ANALYSIS OF DISTRIBUTIONS FOR

INSURANCE CLAIMS DATA

Natakon Nawaratana

A Thesis Submitted in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Applied Mathematics

Suranaree University of Technology

Academic Year 2019

การวิเคราะห์การแจกแจงของข้อมูลข้อเรียกร้องประกันภัย

นายณฐกร นวรตน

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
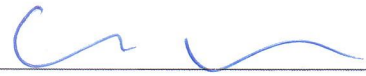
สาขาวิชาคณิตศาสตร์ประยุกต์

มหาวิทยาลัยเทคโนโลยีสุรนารี

ปีการศึกษา 2562

ANALYSIS OF DISTRIBUTIONS FOR INSURANCE CLAIMS DATA

Suranaree University of Technology has approved this thesis submitted in partial fulfillment of the requirements for a Master's Degree.
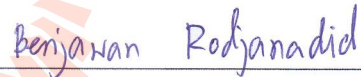
Thesis Examining Committee

_____

(Assoc. Prof. Dr. Eckart Schulz)

Chairperson

_____

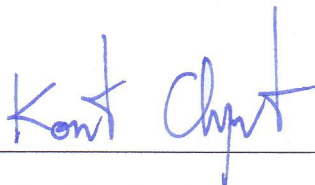(Asst. Prof. Dr. Jessada Tanthanuch)

Member (Thesis Advisor)

_____

(Asst. Prof. Dr. Benjawan Rodjanadid)

Member

_____
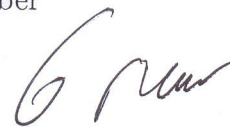
(Asst. Prof. Dr. Tidarut Areerak)

Member

_____   _____

(Assoc. Prof. Flt. Lt. Dr. Kontorn Chamniprasart)   (Assoc. Prof. Dr. Worawat Meevasana)

Vice Rector for Academic Affairs   Dean of Institute of Science

and Internationalization

ณฐกร นวรตน : การวิเคราะห์การแจกแจงของข้อมูลข้อเรียกร้องประกันภัย (ANALYSIS OF DISTRIBUTIONS FOR INSURANCE CLAIMS DATA) อาจารย์ที่ปรึกษา : ผู้ช่วยศาสตราจารย์ ดร.เจษฎา ตัณฑนุช, 62 หน้า.

ประกันภัย/การแจกแจงความน่าจะเป็น/ตัวแบบเชิงเส้นวางนัยทั่วไป

วิทยานิพนธ์นี้มีจุดมุ่งหมายต้องการประยุกต์ใช้องค์ความรู้พื้นฐานด้านสถิติและความน่าจะ เป็นในการสร้างตัวแบบทางคณิตศาสตร์สำหรับข้อมูลทางด้านประกันภัย ข้อมูลสำหรับการทำงาน วิจัยนี้นำจากฐานข้อมูลสาธารณะที่ให้สำหรับงานวิจัยด้านข้อมูล ได้แก่ "Sample Insurance Claim Prediction Dataset" จาก "[Medical Cost Personal Datasets][1]" ข้อมูลโดย Eason สร้างข้อมูลเมื่อ 2018-05-14 ปรับปรุงข้อมูลล่าสุด 2018-06-04 รุ่น 2 นอกจากนี้ได้ใช้โปรแกรมสำเร็จรูป RStudio ในการช่วยวิเคราะห์ข้อมูลทางสถิติและแผนภาพฮิสโทแกรม

ผลการดำเนินการวิจัยพบว่า ในการทดสอบทางด้านสถิติเบื้องต้น การแจกแจงแกมมาและ การแจกแจงล็อกปรกติ ให้ผลลัพธ์ที่ดีในการสร้างตัวแบบ อย่างไรก็ตาม นอกจากการแจกแจงทั้ง สองแล้ว ได้ใช้การแจกแจงทวินามนิเสธในการสร้างตัวแบบเชิงเส้นวางนัยทั่วไปอีกด้วย

จากการสร้างตัวแบบเชิงเส้นวางนัยทั่วไปโดยการแจกแจงทั้งสาม พบว่าตัวแบบที่ใช้การ แจกแจงแกมมา ตัวแปร charges ไม่ขึ้นกับตัวแปร sex region และ insuranceclaim ขณะที่ตัวแบบที่ ใช้การแจกแจงล็อกปรกติ ตัวแปร charges ไม่ขึ้นกับตัวแปร steps เท่านั้น และ ตัวแบบที่ใช้การแจก แจงทวินามนิเสธ ตัวแปร charges ไม่ขึ้นกับตัวแปร sex และ step

จากการทดสอบวัดประสิทธิภาพพบว่าการสร้างตัวแบบเชิงเส้นวางนัยทั่วไปโดยใช้การแจก แจงล็อกปรกติ ให้ผลลัพธ์ที่ดีที่สุดสำหรับข้อมูลข้างต้น

สาขาวิชาคณิตศาสตร์	ลายมือชื่อนักศึกษา_____ณฐกร_____
ปีการศึกษา 2562	ลายมือชื่ออาจารย์ที่ปรึกษา___J.Tenthanuch___

NATAKON NAWARATANA : ANALYSIS OF DISTRIBUTIONS FOR
INSURANCE CLAIMS DATA. THESIS ADVISOR : ASST. PROF.
JESSADA TANTHANUCH, Ph.D. 62 PP.

INSURANCE/PROBABILITY DISTRIBUTION/GENERALIZED LINEAR
MODEL

This thesis aims to apply fundamental statistics and probability to make a mathematical model for the sample insurance data. Source of insurance data for testing our assumption was obtained from the free standard dataset, "Sample Insurance Claim Prediction Dataset" which based on "[Medical Cost Personal Datasets][1]", Dataset owner *Eason*, date created 2018-05-14, last updated 2018-06-04, version 2. RStudio program was used for statistical analysis and to show the data histogram.

The results showed that gamma and log-normal distributions were better fit in modelling than logistic and normal distributions. However, the negative binomial distribution was also another distribution using in this thesis. In the process of GLM modelling, all three proposed distributions were applied. For the model using gamma distribution, variable *charges* did not depend on *sex*, *region* and *insuranceclaim*. For the model using Log-normal distribution, variable *charges* did not depend on *steps* only. In the case of using negative binomial distribution, variable *charges* did not depend on *sex* and *step*. According to the assessment, log-normal distribution was the best in GLM model fitting.

School of Mathematics	Student's Signature _____
Academic Year 2019	Advisor's Signature _____

ACKNOWLEDGEMENTS

CONTENTS

CONTENTS (Continued)

CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Most Thai families have a typical plan for their children as follows: 1) study in a good school and a good university; 2) work in a good place; 3) buy a good car; and 4) have a good marriage. However, different circumstances make people have different necessities. Therefore, life planning provides an appropriate approach for an individual. Nowadays, throughout the world life style changes to that of an ageing society. The statistical records show that the birth rate continue to decrease significantly. The Thailand National Statistical Office informs that Thailand has begun to be an ageing society in the year of 2005 and will become a complete ageing society in 2021. Thailand's urban life style has changed to that of a small individual family with few or no children at all. Hence the search for a systematical model for life style forecasting, which is fit for present day, is still the major work of scientists.

Insurance is one of many good ways to organize one's life. It provides life and non-life risk management. A life insurance is a contract between an insurer and a policyholder in which the insurer guarantees payment of a death benefit to named beneficiaries upon the death of the insured. The insurance company promises a death benefit in consideration of the payment of premium by the insured. The risks that are covered by life insurance include premature death, income during retirement, and illness. Life insurance products mostly consist of whole life, endowment, term, medical and health, and life annuity plan. On the other hand, non-life insurance covers things apart from what is covered in life insurance.

That is, a non-life insurance policy aims to protect an individual against losses and damages other than those covered by life insurance. The risks that are covered by non-life insurance are property loss (for example stolen car, burnt house) liabilities arising from damage caused by an individual to a third party, accidental death or injury. The main products of non-life insurance include motor insurance, fire/house owners/householders insurance, personal accident insurance, medical and health insurance and travel insurance. This research is focusing on the mathematical and statistical model for life insurance.

In order to make an insurance contract, the insurance company may ask the client for much information. Because of the variety of factors, the company then has a process for offering a suitable insurance product to the client. Hence, each insured has an insurance contract corresponding to his/her personal circumstances. However, there is no mathematical analysis of common and differentiating factors of the health insurance in Thailand. The relation of insured data and value of insurance contract should be analyzed. The knowledge obtained will help an insurance company to design appropriate insurance products tailored to the insured. It helps in transforming the insurance data to the information that is useful for the data analysis in the future. The result can reinforce background in many subjects, e.g. mathematics, statistics, economics, actuarial science, data science, etc. It may also support an insurance company in making insurance products for variety customers. The understanding by mathematical and statistical models for the life insurance provides a lot of benefits to both insurance companies and common people. This would help people in finding a suitable life insurance product for life planning, strategically.

This thesis aims to apply fundamental statistics and probability to make a mathematical model for the sample insurance data. Many of probability distribu-

tions related to the health insured were reviewed and studied. Source of insurance data for testing our assumption was obtained from a free standard dataset. The RStudio program was used to analyse statistically and show the histogram of data. Also curve fitting was done by some library function in RStudio. Types of probability distributions which deserve for fitting the model were scoped. Generalized linear models for selected probability distributions were done. All models were tested for accuracy. The process in this research may be applied and extended to further related works.

1.1 Research objectives

The objectives of this thesis were to find and analyze distributions which are appropriate for obtained insurance claims data.

1.2 Scope and limitations

1. Insurance claims data from "Sample Insurance Claim Prediction Dataset" which based on "[Medical Cost Personal Datasets][1]", Dataset owner *Eason*, date created 2018-05-14, last updated 2018-06-04, version 2, available on **https://www.kaggle.com/easonlai/sample-insurance-claim-prediction-dataset**.

2. Statistics calculation based on RStudio software version 1.2.1335 © 2009-2019 RStudio, Build 1379 (f1ac3425) Inc., working on Microsoft Windows 10.

1.3 Research procedure

The research work proceeded as follows:

1. study the theory of mathematics and statistics related with the insurance work

2. study the application of general linear model (GLM)

3. study the RStudio software

4. analyze the obtained life insurance claim data preliminarily

5. develop GLM models for the insurance claim data

6. measure accuracy for the predicting models

1.4   Expected results

It is expected to have a concept to find an appropriate distributions and general linear model for insurance claims data.
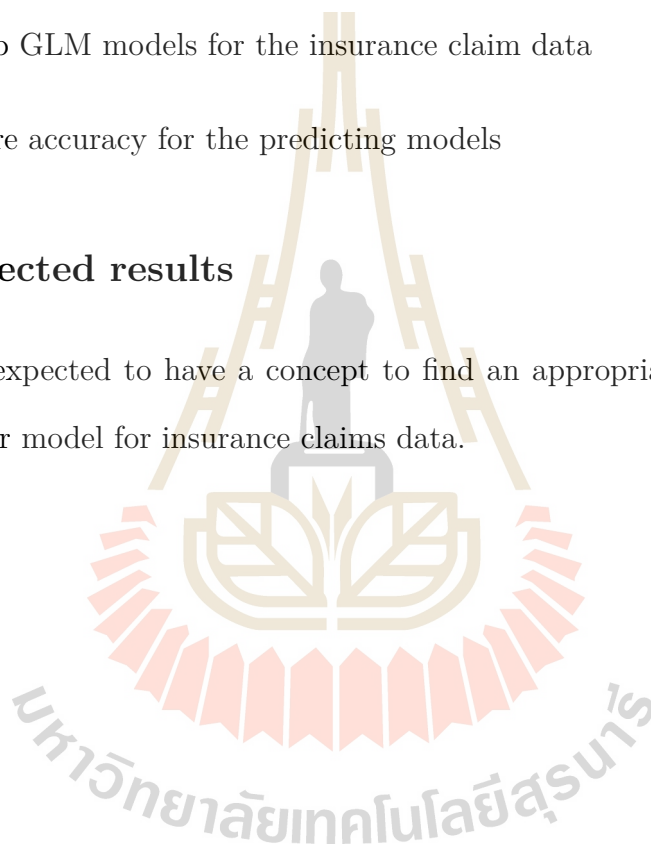
CHAPTER II

LITERATURE REVIEW

In this chapter, the knowledge of basic mathematics and statistics related with actuarial science is reviewed. The following sections consist of the topics in statistical distributions and the theory to model actuarial claim data. Most contents of probability and random variables comes from Grimmett and Stirzaker (2001), Grimmett and Welsh (1986) and Adewale (2017).

2.1 Probability

The mathematical theory of probability starts with the idea of an experiment (or trial), being a course of action whose consequence is not predetermined; this experiment is reformulated as a mathematical object called a *probability space.*

Definition 2.1. If A is some event, the occurrence or non-occurrence of A depends upon the *chain* of circumstances involved. This chain is called an *experiment* or *trial*; the result of an experiment is called its *outcome.* The set of all possible outcomes of an experiment is called the *sample space* and is denoted by Ω.

Definition 2.2. A *Bernoulli trial* or *binomial trial* is a random experiment/trial with exactly two possible outcomes, "success" and "failure", i.e.

$$\Omega = \{\text{success}, \text{failure}\}.$$

Definition 2.3. A non-empty collection $\mathcal{F}$ of subsets of the sample space Ω is called an *event space* of $\mathcal{F}$.

Definition 2.4. A collection $\mathcal{F}$ of subsets of Ω is called a σ-*field* if it satisfies the following conditions:

1. $\emptyset \in \mathcal{F}$;

2. if $A_1, A_2, \ldots \in \mathcal{F}$ then $\overset{\infty}{\underset{i=1}{\cup}} A_i \in \mathcal{F}$;

3. if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$, where A^c is the complement of A.

Definition 2.5. A *probability measure* $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying

1. $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$;

2. if $A_1, A_2, \ldots$ is a collection of disjoint members of $\mathcal{F}$, in that $A_i \cap A_j = \emptyset$ for all pairs i, j satisfying $i \neq j$, then

$$\mathbb{P} \left(\overset{\infty}{\underset{i=1}{\cup}} A_i \right) = \overset{\infty}{\underset{i=1}{\Sigma}} \mathbb{P} \left(A_i \right).$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$, comprising a set Ω, a σ-field $\mathcal{F}$ of subsets of Ω, and a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$, is called a *probability space*.

The above definitions give more efficient tools to measure the likelihoods of the occurrences of events. Compared to the classical concept, the experience of most scientific experimentation is that the proportion of times that A occurs settles down to some value as N becomes larger; that is, writing $N(A)$ for the number of occurrences of A in the N trials, the ratio $N(A)/N$ appears to converge to a constant limit as N increases. The ultimate value of this ratio as being the probability $\mathbb{P}(A)$ that A occurs on any particular trial. In practice, N may be taken to be large but finite, and the ratio $N(A)/N$ may be taken as an approximation to $\mathbb{P}(A)$. Some individuals refer informally to $\mathbb{P}$ as a *probability distribution*.

Definition 2.6. *Conditional Probability* is a measure of the probability of one event occurring with some relationship to one or more other events. The conditional probability of A given B, or the probability of A under the condition B, is usually written as $\mathbb{P}(A|B)$, or sometimes $\mathbb{P}_B(A)$ or $\mathbb{P}(A/B)$, defined by

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)},$$

where $\mathbb{P}(A \cap B)$ is the probability that both events A and B occur.

2.2 Random variables and their distribution

Quantities governed by randomness correspond to functions on the probability space called *random variables*. The value taken by a random variable is subject to chance, and the associated likelihoods are described by a function called the *distribution function*.

Definition 2.7. *A random variable* is a function $X : \Omega \to \mathbb{R}$ with the property that

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$$

for each $x \in \mathbb{R}$. Such a function is said to be $\mathcal{F}$-*measurable*.

Distribution is a statistical concept used in data research. It is a listing or function showing all the possible values of the statistical data and how often they occur. Every random variable has a *distribution function*. Distribution functions are very importance and useful.

Definition 2.8. The *distribution function* of a random variable X is the function

$$F : \mathbb{R} \to [0, 1]$$

given by $F(x) = \mathbb{P}(X \leq x)$.

The distribution function satisfies the following conditions.

Theorem 2.1. *A distribution function F has the following properties:*

1. $\lim\limits_{x \to -\infty} F(x) = 0$,

2. $\lim\limits_{x \to \infty} F(x) = 1$,

3. F is nondecreasing, i.e. if $x < y$ then $F(x) \leq F(y)$,

4. F is right-continuous, i.e. $F(x + h) \to F(x)$ as $h \to 0^+$.

Moreover, some conditions for the relation of the distribution function of a random variable X and the probability measure $\mathbb{P}$ are as follows:

Theorem 2.2. *Let F be the distribution function of X. Then*

1. $\mathbb{P}(X > x) = 1 - F(x)$,

2. $\mathbb{P}(x < X \leq y) = F(y) - F(x)$,

3. $\mathbb{P}(X = x) = F(x) - \lim\limits_{y \to x^-} F(y)$.

2.3 Discrete and continuous random variables

Random variables can be classified into two basic categories, *discrete* and *continuous*.

Definition 2.9. The random variable X is called *discrete* if it takes values in some countable subset $\{x_1, x_2, \ldots\}$, only, of $\mathbb{R}$. The discrete random variable X has *(probability) mass function* $f : \mathbb{R} \to [0,1]$ given by $f(x) = \mathbb{P}(X = x)$.

Definition 2.10. The random variable X is called *continuous* if its distribution function can be expressed as

$$F(x) = \int_{-\infty}^{x} f(u)du \qquad x \in \mathbb{R},$$

for some integrable function $f : \mathbb{R} \to [0, \infty)$ called the *(probability) density function* of X.

2.4 Expectation

Let $x_1, x_2, \ldots, x_N$ be the numerical outcomes of N repetitions of some experiment. The average of this outcomes is

$$m = \frac{1}{N} \sum_i x_i.$$

Consider the N discrete random variables with a common mass function f. For each possible value x, about $Nf(x)$ of the outcome X_i, $i = 1, \ldots, N$, will take that value x. So the average value is

$$m = \frac{1}{N} \sum_x xNf(x) = \sum_x xf(x),$$

where the summation is over all possible values of the X_i. This average is called the *expectation* or *mean value* of the underlying distribution with mass function f.

Definition 2.11. The *mean value*, or *expectation*, or *expected value* of a discrete random variable X with mass function f is defined to be

$$\mathbb{E}(X) = \sum_{x:f(x)>0} xf(x), \tag{2.1}$$

whenever this summation is absolutely convergent.

The expectation (2.1) of a discrete variable X or an average of the possible values of X may be written in form

$$\mathbb{E}(X) = \sum_x x\mathbb{P}(X = x),$$

which means each value being weighted by its probability.

Lemma 2.3. *If X has a mass function f and $g \colon \mathbb{R} \to \mathbb{R}$, then*

$$\mathbb{E}(g(X)) = \sum_x g(x)f(x),$$

whenever this sum is absolutely convergent.

Definition 2.12. If k is a positive integer, the kth *moment* m_k of X is defined to be

$$m_k = \mathbb{E}(X^k) = \sum_x x^k \mathbb{P}(X = x).$$

The kth *central moment* σ_k is

$$\sigma_k = \mathbb{E}((X - \mathbb{E}(X))^k).$$

Definition 2.13. The two moments of most use are

- $m_1 = \mathbb{E}(X)$, called the *mean* (or *expectation*) of X, and

- $\sigma_2 = \mathbb{E}((X - \mathbb{E}(X))^2)$, called *variance* of X.

These two quantities are measures of the mean and dispersion of X; that is, m_1 is the average value of X, and σ_2 measures the amount by which X tends to deviate from the average. The mean m_1 is often denoted μ, and the variance of X is often denoted $\mathrm{var}(X)$. The positive square root $\sigma = \sqrt{\mathrm{var}(X)}$ is called the standard deviation.

For continuous variables, expectations are defined as integrals.

Definition 2.14. The *expectation* of a continuous random variable X with density function f is given by

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx,$$

whenever the integral converges absolutely.

By the definition (2.12) of moment for a discrete random variable, we define the kth moment of a continuous variable X as the following.

$$m_k = \mathbb{E}(X^k) = \int_{-\infty}^{\infty} x^k f(x) dx, \tag{2.2}$$

whenever the integral converges absolutely.

Therefore, we can define *mean* and *variance* for a continuous random variable similar to definition 2.13 as $\mu = \mathbb{E}(X)$ and $\sigma_2 = \mathbb{E}((X - \mathbb{E}(X))^2)$, respectively, where the kth moment is defined by (2.2).

2.5 Distributions related insurance claims data

Well known distributions and distributions used in insurance data analysis and generalized linear modelling are mentioned in this section.

2.5.1 Bernoulli distribution

The *Bernoulli distribution* admits only two possible outcomes, for examples $\Omega = \{\text{Yes}, \text{No}\}$ and $\Omega = \{\text{True}, \text{False}\}$. It is usually considered as $\Omega = \{0, 1\}$, where the event "1" is often called a *success*, the other "0", a *failure*. Further $f(1) = p$ and $f(0) = 1 - p$, where p is the probability of the event occurring and $0 \leq p \leq 1$. The mean and variance of a Bernoulli random variable are p and $p(1-p)$, respectively. The variance is largest when $p = 0.5$. The probability function is

$$f(k) = p^k (1-p)^{1-k}, \quad k = 0, 1. \tag{2.3}$$

The examples of using this distribution in insurance works are

- a claim or no claim on a policy in a given year;

- a person dying or surviving over a given year;

- a claim or no claim on the vehicle insurance.

2.5.2   Binomial distribution

Let $0 \leq k \leq n$, and consider $f(k)$. Exactly $\binom{n}{k}$ points in Ω give a total of k wanted evens; each of these points occurs with probability $p^k(1-p)^{n-k}$, and so

$$f(k) = \binom{n}{k}p^k(1-p)^{n-k} \quad \text{if} \quad 0 \leq k \leq n. \tag{2.4}$$

The random variable X is said to have the *binomial distribution* with parameters n and p. This distribution is the discrete probability distribution of the number of successes in a sequence of n independent experiments. If there are n independent Bernoulli random variables, each with success probability p, then the total number of successes has the binomial distribution.

In insurance claim, policies may be grouped according to geographical area and socioeconomic indicators. For the number of policies n and varying probabilities p, number of claims arising from each area may be able to be described by binomial of the form (2.4).

A binomial random variable is often transformed into a proportion by dividing by n. The resulting random variable k/n is called the binomial proportion and the probability function (2.4) shifted on to $0, 1/n, 2/n, \ldots, 1$.

The binomial distribution is the basis for the popular binomial test of statistical significance, where the binomial test is an exact test of the statistical significance of deviations from a theoretically expected distribution of observations into two categories. The binomial distribution is practically and historically important and leads directly to Poisson distribution as discussed after this.

2.5.3 Poisson distribution

If a random variable X takes values in the set $\{0, 1, 2, \ldots\}$ with mass function

$$f(x) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots,$$

where $\lambda > 0$, then X is said to have the *Poisson distribution* with parameter λ.

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or specified intervals such as distance, area, volume or space, if these events occur with a known constant rate and independently of the time since the last event.

Suppose, in the binomial distribution n becomes large while p becomes small, by (2.4), let $n \to \infty$ and $p \to 0$ in such a way that np approaches a non-zero constant λ. Then,

$$\binom{n}{k} p^k (1-p)^{n-k} \sim \frac{1}{k!} \left(\frac{np}{1-p} \right)^k (1-p)^n \to \frac{\lambda^k}{k!} e^{-\lambda}, \quad \text{for} \quad k = 0, 1, 2, \ldots$$

For the numbers of non-life insurance claims, a Poisson distribution is usually more appropriate to represent some uncertainty. This is because most non-life policies provide cover for a fixed period of time with no limit on the number of claims. However, there are also exceptions to this, because the Poisson distribution has variance equal to its mean.

2.5.4 Geometric distribution

A *geometric variable* is a random variable with the geometric mass function

$$f(k) = p(1-p)^{k-1}, \quad k = 0, 1, 2, \ldots,$$

for some number p in $(0, 1)$.

Suppose that independent Bernoulli trials (parameter p) are performed in times $1, 2, \ldots$ Let X be the time which elapses before the first success; X is called a *waiting time*. Then $\mathbb{P}(X > k) = (1-p)^k$ and thus

$$\mathbb{P}(X = k) = \mathbb{P}(X > k - 1) - \mathbb{P}(X > k) = p(1-p)^{k-1}.$$

The geometric distribution gives the probability that the first occurrence of success requires k independent trials, each with success probability p.

2.5.5 Negative binomial distribution

Let X_r be the waiting time for the rth success of Bernoulli trials of a random variable X_r, which takes values 1 and 0 with probabilities p and $q(=1-p)$, respectively. It is easy to check that X_r has mass function

$$\mathbb{P}(X_r = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad k = r, r + 1, \ldots;$$

it is said to have the *negative binomial distribution* with parameters r and p. The random variable X_r is the sum of r independent geometric variables. Note that if $r = 1$, it becomes the geometric distribution.

The negative binomial distribution is a discrete probability distribution of the number of successes in a sequence of independent and identically distributed *Bernoulli trials* before a specified (non-random) number of failures r occurs, in which the probability of success is the same every time the experiment is conducted.

Panjer (1980) and Heckman & Meyer (1982) (cited in Wright, 2007) have developed the work of forecasting the number of non-life insurance claims with the negative binomial distribution. They believe that the model require two parameters, which may possibly be extracted from mean and variance.

2.5.6 Continuous uniform distribution

A random variable X is uniform on $[a, b]$ if it has distribution function

$$
F(x) = \begin{cases} 0 & \text{if } x \leq a, \\ \dfrac{x-a}{b-a} & \text{if } a < x \leq b, \\ 1 & \text{if } x > b. \end{cases}
$$

Roughly speaking, X takes any value between a and b with equal probability. The *continuous uniform distribution* or *rectangular distribution* is a family of symmetric probability distributions such that all members of the family are equally probable.

2.5.7 Exponential distribution

A random variable X is *exponential* with parameter $\lambda > 0$ if it has distribution function

$$
F(x) = 1 - e^{-\lambda x}, \qquad x \geq 0.
$$

The exponential distribution is the probability distribution that describes the time between events in a *Poisson point process*. The Poisson point process is a type of random mathematical object that consists of points randomly located on a mathematical space. The Poisson point process is often defnned on the real line, where it can be considered as a stochastic process. This distribution proves to be the cornerstone of the theory of *Markov processes* in continuous time.

2.5.8 Normal distribution

The most important continuous distribution is the *normal* (or *Gaussian*) distribution, which has two parameters μ and σ^2 and density function

$$
f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty.
$$

The normal distribution arises in many ways. In particular it can be obtained as a continuous limit of the binomial distribution as $n \to \infty$.

Insurance companies use normal distributions to model certain average cases.

2.5.9 Log-normal distribution

A log-normal distribution is a statistical distribution of logarithmic values from a related normal distribution, i.e. for a variable x, $y = \ln(x)$ is normal distributed,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0.$$

Some research proposes that there is the complexity to obtain the distribution function of the total amount of incurred claims thus a log-normal distribution model is used to for insurance claims data (Zuanetti, Diniz and Leite, 2006).

2.5.10 Gamma distribution

The random variable X has the *gamma* distribution with parameters $\lambda, t > 0$, if it has density

$$f(x) = \frac{1}{\Gamma(t)} \lambda^t x^{t-1} e^{-\lambda x}, \quad x \geq 0.$$

Here $\Gamma(t)$ is the *gamma function*

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx.$$

The gamma distribution is a two-parameter family of continuous probability distributions. The exponential distribution can be considered as a special case of the gamma distribution.

2.5.11   Pólya Distribution

The Pólya distribution is similar to the negative binomial distribution but it works with a continuous time. A Pólya Model is a type of statistical model used to model a variety of contamination processes, including the spread of contagious diseases. This model was proposed by the Hungarian mathematician George Pólya (1887-1985). In the Pólya model, an urn initially contains w white balls and b black balls. A trial consists of drawing one ball at random, noting its color, and then replacing it together with c additional balls of the same color. Obtaining a white ball on the first trial therefore increases the probability of selecting a white ball on the next trial. The probability function for the number W_m of white balls obtained in m trials, is derived by conventional combinatorial methods:

$$
\begin{aligned}
Pr\left\{W_m = n\right\} &= P_n\left(w, b, c; m\right) \\
&= \binom{m}{n} \frac{\Pi_{i=0}^{n-1}\left(w + ic\right) \Pi_{i=0}^{m-n-1}\left(b + ic\right)}{\Pi_{i=0}^{m-1}\left(w + b + ic\right)}.
\end{aligned}
$$

A distribution with probabilities $P_n\left(w, b, c; m\right)$ is known as a Pólya distribution. The ratio $\gamma = c/w$ is customarily called the degree of contagion. When there is no contagion ($c = \gamma = 0$), the Pólya distribution is identical to the simpler binomial distribution for which the probability of drawing a white ball remains constant throughout successive trials (Bahnemann, 2015).

In the case that there is an outbreak by some contagion infection, the probabilities of claims are inconsistent with respect to time. Thus some types of probability, e.g. Poisson distribution, do not explain precisely because they were used to explain phenomena which have a uniform probability of occurrence claims on the considered time. Moreover, each claim must be independent. The negative binomial distribution satisfies that requirement but it is a discrete probability distribution. However, it was found that the Pólya distribution is more appropriate

to model the claim according to our requirements.

2.6 Exponential family

The exponential family is a class of probability distributions. All probability function of the general form

$$f(y) = c(y, \phi) \exp \left\{ \frac{y\theta - a(\theta)}{\phi} \right\}, \tag{2.5}$$

where θ and ϕ are parameters. The parameter θ is called the *canonical parameter* and ϕ is called the *dispersion parameter*. Probability functions which can be written as equation (2.5) are said to be members of the *exponential family*. In terms of $a(\theta)$,

$$\mathbb{E}(y) = \dot{a}(\theta), \quad \text{Var}(y) = \phi \ddot{a}(\theta), \tag{2.6}$$

where $\dot{a}(\theta)$ and $\ddot{a}(\theta)$ are the first and second derivatives of $a(\theta)$ with respect to θ, respectively. Equation (2.5) and equations (2.6) provides two important properties:

1. The distribution can be written as a function of mean and variance.

2. The variance is a function of mean.

Examples of exponential family distributions and their parameters are shown in Table 2.1.

2.7 Generalized Linear Models (GLM)

Regression modeling deals with explaining how one variable is generally thought of as being caused or explained by another or more other variables. The *simple linear model* (*classical linear model* or *normal linear model*) forms the basis of generalized linear modeling.

Table 2.1: Examples of exponential family distributions and their parameters.

Distribution	θ	$a(\theta)$	ϕ	$\mathbb{E}(y)$	$\mathrm{Var}(y)$
Binomial $\mathbf{B}(n,p)$	$\ln\left(\frac{p}{1-p}\right)$	$n\ln\left(1+e^{\theta}\right)$	1	np	$np(1-p)$
Poisson $\mathbf{P}(\lambda)$	$\ln\lambda$	e^{θ}	1	λ	λ
Negative binomial $\mathbf{NB}(\mu,r)$	$\ln\left(\frac{r\mu}{1+r\mu}\right)$	$-\frac{1}{r}\ln\left(1-re^{\theta}\right)$	1	μ	$\mu\left(1+r\mu\right)$
Normal $\mathbf{N}(\mu,\sigma^2)$	μ	$\frac{\theta^2}{2}$	σ^2	μ	σ^2
Gamma $\mathbf{G}(\lambda,\nu)$	$-\frac{1}{\lambda}$	$-\ln\left(-\theta\right)$	$\frac{1}{\nu}$	λ	$\nu\lambda^2$

2.7.1 History and terminology of linear modeling

1. **Simple linear modeling.** The model obtained is able to explain an observed variable y by a another observed variable x. The variable y is called the *response variable*, which may be called in alternative names *dependent variable* or *outcome*. Whereas the variable x is called the *explanatory variable*, which alternative names are *factor, covariate, independent, predictor, driver, risk factor, regressor* or simply the "x" variable.

2. **Multiple linear modeling.** This model extends the previous model by supposing more than one explanatory variable to explain the response variable y.

3. **Transforming the response.** In this case, the model aims to use the observed variables x to explain the transformation of the response variable y, $g(y)$, where g is a monotonic transformation. The most used transformations are *logarithm* and *logit function*[*].

---

[*]$\mathrm{logit}(x) = \ln\left(\frac{x}{1-x}\right) = \ln(x) - \ln(1-x), x \in (0,1)$.

4. **Classical linear modeling.** For this model, the statistical average of y is modeled in terms of x, i.e. the response variable y is replaced by it expected value $\mathbb{E}(y)$.

5. **Generalized linear modeling.** Here $g\left(\mathbb{E}(y)\right)$ is explained in terms of variables x, where g is a monotonic function. Function g is called the *link*.

Table 2.2: The most used link functions.

link functions	$g(\mu)$	$g^{-1}(\mu)$
identity	μ	μ
logarithm	$\ln \mu$	e^{μ}
logit function	$\ln\left(\dfrac{\mu}{1-\mu}\right)$	$\dfrac{e^{\mu}}{1+e^{\mu}}$
reciprocal	$\dfrac{1}{x}$	$\dfrac{1}{x}$

Here, the word "*linear*" in linear modeling means that the variables in x are linearly combined to arrive at the explanation of y, $g(y)$, $\mathbb{E}(y)$ or $g\left(\mathbb{E}(y)\right)$. "*Linearly combined*" means as the followings:

i) If $x_1, x_2, \ldots, x_m$ are the explanatory variables, the linear combination of the explanatory variables is

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m,$$

where β_i is a parameter, $i = 1, \ldots, m$.

ii) The linearity refers to linearity in the coefficients, β_i not the x variables, for examples $\beta_0 + \beta_1 x_1 + \beta_2 (x_2)^2$ and $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$.

2.7.2 The generalized linear model

The **generalized linear model (GLM)** is a model which can be specified to include a wide range of different models. Note that ANOVA and multiple linear regression models are just special cases of GLM. Generalized linear modeling is used to assess and qualify the relationship between a response variable and explanatory variables. GLM is able to describe pattern of the interaction variables. It can be also used for prediction. The modeling differs from an ordinary modelling in two main respects:

- The distribution of the response variable is chosen from the exponential family.

- A transformation of the mean of the response is linearly related to the explanatory variables.

Given a response variable y, the GLM is

$$f(y) = c(y, \phi) \exp \left\{ \frac{y\theta - a(\theta)}{\phi} \right\}, \qquad g(\mu) = x^t \beta, \qquad (2.7)$$

where

- x^t is the transpose of the explanatory variable x, $[1, x_1, \cdots, x_m]$,

- β is a vector of parameters, whose transpose is $[\beta_0, \beta_1, \cdots, \beta_m]$,

- μ is a mean,

- ϕ is the dispersion parameter,

- θ is the canonical parameter,

- a and c are some functions depended on the distribution of variables,

- g is a monotonic differentiable function which is linearly related to explanatory variables contained in x.

Equations (2.7) show that

- the distribution of the response y is in the exponential family;

- observations on y are assumed to be independent;

- given x, μ is determined through $g(\mu) = x^t\theta$;

- the choice of $g(\mu)$ determines how the mean is related to the explanatory variables x, for examples

  - for multiple linear regression models (MLRMs), g is identity,

  $$g(\mu) = \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m,$$

  - for count data, the logarithm is used,

  $$g(\mu) = \ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m,$$

  which is called *loglinear model*,

  - for binary data, the logit function is often used,

  $$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m;$$

- given μ, θ is determined through $\dot{a}(\theta) = \mu$;

- $\mathbb{E}(y) = \dot{a}(\theta)$ and $\text{Var}(y) = \phi\ddot{a}(\theta)$;

- given θ, y is determined as a draw from the exponential density specified in $a(\theta)$;

- the choice of the function $a(\theta)$ determines the response distribution; and

- the choice of the function $c(y, \phi)$ determines the actual probability function.

2.8 Maximum Likelihood Estimation

The maximum likelihood method is a good tool for estimating parameters (Vapnik, 1998). This method was first introduced by Fisher in 1922. He described different problems or estimating functions from given data as the problems of parameter estimation of specific (parametric) models and suggested the maximum likelihood method for estimating the unknown parameters in all these models. The concept of maximum likelihood is the making of the world of ideas and nomenclature including "parameter," "statistic," "likelihood," "sufficiency," "consistency," "efficiency," "information" and "estimation" (Aldrich, 1992). The maximum likelihood method was suggested to be a good tool for estimating parameters of models even for small sample sizes. The following contents are based on the Lecture Notes in Stat 378 by Dr. Karen Buro, Department of Mathematics and Statistics, MacEvan University (Buro, n.d.)

2.8.1 Simple Linear Regression

Definition 2.15. A random variable y fits a *Simple Linear Regression Model*, if and only if there exist $\beta_0, \beta_1 \in \mathbb{R}$ so that for all $x \in \mathbb{R}$

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where ϵ is normal distributed, with zero mean and σ^2 variance.

Theorem 2.4. *If y fits a Simple Linear Regression Model, then for a fixed value of $x \in \mathbb{R}$ the conditional expectation of y given x equals*

$$\mathbb{E}(y|x) = \mathbb{E}(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x$$

and the conditional variance of Y given x equals

$$\mathrm{var}(y|x) = \mathrm{var}(\beta_0 + \beta_1 x + \epsilon) = \mathrm{var}(\epsilon) = \sigma^2.$$

β_0 and β_1 are called the regression coefficients, and are the parameters of the model.

If one has n sample data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, they can be used to estimate the value of β_0 and β_1 of the simple linear regression model.

The least-squares estimators have the property that the total squared vertical distances of the measurements to the least squares line are minimal, i.e. the function

$$\sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

assumes its minimum for the least-square estimates β_0 and β_1.

Theorem 2.5. *The least square estimates for the simple linear regression model are*

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where

$$SS_{xx} = \left(\sum_{i=1}^{n} x_i^2 + \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \right),$$

$$SS_{xy} = \left(\sum_{i=1}^{n} x_i y_i + \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} \right),$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n},$$

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

$\hat{\beta}_0$ and $\hat{\beta}_1$ are called the least-squares estimators of β_0 and β_1 and

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

is called the least-squares regression line.

Some properties of least square estimators are

- The estimators are linear in the random variable $y_i, i = 1, \ldots, n$:

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i, y_i, \quad \text{where } c_i = \frac{x_i - \bar{x}}{SS_{xx}}, i = 1, \ldots, n,$$

and

$$\hat{\beta}_0 = \sum_{i=1}^{n} d_i, y_i, \quad \text{where } d_i = \frac{1}{n} - c_i \bar{x}, i = 1, \ldots, n;$$

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators for β_0 and β_1, respectively, i.e.

$$\mathbb{E}(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad \mathbb{E}(\hat{\beta}_1) = \beta_1;$$

- $\text{var}\left(\hat{\beta}_0\right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}\right) \quad \text{and} \quad \text{var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{SS_{xx}}.$

Definition 2.16. For sample data $(x_i, y_i), i = 1, \ldots, n$, the ith residual is defined as

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right).$$

Some properties of the residuals are

- $\sum_{i=1}^{n} e_i = 0;$

- $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i;$

- $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x};$

- $\sum_{i=1}^{n} x_i e_i = 0;$

- $\sum_{i=1}^{n} \hat{y}_i e_i = 0.$

Theorem 2.6. *Let* $SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$, $SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$ *and* $SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$. *Then*

- *if* $\mathbb{E}(SS_{res}) = \sigma^2$ *then*

$$\hat{\sigma}^2 = \frac{SS_{res}}{n-2},$$

where $\hat{\sigma}^2$ *is a variance of* $\hat{y}$;

- $SS_T = SS_R + SS_{res}$, *i.e.*

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

The second part of the theorem shows that the total sum of squares measuring of the total variation present in the measurements for the response variable y (SS_T) is equal to the addition of the sum for measuring how much of the variation can be accounted for through the regression model (SS_R), and the residual variation (SS_{res}).

Definition 2.17. (Model fit) The way to measure model fit is through the Coefficient of Determination

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{res}}{SS_T}.$$

The SS_{res} is the amount in the total Sum of Squares in y, which remains unexplained by the model. Therefore the R^2 is the proportion of the variance in y which can be explained by the model. A value of R^2 is close to 1 shows that the model is a very good estimator, on the other hand, the model does not fit if R^2 is close to 0.

2.8.2 Maximum Likelihood

Assuming that the *error* for the n data points, $(x_i, y_i), i = 1, \ldots, n$, is independent and identically normal distributed, with zero mean and σ^2 variance. the density function of the normal distribution is needed.

A Maximum-Likelihood Estimator (MLE) for a parameter is chosen, such that the chance of the data occurring is maximal if the true value of the parameter is equal to the value of the Maximum Likelihood Estimator.

The likelihood function L, $L : \Theta \times \mathbb{R}^n \to [0,1]$, assigns to parameter value $\theta \in \Theta$ and sample data $\tilde{y} \in \mathbb{R}^n$ the likelihood to observe the data $\tilde{y}$, if θ is the true parameter value describing the population.

Let $f(\tilde{y}|\theta)$ be the density function for random variable $\tilde{Y}$ (representing the random sample) with parameter θ. Then the Likelihood function is:

$$L(\theta|\tilde{y}) = f(\tilde{y}|\theta)$$

The MLE for θ based on data $\tilde{y}$ is the value $\tilde{}$ which maximizes the likelihood function for the given $\tilde{y}$. In most cases it will be easier to maximize the function $l := \ln(L)$. This is valid because the "ln" function is monotonic increasing.

Maximum likelihood estimation is one of the methods for estimating the parameters of a probability distribution by maximizing a likelihood function.

Two examples of applications of maximum likelihood estimations for discrete distribution and continuous distribution are as follows (Hogg).

The maximum likelihood estimation for Bernoulli distribution

Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample with Bernoulli distribution (2.3). In this case $\theta = p$. The probability that $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$ is the joint probability mass function

$$p^{y_1} (1-p)^{1-y_1} p^{y_2} (1-p)^{1-y_2} \cdots p^{y_n} (1-p)^{1-y_n} = p^{\sum y_i} (1-p)^{\sum y_i},$$

where y_i equals 0 or 1, $i = 1, 2, \ldots, n$. This probability is the likelihood function

$$L(p) = p^{\sum y_i} (1-p)^{\sum y_i}, \quad 0 \le p \le 1. \tag{2.8}$$

Let $l(p) = \ln L(p)$, then

$$l(p) = \left(\sum_i^n y_i \right) \ln p + \left(n - \sum_i^n y_i \right) \ln (1-p).$$

Since function "ln" is a monotonic increasing differentiable function thus the likelihood function $L(p)$ and its logarithm $l(p)$ are maximized for the same value of

p. In order to find the maximum value p, let

$$\frac{d}{dp}\left[l(p)\right] = \frac{\sum y_i}{p} - \frac{n - \sum y_i}{1 - p} = 0. \tag{2.9}$$

Equation (2.9) provides that p is not equal to 0 or 1, which is equivalent to the equation

$$\frac{\sum y_i}{p} = \frac{n - \sum y_i}{1 - p}.$$

The solution of equation (2.9) is $p = \dfrac{\sum y_i}{n}$, which also maximizes equation (2.8). The corresponding statistic

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is called the *maximum likelihood estimator of p*.

The maximum likelihood estimation for Normal distribution

Support $Y_1, \ldots, Y_n$ are independent and identically distributed random variables $\mathbf{N}(\mu, \sigma^2)$. In this case $\theta = (\mu, \sigma^2)$, the probability that $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$ and the common probability density function is

$$
\begin{aligned}
L(\mu, \sigma^2) &= \prod_{i}^{n}\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right)\right] \\
&= \left(\sqrt{2\pi\sigma^2}\right)^{-\frac{n}{2}}\exp\left(-\frac{1}{2\sigma^2}\sum (y_i - \mu)^2\right).
\end{aligned}
$$

The natural logarithm of the likelihood simplifies to

$$l(\mu, \sigma^2) = -\frac{n}{2}\ln 2\pi - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2. \tag{2.10}$$

Taking partial derivatives of equation (2.10) with respect to μ and σ and setting them to 0, which provides

$$
\begin{aligned}
\frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) = 0, \\
\frac{\partial l}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \mu)^2 = 0.
\end{aligned}
$$

Solving the above equations, one obtain

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum \left(Y_i - \hat{\mu}\right)^2}{n}$$

as the *maximum likelihood estimator of* (μ, σ^2).

2.9   Akaike Information Criterion

The Akaike information criterion (AIC) is an estimator of out-of-sample prediction error and thereby relative quality of statistical models for a given set of data. The main concept is to approximate the out-sample prediction loss by the sum of the in-sample prediction loss and a correction term (Ding, Tarokh and Yang, 2018). Suppose that we have a statistical model of some data with k numbers of estimated parameters in the model. The AIC value of the model is the following,

$$\text{AIC} = 2k - 2\ln\left(\hat{L}\right),$$

where $\hat{L}$ is the maximum value of the likelihood function for the model which is defined by $\hat{L} = p(x|\hat{\theta}, M)$, x is the observed data, $\hat{\theta}$ are parameter values that maximize the likelihood function and M is the model.

2.10   Bayesian Information Criterion

The Bayesian Information Criterion (BIC) is an index used in Bayesian statistics to select among a finite set of models. It is another popular model selection principle (Ding, Tarokh and Yang, 2018). The BIC is also known as the Schwarz information criterion or the Schwarz-Bayesian information criteria. When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. Both BIC and AIC attempt to resolve this

problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC. The BIC value of the model is defined as

$$\text{BIC} = \ln(n)k - 2\ln\left(\hat{L}\right),$$

where n is the number of data points in x, the number of observations, or equivalently, the sample size.

2.11 Researches Related with the Applications of Statistical Models to Insurance Claims

In this section, the researches related with the applications of statistical models to insurance claims are reviewed.

Boucher, Denuit and Guillén (2008) modeled insurance claim counts with time dependence based on generalization of Poisson and Negative Binomial Distributions. They found that some intuitive models involving time dependence cannot be used to model the number of reported claims. Also random effect models have a better fit than some other models.

Edwards (2004) calculated the moments of the distribution of aggregate life insurance claims from seriatim inforce data. He approximated the aggregate claims distribution with a mixture of a gamma distribution plus an exponential distribution with parameters chosen.

Boucher and Davidov (2011) presented an application of Tweedie Distribution to Claims Reserving Model. They considered Tweedie's compound Poisson model in a claims reserving triangle in a generalized linear model framework.

Smolárová (2017) proposed applications of Tweedie compound Poisson model in non-life insurance pricing and claims reserving. The model was applied

on the real data.

Gómez-Déniz et al. (2011) proposed a new distribution which is applicable to actuarial works, including short and long tailed count data. They considered the compound version of the geometric distribution.

Achieng (n.d.) studied industrial statistical distributions used in actuarial analysis of insurance claim amounts and more specifically in motor policy, which are Exponential, Gamma, Log-normal and Weibull distributions.

Kumar, Ghani and Mei (2010) claimed that health insurance costs across the world have increased alarmingly in recent years. The cause are payment errors made by the insurance companies while processing claims. The errors result in extra administrative effort to reprocess (or rework) claims, which accounts for up to 30% of the administrative staff in a typical health insurer. They applied data mining to describe a system that helps reduce these errors. The machine learning techniques was used to predict claims that will need to be reworked, generating explanations to help the auditors correct these claims, and experiment with feature selection, concept drift and active learning to collect feedback from the auditors to improve over time.

T. L. Oshini Goonetilleke and H. A. Caldera (2013) analyzed customer attrition by classifying all policy holders who are likely to terminate their policies. Retaining customers who purchase life insurance policies is an even bigger challenge since the policy duration spans for more than twenty years. Thus companies are eager to reduce these attrition rates in the customer-base by analyzing operational data. Data mining techniques play an important role in facilitating these retention efforts.

M. Durairaj and V. Ranjani (2013) studied reports of different types of data mining applications in the health care sector to reduce the complexity of the

study of the health care data transactions.

David, Marcus and Celynda (2016) demonstrated the ability of this state-of-the-art predictive analysis to find potential rare-disease patients in a large and complex database. Machine-learning techniques applied to a de-identified claims database are clearly capable of identifying these undiagnosed and inappropriately treated patients. This information could be valuable to claims managers and employers who may realize savings by helping physicians bring these patients to appropriate treatment sooner. The potential exists to apply this technique to other diseases that are rare.

Spedicato, Dutang and Petrini (2017) explored the applicability of novel machine learning techniques such as tree boosted models to optimize the proposed premium on prospective policyholders. Given the predictive gain over GLMs, they carefully analysed both the advantages and disadvantages induced by their use. As the level of competition increases, pricing optimization is gaining a central role in most mature insurance markets, forcing insurers to optimize the rating and consider customer behaviour.

Noorhannah and Manoj (2018) presented risk prediction in life insurance industry using supervised learning algorithms. Risk assessment is a crucial element in the life insurance. Companies perform underwriting process to make decisions on applications and to price policies accordingly. With the increase in the amount of data and advances in data analysis, the underwriting process can be automated for faster processing of applications. They aim at providing solutions to enhance risk assessment among life insurance firms using predictive analysis. The real world data set with over hundred attributes has been used to conduct the analysis. The dimensionality reduction has been performed to choose prominent attributes that can improve the prediction power of the models. The data dimension has been

reduced by feature selection techniques and feature extraction namely.

CHAPTER III

RESEARCH METHODOLOGY

This chapter presents the process used in this research. The process comprises of 8 parts as follows:

1. data import and packages installation;

2. histogram plots and descriptive analysis;

3. train and test data splitting;

4. fit of distributions;

5. modelling by generalized linear models (GLM);

6. feature selection by variable selection methods;

7. method prediction for GLM fitting;

8. accuracy measurement for the predicting model.

3.1 Data Import and Packages Installation

The data used in this thesis was obtained from "Sample Insurance Claim Prediction Dataset" based on "[Medical Cost Personal Datasets][1]", Dataset owner *Eason*, date created 2018-05-14, last updated 2018-06-04, version 2, available on **https://www.kaggle.com/easonlai/sample-insurance-claim-prediction-dataset**.

Table 3.1: Variables of insurance data using in this thesis.

variable	description
age	age of policyholder
sex	gender of policy holder (female=0, male=1)
bmi	body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weight, ideally 18.5 to 25
steps	average walking steps per day of policyholder
children	number of children / dependents of policyholder
smoker	smoking state of policyholder (non-smoke=0;smoker=1)
region	the residential area of policyholder in the US (northeast=0, northwest=1, southeast=2, southwest=3)
charges	individual medical costs billed by health insurance
insuranceclaim	yes=1, no=0

The insurance data claim is composed of variables as shown in Table 3.1 The data was saved in CSV file which is available for viewing and working by Excel, SPSS, MATLAB, RStudio, etc. However, this thesis focused on using RStudio for the mathematical and statistical study.

R is a language and free software environment for statistical computing and graphics. The R project was first developed by Robert Clifford Gentleman and Ross Ihaka in the early 1990s. Robert Gentleman is a Canadian statistician and bioinformatician and Ross Ihaka retired as an associate professor of statistics at the University of Auckland, New Zealand, in 2017. RStudio software uses the R language to develop statistical programs.

Detail of insurance data import to R and packages usage are shown in section A.1 of Appendix A.

3.2 Histogram Plots and Descriptive Analysis

A histogram is a diagram consisting of rectangles whose area is proportional to the frequency of a variable, which is a visual representation of the distribution of numerical data. It presents an estimate of the probability distribution of a continuous variable.

All numerical values of the nine variables shown in Table 3.1 were considered in histogram charts. The shapes diagram presented in each histogram chart provides us an statistical informatics which leads to the related probability distribution. Note that examples of histogram command in RStudio software are in section A.2 of Appendix A.

3.3 Train and Test Data Splitting

This part divides the data set into two subsets:

- **training set:** This is a subset that we use to train the model.

- **test set:** This is a subset that we use to provide an unbiased evaluation of the final model fit on the training data.

In this thesis, the dataset was randomly split, 70% into the training set and the remaining 30% into the test set. It was done by using caret, R library, an example of which is shown in section A.3 of Appendix A.

3.4 Fit of Distributions

Before fitting one or more distributions to a data set, a predefined set of some distributions is required. The choice is scoped by the knowledge or the characteristic of the data. Even though there is a theory for popular parametric claim size distributions (Wüthrich, 2017), one need to confirm that what the distribution is fit for our data. The RStudio software has the fitdistrplus package which provides a tool for the basic evaluation for that propose.

Some commands concerned fitting the proper distribution are shown in section A.4 of Appendix A.

3.5 Modelling by Generalized Linear Models (GLM)

The purpose of this thesis is to explain how the variable *charges* depends on other variables, i.e. *age, sex, bmi, step, children, smoker, region* and *insurance-claim*. Since our data may not be normal distributed, generalized linear model (GLM) is appropriate for our variables. The family types used in GLM modelling were considered from the result of the previous section (fit of the distributions). Examples of using glm() function of RStudio software are shown in section A.4 of Appendix A.

3.6 Feature Selection by Variable Selection Methods

The Akaike Information Criterion (AIC) value helps in finding the subset of variables in our data set which makes a model has lowest prediction error. Thus, AIC provides a means for feature selection.

There are three strategies of stepwise AIC:

1. **Forward selection**: The selection process is started with an empty model

and variables are added sequentially.

2. **Backward selection**: The selection process is started with the full model and variables are excluded sequentially.

3. **Both (forward and backward) selection**: This selection process combines both forward and backward selection. It starts like the forward model, no predictors, then one sequentially adds the most contributive predictors. However, after adding each new variable, some variables may be removed if they no longer provide improvement in the model fit.

All strategies were applied to our data set. The subset of variables, which provide best performing model (least AIC value), were selected. Note that, for each fitting distribution, the model has a different appropriate subset of variables. RStudio software is available of all three strategies, an example of which is shown in section A.6 of Appendix A.

3.7 Accuracy Measurement for the Predicting Model

The test data set is applied to the obtained model. The following values of the our results are evaluated:

- **root mean square error (RMSE)**

  The root mean square error is the standard deviation of the prediction errors of a model with respect to a test set. RMSE is a measure of how spread out these errors are.

- **mean square error (MSE)**

  The mean squared error (MSE) or mean squared deviation (MSD) of an estimator measures the average of the squares of the errors of a model with respect to a test set.

- **mean absolute error (MAE)**

  The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors on over all instances in the test set.
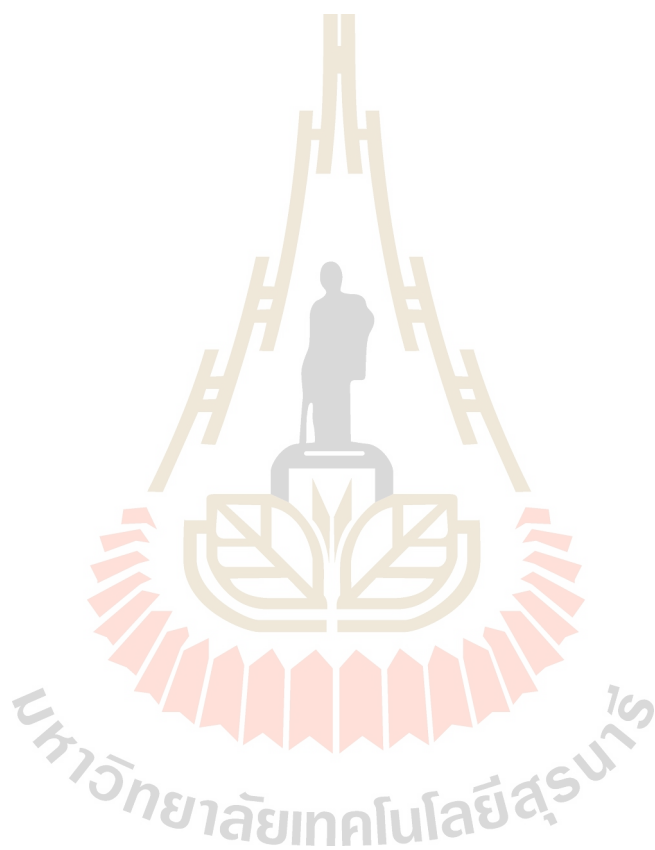
CHAPTER IV

RESULTS AND DISCUSSION

This chapter presents the results from the process proposed in Chapter III Research Methodology. The main goal in this section is about the output from R software manipulated on data set "Sample Insurance Claim Prediction Dataset" ($n = 1,338$) which based on "[Medical Cost Personal Datasets][1]", Dataset owner *Eason*, date created 2018-05-14, last updated 2018-06-04, version 2. available on **https://www.kaggle.com/easonlai/sample-insurance-claim-prediction-dataset**.

4.1   Exploratory Analysis of the Data Set

Some properties of the variables presented in table 4.1 are as follows:

1. *sex, smoker, region* and *insuranceclaim* are normal scale data type.

2. *age, bmi, steps, children* and *charges* are ratio scale data type.

3. The unit of *age* variable is *year*.

4. For *sex* variable, "0" means "female" and "1" means "male".

5. The unit of *bmi* variable is (kg/m^2).

6. The unit of *step* variable is *average walking steps per day*.

7. The unit of *children* variable is *person*.

8. For *smoker* variable, "0" means "non-smoke" and "1" means "smoker".

Table 4.1: Some statistical values of sample insurance claim prediction dataset ($n = 1,338$).

variable	mean	variance	skewness	kurtosis
age	39.20702541	197.4013867	0.055672516	-1.245087653
sex	0.505231689	0.250159595	-0.020951397	-2.002556636
bmi	30.66339686	37.18788361	0.284047111	-0.050731531
steps	5328.623318	6020365.13	0.662112022	-1.149448629
children	1.094917788	1.453212746	0.93838044	0.202454147
smoker	0.204783259	0.162968876	1.46476616	0.145755539
region	1.515695067	1.220770683	-0.038100508	-1.32770195
charges	13270.42227	146652372.2	1.515879658	1.606298653
insuranceclaim	0.585201794	0.242922211	-0.346253982	-1.882924956

9. For *region* variable, "0" means "northeast", "1" means "northwest", "2" means "southeast" and "3" means "southwest".

10. The unit of *charges* variable is *dollar*.

11. For *insuranceclaim* variable, "0" means "no" and "1" means "yes".

4.2 Histogram Plots

The distributions of the data set of each variable were considered by histogram plots. The shape of each histogram shown in figure 4.1 presents informative characteristics of the variable. Obviously, only some variables are suitable to study, e.g. *bmi* and *charges*. However, the study of *charges* provides more benefit for to both insurance companies and common people. The shape of histogram plot for variable *charges* is explored according to some family of distributions. Here we

scope on logistic, negative binomial, normal, log-normal and gamma distributions
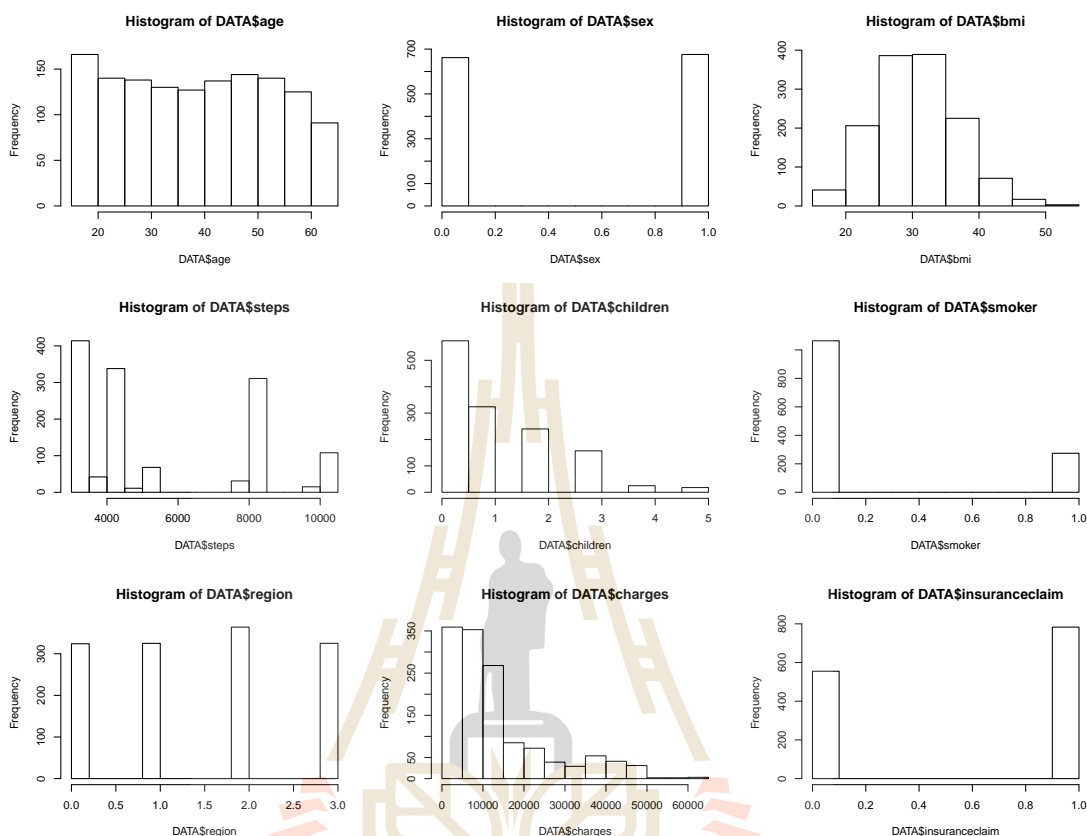since the shape of the histogram plot of *charges* is skewed to the right.

Figure 4.1: Histograms of data of each variable.

4.3 Train and Test

The data set was divided into 2 groups: 70% of the data was selected
randomly into the training set and the remaining 30% samples into test data set.
Here there are 937 units in the training set and 401 units in the test set. Some
statistic values of the training data set are shown in Table 4.2. Compare to the
statistic values of the data set in Table 4.1, the characteristic of the training set
is similar to one of the data set. Our sampling is good to be the representative of

the data set.

Table 4.2: Some statistical values of the training set ($n = 937$).

variable	mean	variance	skewness	kurtosis
age	39.26254002	198.1104863	0.031374321	-1.262186535
sex	0.500533618	0.250266809	-0.002137897	-2.004278075
bmi	30.72647279	37.18961674	0.287976056	-0.152925079
steps	5373.33191	6138814.427	0.619908572	-1.215203543
children	1.058697972	1.395055141	0.962238237	0.291107676
smoker	0.194236926	0.156676153	1.548251684	0.397928085
region	1.50266809	1.239576207	-0.027666038	-1.348435789
charges	13101.42315	148085510.6	1.559048197	1.731307595
insuranceclaim	0.5773746	0.244273869	-0.313774673	-1.905617511

4.4 Fit of Distributions

This process, it is to determine which distributions fits our train set best. Histogram and theoretical densities plot for *charges* variables are shown in Figure 4.2. Log-likelihood was a tool used in this process to find out the suitable distribution roughly.

By the given table, log-likelihood value for log-normal distribution is the highest, -5448.2398, and log-likelihood value for gamma distribution is the second highest, -5477.4609. AIC and BIC values of both distributions are also lower; the lower value provides that it is better in model fitting. Hence both distributions are considered for using in the GLM modelling.

However, negative binomial distribution is also another distribution used
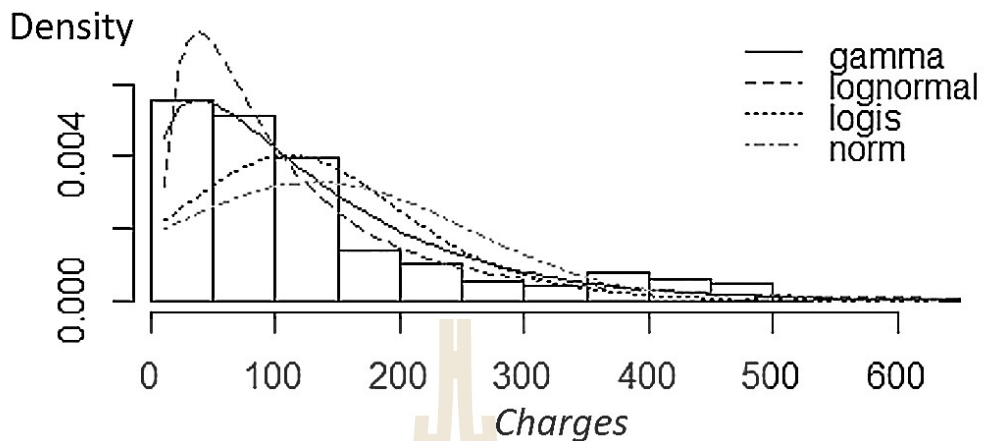
Histogram and theoretical densities

Figure 4.2: Histogram and theoretical densities plot for *charges* variables.

Table 4.3: Parameter Estimates for the Distributions.

Distribution	Log-Likelihood	AIC	BIC
Gamma	-5477.4609	10958.9217	10968.6071
Lognormal	-5448.2398	10900.4796	10910.1649
Logistic	-5774.7823	11553.5646	11563.2500
Normal	-5828.0319	11660.0639	11669.7492

in the GLM modelling because Boucher, Denuit, and Guillén (2008) proposed that insurance claim counts with time dependence can be modelled by negative binomial distribution. Negative binomial distribution is a two parameter discrete distribution which is skew right, its variance value is greater than its mean.

4.5   Generalized Linear Model

4.5.1   Gamma distribution

The R software performed the relation that *charges* depends on 5 variables, i.e. *age*, *bmi*, *steps*, *children* and *smoker*, which **AIC** = 18786,

$$
\begin{aligned}
\frac{1}{charges} = \quad & (1.670 \times 10^{-4}) + (-9.515 \times 10^{-7})age + (-6.734 \times 10^{-7})bmi \\
& + (3.038 \times 10^{-9})steps + (-2.277 \times 10^{-6})children \\
& + (-8.104 \times 10^{-5})smoker.
\end{aligned}
$$

The coefficients of variables in the model are in table 4.4.

Table 4.4: Coefficients of variables in the GLM model based on gamma distribution.

Variables	Coefficient Values
Intercept	1.670×10^{-4}
age	-9.515×10^{-7}
bmi	-6.734×10^{-7}
steps	3.038×10^{-9}
children	-2.277×10^{-6}
smoker	-8.104×10^{-5}

4.5.2 Log-normal distribution

The R software performed the relation that *charges* depends on 7 variables, i.e. *age, sex, bmi, children, smoker, region,* and *insuranceclaim*, which **AIC** = 18184 and

$$
\begin{aligned}
\ln(charges) = \quad & 6.9466 + (0.0349)age - (0.0583)sex + (0.0169)bmi \\
& + (0.0815)children + (1.6123)smoker + (-0.0416)region \\
& + (-0.1077)insuranceclaim.
\end{aligned}
$$

Here σ coefficient value is -0.817 and the coefficients of variables in the model are presented in table 4.5.

Table 4.5: Coefficients of variables in the GLM model based on log-normal distribution.

Variables	μ Coefficient Values
Intercept	6.9466
age	0.0349
sex	-0.0583
bmi	0.0169
children	0.0815
smoker	1.6123
region	-0.0416
insuranceclaim	-0.1077

4.5.3 Negative binomial distribution

The R software performed the relation that *charges* depends on 6 variables, i.e. *age*, *bmi*, *children*, *smoker*, *region*, and *insuranceclaim*, which **AIC** = 18469, and

$$
\begin{aligned}
\ln\left(charges\right) = \ & 7.31797 + (0.02846)age - (0.01756)bmi + (0.06663)children \\
& + (1.54337)smoker - (0.05260)region \\
& - (0.10264)insuranceclaim.
\end{aligned}
$$

The coefficients of variables in the model are presented in table 4.6.

Table 4.6: Coefficients of variables in the GLM model based on negative binomial distribution.

Variables	Coefficient Values
Intercept	7.31797
age	0.02846
bmi	0.01756
children	0.06663
smoker	1.54337
region	−0.05260
insuranceclaim	−0.10264

4.5.4   Evaluation of the Models Obtained

The three prediction models obtained can be evaluated according to the measurements by RMSE, MAE and MSE. The lower value in each measurement means the better model. All values are presented in table 4.7.

Table 4.7: RMSE, MAE and MSE of the models obtained.

| Model based on distribution	RMSE	MAE
Gamma	1.052	5378
Log-Normal	0.3936	4337
Negative binomial	0.5328	4491

CHAPTER V

CONCLUSION AND RECOMMENDATION

In this thesis, we proposed an analysis of distributions for insurance claims data. By the review literature, the application of GLM to model data set was used as a tool in analysis. Many contents in mathematics and statistics of data were studied. By the study, the well known distributions relating to claims data are logistics, normal, gamma and log-normal distributions. The histogram plots of each variable of insurance data provided an information to scope of our study. Here variables *charges* was chosen to analyze. The data obtained was split into 2 groups, training set and test set. The training set was composed of 937 items and the test set composed of 401 items. The histogram of train set was estimated by the distributions which we have proposed. The results in Table 4.3 implied that gamma and log-normal distributions were better fit in modelling. However, the negative binomial distribution was also another distribution used in this thesis. In the process of GLM modelling, all three proposed distributions were applied. For the model using gamma distribution, variable *charges* did not depend on *sex*, *region* and *insuranceclaim*. For the model using Log-normal distribution, variable *charges* did not depends on *steps* only. In the case of using negative binomial distribution, variable *charges* did not depend on *sex* and *step*.

The models obtained were measured the accuracy by the test data set. The result in Table 4.7 shows that log-normal distribution is the best in GLM model fitting, negative binomial distribution is the second best performance and gamma distribution performed worst.
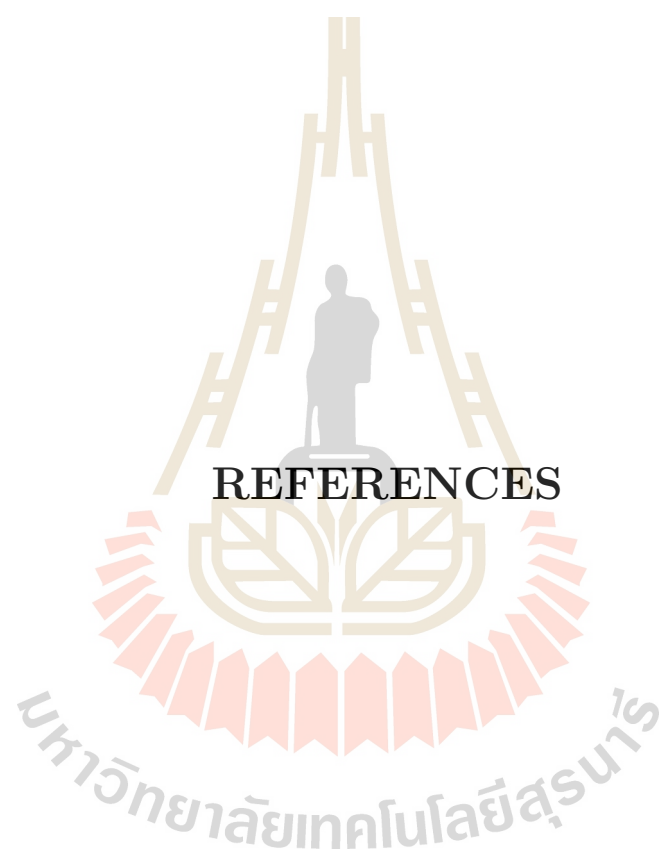
REFERENCES

REFERENCES

Achieng, O. M. (ND.) **Actuarial Modeling for Insurance Claim Severity in Motor Comprehensive Policy Using Industrial Statistical Distributions**, Online: https://www.researchgate.net/profile/.../22_final+paper_Oyugi.pdf.

Adewale, S. D. (2017). A Statistical Analysis to Determin The Distribution and Pattern for An Insurance Health Claim Data, **Thesis**, June 2017, DOI: 10.13140/RG.2.2.20068.42888.

Aldrich, J. (1997). R. A. Fisher and the Making of Maximum Likelihood 1912 – 1922, **Statistical Science**, Vol. 12.(3),pp 162-176.

Antonis, D. (2002). **Notes on Probability Theory and Statistics** Athens University of Economics and Business.

Bahnemann, D. (2015). **Distributions for Actuaries, Casualty Actuarial Society.**, Electronic Edition. https://www.casact.org/pubs/monographs/papers/02-Bahnemann.pdf.

Benckert, L. G., and Jung, J. (1974). Statistical Models of Claim Distributions in Fire Insurance. **ASTIN Bulletin: The Journal of the IAA.** Vol. 8, Issue 1, September 1974 , pp. 1-25, Published online: 01 August 2014, https://doi.org/10.1017/S0515036100009144.

Billingsley P. (1995). **Probability and Measure. 3rd edition.** New York: Wiley.

Boucher, J.P., Denuit, M., and Guillén, M. (2008). Models of Insurance Claim Counts with Time Dependence Based on Generalization of Poisson and

Negative Binomial Distributions, **Casualty Actuarial Society**, Vol. 2, Issue 1, http://www.variancejournal.org/issues/02-01/135.pdf.

Boucher, J.P., and Davidov D. (2011). **On the Importance of Dispersion Modeling for Claims Reserving: An Application with the Tweedie Distribution, Variance 5:2**, pp. 158-172.

Buro, K (ND.) **the lecture note in Stat 378**, Department of Mathematics and Statistics, MacEvan University. https://academic.macewan.ca/burok/Stat378/notes/glm.pdf.

Ding, J., Tarokh, V. and Yang, Y. (2018). Model Selection Techniques: An Overview, **IEEE Signal Processing Magazine**, Vol. 35 (6), pp 16-34.

Durairaj, M., and Ranjani, V. (2013). Data Mining Applications In Healthcare Sector: A Study **International Journal of Scientific and Technology Research** Vol. 2 ,Issue 10.

Edwards, T. (2004). **The Distribution of Aggregate Life Insurance Claims**, Online: https://www.soa.org/research/arch/2004/arch04v38n1_11.pdf.

Frees, E. W., Shi, P. and Valdez, E. A. (2009). Actuarial Applications of a Hierarchical Insurance Claims Model. **ASTIN Bulletin: The Journal of the IAA.** Vol. 39, Issue 1, May 2009, pp. 165-197, Published online by Cambridge University Press: 09 August 2013 :https://doi.org/10.2143/AST.39.1.2038061.

Goldburd, M., Khare, A. and Tevet, D. (2016). Generalized Linear Models for Insurance Rating, **the Casualty Actuarial So-**

ciety. https://www.casact.org/pubs/monographs/papers/05-Goldburd-Khare-Tevet.pdf.

Gómez-Déniz, E.,Sarabia, J. M. and Calderín-Ojeda, E. (2011). A new discrete distribution with actuarial applications, **Insurance: Mathematics and Economics.** Vol. 48, pp 406–412.

Grimmett, G. and Stirzaker, D. (2001). **Probability and Random Processes.** (3rd Ed.) New York: Oxford University Press Inc.

Grimmett, G. and Welsh, D. (1986). **Probability: An Introduction.** Northern Ireland: The Universities Press (Belfast) Ltd.

Hogg, R. V., McKean, J. W. and Craig, A. T. (2005). **Introduction to Mathematical Statistics.** (6th Ed.) Pearson Education, Inc. USA.

Janssen, J. and Manca, R. (2007). **Semi-Markov Risk Models for Finance, Insurance and Reliability.** New York: Springer Science+Business Media, LLC.

Jong, P. D. and Heller, G. Z. (2008). **Generalized Linear Models for Insurance Data.** New York: Cambridge University Press.
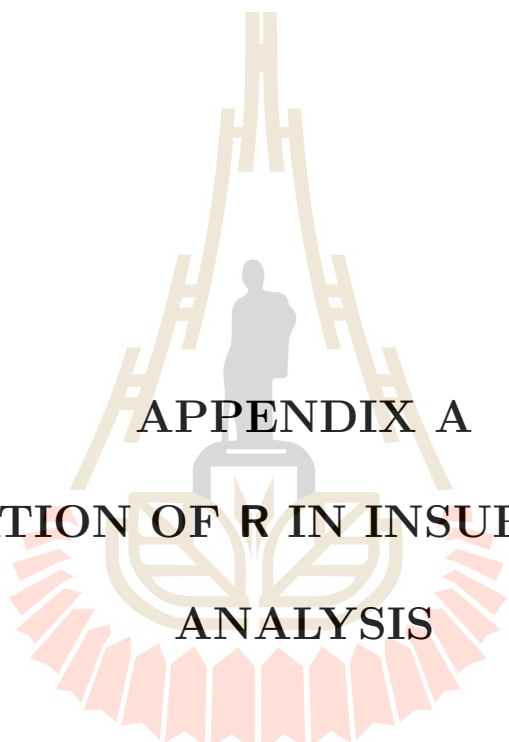
Kumar, M., Ghani, R. and Mei, Z. (2010). Data Mining to Predict and Prevent Errors in Health Insurance Claims Processing. **Conference: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, Washington, DC, USA, July 25-28, 2010, https://10.1145/1835804.1835816.

Leda D. M. (2010). **Lecture Notes in Insurance Risk Theory.** Available online: https://www.kent.ac.uk/smsas/personal/lb209/files/risk-notes-10.pdf.

Lin, X. S. (2006). **Introductory Stochastic Analysis for Finance and Insurance.** Hoboken, New Jersey: John Wiley & Sons.

Noorhannah Boodhun and Manoj Jayabalan (2018). **Risk Prediction in Life Insurance Industry Using Supervised Learning Algorithms** Electronic copy available at:https://www.springerprofessional.de/en/risk-prediction-in-life-insurance-industry-using-supervised-lear/15599544?fulltextView=true.

Oshini, T. L. G., and Caldera, H. A. (2013). Mining Life Insurance Data for Customer Attrition Analysis. **Journal of Industrial and Intelligent Information**, Vol. 1 (1), March 2013.

Panlilio, A. et. al. (2018). **Practical Application of Machine Learning Within Actuarial Work by Modelling, Analytics and Insights in Data Working Party.** Institute and Faculty of Actuaries. Electronic copy available at:https://www.actuaries.org.uk/documents/practical-application-machine-learning-within-actuarial-work.

Pozzolo, A. D. (2010). Comparison of Data Mining Techniques for Insurance Claim Prediction. **Thesis of Università degli Studi di Bologna**, Academic Year 2010/2011. Electronic copy available at: https://dalpozz.github.io/static/pdf/Claim_prediction.pdf KDD'10, July 25-28, 2010, Washington, DC, USA.

Shalizi, C. (2015). Lecture 6: The Method of Maximum Likelihood for Simple Linear Regression, **Modern Regression**. http://www.stat.cmu.edu/ cshalizi/mreg/15/lectures/06/lecture-06.pdf.

Smolárová, T. (2017). Tweedie Models for Pricing and Reserving, **Master Thesis**, Faculty of Mathematics and Physics, Charles University, Prague.

Spedicato, G. A., Dutang, C., and Petrini, L. (2017). **Machine Learning Methods to Perform Pricing Optimization. A Comparison with Standard GLMs**. Electronic copy available at:https://www.variancejournal.org/articlespress/articles/Machine-Spedicato.pdf.

Tse, Y.-K. (2009). **Nonlife Actuarial Models.** New York: Cambridge University Press. ASTIN Bulletin: The Journal of the IAA, Vol. 39, Issue 1, May 2009, pp. 165-197. Electronic copy available at: https://doi.org/10.2143/AST.39.1.2038061.

Vapnik, N.V (1998). **Statistical Leaning Theory.**, John Wiley & Sons, Inc: USA.

Wüthrich, M. V. (2017). **Lecture Notes in Non-Life Insurance: Mathematics and Statistics**, RiskLab Switzerland, Department of Mathematics, ETH Zurich. https://ssrn.com/abstract=2319328.

Zuanetti, D.A., Diniz, C.A.R and Leite, J.G. (2006). A lognormal model for insurance claims data., **Statistical Journal.** Vol. 4, Number 2, June 2006: pp 131–142.
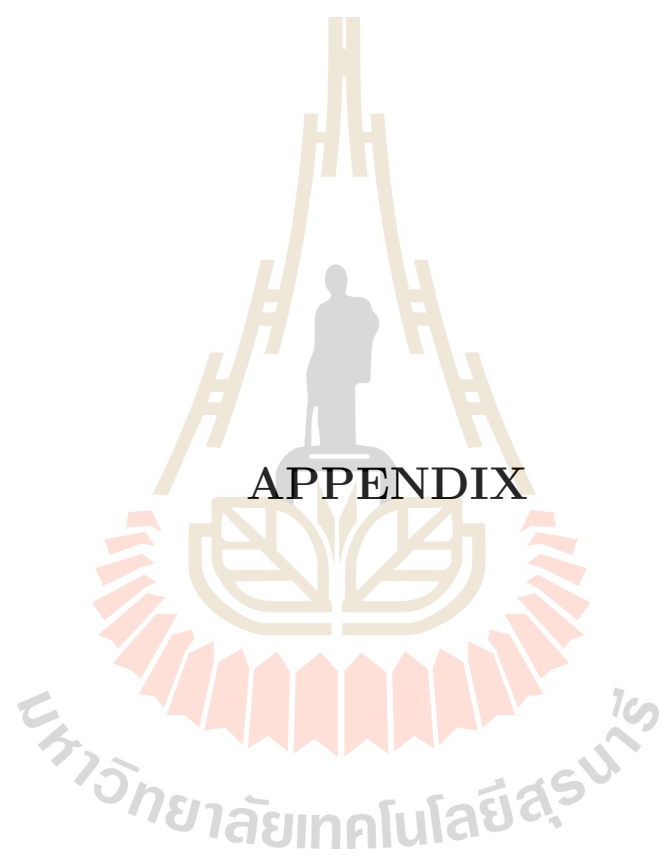
APPENDIX

APPENDIX A

APPLICATION OF **R** IN INSURANCE DATA

ANALYSIS

This chapter presents some R commands using in this thesis.

A.1 Loading data and using related package in **RStudio** software

In order to load data into RStudio software and use libraries, the process is the following:

- Set and Get working directory in R:

wd<-"D:/GLM/n/3/"

setwd(wd)

getwd()

- Reading a CSV File in R:

mydata = read.csv(Data input.csv)

- Using packages:

library(MASS) #Support Functions and Datasets for Venables and Ripley's MASS

library(gamlss) #Generalised Additive Models for Location Scale and Shape

library(fitdistrplus) #Help to Fit of a Parametric Distribution to Non-Censored or Censored Data

library(caret) #Classification and Regression Training

library(kernlab) #Kernel-Based Machine Learning Lab

A.2 Histogram Plots and Descriptive Analysis

To show histogram of input data, some examples are as follows:

hist(mydata$age)

hist(mydata$charges)

hist(mydata$insuranceclaim)

A.3 Train and Test Data Splitting

The following code uses library dplyr to split 70% of the data selected randomly into training set and the remaining 30% sample into test data.

```
train<-sample_frac(mydata,0.7)
sid<-as.numeric(rownames(train))
test<-mydata[-sid,]
```

A.4 Fit of Distributions

Here, library fitdistrplus was used in the thesis. The package fitdistrplus provides functions for fitting univariate distributions to different types of data and allowing different estimation methods. Example of using fitdistrplus package for fitting distribution is the followings.

```
library(fitdistrplus)
x<-train$charges
fg <-fitdist(x, "gamma")
fn <-fitdist(x, "norm")
fln <-fitdist(x, "lnorm")
plot.legend <- c("gamma","norm","lnorm","logis")
hist(x)
denscomp(list(fg,fn,fln,flg), legendtext = plot.legend)
```

A.5 Generalized Linear Model

Generalized linear model are fit using the glm() function. The form of the glm function is

$$\text{glm}(formula, \textbf{family}=familytype\ (\textbf{link}=linkfunction), \textbf{data}=)$$

For log-normal distribution, the commands are

library(gamlss)

glm.lognormal <-gamlss(charges~ages+sex+bmi+step+children+smoker+region
+insuranceclam, family = LOGNO() ,data=train)

In the case of the negative binomial, the commands are

library(MASS)

glm.negbi <-glm.nb(charges~ages+sex+bmi+step+children+smoker+region
+insuranceclam, data=train)

With the library MASS, for other distributions in the exponential family, the command can be changed to

glm.model <-glm.nb(charges~ages+sex+bmi+step+children+smoker+region
+insuranceclam, family = XXXX(link="YYYY"), data=train)

Here XXXX is a family and YYYY is a link function in table A.1

A.6 Feature Selection by Variable Selection Methods

Finding the appropriate subset of variables which makes model performing well via AIC for all three strategies is available in the following process:

library(MASS)

glmmodel <-glm(charges~.,family =XXXX(link="YYYY"), data=train)

modelf <-stepAIC(glmmodel, direction = "forward")

modelb <-stepAIC(glmmodel, direction = "backward")

modelboth <-stepAIC(glmmodel, direction = "both")

Table A.1: Table of family distribution names and link functions used in the glm command.

Family	DefaultLinkFunction
binomial	(link="logit")
gaussian	(link="indentity")
Gamma	(link="inverse")
inverse.gaussian	(link="1/mu^2")
poisson	(link="log")
quasi	(link="indentity",variance="constant")
quasibinomial	(link="logit")
quasipoisson	(link="log")

A.7 Accuracy Measurement for the Predicting Model

Root mean square error (RMSE), mean square error (MSE) and mean absolute error (MAE) can be obtained according to the following commands respectively:

predictmodel <- predict(bestmodel,newdata=test,type = "response")

RMSE <-sqrt(mean((test$ charge - predictmodel)2/test$ charge2))

MAE <-mean(abs(test$ charge - predictmodel))

MSE <-sqrt(mean((test$ charge - predictmodel)2))

CURRICULUM VITAE

NAME : Natakon Nawaratana　　　　　**GENDER :** Male

EDUCATION BACKGROUND:

- Bachelor of Science (Mathematics), Suranaree University of Technology, Thailand, 2016

SCHOLARSHIP:

- His Majesty the King's 7th Cycle Birthday Anniversary Suranaree University of Technology Scholarship

- Development and Promotion of Science and Technology Talents Project (DPST)

CONFERENCE:

- The 14th IMT-GT International Conference on Mathematics, Statistics and Their Applications, Thailand, 8-10 December, 2018

EXPERIENCE:

- Teaching Assistant in 103105 Calculus III, Term 3/2018 and Term 1/2019

- Teaching Assistant in 103001 Foundations for Calculus, Term 3/2018 and Term 1/2019