<-mean(tl)
```

```
#Chain Ladder
require("ChainLadder")
require("XLConnect")
triangle<-readWorksheetFromFile("C:/Users/Kika/Cloud_
   kika/diplomka/data/simulace/2mydata1analysis.xlsx",
   sheet="ct")
M<-MackChainLadder(triangle, est.sigma="Mack")
M
```