COLLEGE OF SCIENCE AND TECHNOLOGY
SCHOOL OF SCIENCES
DEPARTMENT OF MATHEMATICS

MODELLING EXTREME HEALTH INSURANCE CLAIMS USING GENERALIZED PARETO DISTRIBUTION

By

UWINGABIRE Jeannette
Student Number: 217130488

Project submitted in partial fulfillment of the academic requirements for the degree of

Master of Science
in
Applied Mathematics
Option of Statistical modeling and actuarial sciences

Supervisor: Dr. Joseph NZABANITA

June 2018

Declaration

I declare that this thesis has been composed solely by me and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgement, the work presented is entirely my own.

Student: UWINGABIRE Jeannette

Supervisor: Dr. Joseph NZABANITA

Acknowledgement

I would like to express my deep gratitude to my supervisor Dr. Joseph NZABANITA for his patient guidance, encouragement, useful comments and remarks during the planning and development of this master thesis.

Furthermore I would like to thank all the members in the Department of Mathematics at University of Rwanda (UR). In particular, all the lecturers from UR-Sweden program who fully contributed to my academic studies. I thank the University of Rwanda (UR) and Linköping University (LiU) for your partnership. I would like to thank RSSB institute, especially the unit of planning and statistics for their help in offering me the data.

I like to thank all the postgraduate students in the department for their support and help. I would also like to thank my parents for their endless support, encouragement and sympathetic ear through my study. Finally, I offer great thanks to our Almighty God, for protection.

Abstract

Extreme value theory has been used to develop models for describing the distribution of rare events. The generalized Pareto distribution is a very popular two parameter model for extreme events. It was first introduced by Pikands (1975). It is a family of continuous probability distributions used to model extreme value above a given threshold. In this study, we determined the extreme health insurance claims from RSSB and its behavior (distribution). In the methodology the project shows how to choose a threshold. After choosing appropriate threshold, maximum likelihood estimation method was used to estimate parameters because of its efficiency. In application, we used the diagnostic plots to show that the generalized Pareto distribution fit well extreme claims. Estimation of return level gives estimate of the amount of claims RSSB would pay in a given period of time. In data set, the average time elapsing between two successive realizations of the highest value itself is between 10 and 12 years with a probability between $\frac{1}{10}$ and $\frac{1}{12}$. By comparing GPD and Exponential distribution, the result showed that the Exponential distribution fit data better than GPD.

Contents

Contents **v**

List of Figures **vii**

List of Tables **ix**

1 INTRODUCTION **1**

1.1 Background 1

1.2 Problem statement 3

1.3 Objectives of thesis 3

1.4 Structure of thesis 3

2 LITERATURE REVIEW **5**

2.1 Extreme value theory 5

2.2 Peaks-over-threshold 6

2.3 Generalized Pareto distribution model 7

3 METHODOLOGY **9**

3.1 Data description 9

3.2 Threshold Value selection 9

3.3 Estimation of Parameters 10

3.4 Goodness of fit 11

3.5 Return levels (quantiles) estimation 11

3.6 Stationary Tests 13

4 DATA ANALYSIS AND DISCUSSION OF RESULTS **15**

4.1 Descriptive statistics 15

4.2 Threshold Value selection 17

4.3 Estimation of parameters 20

4.4 Model checking 20

4.5 Return level 22

4.6 Exponential Distribution 22

4.7 Model Selection 23

5 CONCLUSION **27**

A R Codes **31**

List of Figures

4.1 Monthly claims data 16
4.2 Scatter plot of claims against months 16
4.3 Scatter plot of claims against years 17
4.4 Mean residual life plot 18
4.5 Threshold range plots 18
4.6 Diagnostic plots for GP 21
4.7 Diagnostic plots for Exponential distribution 23

List of Tables

4.1 Summary statistics. 17
4.2 Comparison between GP distribution models fitted for 13, 13.5 and 14 billion thresholds. 19
4.3 Estimation of parameters for GP. 20
4.4 Return level for GP. 22
4.5 Estimation of parameter for Exponential. 22
4.6 Return level for Exponential distribution. 23
4.7 Comparison between GP and Exponential distribution. 24
4.8 Comparison between return level for GP and Exponential distribution models. 25

Chapter 1

INTRODUCTION

1.1 Background

Health insurance is one of the insurance services that cover the whole or a part of the risk of a person incurring medical expenses, spreading the risk over a large number of persons. Every human being is exposed to be attacked by different diseases and to make accidents at various work places depending on the structure of his or her field. This introduces the need for every person to have access to one of the insurance institutions to cover some of the charges of his/her treatments. In order to help people in those situations in Rwanda, all Rwandans must be covered by a health insurance according to the law. RAMA, MMI and Community Based Health Insurance are the main structures of health insurance and all included in RSSB (Rwanda Social Security Board). Additionally of those institutions, private insurance companies provide also health insurance products. Extreme events in insurance may obviously correspond to individual (or indeed grouped) claims which by far exceed the capacity of a single insurance company. Extreme claim events cause a financial impact on (re)insurance, they are difficult to predict a long time in the future, difficult to set the premium etc, (Ministry of health (2012) and Matthew (2002)).

Extreme value theory has been used to develop models for describing the distribution of unusual events. Extreme value theory also plays an important role in management of insurance risk, for instance the insurance company desires to consider the potential risk of large claims (e.g. tremor, volcanic eruptions, landslides, diseases, accidents, and other risks). The actuary will estimate the price of the insurance (net premium) to cover the potential risk of insurance. Particularly, it is of great importance for re-insurance. McNeil (1997) studies the classic Danish insurance data by applying the peaks over threshold method. Resnick (1997) acknowledges that McNeils work is of great importance in insurance claim application and an excellent example of adopting the extreme value model. Embrechts (1999) demonstrates that extreme value theory is a very useful tool for management of insurance risk.

The extreme value theory based models can be used for asymptotically approximating the behavior of the tail(s) of the distribution function. The extreme value theory (EVT), traditionally used in fields similar to hydrology, meteorology, finance and insurance like what we do in this thesis. Ren and Giles (2007) described EVT as a theory for assessing the asymptotic probability of extreme values, additional expanding that the theory models the distribution of tail part anywhere the risk exists. The two main methods of the EVT approach are the GEV model (Block Maxima) and The Generalized Pareto Distribution, which can give a good model for the upper tail, providing reliable extrapolation for exceedances above an appropriately high threshold. This is called the Peaks-Over-Threshold (POT) method. The method preferred in this thesis of modeling extreme claims of health insurance in RSSB is POT models. These are models for all large observations which exceed a high threshold. The POT has been demonstrated empirically to efficiently utilize more of the data therefore produce more consistent findings compared to the Block Maxima approach [McNeil and Frey (2000); Matthys and Beirlant (2000); Coles (2001); Blum et al. (2002); Gilli and Kellezi (2006)], also it denes which part of the data can be considered as extreme Bommier (2014).

The POT method has been used in many fields to classify extremal events such as floods, wind velocities, wave heights, insurance claims, etc. This method delivers a model for independent exceedances above a high threshold. It is clear that this method needs the determinations of a threshold which is neither too high nor too low. In application of such a model the choice of threshold is essential, as it defines which part of the data can be considered as extreme. We can use the generalized Pareto distribution in this way, to provide a good fit model to extremes of data. The generalized Pareto distribution allows a continuous range of possible shapes that includes both the exponential and Pareto distributions as special cases. We can use either of those distributions to model a particular data set of exceedences. The generalized Pareto distribution allows us to decide which distribution is appropriate. The generalized Pareto distribution has three basic forms, each corresponding to a limiting distribution of exceedence data from a different class of underlying distributions.

- Distributions whose tails decrease exponentially, leads to shape parameter of zero (called Exponential).
- Distributions whose tails decrease as a polynomial, leads to a positive shape parameter (called Pareto).
- Distributions whose tails are finite, leads to a negative shape parameter (called Beta).

1.2 Problem statement

Health insurance is insurance against loss through illness of the insured, especially is insurance providing compensation for medical expenses. An extreme claim in insurance is the individual or group claims which far exceed the capacity of insurance. This, naturally, concentrates on insurance, for which extreme claim events have clear impacts on reserving, pricing and solvency. Insurance must control and quantify the risks, especially those whose impact is considerable such as the large claims. In addition, gap of sufficient knowledge associated with the extreme claims make some difficulties to predict the occurrence of such extreme events. This study applying a scientific methodology, the extreme value theory, to find an estimate model to explain extremely events and to estimate the optimal threshold of extreme claims in RSSB.

1.3 Objectives of thesis

Main objective
The main objective of this thesis is to model extreme claims of health insurance in RSSB.
Specific objectives are:

1. To select an appropriate threshold.
2. To describe the behavior of model distribution fitting extreme claims.
3. To estimate parameters.
4. To estimate return levels.

1.4 Structure of thesis

This study has been organized and grouped into the following sections.Chapter one consists of the introduction of the study. Chapter two consists of the literature review on statistical topics including extreme value theory, pick-over-threshold and Generalized Pareto distribution. Chapter three covers the methodology of the study. Chapter four covers data analysis and discussion of results and chapter five covers conclusion.

Chapter 2

LITERATURE REVIEW

This chapter reviews the relevant literature regarding the extreme value theory, Peaks-over-threshold and Generalized Pareto distribution model.

2.1 Extreme value theory

Historically, work on extreme value problems can be traced back to as early as when Nicholas Bernoulli discussed the mean largest distance from the origin given n points lying at random on a straight line of a fixed length t by Gumbel (1958). Probably the first paper that described an application of extreme values in flood flows was done by Fuller (1914) and Griffith (1920) brought out an application while discussing the phenomena of rupture and flow in solids. In the paper written by von Bortkiewicz (1922) may have contributed to a systematic development of extreme value theory. His paper dealt with the distribution of range in random samples from a normal distribution. The concept of distribution of largest value was introduced for the first time. Von Mises (1923) evaluated the expected value of this distribution. Dodd (1923) calculated its median and discussed some non-normal parent distributions.

A paper that had more direct relevance to the extreme value theory was written by Frechet (1927) in which he discussed the asymptotic distributions of largest values. Fisher and Tippett (1928) published the results of their research into the same problem. In addition, they showed that extreme limit distributions can only be one of three types. Von Mises (1936) presented sufficient conditions for the weak convergence of the largest order statistic to each of the three types of limit distributions given by Fisher and Tippett. Gnedenko (1943), in this breakthrough paper, presented a solid foundation for the extreme value theory and provided necessary and sufficient conditions for the weak convergence of the extreme order statistics. Gnedenkos work was refined later on by many others that include Mejzler (1949) and de Haan (1970).
Following the theoretical developments of the extreme value theory during 1920s and mid-

1930s, many scholarly papers dealing with the variety of practical applications of the theory were published in late 1930s and 1940s. Gumbel (1958) played a pioneering role during 1940s-1950s and from the application point of view. He made many significant contributions to the extreme value theory. He presented all of these in his statistics of extremes and this work was pivotal in promoting extreme value theory as a tool for modeling the extremal behavior of observed physical processes, Gumbel (1954).

The Generalized Extreme Value (GEV), Gumbel, Frechet, Weibull, and the Generalized Pareto (GP) distributions are just the tip of the iceberg of an entirely new and quickly growing branch of statistics. The Gumbel distribution has light or medium tails. Frechet distribution has heavy tails and Weibull distribution has bounded or short tails. There are two widely used approaches available to analyse extreme data, that are: the block- maxima approach and the peaks-over-threshold (POT) approach. The peaks-over-threshold plays an important role in risk management, finance, insurance, economics, hydrology, material sciences, telecommunications and other industries where risky extreme events occur with very small probability, Hanson and Vogel (2008). The application of extreme-value theory to insurance is discussed by Beirlant et al. (1994), Mikosch (1997), McNeil (1997), McNeil and Saladin (1997) with application to Danish data on large fire insurance losses, Rootzen and Tajvidi (1997) with application to Swedish windstorm insurance claims.

2.2 Peaks-over-threshold

The POT method has been used in many fields to identify extremal events such as loads, wave heights, floods, wind velocities, insurance claims, etc. McNeil and Saladin (1997) looked at peaks over threshold method to estimating high quartiles of loss distribution. They discussed the use of peaks over threshold method high quartiles estimation and the possible relevance to excess loss insurance in high layers. Van Montfort and Witter (1985) studied deeply exponential distribution in GPD to evaluate the adequacy of the exponential distribution for a set of data, Hosking and Wallis (1987) studied about several estimators for the GPD model, Hamed and Rao (1999) estimated the frequency flood of river by the GPD model and related the magnitude of extreme events to their frequency of occurrency. Davison and Smith (1990) pointed out that the GPD might form the basis of a broad modeling approach to high-level exceedances. DuMouchel (1983) applied it to estimate the stable index to measure tail thickness, van Montfort and Witter (1985, 1986) and van Montfort and Otten (1991) applied the GPD to model the peaks over a threshold (POT) streamflows and rainfall series, and Smith (1984, 1987, 1991) applied it to analyze flood frequencies and wave heights.

Extending the concept of EVT to the insurance industry, McNeil (1997) used the Danish insurance data to highlight the relevance of Generalized Pareto Distributions (GPD), as a

subclass of GEV, for EVT. He dealt with the parameter estimation and curve fitting for modeling rare historical losses in non-insurance sector. He also dealt with the concept of loss severity and showed how to model the aggregate payments depending on the number of losses. Davison and Smith (1990) discussed the analysis of the extremes of data by modeling the sizes and occurrence of exceedances over high thresholds. They also used GPD model to evaluate river ows and wave heights. Hanson and Vogel (2008) studied the stochastic daily precipitation in U.S using a probability distribution function (namely, L-Moment diagrams and probability plot correlation coefficient analysis) to describe rainfall amounts on wet-day.

2.3 Generalized Pareto distribution model

The generalized Pareto distribution (GPD) was introduced by Pickands (1975) and has since been applied to a number of areas including socio-economic phenomena, physical and biological processes by Saksena and Johnson (1984), dependability studies and the analysis of environmental extremes. Davison and Smith (1990) pointed out that the GPD might form the basis of a broad modeling approach to high-level exceedances. DuMouchel (1983) applied it for estimating the stable index to measure tail thickness, while Davison (1984a, 1984b) modeled contamination due to long-range atmospheric transport of radionuclides, van Montfort and Witter (1985, 1986) and van Montfort and Otten (1991) applied the GPD to model the peaks over threshold (POT) stream flows and rainfall series. Similarly, Joe (1987) employed it to estimate quantiles of the maximum of observations. Wang (1991) applied it to develop a POT model for flood peaks with Poisson arrival time, while Rosbjerg et al. (1992) compared the use of the 2-parameter GP and exponential distributions as distribution models for exceedances with the parent distribution being a GPD.

In an extreme value analysis of the flow of Burbage Brook, Barrett (1992) used the GPD to model the POT flood series with Poisson inter-arrival times. Davison and Smith (1990) presented a comprehensive analysis of the extremes of data by using the GPD for modeling the sizes and occurrences of exceedances over high thresholds. Methods for estimating the parameters of the 2-parameter GPD were reviewed by Hosking and Wallis (1987). Quandt (1966) used the method of moments (MOM), while Baxter (1980) and Cook and Mumme (1981) used the method of maximum likelihood estimation (MLE) for the Pareto distribution.Van Montfort and Witter (1986) used the MLE to fit the GPD to represent the Dutch POT rainfall series and used an empirical correction formula to reduce bias of the scale and shape parameter estimates. Davison and Smith (1990) used the MLE, PWM, a graphical method and least squares to estimate the GPD parameters.
Wang (1991) derived the PWM for both known and unknown thresholds. Before describing the outcome of insurance, it is of great importance to include the occurrence of extremal events. These rare events can due to their mere size have a detrimental effect on the pricing

for the insurance holders. Specifically, small groups can receive overpriced premiums if a large claim affects a person within it. One of the most useful methods of describing extremal events is fitting the data with a GPD, the Generalized Pareto Distribution to the upper tail. (which is defined as the data above a certain threshold). Embrechts, et al. (2012) Here the thought is that for the threshold that the GPD distribution has its best fit, the data below the threshold is considered non-extreme and can thus be treated regularly.Let u denote the threshold and X be a random variable with distribution function F.
The exceedances distribution function over the threshold u is now given by
$F_u(x) = Pr(X - u \leq x / X > u), x > u$
By the conditional probabilities, F_u can also be defined as

$$F_u(x) := \begin{cases} \frac{F(u+x)-F(u)}{1-F(u)}, & if\ x \geq 0 \\ 0, & else \end{cases} \tag{2.1}$$

Let $Y = X - u$ for $X > u$ and for n observed variables $X_1, X_2, X_3, ..., X_n$, we can write $Y_j = X_i - u$ such that i is the index of the j^{th} exceedances, $j = 1, 2, ..., n_u$. The distribution of the exceedances $(Y_1, Y_2, Y_3, ..., Y_{n_u})$ can be approximated by a Generalized Pareto Distribution Pickands, (1975) defined as:

$$G(x\ u, \delta_u, \xi) := \begin{cases} 1 - \left[1 + \left(\frac{\xi y}{\delta_u}\right)\right]^{-\frac{1}{\xi}}, & if\ \xi \neq 0 \\ 1 - \exp\left[-\left(\frac{y}{\delta_u}\right)\right], & if\ \xi = 0 \end{cases} \tag{2.2}$$

,

where $y \geq 0$ if $\xi \geq 0$ and $0 \leq y \leq -\frac{\delta_u}{\xi}$ if $\xi < 0$. Chiang Lee (2009) applying generalized Pareto distribution to the risk management of commerce fire insurance to model and to estimate the tail parameters. Using extreme value theory, he centralized on the GPD and compared with standard parametric model based on Log-normal, Exponential, Gamma and Weibull distributions. The thresholds of GPD are determined through mean excess plot (mean residual plot) and Hill plot in the empirical study. The empirical results show that the GPD method theoretically is well supported technique for fitting a parametric distribution to the tail of an unknown underlying distribution. It can capture the tail behavior of commercial fire insurance loss very well by Chiang Lee (2009).

Chapter 3

METHODOLOGY

This third chapter gives the reader a clear view of how this thesis was carried out. This chapter defines the thesis design, data source and data analysis technique. This chapter includes the methods and the statistical techniques used in analyzing the data of the study.

3.1 Data description

The data under consideration is a secondary data obtained from the unit of planning and statistics of Rwanda Social Security Board (RSSB). In this thesis the data considered is the monthly health insurance claims spanning from January 2012 up to December 2016, thus constituting 60 observations. A peak over threshold (POT) is used to determine extreme claims and R software used to analyze those extreme claims.

3.2 Threshold Value selection

Threshold selection is a selection method which has three different approaches to select the threshold u: Parameter Stability is based on the selection of the threshold on fitting the generalized Pareto distribution at a range of thresholds and looking for stability of parameter estimates. Behrenset al. (2004) mentioned another way to select the threshold which called Bayes Estimation. This model contains uncertainty because a prior, possibly flat, for u is chosen. He proposed a model to fit data characterized by extremal events where the threshold is defined as another model parameter. The third approach is the Mean Residual Life Plot which is based on the mean of the generalized Pareto distribution.

The MRLP also is used to determine the adequacy of the GPD model in practice. A straight line from bottom left to top right of the MRLP is the characteristic of a fat tailed GPD with positive shape parameter ($\xi > 0$), a plot of a down sloping line from top left to bottom right indicates thin tailed character and a straight horizontal line ME plot indicates

exponential. Consider the generalized Pareto distribution as a good model for the excesses of a threshold u_0 generated by a series $X_1, X_2, X_3, ..., X_n$, where X is any period. If GPD is valid for excesses of the threshold u_0, it should be also valid for all $u > u_0$, choosing an adequate change of scale parameter δ_u.
Therefore we have the expected value of our threshold excesses, conditional on being greater than the threshold equals to

$$E(X - u/X > u) = \frac{\delta_u}{1-\xi} = \frac{\delta_{u_0} - \xi u}{1-\xi},$$

where δ_{u_0} is the GPD scale parameter for excesses over threshold u_0. For all $u > u_0$, $E(X - u/X > u)$ is the mean of the excesses of the threshold u. this expectation is a linear function of u means that the estimates might change linearly with u, at level of u for which the generalized Pareto model is appropriate. In this thesis we use Mean Residual Life Plot to select the threshold. Choice of the threshold u is a key problem in GPD model, being a premise of exact estimating parameter δ and ξ and an insurance premium as well. In threshold determination, we face a tradeoff between bias and variance. If we choose a low threshold, the estimation becomes biased, while a high threshold will produce higher variance.

3.3 Estimation of Parameters

Three of the most popular methods of parameter estimation are the method of moments (MOM), the method of probability-weighted moments (PWM), and the method of maximum likelihood estimation (MLE) used for the GPD parameters estimation. In this thesis we are used maximum likelihood estimation to estimate the GPD parameters because this method is very important for large sample also its asymptotic behavior is the best. The maximum likelihood estimator (MLE) is not always valid and regularity conditions do not always exist. The MLE is valid for $\xi > -1$ but the asymptotically normal properties of the MLE are only valid for $\xi > -\frac{1}{2}$. When $\xi < -1$ maximum likelihood estimators generally do not exist, Gumbel (1954).
Let $X_1, X_2, X_3, ..., X_n$ be a sequence of k exceedances of a threshold u. The likelihood function from (2.1) is

$$L(x_i, \delta, \xi) = \prod_{i=1}^{n} f(x_i, \delta, \xi), \tag{3.1}$$

where $f = \frac{dF}{dx}$
For $\xi \neq 0$ the log-likelihood is derived from (2.2) as

$$l(\delta,\xi) = -k\log\delta - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k}\log\left(1+\frac{\xi y_i}{\delta}\right) \tag{3.2}$$

Provided $(1+\frac{\xi y_i}{\delta})$ for $i = 1,2,3,...,k$ otherwise $l(\delta,\xi) = \infty$. The values δ and ξ are the estimation parameters which maximize the equation (2.2). For $\xi = 0$ the log-likelihood is derived from (2.2) as

$$l(\delta) = -k\log\delta - \left(\frac{1}{\delta}\right)\sum_{i=1}^{k} y_i$$

3.4 Goodness of fit

The problem of goodness-of-fit tests for GPD model was first investigated by Davis and Smith (1990). The Anderson Darling statistic A^2 is a modification of the Cramrvon Mises statistic W^2 giving extra heaviness to observations in the tail of the distribution, which is useful in detecting outliers. Based on Cramer-von Mises statistic W^2 and the Anderson-Darling stastic A^2, the goodness-of-fit test procedure is as follows: We have the random sample $X_1, X_2, X_3, ..., X_n$ come from GPD model. Using the maximum likelihood estimate to find the estimates of unknown parameters δ and ξ in GPD model, and make the transformation $z_i = F(x_i)$ for $i = 1,2,3,...,n$ when $X_1, X_2, X_3, ..., X_n$ be the order statistics.
Calculate statistics $W^2$and A^2 like this:

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left(z_i - \frac{2i-1}{2n}\right)^2 \tag{3.3}$$

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\log(z_i) - \log(1-z_{n+1-i})\right] \tag{3.4}$$

3.5 Return levels (quantiles) estimation

The return level of an extreme event, defined as the value x_T, such that there is a probability of T that x_T is exceeded in any given year, or alternatively, the level that is expected to be exceeded on average once every $\frac{1}{T}$ year ($\frac{1}{T}$ is called the return period); in extreme value terminology, x_T is the return level associated with the return period $\frac{1}{T}$. The return period is the period expressed in years by which the annual observation is expected to return. The amount by which an event exceeds the threshold can be modeled by the generalized Pareto distribution,

$$Pr(X > x/X > u) = \left[1 - \xi\left(\frac{x-u}{\delta}\right)\right]^{\frac{1}{\xi}} \tag{3.5}$$

It follows that

$$Pr(X > x/X > u) = \zeta_u \left[1 - \xi \left(\frac{x-u}{\delta}\right)\right]^{\frac{1}{\xi}}, \tag{3.6}$$

where $\zeta_u = Pr(X > u)$ is the probability of the occurrence of an exceedance of a high threshold u,
Therefore the level x_T that is exceeded on average once every T observations is the solution of

$$\zeta_u \left[1 - \xi \left(\frac{x-u}{\delta}\right)\right]^{-\frac{1}{\xi}} = \frac{1}{T}$$

$T = \frac{1}{Pr(X>x)}$. The return level x_T is the quantile having return period T, i.e. the $1 - \frac{1}{T}$ quantile. Equation (3.4) can be rewritten in terms of the return level x_T, that is if $\xi \neq 0$

$$x_T = u + \frac{\delta}{\xi}\left[(\zeta_u T)^{\xi} - 1\right]$$

Which is valid for T large to ensure that $x > u$.
When $\xi = 0$ the T-year return level is,

$$x_T = u + \delta \log(\zeta_u T) \tag{3.7}$$

x_T is called the T observation return level. If we are interested in the N-year return level, let n_y be the number of observations per year, then $T = N \times n_y$. Therefore, the N-year return level is

$$x_N := \begin{cases} u + \frac{\delta}{\xi}\left[(N n_y \zeta_u)^{\xi} - 1\right], & \xi \neq 0 \\ u + \delta \log(N n_y \zeta_u), & \xi = 0 \end{cases} \tag{3.8}$$

This equation suggests that in order to determine the N-year return level, three parameters need to be fitted δ, ξ and ζ_u. If we assume that the exceedances of a high threshold u are rare events, ζ_u could be expected to follow a Poisson distribution. Here, we deviate slightly from Coles (2001) who suggests that the number of exceedances of u follows the binomial distribution Bin (n, ζ_u). The Poisson distribution is characterized by the parameter λ, which is the mean of threshold exceedances per unit time. Then, ζ_u can be estimated as

$$\zeta = \frac{\lambda}{n^y}$$

Where an unbiased estimate of λ is given by
$\lambda = \frac{n_u}{M}$

with n_u the number of exceedances over the selected threshold u and M the number of years of records. Reformulating equation (3.8) in terms of λ, we get

$$x_N := \begin{cases} u + \frac{\delta}{\xi}\left[(\lambda N)^{\xi} - 1\right], & \xi \neq 0 \\ u + \delta \log(\lambda N), & \xi = 0 \end{cases} \tag{3.9}$$

δ and ξ are estimated using the maximum likelihood method.

3.6 Stationary Tests

The stationary model is the model where the model parameters are fixed constant. Stationary tests including graphic examination, KPSS (Kwiatkowski, Philips, Schmidt and Skin) and non-parametric Mann-Kendall tests should be carried out since the assumption of stationarity is essential for the application of GEV and GPD. The KPSS stationary test (Kwiatkowski, Phillips, Schmidt and Shin) judges whether the trend is stabilized around a constant, a linear line or non-stationary by Hasna and Chung (2010). The test statistics are compared with critical values at different significant levels. Therefore, the null (H_0) and alternative (H_1) hypotheses are:
H_0 : Stationary around a constant or a linear trend,
H_1 : The trend is non-stationary.
The p-value of the null hypothesis is used to determine the tendency of the observations. The null (H_0) and alternative (H_1) hypotheses are:
H_0: There is no trend,
H_1: There is an increasing/ decreasing trend by Ender and Ma (2014).

Chapter 4

DATA ANALYSIS AND DISCUSSION OF RESULTS

In this fourth chapter, Generalized Pareto Distribution is used to model the extreme monthly health insurance claims data from RSSB. It also deals with comparison between Generalized Pareto distribution and Exponential distribution models to see which one is fitting data better than an other.

4.1 Descriptive statistics

Descriptive statistics provide simple summaries about the given data set together with simple graphics analysis. The time series plot and scatter plots of monthly claims are shown in Figure 4.1 and Figure 4.2 respectively.

Figure 4.1 represents the time series plot of monthly claims data. These data show random variation. There are no clear patterns or cycles means that the cycles do not repeat at regular intervals and do not have the same shape.

A scatter plot is an important diagnostic tool in the statistics. It obtained by graphing two variables against each other. A scatter plot displays a relationship between two sets of data. it can also be called a scatter gram or a scatter diagram. In a scatter plot, a dot represents a single data point. With several data points graphed, a visual distribution of the data can be seen.

Figure 4.2 represents the scatter plot of monthly health insurance claims against months from 2012 up to 2016 and shows that many extreme health insurance claims occur from March to June.

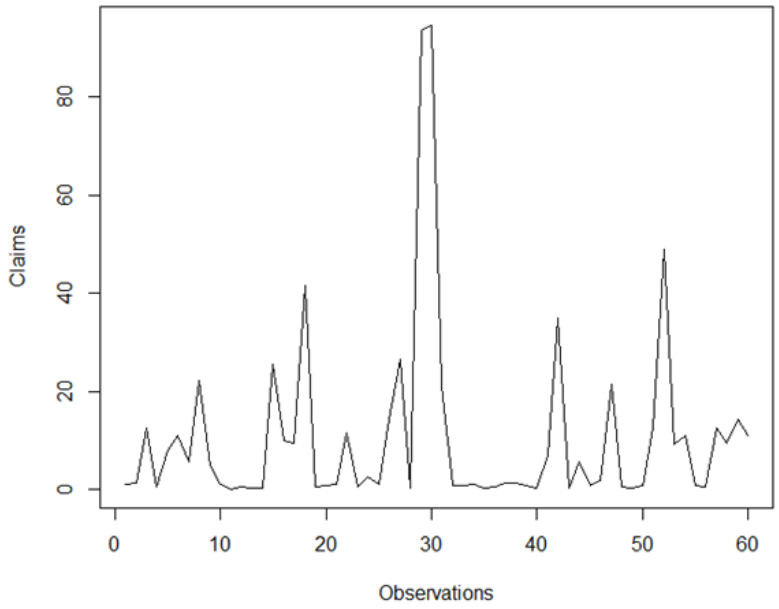

Figure 4.1: Monthly claims data

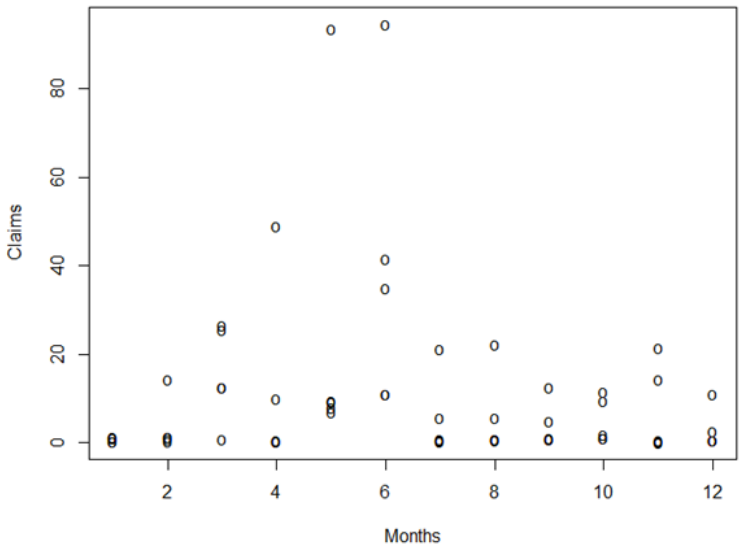

Figure 4.2: Scatter plot of claims against months

Figure 4.3 represents the scatter plot of monthly health insurance claims against years from 2012 up to 2016 and shows that many extreme health insurance claims occur in 2014. Table 4.1 represents the summary statistics of monthly health insurance claims in billions from RSSB for 5 years (2012-2016). The minimum claim is 0.19, maximum claim is 94.5 and the mean of claims is 10.6580.

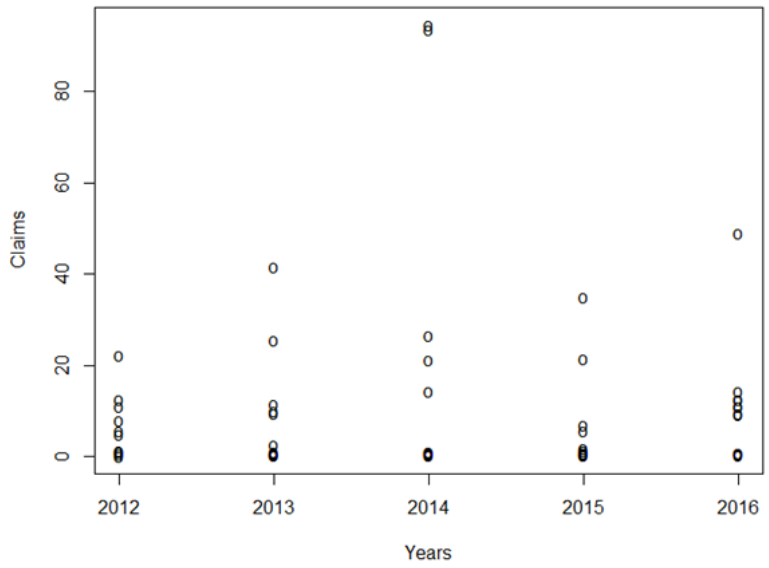

Figure 4.3: Scatter plot of claims against years

Table 4.1: Summary statistics.

Min	1^{st} Quartile	Median	Mean	3^{rd} Quartile	Max
0.19	0.7725	1.7150	10.6580	11.7350	94.5.

4.2 Threshold Value selection

MRLP is the one of the methods used to select (to choose) a value of threshold u because the choice of threshold is a key problem in GPD model. It is very important to select an appropriate threshold before fitting a GP distribution. It should be at least greater than the mean but can not be too high. To determine which threshold is the best, we can use the mean residual life plot and the threshold range plots.

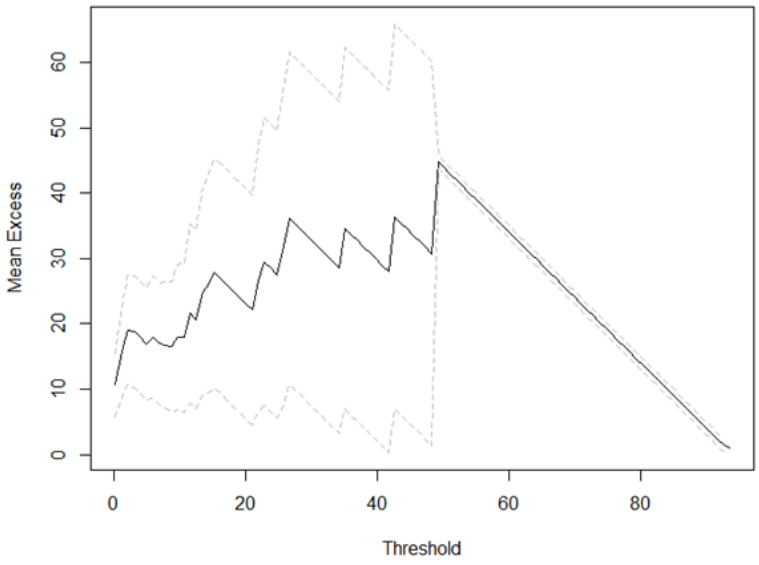

Figure 4.4: Mean residual life plot

The idea behind in Figure 4.4 is to find the range of threshold where the plot is nearly linear, taking into account the 95% confidence bounds. The plot appears roughly linear from about 10 to 28 billions.

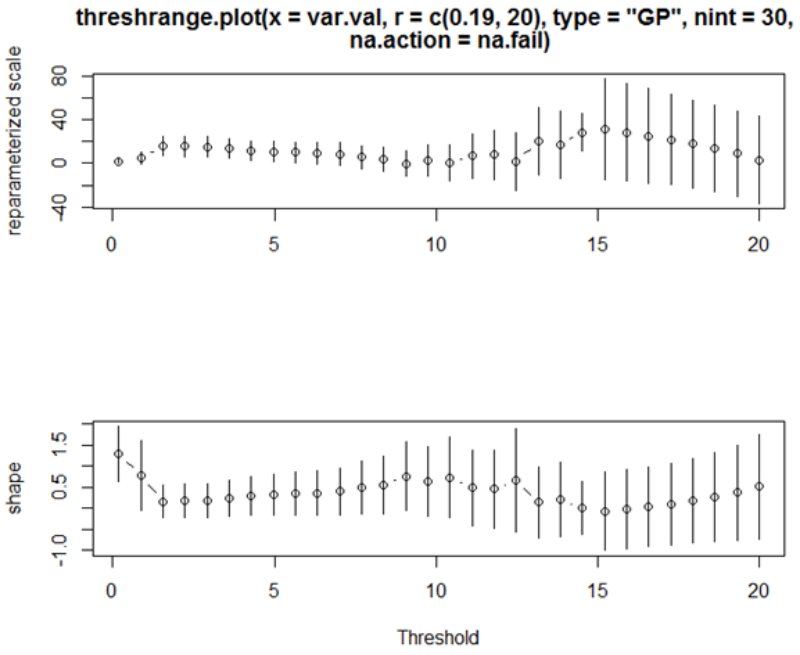

Figure 4.5: Threshold range plots

Figure 4.5 represents where there is a reasonable variation (stability) and the confidence interval bars start to grow large. The value around those situations is the thresholds. By looking at Figure 4.5, it is seen that there is a reasonable variation above u = 14 billion. All this doesn't mean that, we can confirm immediately that 14 billion is the best threshold. It is better to compare the model distributions for different values around those situations and see the best. So let us compare the model distributions for 13 and 13.5 and 14 billion thresholds .

Table 4.2: Comparison between GP distribution models fitted for 13, 13.5 and 14 billion thresholds.

Threshold=13	Threshold=13.5	Threshold=14
Estimation method used: MLE	Estimation method used: MLE	Estimation method used: MLE
Negative Log- Likelihood value: 50.71348	Negative Log- Likelihood value: 50.43524	Negative Log- Likelihood value: 50.13556
Estimated parameters: Scale Shape 22.2524275 0.1225031	Estimated parameters: Scale Shape 20.8021619 0.1676821	Estimated parameters: Scale Shape 19.2347512 0.2210936
Standard Error Estimates: Scale Shape 11.3119832 0.4206947	Standard Error Estimates: Scale Shape 10.8232591 0.4356539	Standard Error Estimates: Scale Shape 10.3022874 0.4556447
Estimated parameter covariance matrix: Scale Shape Scale 127.960965 -3.6785390 Shape -3.678539 0.1769841	Estimated parameter covariance matrix. Scale Shape Scale 117.142938 -3.6083055 Shape -3.608306 0.1897944	Estimated parameter covariance matrix: Scale Shape Scale 106.13713 -3.5609602 Shape -3.56096 0.2076121
AIC=105.427	AIC = 104.8705	AIC=104.2711
BIC=106.3968	BIC = 105.8403	BIC=105.2409

Table 4.2 shows that the scale and shape parameters standard errors, negative log-likelihood, AIC (Akaike information criterion) and BIC (Bayesian information criterion) are small for fit which was calculated using a threshold of 14 billion by comparing to others calculated for 13 and 13.5 billion as thresholds. This implies that 14 billion is a threshold which has a better distribution fit compared to 13 and 13.5 billion. So the best threshold chosen is 14 billion.

4.3 Estimation of parameters

After choosing an appropriate threshold, we estimate parameters (scale and shape). For this case, the maximum likelihood estimate is preferred as method of parameters estimation because this method is very important for large sample also its asymptotic behavior is the best. In order to perform MLE well for the GPD, there are certain criterions to be met:

- The sample size must be large (greater than 50 sample size)
- The values of shape parameter estimate must stay within -0.5 and 0.5; it means that $-0.5 \leq \xi \leq 0.5$.

If these criterions are met the MLE would be preferred due to its effective efficiency with large samples.

Table 4.3: Estimation of parameters for GP.

Parameters	Estimated parameters	95% Lower CI	95% Upper CI
Scale	19.2347512	-0.9573610	39.426863
Shape	0.2210936	-0.6719536	1.114141

Table 4.3 shows that the estimation parameters shape and scale are 0.2210936 and 19.2347512 respectively and 95% confidence interval by performing maximum likelihood estimate method with defined threshold equal to 14 billions. The shape parameter estimate is 0.221093, the $-0.5 \leq \xi \leq 0.5$, the sample size is 60, criterions are satisfied, this means that the maximum likelihood estimator is efficient.

4.4 Model checking

After obtaining the proper threshold of the fitted GPD, we need to assess the quality of the fitted generalized Pareto model. It can be done using probability-probability(pp) plots, Quantile-Quantile (QQ) plots Density plot and return level plots. According to the theory, if the model is correct, both the probability and quantile plots should lie these points approximately linear(approximately around $y = x$ line).

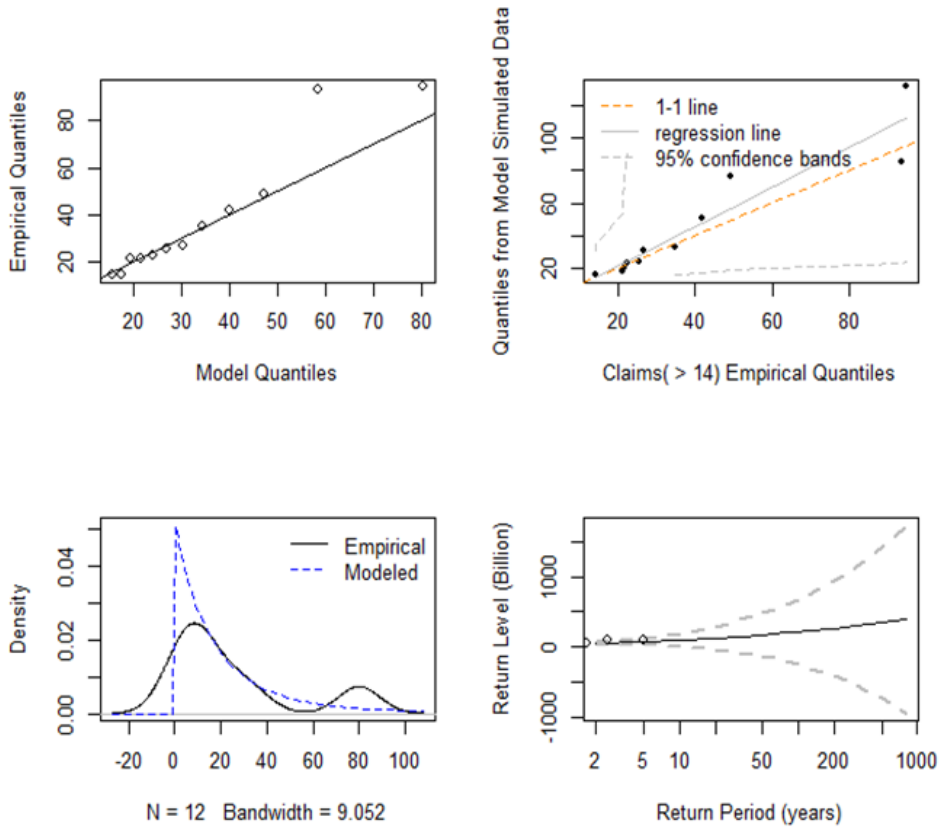

Figure 4.6: Diagnostic plots for GP

Figure 4.6 represents the PP and QQ diagnostic plots for the fitted GPD with the threshold u = 14 billion. Quantile-quantile plot(plot on upper left) with empirical claim excess as y-axis and model as x-axis, probability-probability plot (plot on upper right) with the empirical quantiles as x-axis and quantiles from model simulated data as y-axis with 95% confidence bands. Since out of the diagnostic plots the probability plot and quantile plot are approximately linear and the straight line fits most of the data points, it is safe to conclude that the chosen of GPD with threshold u = 14 billion fits the extreme health insurance claims data points and the model we chose is valid.

Figure 4.6 also represents the return level (value at risk) plot (plot on bellow right), the use of return level plots for model checking is to plot the fitted return level curve and compare it to the data. It gives the idea of the expected return level for each return period. Here "o" points represent the empirical return period of extreme health insurance claims,dashed lines represented the return level curve and the upper and lower confidence intervals represented as the dotted lines. It is seen that all the points lie around the fitted return level curve, this means that GPD is good.

4.5 Return level

By using the terminology of extreme value theory, return levels are often called quantiles or values at risk. For the GPD, one must account for the frequency of exceedances of occurrence in order to define quantiles because GPD is used usually to model the values given that we exceed a threshold u=14 billion.

Table 4.4: Return level for GP.

Return period	2	4	6	8	10
Return level	50.065	70.446	83.9	94.203	102.659
95% CI	(24.38, 75.75)	(27.93, 112.96)	(23.98, 143.82)	(17.43, 170.92)	(10.05, 195.27)

Table 4.4 shows that return levels(quantiles) increase as their corresponding return periods increase, by considering a sequence of years from 2 to 10 with interval of 0.05 years.Then this sequence shows that The return levels of highest claim equal to 94.5 billions in data set of health insurance claims is between of return period of 8 and 10 years.

4.6 Exponential Distribution

Exponential distribution is the special case of Generalized Pareto Distribution depending on shape parameter, if it is zero the GPD is equivalent to the exponential distribution.The shape parameter controls the tail behavior of the distribution and the tendency to produce heavy extremes.

Table 4.5: Estimation of parameter for Exponential.

Parameter	Estimated parameter	95% Lower CI	95% Upper CI
Scale	24.274	10.54	38.0083

Table 4.5 represents the estimation parameter for exponential distribution , and 95% confidence interval.

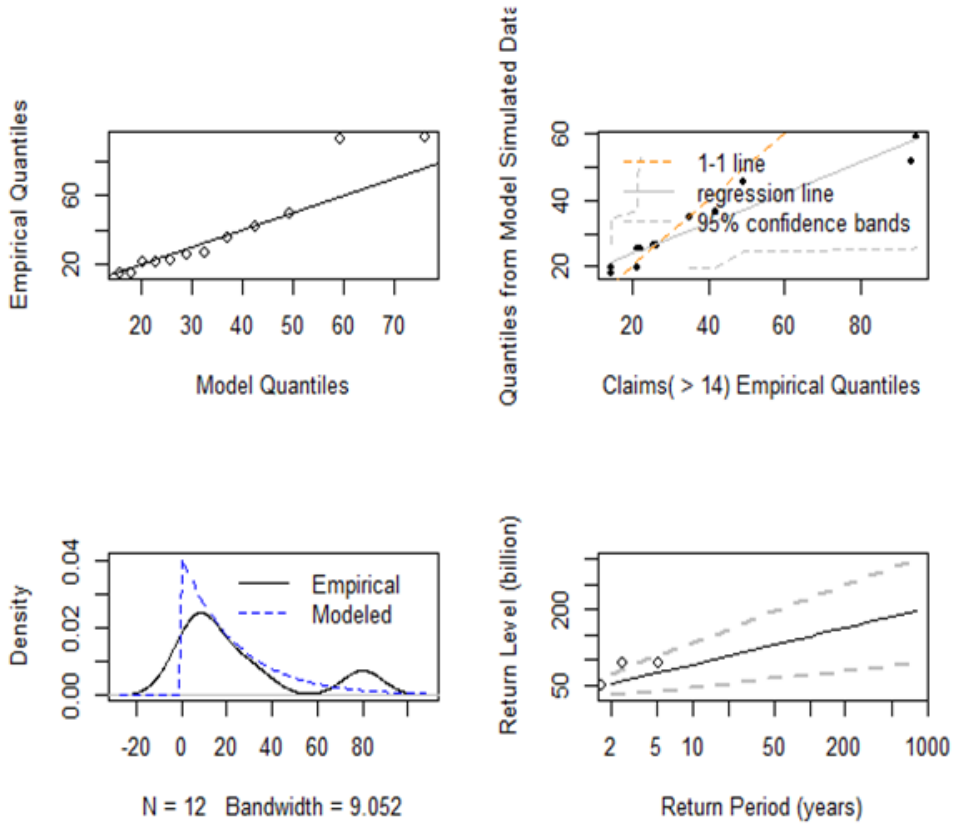

Figure 4.7: Diagnostic plots for Exponential distribution

To check exponential distribution model adequacy, is the same as for GP model to use QQ, PP, density and return level plots. From quantile, empirical and return level plot in Figure 4.7, there are few points which do not fall but closer to the diagonal reference line. It means that exponential distribution fit well the claims excess threshold.

Table 4.6: Return level for Exponential distribution.

Return period	2	4	6	8	10	12
Return level	52.077	68.902	78.745	85.728	91.145	99.312
95% CI	(30.5,73.6)	(37.84,99.9)	(42.1,115.4)	(45.1,126.3)	(47.5,134.8)	(51.1,147.6)

Table 4.6 represents the return periods and the corresponding return levels for exponential distribution , and 95% confidence interval.

4.7 Model Selection

Model selection is the task of selecting a statistical model between GP and exponential distribution models for given data.The goal of model selection is to choose a sparse model that adequately explains the data.

Table 4.7: Comparison between GP and Exponential distribution.

Exponential distribution	Pareto distribution
Estimation method used: MLE	Estimation method used: MLE
Negative Log- Likelihood value: 50.27295	Negative Log- Likelihood value: 50.13556
Estimated parameters: Scale 24.27417	Estimated parameters: Scale Shape 19.2347512 0.2210936
Standard Error Estimates: Scale 7.007348	Standard Error Estimates: Scale Shape 10.3022874 0.4556447
Estimated parameter covariance matrix: Scale Scale 49.10293	Estimated parameter covariance matrix: Scale Shape Scale 106.13713 -3.5609602 Shape -3.56096 0.2076121
AIC=102.5459	AIC=104.2711
BIC=103.0308	BIC=105.2409

Table 4.7 shows that the standard error estimate, ACI and BCI of scale parameter for GP are bigger than ones for exponential distribution. Therefore exponential distribution model fit excess claims better than Generalized Pareto model.Also we can use statistical test like Likelihood-Ratio test to compare the goodness of fit between GP and exponential distribution models,because this LRT is used for testing two nested models, where null model H_0 is a special case of the other alternative model H_1. In our case Exponential distribution model is the special case of generalized Pareto distribution model. Since we know the negative log-likelihood of both models, the test statistics is calculated as ratio between the log-likelihood of simpler model to the model with more parameters.

$$D = -2log\left(\frac{H_0}{H_1}\right)$$

$$= -2\left(log(H_0) - log(H_1)\right)$$

$$= 2\left(50.27295 - 50.13556\right)$$

$$= 0.27478$$

D has a chi-square distribution with one degree of freedom after calculating likelihood ratio test, we compare it with Chi-square critical value with 1 degree of freedom. The model is accepted if $D < \chi^2_{0.05}$. Otherwise, the alternative model is fitted. In our case $D = 0.27478$ is less than $\chi^2_{0.05} = 3.8415$ and the $p - value = 0.6001$ is greater than $\alpha = 0.05$ we accept the null model. Therefore exponential distribution model fitted the extreme claims data better than GP model.

Table 4.8: Comparison between return level for GP and Exponential distribution models.

Return period (Years)	2	4	6	8	10
EXP. Return level (billions)	52.077	68.902	78.745	85.728	91.145
GP. Return level (billions)	50.065	70.446	83.9	94.203	102.659

Table 4.8 represents return level comparison , GP results has higher return level than exponential distribution model. Therefore exponential distribution model is better to predict return period.

Chapter 5

CONCLUSION

The aim of this thesis was to determine the extreme health insurance claims and its behavior. Generalized Pareto distribution has been introduced to check the goodness-of-fit of distribution tails.

In this thesis, we reviewed the literature of extreme value theory, Peaks-over-threshold and Generalized Pareto distribution model. We defined the thesis design, data source and data analysis technique. Also we defined the detail of methods and the statistical techniques used in analyzing the data of the study. In this thesis, we done the conclusions related to each of the theory we introduced at the start of the thesis. The statistical description showed that time series of my data set is stationary. Also the minimum and maximum are 0.19 and 94.5 billion respectively. The scatter plot showed that many extreme claims occur from March to June. The selection of threshold value using mean residual life plot and threshold range plots, showed that was 14 billion. Taking into account the 95% confidence bounds, the extreme claims was 12 observations.

After choosing an appropriate threshold we used the maximum likelihood estimation method to estimate parameters. Scale and shape parameters estimate are 19.2347512 and 0.2210936 respectively. Diagnostic plots showed the goodness of fit of GPD. PP, QQ and density plots showed that the GPD fitted well the extreme claims at 95% confidence interval with $u = 14$. In this thesis we used likelihood ratio test to select the model fitted data better than others. LRT is used for testing two nested models, where null model H_0 is a special case of the other alternative model H_1. LRT compare to Chi-square critical value with 1 degree of freedom showed that $LRT = 0.27478$ is less than $\chi^2_{0.05} = 3.8415$. We accept the null model.

Finally, the extreme claims exist in RSSB. Exponential distribution model fitted extreme claims better than Generalized Pareto distribution model. The predicted return period for the highest value which is 94.5 billion to be exceeded is between 10 and 12 years.

Throughout this study many ideas raised which were not included into this research. Thus, the following are proposals for findings utility and additionally investigation.

1. Based on the project findings, the model that efficiently handle extreme claim of clients, we would like to recommend the insurance companies to rely on this research.

2. Due to the limited time, we have not investigated all possible types of insurance. In line, we would like to recommend further study on extreme claims model for other types of insurance.

References

Hamed Khaled, A. Rao.Ramachandro. (1999). *Flood Frequency Analysis.* CRC Press.

Holger Rootzén & Nader Tajvidi. (1997). Extreme value statistics and wind storm losses. *Taylor & Francis*.

Paul Embrechts, Sidney I. Resnick, and Gennady Samorodnitsky . (1999). Extreme Value Theory as a Risk Management Tool. *North American Actuarial Journal.*

Beirlant, J. Teugels, J.L. and Vynckier, P. (1994). *Extremes in Non-Life Insurance. Extremal Value Theory and Applications,*. Galambos: J., Lenchner, J. and Simiu, E. (eds.), Dordrecht, Kluwer.

Coles, S. (2001). *An introduction to statistical modeling of extreme values.* Springer.

Davison and Smith . (1990). Models for exceedance over high thresholds. *J. R. Stat. Soc., Ser. B, 52,*, 393-442.

Davison, A. C. (1984). Modelling excesses over high thresholds, with an application, in Statistical Extremes and Application. *NA TO ASI Ser. C,vol. 131, edited by J. Tiago de Oliveira*, pp. 461-482, D. Reide Norwell, Mass.

Fisher and Tippett. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 180–290.

Frechet, M. (1927). Sur la loi de probabilit de l'´ecart maximum. *Ann. Soc. Polon. Math. (Cracovie)*, 93–116.

G. Matthys · J. Beirlant. (2000). Adaptive threshold selection in tail index estimation.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une s´erie al´eatoire. *Annals of Mathematics, vol 44,*, 423–453.

Gumbel, E. (1954). *Statistical theory of extreme values and some practical applications.*

Gumbel, E. (1958). *Statistics of Extremes.*

Haan, L. d. (1970). On Regular Variation and its Application to the Weak Convergence of Sample Extremes. *Mathematical Centre Tract 32 (Mathematics Centre, Amsterdam).*

Hanson and Vogel . ((2008)). The Probability Distribution of Daily Rainfall in the United States. *ASCE-EWRI, World Water & Environmental Resources Congress 2008, Hawaii, 2008.*

health, M. o. (October 2012). *Annual Report: Community Based Health Insurance.*

Hosking, J.R.M. and Wallis, J.R. (1987). Parameter and quantile estimators for Generalized Pareto distribution. *Technometrics, 29(3)*, 339-349.

McNeil and Frey. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Empirical Finance*.

MCNEIL, A. (1997). 'Estimating the Tails of Loss Severity Distributions Using Extreme Value Theory. *Astin Bulletin*, 27(2):117–37.

McNeil, A. and Saladin, T. (1997). The Peaks over Threshold Method for Estimating High Quantiles of Loss Distributions. *Proceedings of the 28th International ASTIN Colloquium, Cairns*, 23-43.

McNeil, A.J., and Saladin, T. (1997). *'The Peaks over Threshold Method for Estimating High Quantiles of Loss Distributions.* Proceedings of the XXVIIIth International ASTIN Colloquium.

Mejzler, D. (1949). On a problem of B.V. Gnedenko. *ukrain Math.z*, 67-84.

Ministry of health. (October 2012). *Annual Report: Community Based Health Insurance.*

P. EMBRECHTS, C. KLUPPELBERG, T. MIKOSCH. (1997). *Modelling Extremal Events for Insurance and Finance.* Berlin Heidelberg NewYork, London Paris Tokyo, Hong Kong Barcelona: Springer-Verlag.

Pickands, J. (1975). Statistical inferance using extreme order statistics. *Ann. Stat., 3*, 119-131.

Ren, David E. Giles. (2007). Extreme Value Analysis of Daily Canadian Crude Oil Prices,.

Saksena and Johnson. (1984). Best Unbiased Estimators for the Parameters of a Two-Parameter Pareto Distribution. *Metrika, Volume 31*, 77-83.

Smith, D. a. (1990). Models for exceedance over high thresholds. *J. R. Stat. Soc., Ser. B, 52*, 393-442.

Smith, J. A. (1987). Estimating the upper tail of flood frequency distribution. *Water Resour. Res., 23(8)*, 1657-1666.

Smith, J. A. (1990). *Extreme Value Theory.* Chichester: Wiley,: in Handbook of Applicable Mathematics, Supplement edited by W. Leder.mann.

Smith, R. (1984). Threshold methods for sample extremes. *In: J. Tigao de Oliveria (Editor), Statistical Extremes and Applications Reidel, Dordrecht,*, 621 -638. .

Van Montfort and Otten. (1991). The first and second e of the extreme value distribution EV1. *stochastic hydrology and hydroulics 5*, 69-76.

Van Montfort, M.A.J. and Witter, J.V. (1986). The Generalized Pareto distribution applied to rainfall depths. *Hydrol. Sci. J., 31(2)*, 151-162.

Van Montfort, M.A.J. and Witter, J.V. (1985). Testing exponentially against Generalized Pareto distribution. *J. Hydrol., 78*, 305-315.

von Mises, R. (1936). Probability, Statistics and Truth. *2nd rev. English ed., New York, Dover.*

Appendix A

R Codes

```
install "extremes" package
library (in2extRemes)
in2extRemes ()
file <-read data ("Monthlyclaimsdata.csv")
thresholdrange. plot (x= Monthlyclaims, r=c (0.19, 20), type= "GP", nint=30)
fit1 <- fevd(x = Claims, data = Monthlyclaims$data, threshold = 14, type = "GP")
fit1A <- fevd(x = Claims, data = Monthlyclaims$data, threshold = 12, type = "GP")
fit1B <- fevd(x = Claims, data = Monthlyclaims$data, threshold = 13, type = "GP")
plot (fit1)
ci (fit1, type="parameter")
ci (fit1, type="return. level", method="normal", return. period=c (2, 4, 6, 8, 10),
verbose=TRUE)
fit2 <- fevd(x = Claims, data = Monthlyclaims$data, threshold = 14, type = "Exponential")
Plot (fit2)
ci (fit2, type="parameter")
ci (fit2, type="return. level", method="normal", return. period=c (2, 4, 6, 8, 10, 12),
verbose=TRUE)
lr.test (fit1, fit2)
```