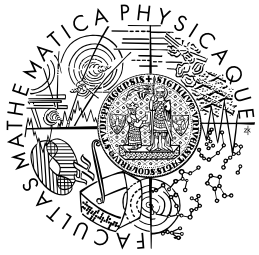

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Dominik Marko

Truncated data and stochastic claims reserving

Department of Probability and Mathematical Statistics

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

Prague 2018

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 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In Prague, 11th May 2018 Dominik Marko

Title: Truncated data and stochastic claims reserving

Author: Dominik Marko

Department: Department of Probability and Mathematical Statistics

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

Abstract: In this thesis stochastic claims reserving under the model of randomly truncated data is presented. For modelling the claims, a compound Poisson process is assumed. Introducing a random variable representing the delay between occurrence and reporting of a claim, a probability model of IBNR claims is built. The fact that some claims are incurred but not reported yet leads to truncated data. Basic results of non-parametric statistical estimation under the model of randomly truncated data are shown, which can be used to obtain an estimate of IBNR claims reserves. Theoretical background is then used for application on real data from Czech Insurers' Bureau.

Keywords: Claims reserving, truncated data, IBNR claims, non-life insurance

I would like to thank my supervisor RNDr. Michal Pešta, Ph.D. for his guidance, valuable advice and time.
I am also very grateful to my family and friends for their support.

Contents

Introduction

Estimation of claims reserves presents an important problem handled in insurance. A lot of methods use aggregate data, in this work information about individual claims development will be used for estimating claims reserves. The purpose of this thesis is to build on available theory for models of randomly truncated data and use this theory on application for claims reserving, demonstrated on the real insurance data.

The thesis is decomposed into four chapters. In the first chapter, the problem of claims reserving in insurance is presented, where the necessity to build reserves for claims that have incurred but have not been reported yet is explained.

Claims are modelled by a compound Poisson process with variable intensity. Reporting delay, which is an important driver of IBNR claims, plays a key role. It is shown that the claims data can be represented in a way that they follow previously mentioned model of randomly truncated data.

The aim of second chapter is to describe the problem arising from truncated data. In the models of randomly truncated data, the estimation of the cumulative distribution function is not that simple because of the fact that the sample does not come from the required distribution. The data are observed conditionally, truncated random variable is observed only if it takes higher values than the truncating random variable. The motivation on deductibles policy in insurance and also physical motivation coming from astronomy is shown. Truncated data are often used in such fields as astronomy, survival studies or demography. Theory of non-parametric statistical estimation under this model is discussed and derivation of maximum likelihood estimates of desired cumulative distribution functions of truncated and truncating variables is shown.

After that, in third chapter we build on the results from first two chapters. We reduced the problem to the issue of determining the joint probability distribution of the delay variable and claim size variable. Using the results from second chapter, we get a non-parametric maximum likelihood estimate of this joint cumulative distribution function. Then joining all the content together, an estimate of IBNR claims reserves claims can be obtained. It is also possible to construct confidence bounds for IBNR claims reserves. If specific assumptions can not be met, a construction of a generalization of the estimate is shown too.

In the last chapter, data provided by Czech Insurers' Bureau are used to demonstrate the application of the theory presented in the thesis. Description and analysis of data is provided, followed by results obtained by methods described in the work. Comparison with a Chain-ladder method, which uses aggregated data, is shown. In the end, results of the practical application are discussed.

1. Claims Reserving Problem

1.1 Introduction to claims reserving

Claims reserving is one of the most important and crucial problems handled in non-life insurance. Non-life insurance, also known as General Insurance in UK or Property and Casualty Insurance in USA, contains usually all kinds of insurance products not covered by life insurance. It is split into different Lines of Business such as motor/car insurance, accident insurance, property insurance, liability insurance, etc. These Lines of Business can further be divided into more sub-classes.

Claims in non-life insurance have a typical feature that they cannot be settled right after occurrence. There are three main reasons, why the insurance company is not able to settle the claim immediately. Firstly, there is usually a reporting delay, which is the time-lag between the day of the claim occurrence and the day when claim is reported to the insurer. The length of the reporting delay depends on the type of the claim, it can be just a few days, but also several years. Second reason is that after reporting, it also takes some time until the claim is settled. Third reason is that the closed claim can be reopened again due to new unexpected developments.

As a consequence of these delays, insurance companies have to make claims reserves. Claims reserves are usually the biggest item on the liability side of a typical non-life insurance company. Because of this, the task of their proper determination is of great importance.

It is necessary for insurance companies to build both premium reserves for future exposures and claims reserves for unsettled claims of past exposures. We distinguish between two different types of claims reserves for past exposures:

1.IBNR reserves (Incurred But Not yet Reported): These are the claims reserves for claims that have already occurred, but have not yet been reported to the insurance company.

2.RBNS reserves (Reported But Not Settled): Claims reserves for claims that have been reported to the insurer, but have not yet been settled. During the settlement period (the period between reporting date and settlement date), more information is being received and successive claims payments are happening.

Until 1970s, the common approach to claims reserves related to the area of reserves for known claims. The practice was that an expert estimator looked at individual claims and made assessments of their value. These estimated values might have been revised with time passing and more information coming up. The sum of all of these individual case estimates provided the estimated claims reserve. Later, from the early 1970s, the problem of estimating claims reserves started to be solved with algorithmic methods. Methods like Chain-ladder or Bornhuetter-Ferguson method have been developed. The downside of these methods was that they didn't provide information about the uncertainty of the estimate. While algorithmic methods provide only an expected value of the reserve, stochastic methods give also higher moments of the reserve, sometimes even the full distribution. Therefore actuaries started to develop and analyze underlying stochastic models that justified these algorithms. One of the first stochastic models was de-

veloped by Mack in 1993, which reproduced chain-ladder estimates. For a broader overview of methods used for calculating claim reserves estimates, we recommend looking at [Wüthrich and Merz, 2008].

1.2 IBNR claims reserving

In our work, we will deal with the first mentioned type of claims reserves for past exposures and that is IBNR reserves. Estimation of loss reserves for IBNR claims is an important task for any insurance company to have the right picture of its liabilities. Improper handling of the IBNR claims can lead and has led to the ruin of insurance companies and placed others in serious financial jeopardy. There are many approaches to the task of calculating estimates of IBNR claims reserves, in this thesis we show one proposed in [Herbst, 1999]. Remainder of this chapter is based on this article.

Nonhomogeneous Poisson process

Before we start with stating model assumptions, let us first define nonhomogeneous Poisson process we are using for modelling the claims. We take the definition of this process from [Ross, 2010].

Definition 1. *Counting process* $\{N(t), t \geq 0\}$ *is a nonhomogeneous Poisson process with intensity function* $\lambda(t),\ t \geq 0$*, if*

$$\begin{aligned} &(i)\ \ N(0) = 0 \\ &(ii)\ \{N(t), t \geq 0\} \textit{ has independent increments.} \\ &(iii) P[N(t+h) - N(t) \geq 2] = o(h) \\ &(iv)\ P[N(t+h) - N(t) = 1] = \lambda(t)h + o(h) \end{aligned} \tag{1.1}$$

for all $t \geq 0$ *and* $h > 0$*. The function* $o(h)$ *satisfies*

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

1.2.1 Model assumptions

Now we will state a few assumptions we will work with throughout this work. Assume that the ith claim of the size Y_i occurred at the moment Σ_i and that the delay between occurrence and reporting this claim is D_i, which means the reporting date is $\Sigma_i + D_i$. In addition, we suppose that Σ_i, Y_i and D_i are non-negative random variables. Another assumption we make is that the number of claims occurred up to time $t > 0$ is modelled by a nonhomogeneous Poisson process $N(t)$ with intensity function $\lambda(t)$. We may interpret the intensity $\lambda(t)$ as a measure of risk exposure at the moment t. One can easily realize that $N(t) = \sum_{i=1}^{\infty} I[\Sigma_i \leq t]$. In this work, $I[.]$ denotes indicator function. Finally, our last assumption is that the pairs (Y_i, D_i), $i = 1, 2, ...$, are independent identically distributed random vectors, independent of the process $\{N(t), t \geq 0\}$, with the joint cumulative distribution function F_{YD}.

Given the times $0 < \tau_1 < \tau$ and the observed values (Y_i, D_i, Σ_i) for the claims reported up to time τ, the goal is to estimate the value $\mathrm{IBNR}(\tau_1, \tau)$, which represents the volume of claims which occurred in the interval $[0, \tau_1]$, but were not reported up to time τ. We suppose that $0 < D_i < \tau$ with probability 1. The value $\mathrm{IBNR}(\tau_1, \tau)$ might be expressed as

$$\mathrm{IBNR}(\tau_1, \tau) = \sum_{i=1}^{\infty} I[\Sigma_i \leq \tau_1, \Sigma_i + D_i > \tau] Y_i. \tag{1.2}$$

When estimating the value of $\mathrm{IBNR}(\tau_1, \tau)$, we can only use information available up to the time τ, which means using only those observations of triples (Y_i, D_i, Σ_i) for which $\Sigma_i + D_i \leq \tau$. Suppose there exist n such triples and denote them by $(Y_i^*, D_i^*, \Sigma_i^*), i = 1, ..., n$.

1.2.2 Representation of the data

Now we will show how the data can be represented in the form of sorted independent identically distributed random vectors. Assume that $N(\tau) = N$ is given. It can be verified by direct calculation of the corresponding distribution functions, that the joint distribution of the random vector $(\Sigma_1, ..., \Sigma_N)$ coincides with the distribution of the vector of order statistics $(S_{(1)}, ..., S_{(N)})$ based on N independent identically distributed random variables $S_1, ..., S_N$, independent of (Y_i, D_i), with the cumulative distribution function G given by

$$G(t) = \begin{cases} 0 & \text{for } t < 0, \\ \frac{\int_0^t \lambda(s)ds}{\int_0^\tau \lambda(s)ds} & \text{for } t \in [0, \tau], \\ 1 & \text{for } t > \tau. \end{cases} \tag{1.3}$$

Put $p(i) = k$ if and only if $S_k = S_{(i)}$. The vector $(p(1), ..., p(N))$ represents a random permutation of $(1, ..., N)$, independent of the pairs $(Y_i, D_i), i = 1, ..., N$. As a consequence, vectors $(Y_i, D_i), i = 1, ..., N$ and $(Y_{p(i)}, D_{p(i)}), i = 1, ..., N$ are identically distributed and it is easy to realize that also the distribution of $(Y_i, D_i, \Sigma_i), i = 1, ..., N$ is the same as the distribution of $(Y_{p(i)}, D_{p(i)}, S_{(i)})$, $i = 1, ..., N$. One can realize that $(Y_{p(i)}, D_{p(i)}, S_{(i)})$, $i = 1, ..., N$ are the triples (Y_i, D_i, S_i) sorted in ascending order by the coordinate S_i.

From the above-mentioned considerations, it follows that the observed triples $(Y_i^*, D_i^*, \Sigma_i^*)$ are identically distributed as the triples (Y_i, D_i, S_i), for which $S_i + D_i \leq \tau$, sorted by the third coordinate S_i in ascending order. In the previous fashion, let us denote these triples by (Y_i^*, D_i^*, S_i^*), $i = 1, ..., n$.

It follows from the previously mentioned arguments that, given $N(\tau) = N$, the claims data can be represented in the form of independent identically distributed triples $(Y_i, D_i, S_i), i = 1, ..., N$. It is easily seen that, given $N(\tau) = N$, the formula for IBNR defined earlier in (1.2) is identically distributed as

$$\mathrm{IBNR}(\tau_1, \tau) = \sum_{i=1}^{N} I[S_i \leq \tau_1, S_i + D_i > \tau] Y_i. \tag{1.4}$$

1.2.3 Estimation of IBNR claims when both G and F_{YD} are known functions

Firstly, we will assume that we know both cumulative distribution functions G and F_{YD} and that $N(\tau) = N$ is fixed, but unknown. This means we deal with conditional distribution given N. Although these assumptions are rarely met in practice and getting the results under such circumstances would not be of great importance, it is a good first step towards our goal of estimating IBNR claims under more realistic assumptions, which will be stated later.

Our aim is to estimate IBNR claims (1.4) given the observed triples $(Y_i^*, D_i^* S_i^*)$, $i = 1, ..., n$. The best estimate of (1.4) in the sense of least squares is its conditional expectation

$$\mathsf{E}\left[\mathrm{IBNR}(\tau_1, \tau) | (Y_i^*, D_i^*, S_i^*), i = 1, ..., n\right] = (N - n)\,\mathsf{E}\left[Y_i\ I[S_i \leq \tau_1] | S_i + D_i > \tau\right]. \tag{1.5}$$

By n we denote a random variable that represents the number of those triples (Y_i, D_i, S_i) for which $S_i + D_i \leq \tau$. Obviously it has binomial distribution with parameters N and $P[S_i + D_i \leq \tau]$. This means that using a maximum likelihood argument we can estimate N by

$$\widehat{N} = \frac{n}{P[S_i + D_i \leq \tau]}.$$

We have

$$\mathsf{E}\left[Y_i I[S_i \leq \tau_1] | S_i + D_i > \tau\right] = \frac{\int_0^\infty \int_{\tau - \tau_1}^{\tau} s(G(\tau_1) - G(\tau - t)) dF_{YD}(s, t)}{P[S_i + D_i > \tau]}$$

and

$$P[S_i + D_i \leq \tau] = \int_0^\tau G(\tau - t) dF_{YD}(\infty, t),$$

which leads us to the following estimate of IBNR:

$$\widehat{\mathrm{IBNR}_0}(\tau_1, \tau) = n \frac{\int_0^\infty \int_{\tau - \tau_1}^{\tau} s(G(\tau_1) - G(\tau - t)) dF_{YD}(s, t)}{\int_0^\tau G(\tau - t) dF_{YD}(\infty, t)}. \tag{1.6}$$

Since it is too optimistic to assume that the joint cumulative distribution function F_{YD} is known, this assumption will be dropped. In this case, in order to construct an estimate of $\mathrm{IBNR}(\tau_1, \tau)$, one has to find an estimate of F_{YD} based on the observed data and insert it into formula (1.6). It is necessary to realize that the observed claim sizes and delays (Y_i^*, D_i^*) do not follow the joint distribution F_{YD}, since they are only observed conditionally under $\Sigma_i + D_i \leq \tau$.

This fact leads us to using the theory of non-parametric statistical estimation under the model of randomly truncated data. We will discuss this in the following chapter.

2. Truncated data

2.1 Description and motivation

In this chapter, we will talk about the problem of estimating a distribution function under the model of randomly truncated data. Let us first show the definition and motivation.

In non-life insurance, it is a common policy to use deductibles, which means if the loss is X and the deductible is d, the insurance company covers only $\max(X - d, 0)$. Losses below d are not reported to the company, so the data contain only those losses, which are bigger than the deductible, i.e. $X \geq d$. Such data are called left truncated data. Data can be also right truncated. This kind of truncation, when deductible d is not a random variable, is non-random. If deductible was a random variable, we would have randomly truncated data.

In this thesis, we work with randomly truncated data (truncated from the left). Let X, V be independent positive random variables with cumulative distribution functions F and G, respectively. We suppose that we observe only those pairs (X, V), for which $X \geq V$, otherwise the observation is lost or truncated. Our aim is to find the estimators of cumulative distribution functions F and G. Although we used notation of G before, later we will see that it is always used for denoting cumulative distribution function of truncating variable.

2.1.1 Physical motivation

Now we are going to show a physical motivation taken from [Lynden-Bell, 1971], which leads to truncated data. In his work, Lynden-Bell developed a method used for deriving luminosity functions and density evolutions from data subject to observational selection. The distribution of ratios of radio power to optical power for quasars is derived from the 3CR quasars. The density evolution is derived from the data and is then used to determine the optical luminosity of quasars.

Dim objects can be only seen nearby and bright objects are rare, so in a sample complete to a given apparent magnitude, both classes may contribute significantly. Even when having the observed distribution of intrinsic luminosities, there is a problem in deriving the luminosity function. This function describes how many objects of each luminosity there are in a given volume rather than in the observed sample, which has been biased by the cut off in apparent magnitude.

It is shown how, on the assumptions that the luminosity function is of the same form at all points along the line of sight but with a normalization that varies as the number density, the luminosity function and the number density can be deduced from the observations of the sample.

Let us define:
i) f_0 the optical flux f_V at the wavelength to which 2500 $\mathring{A}$ would be red-shifted due to the recession of the object, with the observer being at the Earth.
ii) f_R the radio flux that would be observed under similar circumstances at the red-shifted 500 MHz frequency.
iii) F_0 and F_R the total fluxes emitted by the object at 2500 $\mathring{A}$ and 500 MHz

respectively.
iv) P the log radio to optical power ratio, $\log_{10} f_R/f_0$. On the assumptions that the fluxes are emitted isotropically from the source and fluxes f_0 are freed from absorption, the diminution of flux with distance is the same in the radio and optical region, so

$$P = \log f_R/f_0 = \log F_R/F_0.$$

v) S(178) is the actually observed 3CR flux at 178 MHz.
vi) z is the redshift of the source.

The main problem is determining the distribution of P among the quasars. The data are represented by observed fluxes of 40 3CR quasi-stellar sources. They do not form an unbiased sample of quasars in any given volume, but are bright enough in the radio region to be included in 3CR and bright enough to have their optical identifications confirmed with redshifts. We cannot remove the bias caused by selection of apparently bright objects without further assumption.

Schmidt [1968] has found that the distribution of the number of quasars with P is independent of their absolute optical flux F_0. His finding on the distribution of P, joined with assumption on the normalization of the luminosity function, leads to the following luminosity function Φ:

$$\Phi = D(z)\phi_1(P)\phi_2(\log F_0), \tag{2.1}$$

where ϕ_1 and ϕ_2 are distribution functions of their arguments, $D(z)$ is the co-moving number-density of quasars as a function of redshift z.

We define for each source V, the comoving volume of that part of the Universe nearer than the source at the current epoch. The comoving volumes V nearer than each quasar are calculated from the redshifts. The comoving number-density of quasars is the number per unit comoving volume.

Coming back to 2.1, for objects situated at a distance corresponding to redshift z the comoving number density of quasars with log power ratios in the range P to $P + dP$ and with log optical flux between $\log F_0$ and $\log F_0 + d\log F_0$ is

$$\Phi dP d\log F_0 = D(z)\Phi_1(P)\Phi_2(\log F_0) dP d\log F_0. \tag{2.2}$$

This product form of luminosity function implies that the distribution of P is uncorrelated with apparent optical magnitude (it is a function of $\log F_0$ and z, both of them uncorrelated with P in the above equation). Independence of the distribution of P to a selection by optical apparent magnitude means that the only bias in the observed sample is through the radio selection that the source must be bright enough to be included in $3CR$. The selection is quantified with factor β by which each source's radio flux has been reduced before its S(178) flux would be reduced below the 9 flux unit limit of 3CR. Such object, that had its radio flux reduced in this way, would have log power ratio $P_m = P - \log\beta$, therefore the selection on the value of P of each object is $P \geq P_m$.

Given a number of observed values of P each associated with a quasar and a known limit P_m such that only those quasars, for which $P \geq P_m$, are included in the observed sample, Lynden-Bell in his paper deduced the true distribution of P among quasars. The only bias in the sample is $P \geq P_m$.

The estimation is solved with different nomenclature for the variables P and P_m. Let $L = P = \log \frac{F_R}{F_0}$ and denote by M the limiting value of L beyond which an object of the same optical flux is rejected from 3CR because of being too radio weak. The question is to find the distribution ϕ of an observable L from the known pairs of values (L, M), for which $L \geq M$, with no further bias. This problem is exactly the one we are dealing with in our thesis, which is estimating a distribution function under the model of randomly truncated data. Lynden-Bell in his work first introduced the product limit estimate of the aforementioned distribution function.
The derivations in [Lynden-Bell, 1971] leading to the estimators are not shown in our work, instead we choose to show the derivations from [Woodroofe, 1985], where he used another motivation, which arises from astronomy too.

The absolute and apparent luminosities of an astronomical object are defined as the brightness of this astronomical object at a fixed distance and as observed on earth. Magnitude is defined as the negative logarithm of luminosity. In some models it is assumed that the redshift z and the absolute magnitude M of astronomical objects are independent random variables related to the apparent magnitude m by the equation

$$m = f(z) + M, \tag{2.3}$$

where f is a known or at least a nearly known function. It is natural to expect, that it is possible to detect only those objects which are sufficiently bright, say $m \leq m^*$, which gives $M - m^* \leq -f(z)$. Putting $X = \exp[-f(z)]$ and $V = \exp[M - m^*]$ gives the model of truncated data.

Now we proceed with an approach proposed by [Woodroofe, 1985].

2.2 Formulation of the problem

Denote by H_* the joint distribution function of (X^*, V^*), i.e. the distribution of (X, V) given $X \geq V$, where X and V are independent, positive random variables with distribution functions F and G, taken to be continuous from the right, and by F_*, G_* the marginal distribution functions of X and V given $X \geq V$. Obviously

$$\begin{gathered} H_*(x, v) = P(X \leq x, V \leq v | X \geq V) = \alpha^{-1} \int_0^x G(v \wedge z) dF(z), \\ F_*(x) = H_*(x, \infty) \text{ and } G_*(v) = H_*(\infty, v), \ \ 0 \leq x, v < \infty, \end{gathered} \tag{2.4}$$

where $v \wedge z$ denotes the minimum of v and z, $\alpha = P(X \geq V) = \int_0^\infty G(z) dF(z) = \int_0^\infty [1 - F(z-)] dG(z)$ and $F(z-) = P(X < z)$. We assume that $\alpha \in (0, 1)$. Given the data in form of observed pairs (X_i^*, V_i^*), we want to find consistent estimators of F and G.

For an arbitrary distribution function W defined on $[0, \infty)$, define

$$a_W = \inf\{z > 0 : W(z) > 0\} \geq 0$$

and

$$b_W = \sup\{z > 0 : W(z) < 1\} \leq \infty.$$

Then (a_W, b_W) is the interior of the convex support of W. We assume that $a_G < b_F$ for $\alpha > 0$. If $\alpha > 0$ and if the relationship between F_* and G_* and F and G is defined by 2.4, then $a_{F_*} = \max\{a_F, a_G\}$, $b_{F_*} = b_F$, $a_{G_*} = a_G$ and $b_{G_*} = \min\{b_F, b_G\}$. Let us denote

$$\begin{aligned} K &= \{(F,G) : F(0) = 0 = G(0),\ \ \alpha(F,G) > 0\}, \\ K_0 &= \{(F,G) \in K : a_G \le a_F,\ \ b_G \le b_F\}, \\ T(F,G) &= H_*,\ \ (F,G) \in K. \end{aligned}$$

Lemma 1. *(i) Assume that $(F,G) \in K$ and let F_0 and G_0 denote the conditional distributions of X and V given $X \ge a_G$ and $V \le b_F$. Then $(F_0, G_0) \in K_0$ and $T(F_0, G_0) = T(F,G)$.*
ii) $T(K) = T(K_0)$.

Proof. (from Woodroofe [1985]) $V \le X$ implies $X \ge a_G$ and $V \le b_F$ w.p. 1, thus $T(F,G) = T(F_0,G_0)$. Since $(F,G) \in K$, $a_{G_0} = a_G$ and $a_{F_0} = \max(a_F, a_G) \ge a_G = a_{G_0}$. By the same argument, one can show that $b_{G_0} \le b_{F_0}$, which completes the proof of (i). Since $K \subset K_0$, (ii) follows.

□

We define hazard rate of a random variable X with distribution function F and density f as

$$\begin{aligned} \lambda(x) &= \lim_{h \downarrow 0} \frac{1}{h} P(x < X \le x + h | X > x) \\ &= \frac{\lim_{h \downarrow 0} \frac{1}{h} P(x < X \le x + h)}{P(X > x)} = \frac{f(x)}{1 - F(x)}. \end{aligned} \tag{2.5}$$

It can be interpreted as the intensity that the random variable X is x, given that it is at least equal to x. The cumulative hazard function of a distribution function F(with $F(0) = 0$) is defined as

$$\Lambda(x) = \int_0^x \lambda(z) dz \ \ 0 \le x < \infty.$$

One can realise that

$$\lambda(x) = -\frac{d}{dx} \log[1 - F(x)]$$

and integrating from 0 to x, taking into account that $F(0) = 0$, gives us

$$\int_0^x \lambda(z) dz = -[\log[1 - F(z)]]_{z=0}^x = -\log[1 - F(x)],$$

which leads us to

$$1 - F(x) = e^{-\Lambda(x)}.$$

Cumulative hazard function can be decomposed into a continuous and a discrete parts. Let D denote the set of x for which $0 \le x < b_F$ and $\lambda(x) = \Lambda(x) - \Lambda(x-) > 0$, and $\Lambda_c(x) = \Lambda(x) - \sum_{s \in D, s \le x} \lambda(s)$, $0 \le x < b_F$. Then the distribution function F is uniquely determined with the cumulative hazard function Λ by the relationship

$$1 - F(x) = (\prod_{s \in D, s \le x} [1 - \lambda(s)]) \exp[-\Lambda_c(x)],\ \ 0 \le x < b_F. \tag{2.6}$$

Theorem 2. *Suppose that $H_* \in T(K)$. Then there is a unique pair $(F, G) \in K_0$ for which $T(F, G) = H_*$. The pair (F, G) is determined by following conditions*

$$\Lambda(x) = \int_0^x \frac{dF_*(z)}{C(z)}, \ \ 0 \leq x < \infty, \tag{2.7}$$

and

$$\int_v^\infty \frac{dG(z)}{G(z)} = \int_v^\infty \frac{dG_*(z)}{C(z)}, \ \ 0 \leq v < \infty,$$

where

$$C(z) = G_*(z) - F_*(z), \ \ 0 \leq z < \infty.$$

Proof of the theorem can be found in [Woodroofe, 1985].

Let $(F, G) \in K$ and let F_0, G_0 be the same conditional distributions as defined in Lemma 1. From Lemma 1, we have $(F_0, G_0) \in K_0$ and $T(F_0, G_0) = T(F, G)$. Theorem 2 also says, that there is only one such pair, which gives us that (F_0, G_0) is the only pair in K_0 for which $T(F_0, G_0) = T(F, G)$.

2.3 Estimation

Again, we assume that X and V are independent, positive random variables with distribution functions F and G for which $(F, G) \in K$. We also assume that $(X_1, V_1), ..., (X_N, V_N)$ are i.i.d. as (X, V). Suppose that only those pairs (X_i, V_i), for which $i \leq N$ and $V_i \leq X_i$, can be observed and that there is at least one such pair. Denote these observed pairs $(X_1^*, V_1^*), ..., (X_n^*, V_n^*)$, so labeled that observed pairs $(X_1^*, V_1^*), ..., (X_n^*, V_n^*)$, are conditionally i.i.d. given n. Given this sample of observed pairs, our aim is to find the estimators of F and G.

Let F_n^* and G_n^* denote the empirical distribution functions of $X_1^*, ..., X_n^*$ and $V_1^*, ..., V_n^*$:

$$\begin{aligned} F_n^*(z) &= \frac{1}{n} \#\{i \leq n : X_i^* \leq z\}, \\ G_n^*(z) &= \frac{1}{n} \#\{j \leq n : V_j^* \leq z\}, \ \ 0 \leq z < \infty, \end{aligned}$$

where $\#A$ denotes the cardinality of a set A. These empirical distribution functions estimate the conditional distribution functions F_* and G_*. Since we have these estimates, we can now find estimators of F and G by using the inversion formula of Theorem 2. In $C(z)$ defined in Theorem 2, we substitute F_*, G_* with their estimates F_n^*, G_n^* to get the following estimate of $C(z)$:

$$C_n(z) = G_n^*(z) - F_n^*(z-), \ \ 0 \leq z < \infty.$$

It is obvious that $C_n(X_i^*) \geq \frac{1}{n} \ \forall \ i \leq n$. Looking back at Theorem 1, the cumulative hazard function Λ might be estimated by

$$\widehat{\Lambda}_n(z) = \int_0^z \frac{dF_n^*(x)}{C_n(x)} = \sum_{i: X_i^* \leq z} \frac{1}{nC_n(X_i^*)}, \ \ 0 \leq z < \infty. \tag{2.8}$$

$\widehat{\Lambda}_n$ is a step function with discontinuities at $X_1^*, ..., X_n^*$. Looking at equation 2.6, we come to the conclusion that F can be estimated by

$$\widehat{F}_n(z) = 1 - \prod_{i:X_i^* \leq z} \Big[1 - \frac{r(X_i^*)}{nC_n(X_i^*)}\Big], \;\; 0 \leq z < \infty, \tag{2.9}$$

where $r(X_i^*) = \sum_{k=1}^n I[X_k^* = X_i^*]$ for $1 \leq i \leq n$. The product extends over distinct values of $X_1^*, ..., X_n^*$ and an empty product is defined to be 1.

Similar construction is possible for the estimation of G and after some algebra we arrive to the following estimator of G

$$\widehat{G}_n(z) = \prod_{j:V_j^* > z} \Big[1 - \frac{s(V_j^*)}{nC_n(V_j^*)}\Big], \;\; 0 \leq z < \infty, \tag{2.10}$$

where $s(V_j^*) = \sum_{k=1}^n I[V_k^* = V_j^*]$ for $1 \leq j \leq n$. Let us note, that these product limit estimators of distribution functions F and G were first introduced by Lynden-Bell [1971].

In section 4 in [Keiding and Gill, 1990] it is shown that $(\widehat{F}_n, \widehat{G}_n)$ is the non-parametric maximum likelihood estimator of (F, G), which is a direct consequence of the embedding of the left truncation into the Markov process model, for which there exist results on non-parametric maximum likelihood estimation. Firstly, it is shown that previously derived estimate $\widehat{F}_n$ (defined in 2.9) of the distribution function F maximizes the conditional likelihood of X^* given V^*, i.e. it maximizes

$$L_{X^*|V^*} = \prod_{i=1}^n \frac{dF(X_i^*)}{1 - F(V_i^*)},$$

so it is the conditional non-parametric maximum likelihood estimator of F.

Similarly, also estimate $\widehat{G}_n$ (defined in 2.10) of the distribution function G maximizes the conditional likelihood of V^* given X^*, hence it is the conditional non-parametric maximum likelihood estimator of G.

For simplicity, suppose that there are no ties among $X_1^*, ..., X_n^*$, $V_1^*, ..., V_n^*$ and consider estimating distribution functions F and G by distributions that are supported by $\{X_1^*, ..., X_n^*\}$ and $\{V_1^*, ..., V_n^*\}$. For such distributions, provided that 2.11 (defined below) does not happen, estimators $\widehat{F}_n$ and $\widehat{G}_n$ maximize the full likelihood

$$L_{X^*,V^*} = \alpha^{-n} p_1 \times ... \times p_n \times q_1 \times ... \times q_n$$

with respect to $p_1, ..., p_n$ and $q_1, ..., q_n$, where $p_1, ..., p_n$ and $q_1, ..., q_n$ are the masses assigned to $X_1^*, ..., X_n^*$, $V_1^*, ..., V_n^*$. Since $\widehat{F}_n$ and $\widehat{G}_n$ maximize the full likelihood, they are unconditional maximum likelihood estimates of F and G.

The estimators $\widehat{F}_n$ and $\widehat{G}_n$ may be supported by proper subsets of $\{X_1^*, ..., X_n^*\}$ and $\{V_1^*, ..., V_n^*\}$. Order the values of $X_1^*, ..., X_n^*$ and $V_1^*, ..., V_n^*$ and get $X_{(1)}^* < ... < X_{(n)}^*$ and $V_{(1)}^* < ... < V_{(n)}^*$.

It may happen that

$$nC_n[X_{(k)}^*] = 1, \;\; \text{for some } k, \;\; 1 \leq k < n. \tag{2.11}$$

in which case

$$\widehat{F}_n[X_{(k)}] = 1,$$

which is an undesirable property of the estimator, that can lead to unreasonable estimates.

We can solve this problem in following if ad hoc way. Let k_n be a nonincreasing function for which $k_n(x) > k_n[X^*_{(n)}] = \frac{1}{n}$ for all $x < X^*_{(n)}$. If C_n is replaced by

$$C^{\#}_n(z) = \max\{C_n(z), k_n(z)\}, \;\; 0 \leq z \leq x_{(n)},$$

in 2.9, then the resulting estimator, which we denote by $F^{\#}_n$, is not supported by any proper subset of $\{X^*_1, ..., X^*_n\}$, with $\frac{1}{nk_n[X^*_{(i)}]}$ being the maximum proportion of the estimated probability $1 - F^{\#}_n[X^*_{(i)}-]$ which we are willing to assign to $X^*_{(i)}$ for $i = 1, ..., n$.

One can estimate $\alpha = P[V \leq X]$ from the observed pairs (X^*_i, V^*_i), $i = 1, ..., n$ by following non-parametric maximum likelihood estimator of α:

$$\widehat{\alpha}_n = \int_0^\infty \widehat{G}_n d\widehat{F}_n.$$

It is not hard to realize that $\widehat{\alpha}_n > 0$ if $nC_n[X^*_{(i)}] > 1$ for all $i \leq n-1$. Otherwise we replace $\widehat{F}_n$ and $\widehat{G}_n$ with $F^{\#}_n$ and $G^{\#}_n$. From the estimated α, it is possible to estimate N by

$$\widehat{N} = \frac{n}{\widehat{\alpha}_n}.$$

If the distribution function G cannot vary freely, Vardi[Vardi, 1985] has shown that if G is known, the non-parametric maximum likelihood estimate of F is

$$\tilde{F}(t) = \sum_{i=1}^{n} G(X^*_i)^{-1} I[X^*_i \leq t] \Bigg(\sum_{i=1}^{n} G(X^*_i)^{-1} \Bigg)^{-1}. \tag{2.12}$$

Vardi's result was further generalized by [Wang, 1989], who noticed that if G varies across a parametric family $G = \{G_\theta, \theta \in \Theta\}$, then the unconditional nonparametric maximum likelihood estimate of F can be obtained from 2.12 by replacing G by an estimate $G_{\widehat{\theta}}$, where $\widehat{\theta}$ is the maximum likelihood estimate of θ derived from the conditional distribution of V^* given X^* (based on observations of V^*_i's given X^*_i's). These results suggest an idea to alternatively obtain the maximum likelihood estimate of F under unknown G by replacing G in 2.12 by its maximum likelihood estimator $\widehat{G}_n$ defined in 2.10 (justification of this approach is shown in [Keiding and Gill, 1990]). Then we have the following estimator of F:

$$\widehat{F}(t) = \sum_{i=1}^{n} \widehat{G}(X^*_i)^{-1} I[X^*_i \leq t] \Bigg(\sum_{i=1}^{n} \widehat{G}(X^*_i)^{-1} \Bigg)^{-1}. \tag{2.13}$$

It is shown in the references given in this chapter that all the estimates are consistent and asymptotically normal.

3. Estimation of IBNR claims under the model of randomly truncated data

In this chapter we join content from the previous two chapters. In the first chapter, we deduced a formula for calculating the IBNR claims reserves. In the second chapter, the nonparametric maximum likelihood estimators of unknown distribution functions under the model of randomly truncated data were derived. Now we can join these results to show how to use truncated data for calculating the estimate of IBNR claims. In subsection 1.2.3 we assumed that the joint cumulative distribution function F_{YD} of random vectors $(Y_i, D_i), i = 1, 2, ...$ is known. This assumption is now dropped and estimation of IBNR claims under unknown F_{YD} will be shown. We will separate cases when we assume that the distribution function G is known and when we assume it is unknown. This chapter uses results derived by [Herbst, 1999].

3.1 Estimation of IBNR claims under known G

Many times we may assume that the intensity function $\lambda(t)$ of the Poisson process, which we are using for modelling the claims occurrences is of the following form

$$\lambda(t) = \lambda\eta(t), \tag{3.1}$$

where $\eta(t)$ denotes known number of insured risks at the time t, while λ is an unknown constant such that $\lambda > 0$. Then the cumulative distribution function G given in 1.3 is known and is of the form

$$G(t) = \frac{\int_0^t \eta(s)ds}{\int_0^\tau \eta(s)ds}, \ \ t \in [0, \tau]. \tag{3.2}$$

Function G of form 3.2 can be interpreted as a ratio of cumulative time of risk exposure in the interval $[0, t]$ (sum of exposure times for all the insured risks) to the cumulative time of risk exposure in the whole observed interval $[0, \tau]$.

Now the only thing we are missing in order to finish the estimation of IBNR claims is finding an estimate of joint cumulative distribution function F_{YD}, which we could insert into 1.6. Estimate of F_{YD} has to be based on the claims data available up to time τ, which consist of the observed triples $(Y_i^*, D_i^*, \Sigma_i^*)$, $i = 1, ..., n$, or equivalently (Y_i^*, D_i^*, S_i^*), $i = 1, ..., n$. Using results derived in previous chapter, a maximum likelihood estimate of F_{YD} can be obtained as an easy generalization of the estimate 2.12.

Let $X_i^* = \tau - D_i^*$ and $V_i^* = S_i^*$. From the results by [Vardi, 1985], it follows that the maximum likelihood estimate of the cumulative distribution function of $X_i = \tau - D_i$ (see 2.12) is a step function with atoms of the size $G(\tau - D_i^*)^{-1}\left(\sum_{i=1}^n G(\tau - D_i^*)^{-1}\right)^{-1}$ at the points $\tau - D_i^*$, $i = 1, ..., n$. Consequently, the maximum likelihood estimate of the distribution function of D_i,

i.e. the cumulative distribution function $F_{YD}(\infty, t)$, has atoms of the size $G(\tau - D_i^*)^{-1}\Big(\sum_{i=1}^n G(\tau - D_i^*)^{-1}\Big)^{-1}$ at the points D_i^*, $i = 1, ..., n$.

It follows that the likelihood function

$$L_{D^*,S^*} = \prod_{i=1}^{n} \frac{dF_{YD}(\infty, D_i^*)dG(S_i^*)}{\int_0^\tau G(\tau - t)dF_{YD}(\infty, t)}$$

is maximized by a function with jumps of the size $G(\tau - D_i^*)^{-1}\Big(\sum_{i=1}^n G(\tau - D_i^*)^{-1}\Big)^{-1}$ at the points D_i^*, $i = 1, ..., n$. One can easily realize that as a consequence of this, the likelihood function

$$L_{Y^*,D^*,S^*} = \prod_{i=1}^{n} \frac{dF_{YD}(Y_i^*, D_i^*)dG(S_i^*)}{\int_0^\tau G(\tau - t)dF_{YD}(\infty, t)}$$

is maximized by a distribution function $\tilde{F}_{YD}$ with atoms of the size $G(\tau - D_i^*)^{-1}\Big(\sum_{i=1}^n G(\tau - D_i^*)^{-1}\Big)^{-1}$ at the points (Y_i^*, D_i^*), $i = 1, ..., n$. Thus, the estimate of F_{YD} takes form

$$\tilde{F}_{YD}(s, t) = \sum_{i=1}^{n} G(\tau - D_i^*)^{-1} I[Y_i^* \le s, D_i^* \le t]\Big(\sum_{i=1}^{n} G(\tau - D_i^*)^{-1}\Big)^{-1}. \tag{3.3}$$

To get an estimate of $\mathrm{IBNR}(\tau_1, \tau)$ we replace F_{YD} in 1.6 by its estimate $\tilde{F}_{YD}$. We get

$$\begin{aligned} \widehat{\mathrm{IBNR}}_1(\tau_1, \tau) &= n\frac{\int_0^\infty \int_{\tau-\tau_1}^\tau s(G(\tau_1) - G(\tau - t))d\tilde{F}_{YD}(s, t)}{\int_0^\tau G(\tau - t)d\tilde{F}_{YD}(\infty, t)} \\ &= \sum_{i:\tau - D_i^* < \tau_1} Y_i^*(G(\tau_1)G(\tau - D_i^*)^{-1} - 1). \end{aligned} \tag{3.4}$$

One can realize that the summands in 3.4 are independent and identically distributed. Thus, we may investigate asymptotic behaviour of $\widehat{\mathrm{IBNR}}_1$. We get

$$n^{-1/2}(\mathrm{IBNR}(\tau_1, \tau) - \widehat{\mathrm{IBNR}}_1(\tau_1, \tau)) = n^{-1/2}\sum_{i=1}^{N} Y_i I_i, \tag{3.5}$$

where

$$\begin{aligned} I_i =& I[S_i + D_i > \tau, S_i \le \tau_1] - (G(\tau_1)G(\tau - D_i)^{-1} - 1) \\ & I[S_i + D_i \le \tau, \tau - D_i < \tau_1]. \end{aligned}$$

It can be easily checked, that $\mathsf{E}[Y_i I_i] = 0$ and

$$\begin{aligned} \mathsf{E}[Y_i I_i]^2 &= \int_0^\infty \int_{\tau-\tau_1}^\tau s^2\Big((G(\tau_1) - G(\tau - t)) + (G(\tau_1)G(\tau - t)^{-1} - 1)^2 G(\tau - t)\Big) \\ & dF_{YD}(s, t) = \int_0^\infty \int_{\tau-\tau_1}^\tau s^2 G(\tau_1)\Big(G(\tau_1)G(\tau - t)^{-1} - 1\Big)dF_{YD}(s, t). \end{aligned}$$

If n is large, 3.5 has approximately normal distribution with zero mean and the variance

$$\begin{aligned}\sigma^2 =&P[S_i + D_i \leq \tau]^{-1} \int_0^\infty \int_{\tau-\tau_1}^\tau s^2 G(\tau_1)\Big(G(\tau_1)G(\tau-t)^{-1} - 1\Big) dF_{YD}(s,t) \\ =&\Big(\int_0^\tau G(\tau-t) dF_{YD}(\infty,t)\Big)^{-1} \int_0^\infty \int_{\tau-\tau_1}^\tau s^2 G(\tau_1)\Big(G(\tau_1)G(\tau-t)^{-1} - 1\Big) \\ & dF_{YD}(s,t).\end{aligned} \tag{3.6}$$

When the distribution function F_{YD} is not known, we can replace it by its estimate 3.3 and estimate the variance σ^2 by

$$\widehat{\sigma}^2 = n^{-1} \sum_{i:\tau-D_i^*<\tau_1} Y_i^{*2} G(\tau_1)G(\tau-D_i^*)^{-1}\Big(G(\tau_1)G(\tau-D_i^*)^{-1} - 1\Big).$$

As a consequence of above-mentioned facts, we can construct the confidence bounds for IBNR claims. For large n we have

$$P[\mathrm{IBNR}(\tau_1,\tau) \leq \widehat{\mathrm{IBNR}}_1(\tau_1,\tau) + \widehat{\sigma} n^{1/2} u_\beta] = \beta, \tag{3.7}$$

where β is the chosen probability and u_β is the corresponding β-quantile of the standard normal distribution. It can be interpreted that with probability β, the value of $\mathrm{IBNR}(\tau_1,\tau)$, i.e. the total size of claims incurred up to time τ, but not reported until time τ does not exceed the value $\widehat{\mathrm{IBNR}}_1(\tau_1,\tau) + \widehat{\sigma} n^{1/2} u_\beta$.

3.2 Estimation of IBNR claims when G is unknown

Until now, we have assumed that G is a known distribution function. This assumption is now dropped. By applying the same logic that led to an equivalent calculation of the estimate of the cumulative distribution function F in formula 2.13, we replace the cumulative distribution function G in 3.4 by its conditional maximum likelihood estimate based on $\Sigma_1^*, ..., \Sigma_n^*$, given $D_1^*, ..., D_n^*$, in the form (see 2.10)

$$\widehat{G}_n(t) = \prod_{j:\Sigma_j^*>t} \left[1 - \frac{s(\Sigma_j^*)}{nC_n(\Sigma_j^*)}\right], \;\; 0 \leq z < \infty, \tag{3.8}$$

where $s(\Sigma_j^*) = \sum_{k=1}^n I[\Sigma_k^* = \Sigma_j^*]$ for $1 \leq j \leq n$. The product extends over distinct values of Σ_j^*.

Having this estimate of G, we can replace unknown G and get the following estimate of $\mathrm{IBNR}(\tau_1,\tau)$:

$$\begin{aligned}\widehat{\mathrm{IBNR}}_2(\tau_1,\tau) =& \sum_{i:\tau-D_i^*<\tau_1} Y_i^* \Bigg(\Bigg(\prod_{\Sigma_j^*>\tau_1} \Big(1 - \frac{s(\Sigma_j^*)}{nC_n(\Sigma_j^*)}\Big)\Bigg) \\ & \Bigg(\prod_{\Sigma_j^*>\tau-D_i^*} \Big(1 - \frac{s(\Sigma_j^*)}{nC_n(\Sigma_j^*)}\Big)\Bigg)^{-1} - 1\Bigg) \\ =& \sum_{i:\tau-D_i^*<\tau_1} Y_i^* \Bigg(\Bigg(\prod_{\tau_1\geq\Sigma_j^*>\tau-D_i^*} \Big(1 - \frac{s(\Sigma_j^*)}{nC_n(\Sigma_j^*)}\Big)\Bigg)^{-1} - 1\Bigg).\end{aligned} \tag{3.9}$$

The product also extends over distinct values of Σ_j^*.

4. Practical numerical example on real data

In this chapter we will demonstrate the use of the methods for calculating IBNR claims reserves estimates derived in the previous chapters on the real data. Firstly we introduce the data, describe and analyze them and then follow with calculations of reserves estimates.

4.1 Data

4.1.1 Description of the data

Data are provided by Czech Society of Actuaries. They collected the data from the database of Czech Insurers' Bureau (CIB), which include the records about claims that are paid from Guarantee Fund of Czech Insurers' Bureau. The compensations for personal injury or death caused by operation of an unidentified vehicle for which an unidentified person is liable, damage caused by operation of a vehicle without liability insurance and similar claim occurrences are provided from this fund. The reason for publishing this data was to provide data for educational or scientific purposes because of the strong demand for real data that could be used for bachelor theses, master theses or various seminar works at the universities. There were some aggregated data sets available, but they lack detailed information about individual claims. Getting real data from insurance companies is complicated, since they are not willing to provide data because of business confidentiality and sensitivity of personal data. CIB is exchanging data with other subjects such as insurance companies or the police of Czech Republic. For data given to CIB by other subjects, it is difficult to guarantee full anonymity, aggregation and transformation of the data. On the other hand "own data" of CIB containing information about the development of claims portfolio are suitable for educational purposes, since they are not biased by competitive environment, while not compromising personal data privacy they can be used. CIB uses techniques of claims reserving and claims payments similarly to insurance companies, the sample from their database is just much smaller since there are much more vehicles with compulsory MTPL (Motor Third-Party Liability) insurance than non-insured vehicles. The data are without transformations. We might expect some faulty records since it is stated in the notes coming with the data file that they are kept intentionally.

4.1.2 Data sheet

Data file consists of one MS Office file with 64,298 records about claims, one on each row. There are seven columns for each record, each for:
• ID - Identification number of the loss event in the database, from this number it is not possible to detect the specific case, thus not even the connection to the personal data. There are 44,076 unique claims IDs, which means that if there are more payments associated with one claim, each payment is on a different row in

data file
• Type - We distinguish between two types of claims, bodily injuries are denoted as "Bodily" and material damages as "Material". Obviously this a categorical variable. In this chapter we refer to bodily injuries claims as bodily claims and claims connected to material damage as material claims.
• Accident - Date of the occurrence of claim. Minimal value of this variable, i.e. the first claim occurrence, in the data set is on the date 2.1.2000 and the last claim occurrence is on the date 10.12.2016.
• Reporting - Date on which the claim was registered into the system, which is usually right after reporting. It should be true that Accident $\leq$ Reporting.
• Payment - Date on which the claim loss was paid to the insurer from the Guarantee Fund. It should hold that Reporting $\leq$ Payment. The first payment is made on 28.7.2000 and the last payment is on the date 30.12.2016.
• Amount - Claim amount in Czech crowns, that is paid for the claim loss at the given date. Data contain only positive values, since regress payments (negative values), when the driver operating non-insured vehicle pays back to the Gurantee Fund, are not included in the data set.
• YearP - Year in which the payment was made.

The first thing one should see is that the values for entries Accident, Reporting and Payment are not in traditional date format, they are numeric, probably meaning days passed from some date. Since we know that the minimum value of the year of the claim payment should be 2000 and maximum 2016, it can be easily deduced that the values for entries Accident, Reporting are the days passed from 1.1.1900, starting with value 1 on this date. Maximum value from all the values is value for Payment and this value corresponds to the date 30.12.2016. Minimum value is for value Accident and it is the value corresponding to 2.1.2000. Hence we assume that the range of our data is from 1.1.2000 until 31.12.2016. We transform data so that this date range corresponds with numerical range from 1 to 6210 (representing number of days passed from the beginning day, i.e. starting with value 1 on 1.1.2000).

In the table 4.1 first five rows of the data file are shown. As an example, first row can be interpreted in the following way: the loss event with ID 72183 of material type occurred on the day denoted by 38554, was reported on the day denoted by 38559, payment was made on day 39024 and the amount of the payment was 17,999 Czech crowns paid in year 2006.

ID	Type	Accident	Reporting	Payment	Amount	YearP
72183	Material	38554	38559	39024	17,999	2006
72189	Material	38517	38559	38658	41,170	2005
72194	Material	38522	38559	38658	15,901	2005
72197	Material	38505	38559	38707	17,753	2005
72198	Material	38508	38559	38924	21,159	2006

Table 4.1: First rows of the claims data records

Let us also mention that because of the inconsistencies, we removed from the original data seven records about claims development out of which six records had the occurrence date later than the reporting date and one record had reporting

date later than the payment date. All of these removed records were connected to material damages. Twenty-eight bodily and two hundred one material records with zero values of reporting delay were removed too since it was assumed in the beginning that $0 < D_i < \tau$ with probability 1. Since there are 64,298 records and the number of unique IDs is 43,955, some records are connected to the same loss event. All of these records with common ID have also the same accident date and reporting date, hence we can merge the information about one loss event into one record while summing up the values of paid amount to get the claim size. By claim size we understand the Czech crown cost associated with each claim. It is easy to realize that we have all the necessary information for the calculation of IBNR claims reserve, because from the data we can extract 43,955 triples $(Y_i^*, D_i^*, \Sigma_i^*)$, which represent the observed size, reporting delay and occurrence date of claims. Now let us look at a more detailed analysis of the data.

4.2 Data analysis

Firstly it would be smart to take a look if there are any differences between material claims and bodily claims. Prior to any analysis, it is natural to expect that the time-lag between occurrence and reporting should be higher for bodily claims than for material claims. This is because of the nature of bodily claims, it often takes more time until the consequences of the accident show up. The insured person might feel fine right after the accident, but after some time the symptoms that point to injury show up. On the contrary, material damages are more evident and can be quantified more quickly. This leads to an idea of calculating reserves for bodily and material claims separately. We will validate this idea throughout this chapter when showing the differences between bodily and material claims when it comes to reporting delay, number of occurred claims and claim sizes. Let us note that reporting delay and occurrence time of claims are recorded in days.

4.2.1 Reporting Delay

We start with an analysis of reporting delay. It is important to realize an important thing about the data in the last years, which is that the observed reporting delay cannot be longer than the days remaining to the date 31.12.2016 because we do not have any information about claims incurred before and reported after this day, i.e. for claims with occurrence year 2016 the observed reporting delay is less than 1 year, for claims occurred in 2015 less than 2 years etc. Thus if we worked with records from whole data set, the reporting delay would be biased towards shorter values. One can notice this bias from the table 4.2, where the observed average reporting delay drops drastically during last years. In other words, observed reporting delays come from conditional distribution, conditional on reporting time being less or equal than the time τ, up to which we observe data. This has been discussed in previous parts of the thesis where we mentioned that observations of (Y_i, D_i, Σ_i) are observed conditionally on $\Sigma_i + D_i \leq \tau$.

To get a picture about basic characteristics, descriptive statistics and distribution of reporting delay (and later also claims count), while cancelling out the effect of truncation, we can take a smaller sample from data set, for which the

effect of bias caused by truncation is zero (or at least negligible). Following the assumptions made in the first chapter that $(Y_i, D_i), i = 1, 2, ...,$ are i.i.d. random vectors, leaving out the observations near upper endpoint of time interval $[0, \tau]$ does not affect the analysis, only downside is that we have less number of observations. This approach is used only for the description of the data. For estimation of cumulative distribution functions of G and F_{YD} or calculating estimates of IBNR claims reserves, previously shown approach under model of truncated data will be used. Let us remind that the way how to obtain the maximum likelihood estimate of cumulative distribution function of reporting delay is shown in section 3.1.

The maximum value of reporting delay is 1,512 for material claims and 1,635 for bodily claims, which is more than four years. Because of the fact that the reporting delay longer than four years is observed only for two out of 31,606 material claims and also two out of 4,562 bodily claims observed until the end of the year 2012, the bias towards shorter values of delay is negligible. Thus, the analysis of the reporting delay can be made from data with accident years from 2000 to 2012.

There is also another important thing that we can see in the table 4.2, even when excluding last years biased by larger reporting delays not being observed yet, we see that the distribution of reporting delay might change over the time, since the observed average reporting delay is decreasing over the years, especially in the first few occurrence years. Assuming that the distribution of reporting delay depends on occurrence time would be in contradiction with assumption we made in the beginning, which was that the pairs $(Y_i, D_i), i = 1, 2, ...,$ are independent identically distributed random vectors, independent of process $\{N(t), t \geq 0\}$. Since our sample is not so big, especially for bodily claims, we do not leave the observations from first years out. We will discuss later, if possible differences in distribution of reporting delay in the first years compared to more recent years might cause problems with accuracy of the estimation of IBNR claims reserves.

In the table 4.3 we can see that the value of first quartile is a bit higher for bodily claims, median and mean more than two times higher, while third quartile four times higher.

In the figure 4.1, one can see graphically the aforementioned differences between the quartiles and medians of reporting delay for material and bodily claims. We excluded outliers from the boxplots.

Looking at the histogram of observed reporting delay (figure 4.2) we see that relative frequencies are shifted more towards left for material claims, i.e. there are more claims with smaller values of observed reporting delay, while for bodily claims there are more claims with higher values of observed reporting delay.

4.2.2 Claims count and time occurrence

In this section we take a look at an analysis of claims count. In the tables 4.4 and 4.5 we see observed numbers of claims occurrences in consecutive years for bodily and material claims respectively. We see that the values of claims count per year are much higher for material claims than for bodily claims, hence under the assumption that the claims counts follow Poisson process, they are driven by Poisson processes with different intensity functions. The necessity to calcu-

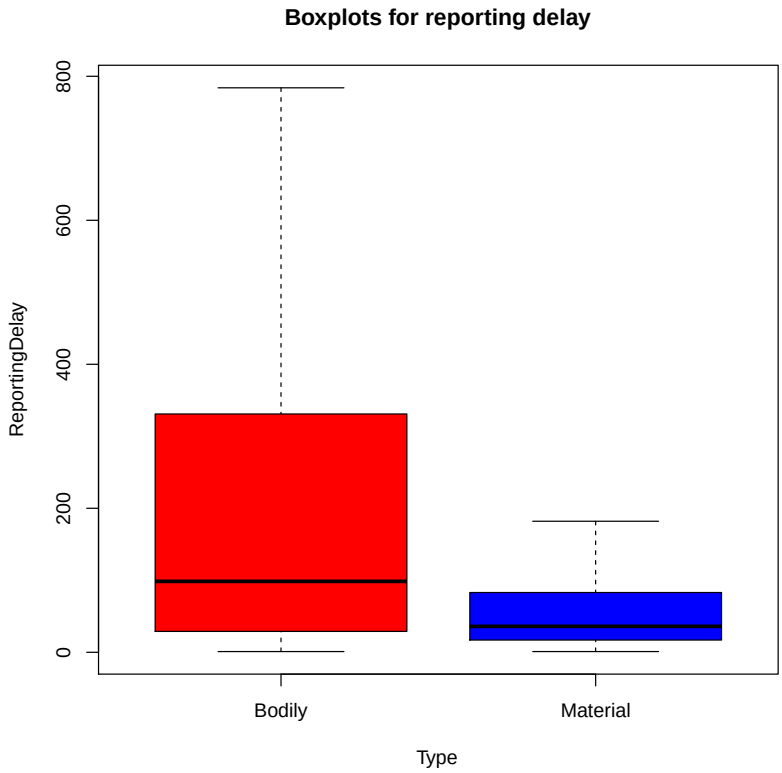

Figure 4.1: Boxplots for bodily and material claims

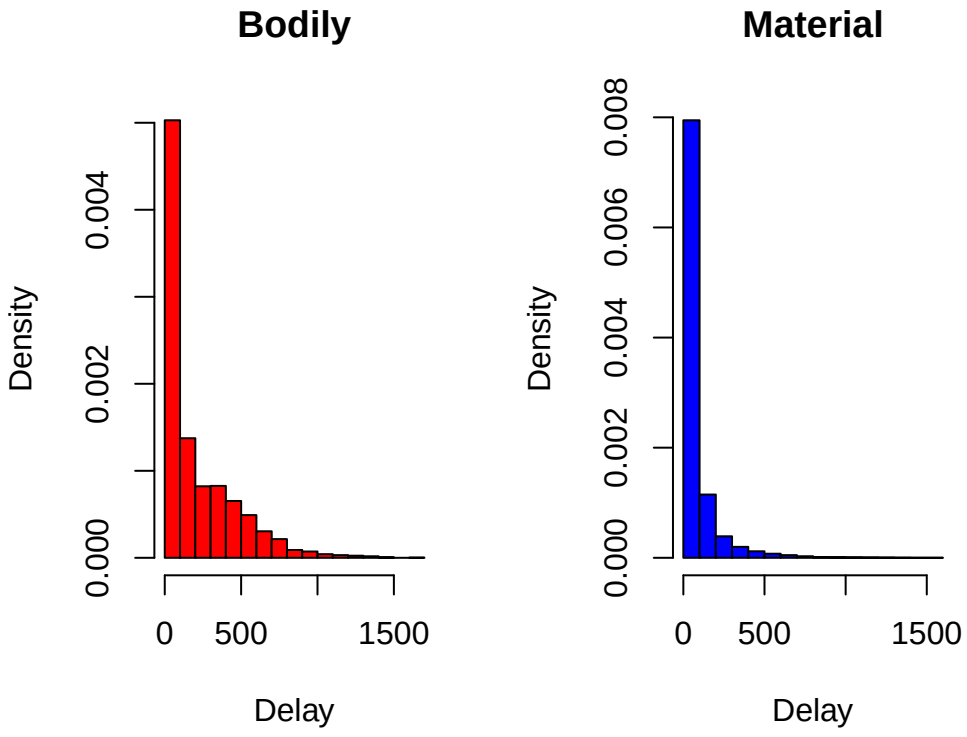

Figure 4.2: Histograms of reporting delay for bodily and material claims

Ocurrence year	Average bodily delay	Average material delay
2000	292.92	131.74
2001	263.36	134.41
2002	228.02	117.79
2003	221.6	98.84
2004	230.34	80.5
2005	219.37	72.87
2006	183.31	70.26
2007	223.19	60.10
2008	204.33	62.72
2009	208.66	60.06
2010	174.58	66.22
2011	144.67	73.85
2012	151.4	60.58
2013	113.22	49.26
2014	102.69	45
2015	90.78	37.06
2016	29.48	23.17

Table 4.2: Observed average reporting delay for occurrence years from 2000 to 2016

type	Min	1st Qu.	Median	Mean	3rd Qu.	Max	St. dev.
Bodily	1	29	98.5	208	331	1635	243.17
Material	1	17	36	78.12	83	1512	125.08

Table 4.3: Descriptive statistics of reporting delay for bodily and material claims

late reserves separately for each type of claims is now fully clear. It would be essential to determine if it is reasonable to assume homogeneous or nonhomogeneous Poisson process. The numbers of occurred claims were much lower in the first two years than in the following years. The value grew significantly in 2002, while being kind of stable until 2008. In 2009 there was an introduction of a fee to the Guarantee Fund from drivers operating vehicles without compulsory MTPL Insurance. This in connection with other efforts to lower the number of non-insured vehicles affected the numbers of yearly claim occurrences, which we see in decreasing values since year 2009. Again we have to mention the effect of reporting delay and truncation, since we have only information about claims occurrences that have been reported until the end of 2016, it surely affects the values of observed occurrences in last years. Indeed, we see a big drop especially in the last year, which is very likely the consequence of many claims not being reported yet. From information available at the present moment, the number of occurred claims actually grew a bit in year 2016 compared to 2015, which confirms this consideration. But even after excluding the last years from our considerations, there is a noticeable variability in the observed numbers of yearly claims occurrences, which suggests using nonhomogeneous Poisson process for modelling the claims count, which confirms what we already assumed.

year	number of claims	year	number of claims
2000	143	2009	354
2001	182	2010	310
2002	340	2011	315
2003	426	2012	273
2004	466	2013	289
2005	417	2014	268
2006	414	2015	244
2007	475	2016	161
2008	447		

Table 4.4: Number of bodily claims occurrences in years from 2000 to 2016

year	number of claims	year	number of claims
2000	736	2009	2551
2001	1279	2010	2457
2002	2321	2011	2155
2003	2667	2012	1984
2004	2834	2013	1970
2005	3264	2014	1850
2006	3064	2015	1725
2007	3049	2016	1280
2008	3245		

Table 4.5: Number of material claims occurrences in years from 2000 to 2016

Now we would like to deduce the form of the intensity function of this Poisson process so that we could estimate IBNR claims under known G.

Determining the intensity function of Poisson process

Previously discussed idea (see 3.1) that the intensity function $\lambda(t)$ of the Poisson process $\{N(t), t \geq 0\}$ could be of the form

$$\lambda(t) = \lambda\eta(t), \tag{4.1}$$

where $\eta(t)$ denotes known number of insured risks at the time t, while λ is an unknown constant such that $\lambda > 0$, has some merit. We might assume that in motor car liability insurance the change of intensity function depends mostly on the number of insured risks, while constant λ might represent the stable nature of car accidents (the rate of claim occurrence for one car could be assumed to stay constant over time), thus the variability of intensity function $\lambda(t)$ over time is caused by changes in $\eta(t)$ in time. In reality this approach might be used, data for vehicles without MTPL insurance should be available, but since we do not have any records of the number of insurance risks (number of vehicles without MTPL insurance), we have to find other way to deduce the intensity function, from which we can also obtain the cumulative distribution function G, which is the desired object of interest. We deviate a bit from the approach we used in

previous chapter, but for demonstrative reasons we want to assume known form of G so that we can show the use of method proposed in section 3.1.

As has been already mentioned before, last years in the data set are biased by truncation and observed numbers of claim occurrences should be lower than true numbers. Therefore we cut off last four years from the sample used for estimation to remove bias caused by incurred but not yet reported claims and estimate intensity function $\lambda(t)$ from data with occurrence years between 2000 and 2012. The effect of truncation should be negligible for this time range. Let us first think what would be a reasonable form of intensity function. We can take a look at figures 4.3 and 4.4 with numbers of monthly claims occurrences in consecutive months. We would like to fit the points representing monthly values with some smooth function which is increasing from the beginning and then starts decreasing from certain point to fit the development of the number of occurrences. The intensity for time period after year 2012 will be made as prediction based on the fitted model.

From more considered models we choose cubic function (for both types of claims), which fits the data quite well and makes reasonable prediction of intensity function for years 2013-2016. We can see this in figures 4.3 and 4.4, the blue line represents the values of fitted cubic model, the vertical line indicates last point of sample we used for estimation. From this point the blue line represents values of prediction based on fitted model. Using this fit is a bit subjective, we want reasonable fit for the observed data and also for predicted values, but we cannot use any forecast criterion, since the values of observed claim occurrences in forecasting time period are biased by truncation. For simplicity, we considered smooth functions described in section 9.1. in [Cipra, 20013].

Since the occurrence time is measured in days, we take time as a discrete variable. We would like to obtain daily intensities, so in the end we fit the daily data instead of monthly (we get similar results and easily solve the problem with different numbers of days in different month). Now when assuming that $\lambda(t)$ is a cubic function, then also cumulative distribution function G can be assumed to be known, taking the form

$$G(t) = \frac{\sum_{s=1}^{t} \lambda(s)}{\sum_{s=1}^{\tau} \lambda(s)}, \;\; t = 1, ..., 6210.$$

Once we assume known form of G, we can plug it into formula 3.4 and get the IBNR claims reserves estimates.

4.2.3 Claims Severity

Finally we take a look at claims severity. We do not adjust payments for inflation and we take claim severity as the sum of all payments associated with the claim. It has the same meaning as claim sizes that we denoted by random variables $Y_i, i = 1, 2, ...$, for which we assume that they are i.i.d. random variables as was supposed in the beginning. We do not have any information if the claim is settled or not, but for simplicity, we will assume that the data contain settled claims.

With these assumptions it is simple to describe the severity of claims from data, since there is no presence of truncation or dependence. Firstly let us look at some descriptive statistics of claims severity in the table 4.6. We see that the

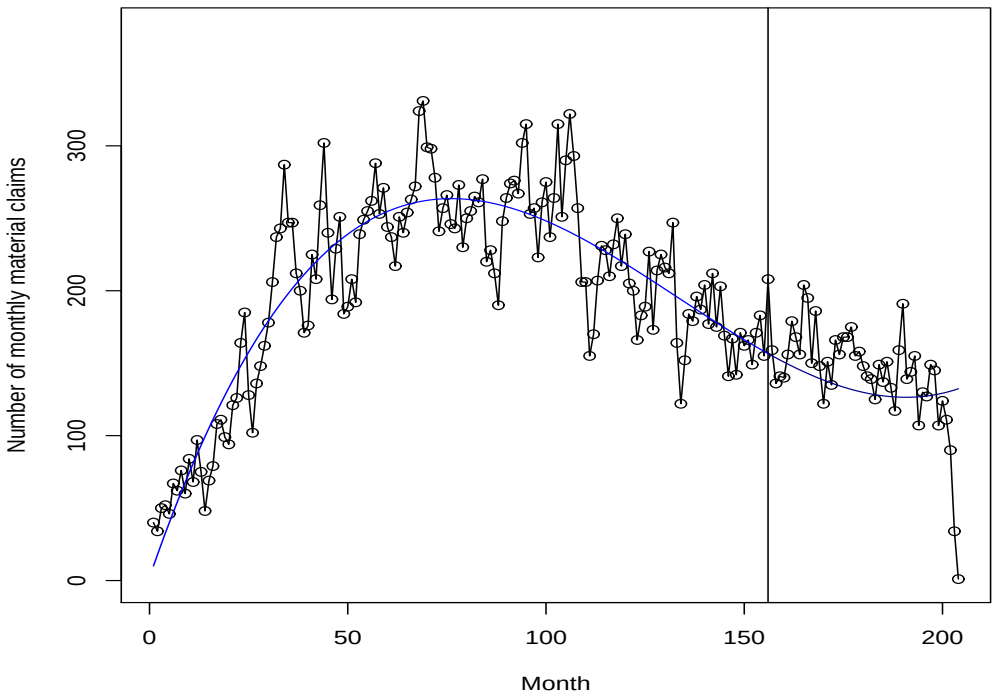

Figure 4.3: Number of monthly material claims occurrences in consecutive months

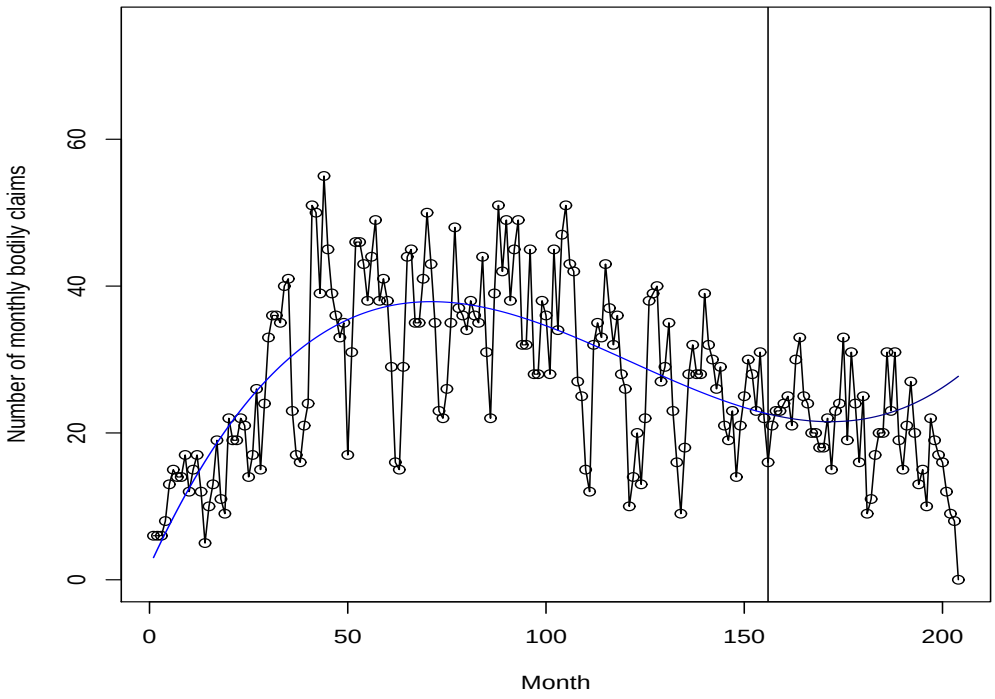

Figure 4.4: Number of monthly bodily claims occurrences in consecutive months

first quartile is bit higher for material claims, but median, mean values and third quartile much higher for bodily claims. We can conclude that the distributions of claims severity should have heavier right tails and since we have have only positive values, the distribution claims severity should follow is non-negative.

Type	Min	1st Qu.	Median	Mean	3rd Qu.	Max	S. dev.
Bodily	100	12,942	48,221	152,560	142,457	10,879,375	395,726
Material	114	14,287	26,265	44,699	50,685	2,679,750	65,551.76

Table 4.6: Descriptive statistics for claims severity

When considering distribution of claims severity, we choose between gamma and log-normal distribution, which are distributions commonly used for modelling claims severity. In case of interest in the properties of these distributions, one can look in [Anděl, 2007]. In the figures 4.5 and 4.6 histograms of logarithm of claim size for material and bodily claims are shown alongside with reference line of theoretical normal distribution density. The parameters of normal distributions were estimated using maximum likelihood method with function fitdistr from package MASS in statistical software R. This function provides maximum likelihood estimates, the likelihood of log-normal distribution was higher for both types of claims. Also the graphical analysis from histograms and Q-Q plots performed better for log-normal distribution, in the figure 4.7 we can see normal Q-Q plots for logarithm of claim size for material and bodily claims. We see that the points fall approximately along the reference line. In the figure 4.6 we can see that relative frequencies for bodily claims are skewed a bit more to the right than the reference line of normal distribution density, but we can still conclude that log-normal model fits the data quite well.

4.3 Estimating the IBNR claims reserves

In this section we will demonstrate how the results derived in previous parts of the work can be applied for estimating the IBNR claims reserves. Firstly we calculate the estimate of claims reserves for claims incurred up to the end of year 2016 not reported until this date. We have 5,524 observed triples of $(Y_i^*, D_i^*, \Sigma_i^*)$ for bodily claims and 38,431 observed triples of $(Y_i^*, D_i^*, \Sigma_i^*)$ for material claims. The values of Σ_i^* and D_i^* are recorded in days.

4.3.1 IBNR claims reserves estimates when G is known

Firstly we estimate the IBNR claims reserve under assumption that cumulative distribution function G is known. Using the formula 3.4 we get the the following estimate of IBNR claims reserves for claims incurred up to end of 2016 not reported until end of 2016

$$\widehat{\mathrm{IBNR}}_1(6210, 6210) = \sum_{i:6210-D_i^*<6210} Y_i^*(G(6210)G(6210 - D_i^*)^{-1} - 1),$$

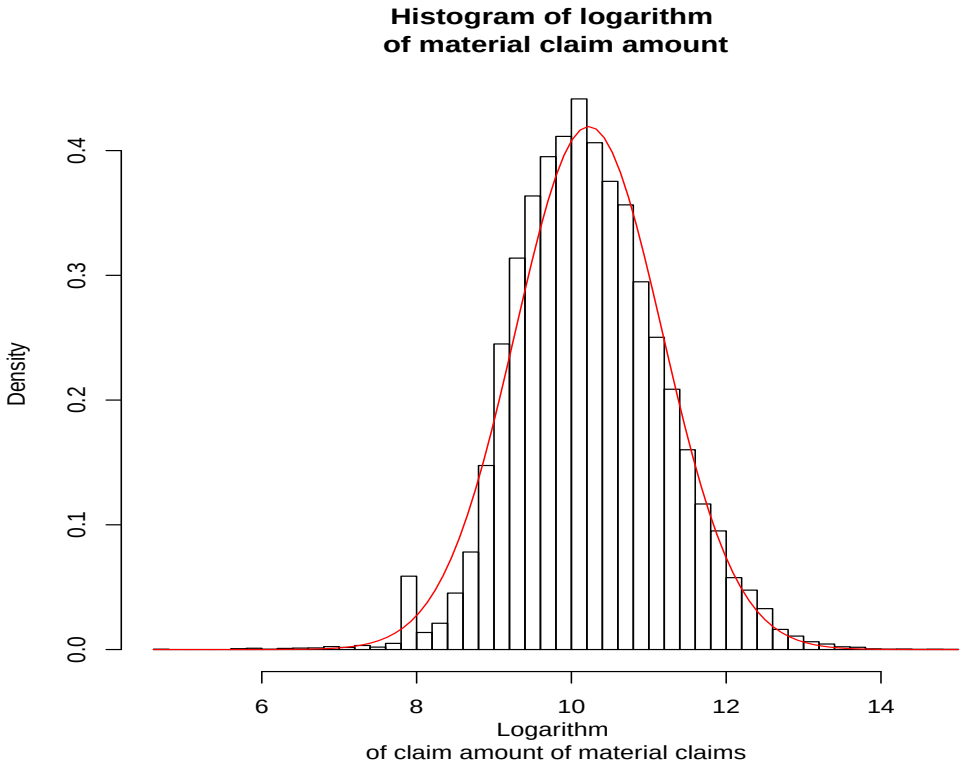

Figure 4.5: Histogram of logarithm of claim size - material claims

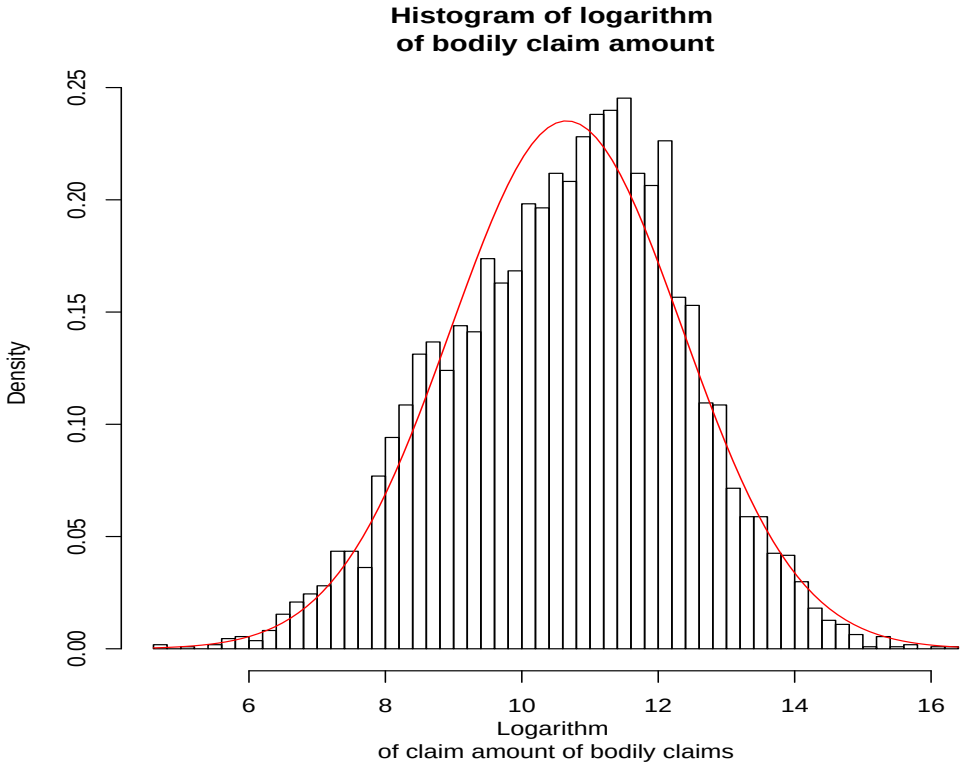

Figure 4.6: Histogram of logarithm of claim size - bodily claims

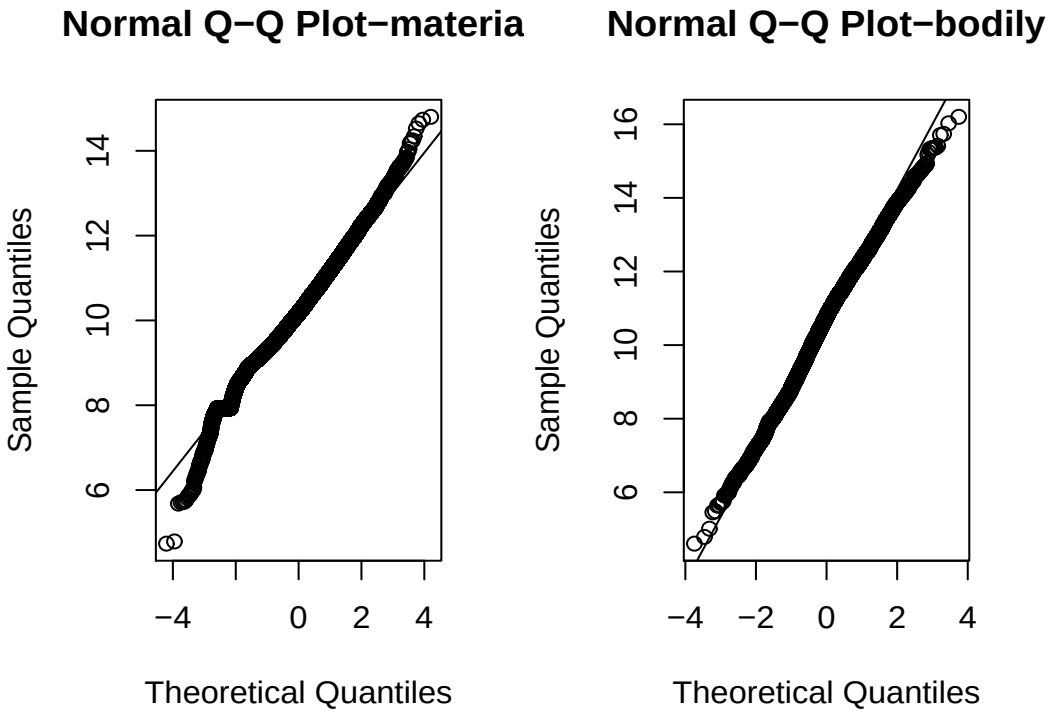

Figure 4.7: Normal Q-Q plots for logarithm of claim size

where we assume that

$$G(t) = \frac{\sum_{s=1}^{t} \lambda(s)}{\sum_{s=1}^{\tau} \lambda(s)}, \;\; t = 1, ..., 6210.$$

with $\lambda(t), t = 1, ..., 6210$, being a cubic function, parameters of which are calculated in software R.

After inserting the values of claim amounts Y_i^* and corresponding reporting delays D_i^* we get the estimate

$$\widehat{\mathrm{IBNR}}_1^{\mathrm{bodily}}(6210, 6210) = 17{,}956{,}109$$

for bodily claims and

$$\widehat{\mathrm{IBNR}}_1^{\mathrm{material}}(6210, 6210) = 13{,}169{,}129$$

for material claims.

We may also construct asymptotic confidence bound for IBNR claims using 3.7. Let us calculate for example 95% confidence bound, since $\widehat{\sigma} n^{1/2} u_{0.95} = 6{,}865{,}887$, we can conclude that the total value of bodily claims incurred but not reported up to year 2016 does not exceed the value $\widehat{\mathrm{IBNR}}_1^{\mathrm{bodily}}(6210, 6210) + \widehat{\sigma} n^{1/2} u_{0.95} = 24{,}821{,}996$ with probability 95%.

We can also calculate estimated asymptotic standard error of the estimate $\widehat{\mathrm{IBNR}}_1(6210, 6210)$, which is $\widehat{\sigma} n^{1/2} = 4{,}174{,}163$.

For material claims we calculate that $\widehat{\sigma} n^{1/2} u_{0.95} = 2{,}238{,}741$ and value of upper limit of confidence bound is then 15,407,871. Estimated asymptotic standard error of the estimate is 1,361,058.

4.3.2 IBNR claims reserves estimates under unknown G

This time we assume that not only the joint cumulative distribution function F_{YD} is unknown, but also the previously mentioned cumulative distribution function G is unknown. The theory of non-parametric statistical estimation under the model of truncated data led us to non-parametric maximum likelihood estimator 3.8, with which we substitute unknown cumulative distribution function G. Applying the formula 3.9 we get the following estimate of IBNR claims reserves

$$\widehat{\mathrm{IBNR}}_2^{\text{bodily}}(6210, 6{,}210) = 13{,}529{,}508$$

for bodily claims and

$$\widehat{\mathrm{IBNR}}_2^{\text{material}}(6210, 6{,}210) = 8{,}271{,}761$$

for material claims.

4.4 Comparison with chain ladder

Interesting comparison might be with method that does not use information about individual claims records but uses aggregated data. One of the most popular and widely used method of this kind, one which uses compressed data in form of development triangles, is chain-ladder method. Firstly we introduce this method and then provide the comparison.

4.4.1 Chain-ladder method

Before describing the method, we should introduce the necessary notation. We take the definition of Chain-ladder method and (slightly adjusted) notation from [Wüthrich and Merz, 2008].

Let C_{ij} denote all payments for claims that occurred in year i and were paid in development period j, i.e. the payment was done in accounting year $i+j$. We assume that $i \in \{0, ..., I\}$ and $j \in \{0, ..., J\}$, where I denotes the last observed accident year and J is the last development period (i.e. the development of claim is finished after J years, meaning $C_{i,j} = 0$ for $j > J$). We assume that $J \leq I$. The data are often structured in claims development triangle (also known as run-off triangle). Usually the accident years are on the vertical line and development periods on the horizontal line. Cumulative claim payments are defined as $D_{ij} = \sum_{k=1}^{j} C_{ik}$. The value $D_{i,J}$ is called the ultimate claims amount of accident year i. The observations of C_{ij} are available only for indices i, j for which $i + j \leq I$. They form the upper triangle, which we can denote as

$$D_I^U = \{C_{i,j}; i + j \leq I\}.$$

The outstanding payments $C_{i,j}$ for $i + j > I$, which form the lower triangle, need to be predicted. We will denote the lower triangle as

$$D_I^L = \{C_{i,j}; i + j > I\}.$$

The outstanding loss liabilities for accident year i at time I are given by

$$R_i = \sum_{j=I-i+1}^{J} C_{i,j} = D_{i,J} - D_{i,I-i}$$

They represent the claims reserves we would like to estimate. Now we have all necessary variables defined and can continue with model.

Model assumptions

Assumptions of Chain-ladder model are
1.) There exist development factors (or alternatively called age-to-age factors) $f_0...., f_{J-1} > 0$ such that for all $0 \leq i \leq I$ and all $1 \leq j \leq J$ we have

$$\mathrm{E}[D_{i,j}|D_{i,0}, ..., D_{i,j-1}] = \mathrm{E}[D_{i,j}|D_{i,j-1}] = f_{j-1}D_{i,j-1}.$$

2.) The cumulative claims $D_{i,j}$ of different accident years are independent, i.e.

$$\{D_{i,0}, ..., D_{i,J}\}, \{D_{k,0}, ..., D_{k,J}\} \text{ are independent for } i \neq k,$$

We make assumptions only about the first moments, but still it is sufficient for estimating (conditionally) expected future claims. From the assumptions above we have

$$\mathrm{E}[D_{i,J}|D_I^U] = \mathrm{E}[D_{i,J}|D_{i,I-i}] = D_{i,I-i}f_{I-i}...f_{J-1}$$

for all $I + 1 - J + 1 \leq i \leq I$.

The resulting Chain-ladder estimator for $\mathrm{E}[D_{ij}|D_I^U]$ is then given by

$$\widehat{D}_{i,j}^{CL} = \widehat{\mathrm{E}}[D_{ij}|D_I^U] = D_{i,I-i}\widehat{f}_{I-i}...\widehat{f}_{j-1},$$

where $\{\widehat{f}_l, l = 0, ..., J-1\}$ are estimated development factors estimated by

$$\widehat{f}_j = \frac{\sum_{i=0}^{I-j-1} D_{i,j+1}}{\sum_{i=0}^{I-j-1} D_{i,j}}, \;\; j = 0, ..., J-1.$$

The estimates of outstanding loss liabilities R_i for accident year i are then

$$\widehat{R}_i^{CL} = \widehat{D}_{i,J}^{CL} - D_{i,I-i} = (\widehat{f}_{J-1}...\widehat{f}_{I-i} - 1)D_{i,I-i}.$$

In [Wüthrich and Merz, 2008] it is proved that $\widehat{f}_0...., \widehat{f}_{J-1} > 0$ are conditionally and also unconditionally unbiased estimators for development factors $f_0...., f_{J-1} > 0$. Estimators $\widehat{f}_0...., \widehat{f}_{J-1} > 0$ are also uncorrelated. Furthermore $\widehat{D}_{i,J}^{CL}$ is given $D_{i,I-i}$ an unbiased estimator for $\mathrm{E}[D_{i,J}|D_I^U] = \mathrm{E}[D_{i,J}|D_{i,I-i}]$. At the same time it is (unconditionally) unbiased estimator for $\mathrm{E}[D_{i,J}]$, i.e. $\mathrm{E}[\widehat{D}_{i,J}^{CL}] = \mathrm{E}[D_{i,J}]$.

Mack in his model of chain-ladder added additional third assumption, which is that there exist variance parameters $\sigma_0^2, ..., \sigma_{J-1}^2$ such that for all $0 \leq i \leq I$ and $1 \leq j \leq J$ we have that

$$\mathrm{Var}(D_{i,j}|D_{i,0}, ..., D_{i,j-1}) = \mathrm{Var}(D_{i,j}|D_{i,j-1}) = \sigma_{j-1}^2 D_{i,j-1}.$$

This allows calculating standard error of the chain-ladder estimates, which is explained in more detail in [Mack, 1993].

From the data available, firstly we have to construct the triangles, which is the input for chain-ladder method. In order to be consistent with other methods we use for calculating IBNR claims reserves, variables $C_{i,j}$ will be adjusted a little bit. Instead of incremental claim payments they will denote all payments that occurred in year i and were reported in development period j. Same adjustment is made for cumulative claim payments. In the tables 4.7 and 4.8 we provide cumulative triangles for observed material and bodily claims. From these cumulative triangles one can see that the biggest volume of the claims is reported during the accident year, quite less the next year, while during next three years the cumulative claims amounts are increasing only slightly. We assume that the development of the claim is finished in 5th year (there are no claims with longer development period in the data set).

Year	0	1	2	3	4
2000	210.9	299.9	308	309.7	310.1
2001	347.0	514.5	547.3	550.2	552.3
2002	648.3	957.6	986.4	990	990
2003	761.5	1,075.5	1,099.7	1,101.3	1,101.5
2004	973.2	1,264.7	1,272.5	1,276.3	1,276.3
2005	1,212.9	1,447.6	1,454.1	1,463.1	1,463.1
2006	1,136.9	1,393.2	1,399.7	1,403.8	1,404.9
2007	1,206.7	1,430.4	1,441.6	1,443	1,443.9
2008	1,325.3	1,520.6	1,540.1	1,542.2	1,542.3
2009	955.4	1,110.2	1,124.9	1,129.4	1,129.7
2010	918.8	1,063.0	1,081.2	1,085.4	1,085.4
2011	834.6	959.2	972.4	975.6	975.8
2012	760.9	868.0	876.9	878.1	879.4
2013	713.8	802.1	809.2	812	
2014	727.7	814.4	818.6		
2015	763.9	826.1			
2016	566.9				

Table 4.7: Cumulative triangle for material claims in hundred thousands of Czech crowns (rounded to the nearest ten thousand)

4.4.2 Comparison of the results

As we previously mentioned, claims development should be finished after four years, thus all claims that occurred until end of year 2012 should be recorded in full data set. This allows us to calculate estimates of outstanding loss liabilities using methods derived earlier using data we would observe at the end of year 2012 (claims occurred until 31.12.2012 reported up to this date) and compare them with results obtained from data. The comparison is provided in tables 4.9 and 4.10 with values of estimation and standard error of estimation (if it is possible to calculate them with certain method). For calculating estimates of outstanding

Year	0	1	2	3	4
2000	50.7	96.2	100.0	106.6	106.6
2001	66.3	191.6	207.1	209.3	209.3
2002	211.0	324.5	347.2	352.2	354.7
2003	352.0	505.4	532.9	536.8	540.4
2004	352.6	538.3	588.8	591.9	593.0
2005	411.8	667.2	690.0	695.9	698.2
2006	489.0	720.1	742.1	743.6	743.6
2007	495.2	723.6	760.2	771.0	773.1
2008	594.9	788.8	828.4	839.0	877.8
2009	328.6	586.1	614.5	623.4	626.7
2010	404.2	502.9	525.3	531.2	532.4
2011	312.1	398.9	408.8	409.1	410.6
2012	322.3	438.9	441.4	455.4	455.4
2013	362.4	396.5	397.4	402.0	
2014	492.4	611.8	612.6		
2015	258.6	285.9			
2016	205				

Table 4.8: Cumulative triangle for bodily claims in hundred thousands of Czech crowns (rounded to the nearest ten thousand)

loss liabilities with chain-ladder method we used the package ChainLadder in software R.

Method	IBNR value	Standard Error
Truncated data under known G	19,014,120	4,539,732
Truncated data under unknown G	21,938,942	–
Chain-Ladder	23,921,042.29	9,583,917.23
Real results	15,524,123	–

Table 4.9: Comparison of IBNR claims reserves calculated using different methods for bodily claims until year 2012

Method	IBNR value	Standard Error
Truncated data under known G	17,959,567	1,710,917
Truncated data under unknown G	23,600,245	–
Chain-Ladder	22,037,702.77	10,111,064.43
Real results	13,968,948	–

Table 4.10: Comparison of IBNR claims reserves calculated using different methods for material claims until year 2012

We see that all the methods overestimated the real results. The estimated standard errors of chain-ladder method are much higher, their value is higher than 40% of the estimated IBNR reserve. One can realise that the differences

between estimated values of IBNR claims reserves under model of truncated data assuming known and unknown G are caused only by differences of this cumulative distribution functions, which is obvious looking at formulas 3.4 and 3.9. Looking at figures 4.8 and 4.9 we see that values of NPMLE (non-parametric maximum likelihood estimate) of G are lower than the values of known cumulative distribution function G taken at the same time (which is graphically represented by blue line of NPMLE of G being below red line of assumed known G). The most important for our estimation is the behaviour of cumulative distribution functions near the upper endpoint of considered time interval, the lower the values of cumulative distribution function near the endpoint of the interval (where most values of $\tau - D_i^*$ are concentrated, since observed average reporting delay is not that big), the higher the value of estimated IBNR reserve.

One can also see the effect of truncation in figures 4.8 and 4.9 since the behaviour of cumulative distribution functions at the upper endpoint of considered time interval indicates higher density of claim occurrences for distribution having non-parametric maximum likelihood estimate of G for cumulative distribution function than the distribution with empirical distribution function for CDF.

The accuracy of the estimates of IBNR claims derived under model of randomly truncated data will be sensitive to violation of assumptions. When we take a sample consisting of records with accident years in 2006-2012 (reported until end of 2012), the values of estimates of IBNR claims reserves decrease for material claims, resulting in smaller deviation from real results. The result was 14,753,615 for material claims, when we used the method under unknown G. Important role in this plays the distribution of reporting delay, which is probably different for claims occurred in older years (see table 4.2). The similar problem might be with inflation, causing the distribution of claim sizes for older years is different than the one of claim sizes for recent years. We do not observe indications of effect of inflation in our data, but in general one should be careful when using other data set. One could come to an idea of using statistical tests to analyse, if the samples from different time periods come from same distribution. But this would be of no help, since for example for reporting delay, using t-test (version without assuming normality, often called z-test) on two samples, one consisting of claims with occurrence year 2011 and second with year 2012, gives us rejected null hypothesis. Same goes for Kruskal-Wallis test if we test more years (see [Anděl, 2007] for description of these tests). We could not use this approach for determining appropriate time range of sample used for calculations.

We also provide results in tables 4.11 and 4.12 containing estimates of IBNR reserves for claims occurred until end of 2016 not reported up to this date.

Method	IBNR value	Standard Error
Truncated data under known G	17,956,109	4,174,163
Truncated data under unknown G	13,529,508	–
Chain-Ladder	13,559,351.37	7,596,750.19

Table 4.11: Comparison of IBNR claims reserves calculated using different methods for bodily claims until year 2016

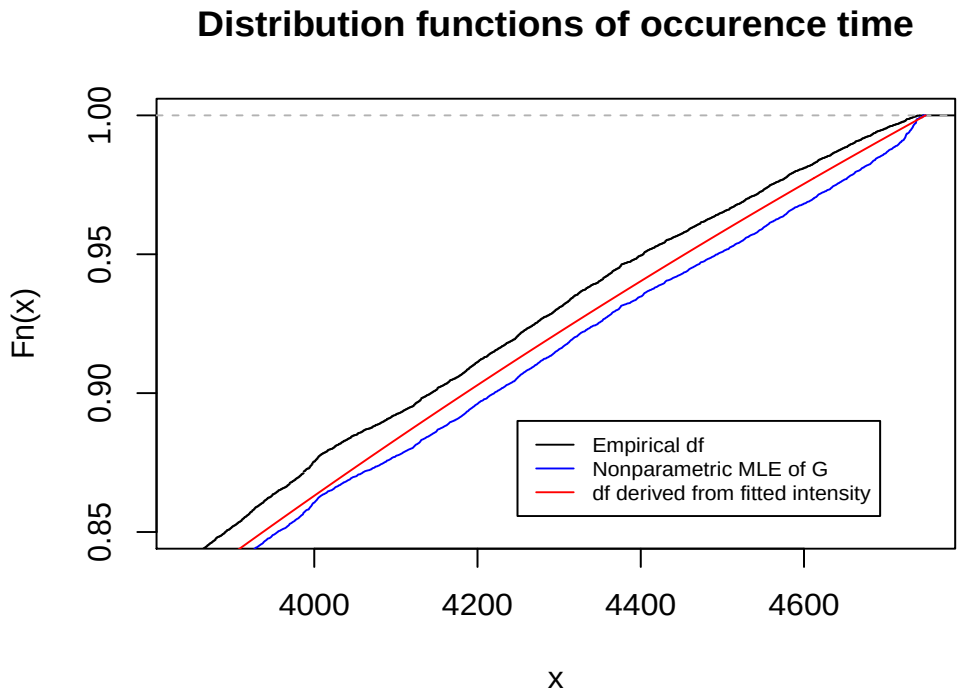

Figure 4.8: Different cumulative distribution functions of occurrence time for material claims using data set with ending year 2012

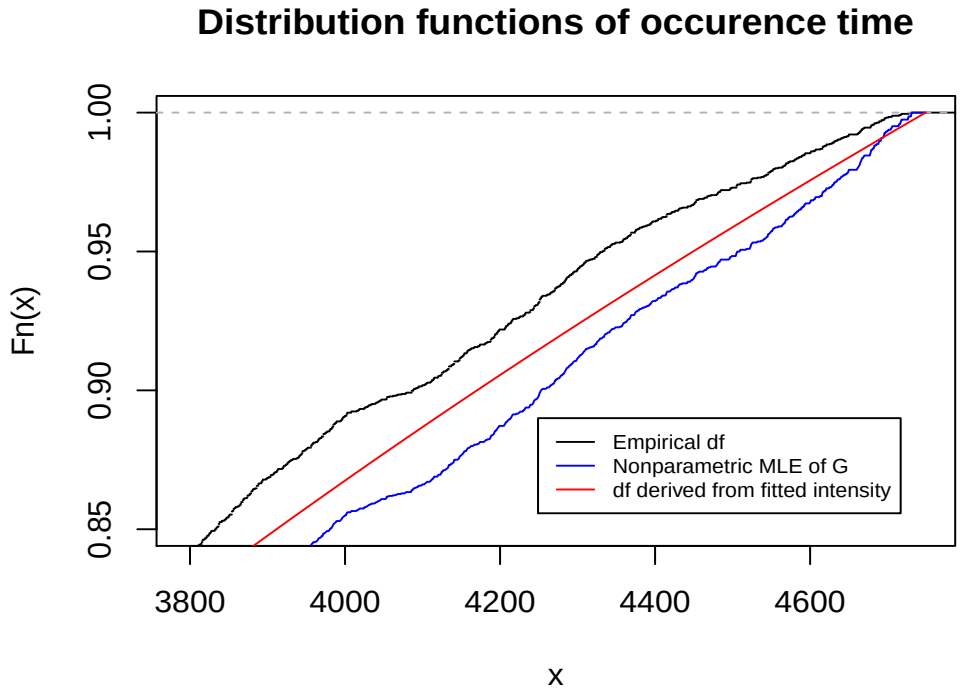

Figure 4.9: Different cumulative distribution functions of occurrence time for bodily claims using data set with ending year 2012

Method	IBNR value	Standard Error
Truncated data under known G	13,169,129	1,361,058
Truncated data under unknown G	8,271,761	–
Chain-Ladder	14,923,891.88	8,285,236.88

Table 4.12: Comparison of IBNR claims reserves calculated using different methods for material claims until year 2016

Conclusion

In the theoretical part we discussed why for insurance companies it is necessary to build IBNR claims reserves. Traditional techniques used for this task work with data aggregated into development triangles. Aim of presented thesis was to build a technique for estimation of IBNR claims, which uses individual claim-by-claim data. We proposed a probability model of IBNR claims which uses this type of data. This model is based on compound Poisson process.

The problem that arose was that the distribution, which observed claims sizes, reporting delays and times of claims occurrences follow, is not the desired distribution we need to know so that we can calculate reserves using proposed probability model of IBNR claims. The problem led to model of randomly truncated data, which was introduced and basic problems and results of non-parametric statistical estimation under this model were shown. Putting these results to use, we could finish the task of deriving estimates of IBNR claims reserves under considered model.

This approach was demonstrated on the real data from Czech Insurer's Bureau. Firstly basic characteristics of the data were discussed. Necessary information about individual claims were available, so we could proceed with calculations of the estimates of IBNR claims reserves. We came to a conclusion that the accuracy of the method is dependent on data meeting the assumptions. We saw that presence of observations with older occurrence years, for which the distribution of reporting delay probably differs, led to overestimation of claims reserves. It is understandable that proposed method is particularly suitable for small but fast developing portfolio observed during shorter periods of time. For large and relatively stable portfolio, the results are similar to more simple methods, like chain-ladder.

Bibliography

J. Anděl. *Základy matematické statistiky*. Druhé opravené vydání. Matfyzpress, Praha, 2007. ISBN 80-7378-001-1.

T. Cipra. *Finanční ekonometrie*. Druhé upravené vydání. Ekopress s.r.o., 20013. ISBN 978-80-86929-93-4.

T. Herbst. An application of randomly truncated data models in reserving IBNR claims. *Insurance: Mathematics and Economics*, 25(1):123–131, 1999.

N. Keiding and R.D. Gill. Random truncation models and markov processes. *The Annals of Statistics*, 18(2):582–602, 1990.

D. Lynden-Bell. A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notices of the Royal Astronomical Society*, 155(1):95–118, 1971.

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

S. M. Ross. *Introduction to Probability Models*. Tenth Edition. Elsevier Inc., 2010. ISBN 978-0-12-375686-2.

M. Schmidt. Space distribution and luminosity functions of quasi-stellar radio sources. *The Astrophysical Journal*, 151:393–409, 1968.

Y. Vardi. Empirical distributions in selection bias models. *The Annals of Statistics*, 13(1):178–203, 1985.

M.C. Wang. A semi-parametric model for randomly truncated data. *Journal of the American Statistical Association*, 84(1):742–748, 1989.

M. Woodroofe. Estimating a distribution function with truncated data. *The Annals of Statistics*, 13(1):163–177, 1985.

M. Wüthrich and M. Merz. *Stochastic claims reserving methods in insurance*. John Wiley & Sons, 2008. ISBN 9780470723463.

List of Figures

List of Tables

Appendix

```
# data manipulation
data<-read.csv("C:/DIPLOMKA/Data/ckp-BM-2016.csv"
,sep=";",header=TRUE)
data<-subset(data, Reporting-Accident>0)
data<-subset(data, Payment-Reporting>=0)
beginning=min(data$Accident)-1
tau=max(data$Payment)+1
data<-transform(data,
Accident = Accident-beginning+1)
data<-transform(data,
Reporting = Reporting-beginning+1)
library(plyr)
library(ChainLadder)
library(MASS)
data<-ddply(data,~Number+Accident+Type
+Reporting,summarise,Amount=sum(Amount))
data<-transform(data,ReportingDelay =
Reporting-Accident)
days=c(366,365,365,365,366,365,365,365,366,
365,365,365,366,365,365,365,366)
dayscumul=cumsum(days)
# year 2000 0...366 year 2 367..731 etc...
dayscumul=c(0,dayscumul)
years=2000:2016
months=c(31,28,31,30,31,30,31,31,30,31,30,31)
months2=c()
for (i in 1:17){
months2=c(months2,months)}
for (i in 1:5)
{months2[2+12*4*(i-1)]=29}
# number of cumulative days separating months
cumulmonths=cumsum(months2)
# mapping accident and reporting year
YearAcc=c()
for (i in 1:length(data$Accident))
{ for (j in 1:17)
  {   if((data$Accident[i]>dayscumul[j])
      & (data$Accident[i]<=dayscumul[j+1]))
      {    YearAcc=c(YearAcc,years[j])
} }   }
YearRep=c()
for (i in 1:length(data$Accident))
{  for (j in 1:17)
  {
      if((data$Reporting[i]>dayscumul[j])
      & (data$Reporting[i]<=dayscumul[j+1]))
      {
      YearRep=c(YearRep,years[j])
```

```
        }
} }
data<-transform(data,YearA=YearAcc)
data<-transform(data,YearR=YearRep)
# separate bodily and material claims
bodily  <- subset(data, Type=="Bodily",)
material <- subset(data, Type=="Material",)
nB=length(bodily$Amount)
nM=length(material$Amount)
# daily numbers of accidents
dailyAccM=c()
for (i in 1:6210)
{dailyAccM=c(dailyAccM,sum(material$Accident==i))}

dailyAccB=c()
for (i in 1:6210)
{dailyAccB=c(dailyAccB,sum(bodily$Accident==i))}

# fit intensities
# material
dailydata=data.frame(x = 1:4749,
                     y = dailyAccM[1:4749])
new.dataDaily= data.frame(x=c(4749:6210))
m1CubicDay<-lm(y ~ I(x)+I(x^2)+I(x^3), data = dailydata)

#bodily
dailydataB=data.frame(x = 1:4749,
                     y = dailyAccB[1:4749])
m1CubicDayB<-lm(y ~ I(x)+I(x^2)+I(x^3), data = dailydataB)


##specify the times and considered data frame

start=1
tau=4749
tau1=4749
range=tau-start+1
bodilyCut=subset(bodily,start<=Accident&Accident<=tau1
& Reporting<=tau)
materialCut=subset(material,start<=Accident&
Accident<=tau1 & Reporting<=tau)
## numbers of observations
nBc=length(bodilyCut$Amount)
nMc=length(materialCut$Amount)

### get ``known'' G
# in case of reserves at end of 2016
# dailyIntensities=c(predict(m1CubicDay),
# predict(m1CubicDay, newdata=new.dataDaily))

dailyIntensities=c(predict(m1CubicDay))
```

```
lambdaM=dailyIntensities[start:tau]
distribM=c()
for (t in 1:range)
{distribM=c(distribM,sum(lambdaM[1:t]))}
Gmat=distribM/distribM[range]

# dailyIntensitiesB=c(predict(m1CubicDayB), #year 2016
# predict(m1CubicDayB, newdata=new.dataDaily))

dailyIntensitiesB=c(predict(m1CubicDayB))
lambdaB=dailyIntensitiesB[start:tau]
distribB=c()
for (t in 1:range)
{distribB=c(distribB,sum(lambdaB[1:t]))}
Gbod=distribB/distribB[range]


## IBNR reserves under known G
IBNRmatG=0
for (i in 1:length(materialCut$Amount))
{IBNRmatG=IBNRmatG+materialCut[i,]$Amount*
((Gmat[range]/Gmat[range-materialCut[i,]
$ReportingDelay])-1)}

IBNRbodG=0
for (i in 1:length(bodilyCut$Amount))
{IBNRbodG=IBNRbodG+bodilyCut[i,]$Amount*
((Gbod[range]/Gbod[range-bodilyCut[i,]
$ReportingDelay])-1)}

##CONFIDENCE BOUNDS
estimvarM=0
for (i in 1:length(materialCut$Amount))
{ estimvarM=estimvarM+(materialCut[i,]$Amount)^2*
(Gmat[range]/Gmat[range-materialCut[i,]
$ReportingDelay])*
((Gmat[range]/Gmat[range-materialCut[i,]
$ReportingDelay])-1)}
estimvarM=estimvarM/nMc
boundM=IBNRmatG+sqrt(estimvarM)*
sqrt(nMc)*qnorm(0.95)

estimvarB=0
for (i in 1:length(bodilyCut$Amount))
{ estimvarB=estimvarB+(bodilyCut[i,]$Amount)^2*
(Gbod[range]/Gbod[range-bodilyCut[i,]
$ReportingDelay])*
((Gbod[range]/Gbod[range-bodilyCut[i,]
$ReportingDelay])-1)}
estimvarB=estimvarB/nBc
boundB=IBNRbodG+sqrt(estimvarB)*
```

```
sqrt(nBc)*qnorm(0.95)

## IBNR reserves under unknown G

Gb <- ecdf(bodilyCut$Accident)
Fb <-ecdf(tau-bodilyCut$ReportingDelay)
bodilyCut<-transform(bodilyCut,
Cn=(Gb(Accident)-Fb(Accident-0.01)))
bodilyCut<-transform(bodilyCut,
nCn=nBc*(Gb(Accident)-Fb(Accident-0.01)))
bodilyCut<-transform(bodilyCut,frequencies=sapply
(bodilyCut$Accident,function(x)sum
(bodilyCut$Accident==x)))
bodilyCut<-transform(bodilyCut,insideprod =
1-(frequencies/nCn))

Gm <- ecdf(materialCut$Accident)
Fm <-ecdf(tau-materialCut$ReportingDelay)
materialCut<-transform(materialCut,
Cn=(Gm(Accident)-Fm(Accident-0.01)))
materialCut<-transform(materialCut,
nCn=nMc*(Gm(Accident)-Fm(Accident-0.01)))
materialCut<-transform(materialCut,frequencies=sapply
(materialCut$Accident,function(x)sum
(materialCut$Accident==x)))
materialCut<-transform(materialCut,insideprod =
 1-(frequencies/nCn))

#### MATERIAL IBNR
## distinct values of value sigma_i
distinctvalM=subset(materialCut, !duplicated(Accident))
npmleGmat=c()
for (t in 1:range)
{npmleGmat=c(npmleGmat,prod(subset(distinctvalM,
Accident>t+start-1)$insideprod)) }

# we multiplicate the expressions inside of the product
IBNRmat=0
for (i in 1:length(materialCut$Amount))
{
IBNRmat=IBNRmat+materialCut[i,]$Amount*
((npmleGmat[range-materialCut[i,]$ReportingDelay]^(-1))-1)
}
### BODILY IBNR
distinctvalB=subset(bodilyCut, !duplicated(Accident))
npmleGbod=c()
for (t in 1:range)
{ npmleGbod=c(npmleGbod,prod(subset
(distinctvalB,Accident>t+start-1)$insideprod))}

IBNRbod=0
```

```
for (i in 1:length(bodilyCut$Amount))
{IBNRbod=IBNRbod+bodilyCut[i,]$Amount*
((npmleGbod[range-bodilyCut[i,]$ReportingDelay]^(-1))-1)}

setEPS()
postscript("distfunctionsM.eps",width=5.4, height=4.3)
plot(Gm,ylim=c(0.85,1),xlim=c(quantile(materialCut
$Accident,0.84)
,tau),main="Distribution␣functions␣of␣occurence␣time")
lines(npmleGmat,col="blue")
lines(Gmat,col="red")
legend(tau-500
,0.89,legend=c("Empirical␣df", "Nonparametric␣MLE␣of␣G",
"df␣derived␣from␣fitted␣intensity" ),col=
c("black","blue","red"), lty=c(1,1,1),cex = 0.65)
dev.off()
setEPS()
postscript("distfunctionsB.eps",width=5.4, height=4.3)
plot(Gb,ylim=c(0.85,1),xlim=c(quantile(bodilyCut$Accident,0.84)
,tau),main="Distribution␣functions␣of␣occurence␣time")
lines(npmleGbod,col="blue")
lines(Gbod,col="red")
legend(tau-500
,0.89,legend=c("Empirical␣df", "Nonparametric␣MLE␣of␣G",
"df␣derived␣from␣fitted␣intensity" ),col=
c("black","blue","red"), lty=c(1,1,1),cex = 0.65)
dev.off()


#CHAIN-LADDER bodily 2012
I=13
J=13
endyear=2012
### incremental triangle
a=matrix(data=NA,nrow=I,ncol=J)

for (i in 1:I)
{ row=c()
  for (j in 1:J)
   {  if(j>=i)
     {  row=c(row,sum(bodily[ which
      ( bodily$YearA == 1999+i &
      bodily$YearR== 1999+j) , ]$Amount))
            }
   }
      if (i>1) {row=c(row,c(rep(0,i-1)))}
        a[i,]=(row)
}

#cumulative triangle
b=matrix(data=NA,nrow=I,ncol=J,dimnames =
```

```
list(c(2000:endyear),c(1:I)))

for (i in 1:I)
{       b[i,1]=a[i,1]
        for (j in 2:J)
        {if(i+j<=I+1)
                {b[i,j]=a[i,j]+b[i,j-1]}
        }
}

tricumulB <- as.triangle(as.matrix(b))
mackB<-MackChainLadder(Triangle = tricumulB, est.sigma = "Mack")
mackB


#CHAIN-LADDER material
I=13
J=13
C=matrix(data=NA,nrow=I,ncol=J)

for (i in 1:I)
{ row=c()
  for (j in 1:J)
 {  if(j>=i)
    {  row=c(row,sum(material[ which
       ( material$YearA ==  1999+i &
        material$YearR== 1999+j) , ] $Amount))
        }
 }
      if (i>1) {row=c(row,c(rep(0,i-1)))}
        C[i,]=(row) }

D=matrix(data=NA,nrow=I,ncol=J,dimnames =
 list(c(2000:endyear),c(1:I)))

for (i in 1:I)
{       D[i,1]=C[i,1]
        for (j in 2:J)
        {if(i+j<=I+1)
                {D[i,j]=C[i,j]+D[i,j-1]}
        }
}

tricumulM <- as.triangle(as.matrix(D))
mackM<-MackChainLadder(Triangle = tricumulM, est.sigma = "Mack")
mackM

####number of claims occurred in consecutive months
cumulmonths2=c(0,cumulmonths)

ClcountMonthB=c()
```

```
for (i in 1:204)
{ClcountMonthB=c(ClcountMonthB,
sum((bodily$Accident>cumulmonths2[i])&
(bodily$Accident<=cumulmonths2[i+1])))}

ClcountMonthM=c()
for (i in 1:204)
{ClcountMonthM=c(ClcountMonthM,
sum((material$Accident>cumulmonths2[i])&
(material$Accident<=cumulmonths2[i+1])))}

### monthly analysis of lambda

dfM <- data.frame(x = 1:156,
                  y = ClcountMonthM[1:156])
new.data <- data.frame(x=c(156:204))
mQuadr<-lm(y ~ I(x)+I(x^2), data = dfM)
mCubic<-lm(y ~ I(x)+I(x^2)+I(x^3), data = dfM)
a_start <-210
b_start <- -1.1
c_start <- 0.95
mGomp <- nls(log(y) ~  I(a+b*c^x), data = dfM,
start=list(a=a_start, b=b_start,c=c_start))

# mModExp <- nls(y ~  I(a+b*c^x), data = dfM,
# start=list(a=a_start, b=b_start,c=c_start))

plot(ClcountMonthM[1:204],ylim=c(0,450),
ylab="Number␣of␣monthly␣claims",xlab="Month")
lines(1:204,ClcountMonthM[1:204])
abline(v=156)
lines(predict(mQuadr),col="brown") #modif expo
lines(predict(mCubic),col="purple") #modif expo
lines(exp(predict(mGomp)),col="blue") #modif expo
lines(156:204,predict(mQuadr, newdata=new.data),col="brown")
lines(156:204,predict(mCubic, newdata=new.data),col="purple")
lines(156:204,exp(predict(mGomp, newdata=new.data)),col="blue")


setEPS()
postscript("claimscountfitM.eps")
plot(ClcountMonthM[1:204],ylim=c(0,380),
ylab="Number␣of␣monthly␣material␣claims",xlab="Month")
lines(1:204,ClcountMonthM[1:204])
abline(v=156)
lines(predict(mCubic),col="blue") #modif expo
lines(156:204,predict(mCubic, newdata=new.data),col="darkblue")
dev.off()

dfB <- data.frame(x = 1:156,
                  y = ClcountMonthB[1:156])
```

```
mCubicB<-lm(y ~ I(x)+I(x^2)+I(x^3), data = dfB)

setEPS()
postscript("claimscountfitB.eps")
plot(ClcountMonthB[1:204],ylim=c(0,75),
ylab="Number␣of␣monthly␣bodily␣claims",xlab="Month")
lines(1:204,ClcountMonthB[1:204])
abline(v=156)
lines(predict(mCubicB),col="blue") #modif expo
lines(156:204,predict(mCubicB, newdata=new.data),col="darkblue")
dev.off()


### data analysis

# reporting delay
for (i in 1:17)
{print(mean(subset(material,YearA==1999+i)$ReportingDelay))}

for (i in 1:17)
{print(mean(subset(bodily,YearA==1999+i)$ReportingDelay))}

sum(bodily$ReportingDelay>1461)
sum(material$ReportingDelay>1461)

summary(material$ReportingDelay[material$Accident<=4749])
summary(bodily$ReportingDelay[bodily$Accident<=4749])
sd(material$ReportingDelay[material$Accident<=4749])
sd(bodily$ReportingDelay[bodily$Accident<=4749])

setEPS()
postscript("histboth2012.eps",width=5.4, height=4.3)
par(mfrow = c(1, 2))
hist(subset(subset(data,Accident<=tau),
Type == "Bodily")$ReportingDelay, prob = TRUE,
 main = "Bodily", col = "red", xlab = "Delay")
hist(subset(subset(data,Accident<=tau),
Type == "Material")$ReportingDelay, prob = TRUE,
 main = "Material", col = "blue", xlab = "Delay")
dev.off()

setEPS()
postscript("delay2012.eps")
plot(ReportingDelay ~ Type,
data = subset(data,Accident<=tau),
col=(c("red","blue")), outline=FALSE,
main="Boxplots␣for␣reporting␣delay")
dev.off()

# severity
```

```
summary(bodily$Amount)
summary(material$Amount)
sd(material$Amount)
sd(bodily$Amount)

fitLN<-fitdistr(material$Amount,"log-normal")$estimate
fitG1<-fitdistr(material$Amount, dgamma,
list(shape = 1, rate = 0.5), lower = 0.01)$estimate

hist(material$Amount,breaks=100,xlim=c(0,350000),
ylim=c(0,0.00003),prob=TRUE)
lines(dlnorm(0:max(material$Amount),fitLN[1] ,
fitLN[2]),col="red")
lines(dgamma(0:max(material$Amount),fitG1[1] ,
fitG1[2]),col="blue")

fitLNb<-fitdistr(bodily$Amount,"log-normal")$estimate
fitGb<-fitdistr(bodily$Amount, dgamma, list(shape = 200,
rate = 0.5), lower = 0.01) $estimate

hist(bodily$Amount,breaks=1000,xlim=c(0,500000),
ylim=c(0,0.00003),prob=TRUE)
lines(dlnorm(0:max(bodily$Amount),fitLNb[1] ,
fitLNb[2]),col="red")
lines(dgamma(0:max(bodily$Amount),fitGb[1] ,
fitGb[2]),col="blue")


setEPS()
postscript("histpaymM.eps")
hist(log(material$Amount), breaks=50,
prob = TRUE, main="Histogram␣of␣logarithm
of␣material␣claim␣amount",xlab="Logarithm
of␣claim␣amount␣of␣material␣claims")
curve(dnorm(x, fitLN[1], fitLN[2]), col = "red", add = TRUE)
dev.off()

setEPS()
postscript("histpaymB.eps")
hist(log(bodily$Amount), breaks=50,
prob = TRUE,main="Histogram␣of␣logarithm
of␣bodily␣claim␣amount",xlab="Logarithm
of␣claim␣amount␣of␣bodily␣claims")
curve(dnorm(x, fitLNb[1], fitLNb[2]), col = "red", add = TRUE,)
dev.off()

setEPS()
postscript("qqplots.eps",width=5.4, height=4.3)
par(mfrow = c(1, 2))
qqnorm(log(material$Amount), pch = 1, main="Normal␣Q-Q
Plot-material")
```

```
qqline(log(material$Amount), pch = 1)
qqnorm(log(bodily$Amount), pch = 1, main="Normal␣Q-Q
Plot-bodily")
qqline(log(bodily$Amount), pch =1)
dev.off()
```