UNIVERSITÉ DU QUÉBEC À MONTRÉAL

INDIVIDUAL LOSS RESERVING IN GENERAL INSURANCE USING NEURAL NETWORKS.

DISSERTATION

PRESENTED

AS PARTIAL REQUIREMENT

TO THE MASTERS IN MATHEMATICS

BY

ERNEST DANKWA

DECEMBER 2020

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MODÉLISATION INDIVIDUELLE DES RÉSERVES EN ASSURANCE I.A.R.D. BASÉE SUR DES RÉSEAUX NEURONES.

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

ERNEST DANKWA

DÉCEMBRE 2020

REMERCIEMENTS

Tout d'abord, je voudrais remercier le Dieu Tout-Puissant pour sa grâce et ses bénédictions divines de vie et de soutien tout au long de ma période d'étude dans cette grande institution. Ensuite, ma profonde gratitude va à mon directeur de recherche, Mathieu Pigeon. Je le remercie pour tout son soutien, son inspiration, ses conseils et, surtout, son engagement envers l'excellence. Ils ont été essentiels non seulement pour le succès de ce travail mais aussi pour tout mon progrès et mon bien-être. Je suis également reconnaissant à tous les membres du corps professoral et au personnel du département de mathématiques qui ont créé un environnement d'étude productif pour moi . De plus, je reconnais la contribution significative de Co-operators qui a fourni les données sans lesquelles la rédaction de ce mémoire n'aurait pas été possible. Enfin, j'ai spécialement dédié ce travail à ma femme bien-aimée, Doris Borketey, à mes fils, Joshua et Mishael, à ma chère mère, Mary Ashyia, et à la mémoire de mon défunt père, S.M. Dankwa, en reconnaissance de leur amour et de leurs précieux sacrifices pour que mes études soient un succès.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

RÉSUMÉ

En assurance I.A.R.D, pratiquement tous les domaines d'activité tels que la conception de produits, la souscription, la tarification et la prise de décisions stratégiques et financières, peuvent être impactés par l'évaluation des réserves actuarielles. Les modèles classiques d'évaluation des réserves utilisent des données agrégées, souvent structurées sous la forme de triangles de développement. Dans de tels modèles, on suppose que les réclamations sont homogènes et que les paiements futurs ne dépendent que des années d'accident et de développement. Un inconvénient majeur de ces modèles est de ne pas intégrer les caractéristiques individuelles des réclamations qui peuvent permettre d'expliquer, en partie, l'hétérogénéité présente dans les données. Ces dernières années, cette hétérogénéité observée dans plusieurs bases de données d'assurance a motivé plusieurs chercheurs et chercheuses pour non seulement développer de nouvelles approches pour l'évaluation des réserves individuelles, mais aussi pour intégrer plusieurs outils d'apprentissage statistique afin de mieux prendre en compte les caractéristiques spécifiques des sinistres. Dans ce travail, nous explorons les réseaux de neurones en tant qu'outil d'apprentissage statistique et les intégrons dans un modèle d'évalua-tion individuelle des sinistres en assurance I.A.R.D. Nous proposons une approche de type fréquence-gravité dans laquelle un modèle linéaire généralisé est utilisé pour modéliser la fréquence des réclamations, et la sévérité des réclamations est modélisée à l'aide de réseaux de neurones. Nous appliquerons notre modèle à une base de données provenant d'une compagnie d'assurance canadienne et comparerons les estimations produites par ce modèle à celles obtenues à l'aide de certains modèles classiques.

Mots clés: Réserves individuelles, Réserve RBNS, Modèles linéaires généralisés, Modèle fréquence-sévérité, Apprentissage automatique, Réseaux de neurones.

ABSTRACT

In general insurance, virtually all areas of activity such as product design, underwriting, pricing, and strategic and financial decision making, can be impacted by the valuation of actuarial reserves. Traditional reserve valuation models use aggregate data, often structured in the form of development triangles. In such models, it is assumed that claims are homogeneous and that future payments depend only on years of accident and development. A major drawback of these models is that they do not integrate the individual characteristics of the claims which may help to explain, in part, the heterogeneity present in the data. In recent years, this heterogeneity observed in several insurance databases has motivated several researchers to not only develop new approaches for the evaluation of individual reserves, but also to integrate several statistical learning tools in order to better take take into account the specific characteristics of claims. In this work, we explore neural networks as a statistical learning tool and integrate them into an individual claim assessment model in general insurance. We propose a frequency-severity approach in which a generalized linear model is used to model the frequency of claims, and the severity of claims is modeled using neural networks. We will apply our model to a database from a Canadian insurance company and compare the estimates produced by this model with those obtained using some conventional models.

Keywords: Individual reserves, RBNS reserves, Generalized linear models, Frequency-severity model, Machine learning, Neural networks.

INTRODUCTION

A unique feature of property and casualty insurance is that claims vary in several ways. While some claims may rarely occur, others may occur frequently. The amount of claims may also range from some few dollars to several billions of dollars. Moreover, while some claims may settle quickly, others may take several years to settle. For instance, following the September 11, 2001 terrorist attack in USA, Larry A. Silverstein, the leaseholder who suffered financial loss due to collapse of the building, claimed over $7 billion from his insurers. This led to a legal issue which was settled on December 2004 at an amount of $4.55 billion.

Apart from the terrorist attack, the occurrence of floods and hurricanes have also had adverse impact on the insurance industry over the years. But irrespective of all these events, it is the primary responsibility of every insurer to ensure that an adequate amount of reserves is made available for the payment of future claims. However, for over several decades, the estimation and maintenance of loss reserves remain a major concern in property and casualty insurance. (Wenck, 1987) and (Leadbetter & Dibra, 2008) indicated that loss reserving is a major cause of insolvencies for several companies in the insurance industry.

This thesis explores a statistical/machine learning approach for loss reserving in the individual framework. It is not meant to be a comprehensive material on individual loss reserving and machine learning. However, it is laid out to give readers useful insights on existing loss reserving models and how the use of statistical/machine learning tools such as neural networks can help incorporate claim-specific features in loss reserving models. To an actuarial audience, it provides

significant contribution to the process of modeling and estimation of insurance data with respect to loss reserving. For a statistical audience, we make the concepts of statistical/machine learning more interesting by applying methods to real data set of an insurance company.

The rest of the work is organized as follows: In the first chapter, we will give a general overview of the property and casualty insurance to expose readers to the nature of the business and some concepts regarding loss reserving. In the second chapter, we shall discuss some classical models used for loss reserving and give an illustrative example. Next to that chapter, we shall discuss some statistical/machine learning tools and give theoretical and technical details of neural networks. We will introduce our proposed model for loss reserving in the fourth chapter, discuss the application of the classical and proposed models to an insurance data in the fifth chapter and draw conclusion of the studies.

CHAPTER I

GENERAL OVERVIEW

In this chapter, we shall expose our readers to property and casualty insurance business and highlight some key aspects of the business as well as some concepts regarding loss reserving.

1.1 The general insurance business

General insurance is a type of insurance that protects individuals and organizations from the financial losses due to the occurrence of unpredictable events such as fire, motor accidents, flood, storm, earthquake, theft, travel mishaps and legal actions. In USA and Canada, general insurance is popularly known as property and casualty insurance while Europe and other parts of the world it is referred to as non-life insurance. By purchasing a general insurance policy, individuals and organizations, known as policyholders, transfer their financial risks to the insurance company, the insurer, in exchange for the payment of periodic (usually monthly or annual) amounts called premiums.

Some of the key operations insurers carry out in business include determining the right insurance products and features that meet their customers' needs, selecting and classifying insured through a process called underwriting, using analytic pricing methods to determine insurance cost and policy premiums, making pay-

ments for insurance claims and managing funds to ensure that adequate funds are made available to cater for future claims. The activities of insurance companies are regulated and governed by state laws. An insurer who fails to accumulate adequate resources for the payment of future claims becomes insolvent and may be required by law to cease operating. Therefore, in general insurance, the ability of an insurance company to meet its future obligations is very important to all stakeholders of the organisation including policyholders, management, investors and regulators.

1.2 Loss reserving and its importance in general insurance

The term loss reserve is described by Actuarial Standard of Practices (ASOP 43) as an insurer's estimate of unpaid claims. It represents the insurer's obligation for future payment resulting from claims due to the occurrence of events covered in the insurance policy. The term "technical provision" is also used in some actuarial literature to mean loss reserve. Loss reserve is usually the largest liability item for property and casualty insurers and it is a very significant factor used to determine the solvency of an insurance company. It is the responsibility of actuaries to develop, quantify, evaluate and monitor loss reserves.

Now, we shall examine the importance of accurate loss reserving practices to the key stakeholders of general insurance companies.

First and foremost, it is in the best interest of policyholders that reserves are estimated properly because they are direct beneficiaries of the payments of claims made by the insurance company. Insolvencies and defaults in the payment of claims adversely affect the standard of living of policyholders which cause the public and potential policyholders to lose faith in insurance as far as the provision of financial security is concerned. In order to prevent such misfortunes, insurers

must ensure they have adequate amount of money available to pay for future claims.

Unlike policyholders, investors are not direct beneficiaries of claim payouts, however, they rely on the financial statements of insurance companies of which the estimate of loss reserves is a key input in most of the financial metrics. If reserves are miscalculated, the figures on balance sheets and income statements of the insurer will be inaccurate. This means that the true financial position of the insurer will be misrepresented. Therefore, investors will not have accurate information to make effective decisions. Worst of it all, investors are likely to loss their investment when an insurer becomes insolvent. Therefore, investors have interest in accurate loss reserving and good solvency measures of insurers.

Apart form policyholders and investors, the managers in the various department of an insurance company also, directly or indirectly, depend on accurate loss reserving in order to make effective decisions on company operations such as product development, pricing, underwriting, strategic and financial decisions.

Accurate loss reserving helps the product development department to determine appropriate features of insurance products such as limits and deductibles of insurance as well as reinsurance needs for an insurance product. Estimates of loss reserves also affect pricing of insurance products. Underestimation of reserves can influence the insurer an insurer to lower premium of insurance products. This can lead to inability to pay for future claims. On the other hand, overestimation of reserves can influence the insurer to increase premium of insurance products. In that case, the insurer may loss its market share to another insurer offering competitive products at more affordable premiums. The claim department also need accurate loss reserving so that adequate amount will be available to carry out their functions such as payment of claims and legal costs. Accurate loss reserving

is also necessary for managers to make financial and strategic decisions such as investment of funds, allocation of capital among lines of business, entry and exit of lines business.

From the fiscal point of view, improper estimates of loss reserves has impact on the amount of taxes the insurance company pays to the government. Overestimation of loss reserves reduces the company's taxable income for the company to pay less taxes to the government. However, the practice of using high level of loss reserve as a tax evasion mechanism is not encouraged since the company may end up having issues with tax authorities. In addition to that, announcements of high reserves may have adverse effects on share values, bond rating and managerial compensations because shareholders and bondholders may have difficulty distinguishing tax-motivated reserve from those that reflect true economic risks of the company.

Last but not least, insurance regulators rely on the financial statements of an insurer to carry out their supervisory role. Inaccurate reserves could result in a misstatement of the true financial position of an insurer. According to (Friedland, 2010), if a financially struggling insurer masks its true state with inadequate reserves, a regulator may not become involved to help the insurer regain its strength until it is too late.

1.3 Legal requirements for loss reserves

Apart from protecting the interests of all stakeholders, it is important to mention that proper estimation of unpaid claims by insurers is a legal requirement. For instance, (Friedland, 2010) stated a New York insurance law passed in the early 1960s which says that "... every insurer shall maintain reserves in an amount estimated in the aggregate to provide for the payment of all losses or claims

incurred on or prior to the date of settlement whether reported or unreported which are unpaid as of such date and for which such insurer may be liable, and also reserves an amount estimated to provide for the expenses of adjustment or settlement of such claims." These days, the role of the Appointed Actuary has been created through insurance legislation in countries around the world to meet the legal requirements for accurate estimation of unpaid claims.

In Canada, the Insurance Companies Act (1991) stipulates federal laws that govern the insurance industry. In this Act, the Office of the Superintendent of Financial Institutions (OSFI) is mandated to be the primary regulator of insurance companies responsible for setting guidelines concerning capital and solvency of insurers as well as standards for financial reporting. The Insurance Companies Act requires all federally regulated insurers to have an appointed actuary. It is a legal responsibility of the appointed actuary to value the actuarial and other policy liabilities of the company at the end of a financial year.

International Accounting Standards Board (IASB), an independent international organization for accounting and financial reporting, in the year 2005, issued International Financial Reporting Standard (IFRS 4) as the first guidance on accounting for insurance contracts. The standard requires the disclosure of information that will help users understand the amounts in the insurer's financial statements and evaluate the nature and extent of risks arising from insurance contracts. As of 1 January 2021, the IFRS 4 will be replaced by IFRS 17 which has more detailed approach to measurement and disclosure of an insurer's obligations from insurance contracts irrespective of the nature of the contracts. OSFI formally recognized the IFRS 17 in May 2018, hence placing emphasis on maintaining capital requirement and solvency of insurers as well as creating comparable financial reporting across international boundaries.

1.4 The claim development process

In general insurance, the final settlement of claims usually takes some time. Every claim goes through a process before it is finally settled. The development process for a claim with number, say k, can be illustrated by a timeline as shown Figure 1.1. The occurrence/accident date is the exact date which the event that caused

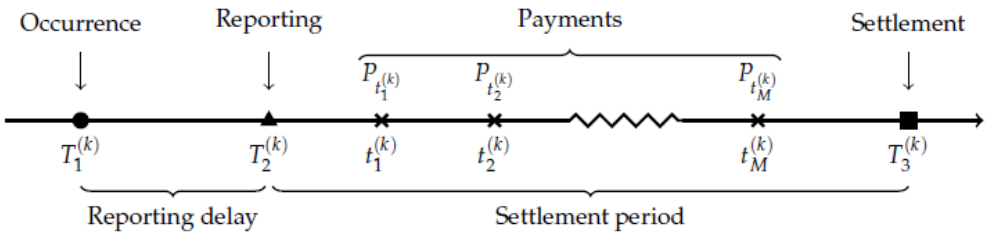

Figure 1.1 Claim development process. Adapted from (Duval & Pigeon, 2019).

the financial loss occurred while the reporting date is the date on which the insured informed the insurer of the occurrence of the accident. It is common that the date of the accident may not be the same as the reporting date and the insurer may not be aware of the existence of a claim until several years later. When the occurrence of the accident is not reported to the insurer on the same date as the accident date, the time-lag between the accident date and the reporting date is called reporting delay. In Figure 1.1, for instance, $T_1^{(k)}$ and $T_2^{(k)}$ represent occurrence and reporting dates, respectively, and the time between $T_1^{(k)}$ and $T_2^{(k)}$ represents reporting delay. This situation is more frequent for liability insurance claims. Between the accident date and reporting date, the claim is said to be Incurred But Not Reported (IBNR).

After the accident is reported the settlement of the claim may take several years due to legal activities. Until a claim finally settles on a closing/settlement date, the insurer may make an initial payment and several other payments (or refunds).

In Figure 1.1, for instance, the final settlement occurred at $T_3^{(k)}$. Before that date, M payments, $P_{t_1^{(k)}}, P_{t_2^{(k)}}, \ldots, P_{t_M^{(k)}}$ were made at times $t_1^{(k)}, t_2^{(k)}, \ldots, t_M^{(k)}$, respectively. Whereas claims from property insurance usually settle faster, claims from liability insurance or bodily injury often take a long time to settle. The time-lag between the reporting date and the settlement date is called closing/settlement delay. From Figure 1.1, the time between $T_2^{(k)}$ and $T_3^{(k)}$ represents closing/settlement delay. During the period between the reporting date and settlement date, the claim is said to be Reported But Not Settled (RBNS). Even after the settlement of a claim, the closed claim can be reopened due to unexpected new development or if a relapse occurs (Wüthrich & Merz, 2008). This will also call for several other payments before the claim finally closes.

When insurance companies do their periodic evaluations, it is expected by law that they estimate the total of unpaid claims. Actuaries, the professionals mandated to estimate reserves, put together Reported But Not Settled (RBNS) claims and Incurred But Not Reported (IBNR) claims to arrive at the total of unpaid claims.

1.5 Individual approach versus collective approach

Based on the granularity of data used to estimate loss reserves, a loss reserving model can be classified as a collective approach or an individual approach. In a collective approach for loss reserving, aggregate data for an entire portfolio is used for loss reserving. Usually, in a collective approach, individual claims are aggregated by accident period and developing delays and organized in a run-off loss triangle. (Denuit & Trufin, 2018) used a collective approach of loss reserving for long-tailed claims data. Classical models such as the chain-ladder and Mack's model are based on the collective approach.

Advantages of the collective loss reserving models include the following.

- They are easy to understand since the loss data is summarized in a run-off triangle.
- They are easy to implement since they apply straightforward formula.
- They can be used with minimum computational requirements.

Despite the above advantages, the collective loss reserving models have the following disadvantages.

- Collective models undermine heterogeneity in claims by their inability to incorporate any information about the individual claims.
- The pattern of claims do not remain the same for the future as it is assumed in some collective models such as the chain-ladder (Friedland, 2010).
- According to (England & Verrall, 2002), the observed data in a run-off triangles is typically small leading to large prediction error.
- (Charpentier & Pigeon, 2016) pointed out that the results of aggregate models will not be enough when modern solvency laws call for full conditional distribution of future cash flows.
- With collective models there is a misalignment of risk functions when ratemaking is based on individual data and reserving is based on aggregate data.
- Some collective models such as the chain-ladder model only give best estimates of reserve without providing any information about the variability of reserve and the level of risk.

On the other hand, the individual approach of loss reserving models the loss reserves of claims on individual policies. The total reserve is the sum of all reserves

of the various policies in the portfolio. The individual approach facilitates the study of each single claim to know its ultimate amount. One can also include claim specific features in the estimation of reserves. (Huang et al., 2015) and (Pigeon et al., 2013) presented examples of individual approach for loss reserving.

Advantages of individual loss reserving models include the following.

- Unlike collective loss reserving models, the individual loss reserving models allow claim specific information to be incorporated in the loss reserving models.
- They offer much simpler way to investigate frequency and severity of claims separately.
- Individual loss reserving models provide information about variability of reserve.
- They help to align loss reserving with ratemaking function which is mostly based on individual data.
- They relax the assumption that the pattern of claim remain the same for future.
- They enable the establishment of reserves that meet modern solvency requirements. Risk of the insurer can be determined from full conditional distribution of future losses.

Despite the above advantages, the individual loss reserving models have the following disadvantages.

- They are data oriented and require huge data to make meaningful predictions of future losses.

- The size of individual loss data tends to be huge which mean they require high computing powers.
- Individual loss reserving models tend to be more complicated than the collective loss reserving models.

After considering the individual and collective approaches to loss reserving, we decide to adopt the individual loss reserving approach in this work for two main reasons:

1. The advantages of the individual loss reserving approach far exceed that of the collective loss reserving approach.
2. Current solvency laws stipulated by the Office of the Superintendent of Financial Institutions (OSFI) require insurers to establish reserves that reflect the true risk of the company. Individual loss reserving approach ensures the compliance of modern solvency laws.

Therefore, we shall point it out that this research will give more attention to individual loss reserving methods. Readers interested in more applications of collective models should see (Wüthrich & Merz, 2008).

CHAPTER II

CLASSICAL MODELS FOR LOSS RESERVING

Loss reserving methodology has developed from very simple models to modern day complex ones. In this chapter, we shall discuss some classical models used for loss reserving in general insurance such as the chain-ladder (CL), Mack's model (MC) and generalized linear models (GLMs).

2.1 Literature review

Over the past 30 years, various researchers have made significant contributions towards the development of models for loss reserving in general insurance. While some of these models are stochastic, others are deterministic. A useful taxonomy of loss reserving models over the years can be found in (Taylor, 2000). In general, there have been increased interest in stochastic claim reserving in recent times. That is, there is a substantial growth in loss reserving models that are defined in probabilistic environments to induce randomness in output rather than models whose output are fully determined by parameter values. However, the stochastic models are only used by a limited number of practitioners (England & Verrall, 2002). The deterministic models were first developed as computational algorithms that focused mainly on obtaining estimates for reserves. Although the stochastic models are less dominant in current loss reserving practices, various researchers

have used them to extend the assumptions of the deterministic models so that uncertainties of reserve estimates and even full predictive distributions of reserve estimates can be provided along with reserves estimates. Overview of the deterministic and stochastic models can be found in (Taylor, 2000), (England & Verrall, 2002), (Wüthrich & Merz, 2008) and (Friedland, 2010).

In loss reserving for general insurance, it is well known that the most fundamental method underlying both deterministic and stochastic methods is the chain-ladder. The basic idea of this method was already known to (Tarbell, 1934). The chain-ladder method is a deterministic approach to loss reserving traditionally based on run-off triangles. However, the stochastic version of the chain-ladder which allows the estimation of mean square error of estimated reserves can be found in (Mack, 1993) and (Mack, 1999).

In recent years, several researchers have contributed to various extensions of the chain-ladder model. For instance, (Gogol, 1993) and (Verrall, 2004b) proposed models that combine prior information derived from expert opinion on total cost of claims or estimate of loss ratio with observations from a database. (Miranda et al., 2012) also proposed a double chain-ladder model that make use of triangles developed from paid and incurred losses.

The chain-ladder method calculates loss reserves using run-off triangles of paid and incurred losses representing the sum of paid losses and case reserves. The approach is considered to be a collective or macro-level loss reserving (Wüthrich & Merz, 2008). In essence, the method works under the assumption that patterns of claims activities in the past will continue to be seen in the future (Friedland, 2010).

The limitations of the chain-ladder method and other macro-level reserving models have been discussed by some researchers. These limitations include over-

parametrization of the chain-ladder method ((Wright, 1990) and (Renshaw, 1994)); unstable predictions for recent accident years (Bornhuetter & Ferguson, 1972); problems with the presence of zero or negative cells in run-off triangles (Kunkler, 2004); difficulties in separate assessment of RBNS and IBNR claims ((Schnieper, 1991) and (Liu & Verrall, 2009)) and difficulties in the simultaneous use of incurred and paid claims (Quarg & Mack, 2004).

Various contributions have been made by researchers to address the issues of the chain-ladder method. (Bornhuetter & Ferguson, 1972) built a loss reserving model that incorporates prior information about the insurers exposure to loss. (Verrall, 2004a) and (Meyers, 2007) advanced the traditional loss reserving methods by combining the chain-ladder and Bornhuetter-Furguson methods with Bayesian methodology. This enabled them to derive predictive distributions for future losses. Other researchers developed methods to obtain variability of estimates and mean square error of prediction. For instance, (Schnieper, 1991) and (Liu & Verrall, 2009) suggested using simulations based on separate RBNS and IBNR run-off triangles.

Some researchers viewed the problem of loss reserving from the perspective of regression models. (Halliwell, 1996) introduced a generalized least-squares approach in loss reserving. Further contributions to this approach were made by other authors including (Hamer, 1999) and (Ludwig et al., 2011).

Since the early 1990s, the use of generalized linear regression in loss reserving has gained increasing interest in actuarial literature. There exist an extensive literature that investigate the statistical basis of the chain-ladder method with a focus on the distributional assumptions of the aggregate data and the use of generalized linear regression models. For example, (Murphy, 1994) and (Barnett & Zehnwirth, 2000) studied the chain-ladder technique within the context

of normal linear regression. (England & Verrall, 2002) revealed that a normal model as an approximation to the negative binomial model provides a link to Mack's model. (Wright, 1990) and (Mack, 1991) also studied the link between the chain-ladder technique and the Poisson distribution. The implementation of the Poisson distribution using standard statistical methodology can be found in (Renshaw & Verrall, 1998). (Verrall, 2000) provided more details on the negative binomial model and its relationship to the over-dispersed Poisson model. The gamma model proposed by (Mack, 1991) also received considerable attention in literature. (Wüthrich, 2003) generalised the gamma cell distribution model in (Taylor, 2000) by using the Tweedie's compound Poisson distribution for claims reserving.

Having reviewed the some of the work done in literature, we shall describe the chain-ladder, Mack's model and generalized linear models in details.

2.2 Loss triangles

Estimates for macro-level models are often based on run-off triangles that summarize loss data based on occurrence/accident years and development years. Table 2.1 gives the format of a loss triangle over I and J accident and development years, respectively. For accident years i, $i = 1, \ldots, I$, and development years j, $j = 1, \ldots, J$, $C_{i,j}$ represents the cumulative amount of payment for the loss. An incremental loss triangle can be obtained from the cumulative loss triangle by taking the difference of cumulative payments for successive development years. That is,

$$\begin{aligned} Y_{i,1} &= C_{i,1}, \quad i = 1, \ldots, I \\ Y_{i,j} &= C_{i,j} - C_{i,(j-1)}, \quad i = 1, \ldots, I, \text{ and } j = 2, \ldots, J. \end{aligned}$$

Table 2.1 A run-off loss triangle

	Development year					
Acc. year	1	2	3	...	$J-1$	J
1	$C_{1,1}$	$C_{1,2}$	$C_{1,3}$	...	$C_{1,(J-1)}$	$C_{1,J}$
2	$C_{2,1}$	$C_{2,2}$	$C_{2,3}$	...	$C_{2,(J-1)}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$		
i	$C_{i,1}$	$C_{i,2}$	...	$C_{i,(J-i+1)}$		
$\vdots$	$\vdots$	$\vdots$	$\ddots$			
I	$C_{I,1}$					

$Y_{i,j}$ is described as incremental payment for accident year i and development year j. The cumulative loss triangle forms the basis of the macro-level models discussed in the subsequent sections.

We shall use Table 2.2 as a fictitious loss data of an insurer to illustrate the concepts in classical loss reserving.

Example 2.2.1. *The information provided in the records of the insurer can be explained as follows: In the year 1995, two accidents occurred in April and May, respectively. For those accidents, five payments have been made by the insurer from the year 1995 through to 1998. In the year 1996, one accident occurred of which 3 payments have been made as of year 1998. Two accidents occurred in the year 1997 and two payments have been made as of year 1998. Finally in the year 1998, one accident occurred of which two payment have been made by the insurer.*

In Example 2.2.1, an incremental loss triangle can be construct by taking note of the following

- The accident years under consideration are the 1995 through to 1998. Thus,

Table 2.2 A data of loss payments made by an insurer

Accident date	Payment date	Amount in $
04/1995	06/1995	2,500
05/1995	08/1995	1,550
04/1995	01/1996	700
05/1995	10/1997	650
04/1995	11/1998	420
03/1996	04/1996	3,900
03/1996	09/1997	950
03/1996	06/1998	600
07/1997	09/1997	2,800
10/1997	10/1998	1,500
02/1998	03/1998	2,900
02/1998	06/1998	400

following the notation in Table 2.2, we have: $i = 1, \ldots, 4$, which corresponds to the years $1995, 1996, 1997$ and 1998, respectively.

- We have four development years since the records indicate payments were made for losses in the years from 1995 through to 1998. Also, following the notation in Table 2.2, we have $j = 1, 2, 3, 4$.

The incremental loss triangle summarizes the total amount of loss paid in the development years, j, for each accident, i. For instance, in accident year 1995, a total payment of 4,050 (i.e. 2,500+1,550) was made in the development year 1; 700 was made in development year 2; 650 was made in development year 3; and 420 was made in development year 4 etc. The summary of incremental losses for

Example 2.2.1 is given in Table 2.3.

Table 2.3 Incremental Loss triangle

	Development year			
Acc. year	1	2	3	4
1995	4,050	700	650	420
1996	3,900	950	600	
1997	2,800	1,500		
1998	3,200			

The cumulative loss triangle sums all loss made up to a development years, j, for each accident, i. For instance, in accident year 1995, a total payment of 4,050 (i.e. 2,500+1,550) was made in the development year 1; 4,750 (i.e. 4,050+700) was paid in development year 2; 5,400 (i.e. 4,750+650) was paid in development year 3; and 5,820 (i.e. 5,400+420) was paid in development year 4 etc. The summary of cumulative losses for Example 2.2.1 is given in Table 2.4.

Table 2.4 Cumulative Loss triangle

	Development year			
Acc. year	1	2	3	4
1995	4,050	4,750	5,400	5,820
1996	3,900	4,850	5,450	
1997	2,800	4,300		
1998	3,200			

2.3 Chain-ladder method (CL)

Among all the loss reserving techniques, the chain-ladder method is the most popular, both in theory and practice. The method is recognized in actuarial literature as a computational algorithm for the estimation of loss reserves. (Mack, 1993) presented a distribution-free derivation of the chain-ladder method. Several extensions of the technique have been made in literature including conditional prediction error of the distribution free CL method and multivariate CL (Wüthrich & Merz, 2008). As proposed by (Mack, 1993), the key assumptions of the distribution free chain-ladder are defined as follows:

Definition 2.3.0.1 (Chain-ladder method)**.**

(CL1): *Cumulative claim amounts for different years of occurrence are independent.*

(CL2): *There exist development factors* λ_j *such that* $C_{i,(j+1)} = \lambda_j C_{i,j}$*, for* $i = 1, \ldots, I$ *and* $j = 1, \ldots, J-1$.

The development factors of the CL method, λ_j, are unknown and estimated by the expression below:

$$\widehat{\lambda}_j = \frac{\sum_{i=1}^{J-j} C_{i,(j+1)}}{\sum_{i=1}^{J-j} C_{i,j}} \quad \text{for} \quad j = 1, \ldots, J-1.$$

Furthermore, the estimator for the total cost of claims for each year of occurrence and that of reserve as given as follows:

$$\begin{aligned} \widehat{C}_{i,J} &= (\widehat{\lambda}_{J-i+1} \times \ldots \times \widehat{\lambda}_{J-1}) C_{i,(J-i+1)} \\ \widehat{R}_i &= \widehat{C}_{i,J} - C_{i,(J-i+1)} \\ \widehat{R} &= \sum_{i=2}^{I} \widehat{R}_i. \end{aligned}$$

The algorithm functions under the assumption that the pattern of claim activities in the past will be the same for future claims. It is shown that the development factors are uncorrelated even though they are not independent. Another useful property under the assumptions given above is that the estimator of the total claim is unbiased hence providing a justification for the reserve estimates. These properties and other related ones are well illustrated in (Wüthrich & Merz, 2008).

Despite the fact that the method is intuitive and simple, it has several flaws:

- The development factor is identical for all years of occurrence which is not the case in practice if there is a change in jurisdiction or management.
- It is impossible to make an estimate of the accuracy of estimate since the method is deterministic. Confidence intervals of estimated reserves cannot be easily obtained using the algorithm.
- Individual claim characteristics are not incorporated in the algorithm.
- Over-parametrization may affect recent estimates of reserve since the evaluation of reserve tends to be a product of many estimates.

2.3.1 Mack's model

The stochastic version of the CL algorithm was proposed by Mack (1993). The advantage of Mack's model is that, the model enables one to estimate the uncertainties of the reserve estimates in addition to estimates for reserves. The Mack's model has interesting properties and has laid the foundation for many research works in loss reserving.

In the Mack's model, the cumulative amount of claims payment for accident year, i, and development year, j, is denoted by $C_{i,j}$, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

Given below are the hypothesis of the Mack model:

Definition 2.3.1.1 (Mack's model)**.**

(MC1): *Cumulative claim amounts for different years of occurrence are independent.*

(MC2): *The development factors, $\lambda_1, \ldots, \lambda_{J-1}$ and cumulative payments satisfy the expression below:*
$E[C_{i,(j+1)}|C_{i,1}, \ldots, C_{i,j}] = \lambda_j C_{i,j}$ *for* $j = 1, \ldots, J-1$.

(MC2b): *The cumulative payments, $(C_{i,j}), j = 1, 2, \ldots$ form a Markov chain and there exist strictly positive variance parameters, $\sigma_1^2, \ldots, \sigma_{J-1}^2$, such that for $1 \leq i \leq J$ and $1 \leq j \leq J-1$, we have the expression below:*
$Var[C_{i,(j+1)}|C_{i,1}, \ldots, C_{i,j}] = \sigma_j^2 C_{i,j}$.

Similar to the chain-ladder algorithm, the development factors are estimated by

$$\widehat{\lambda}_j = \frac{\sum_{i=1}^{J-j} C_{i,(j+1)}}{\sum_{i=1}^{J-j} C_{i,j}} \quad \text{for} \quad j = 1, \ldots, J-1.$$

The variance parameters are also estimated as follows:

$$\begin{aligned} \widehat{\sigma}_j^2 &= \frac{1}{J-j-1} \sum_{i=1}^{J-j} C_{i,j} \Big(\frac{C_{i,(j+1)}}{C_{i,j}} - \widehat{\lambda}_j \Big)^2 \\ \widehat{\sigma}_{J-1}^2 &= \min\Big[\frac{\widehat{\sigma}_{J-2}^4}{\widehat{\sigma}_{J-3}^2}, \min\Big(\widehat{\sigma}_{J-3}^2, \widehat{\sigma}_{J-2}^2 \Big) \Big]. \end{aligned}$$

The hypotheses bring out very useful properties of the Mack's model some of which are given as follows:

- Concerning ultimate cost of claims we have
  $E[C_{i,J}|C_{i,(j-i+1)}] = C_{i,(j-i+1)} \lambda_{J-1} \lambda_{J-2} \ldots \lambda_{j-i+1}$, for $1 \leq i \leq J$.
- The estimator, $\widehat{\lambda}_j$ is an unbiased estimator for λ_j.

- The estimators $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{J-1}$ are uncorrelated.
- Given $C_{i,(j-i+1)}$, $\widehat{C}_{i,J}$ is an unbiased estimator for $E[C_{i,J}|C_{i,(j-i+1)}]$.
- $\widehat{C}_{i,J}$ is an unbiased estimator for $E[C_{i,J}]$.
- The estimator, $\widehat{\sigma}_j^2$, is an unbiased estimator for σ_j^2.

Details about the estimators of the Mack's model, proof of the their properties and other useful remarks can be found in (Wüthrich & Merz, 2008).

Using the Mack's model, the estimate for reserve for accident year, i, is given by

$$E[C_{i,J}|C_{i,(j-i+1)}] - C_{i,(j-i+1)} = C_{i,(j-i+1)}(\widehat{\lambda}_{j-i+1} \times \ldots \times \widehat{\lambda}_{J-1} - 1).$$

Consequently, the reserve estimate for the Mack's model and that of the chain-ladder algorithm are the same.

Example 2.3.1. *In this example, we recall Table 2.4 which is the cumulative loss triangle for Example 2.2.1. We shall give an illustration of how the development factors and reserves are estimated. The result is the same for both the Mack's model and the chain-ladder method.*

By applying the expression of the estimator given in sections 2.2 and 2.3, we can calculate the development factors as:

$$\begin{aligned}
\widehat{\lambda}_1 &= \frac{4{,}750+4{,}850+4{,}300}{4{,}050+3{,}900+2{,}800} = 1.293023 \\
\widehat{\lambda}_2 &= \frac{5{,}400+5{,}450}{4{,}750+4{,}850} = 1.130208 \\
\widehat{\lambda}_3 &= \frac{5{,}820}{5{,}400} = 1.077778.
\end{aligned}$$

The developments factors obtained above are then used to predict the unknown loss

amount as follows;

$$\begin{aligned}
\widehat{C}_{2,4} &= C_{2,3} \times \widehat{\lambda}_3 = 5{,}450 \times 1.077778 = 5{,}873.889 \\
\widehat{C}_{3,3} &= C_{3,2} \times \widehat{\lambda}_2 = 4{,}300 \times 1.130208 = 4{,}859.896 \\
\widehat{C}_{3,4} &= C_{3,2} \times \widehat{\lambda}_2 \times \widehat{\lambda}_3 = 4{,}300 \times 1.130208 \times 1.077778 = 5{,}237.888 \\
\widehat{C}_{4,2} &= C_{4,2} \times \widehat{\lambda}_3 = 4{,}137.674 \\
\widehat{C}_{4,3} &= C_{4,3} \times \widehat{\lambda}_1 \times \widehat{\lambda}_2 = 4{,}676.434 \\
\widehat{C}_{4,4} &= C_{4,4} \times \widehat{\lambda}_1 \times \widehat{\lambda}_2 \times \widehat{\lambda}_3 = 5{,}040.157.
\end{aligned}$$

The above results helps to complete Table 2.4 by filling the unknown values in the lower portion of the table. The complete table of known loss amounts and the predicted loss amounts (in bold characters) for the insurer is shown in Table 2.5.

Table 2.5 The complete cumulative Loss triangle

	Development year			
Acc. year	1	2	3	4
1995	4,050	4,750	5,400	5,820
1996	3,900	4,850	5,450	**5,873.89**
1997	2,800	4,300	**4,859.90**	**5,237.89**
1998	3,200	**4,137.67**	**4,676.43**	**5,040.16**

Once the table is completed, we apply the formulas for estimating reserves, $\widehat{R}_i$. That is, the values on the diagonal, representing the latest loss amounts, are subtracted from the corresponding ultimate loss (found on the last column of the table) to get the reserves for each accident year. The total reserve is the sum of all the

calculated accident year reserves. This is illustrated below:

$$\begin{aligned}
\widehat{R}_2 &= \widehat{C}_{2,4} - C_{2,3} = 5{,}873.89 - 5{,}450 = 423.89 \\
\widehat{R}_3 &= \widehat{C}_{3,4} - C_{3,2} = 5{,}237.89 - 4{,}300 = 937.89 \\
\widehat{R}_4 &= \widehat{C}_{4,4} - C_{4,1} = 5{,}040.16 - 3{,}200 = 1{,}840.157 \\
\widehat{R} &= \sum_{i=2}^{4} \widehat{R}_i = \mathbf{3{,}201.93}.
\end{aligned}$$

Even though the Mack's model for loss reserving is widely applied, easy to understand and produces a measure of uncertainty unlike the traditional chain ladder, it has some drawbacks that have drawn the attention of several authors. For example:

- Over-parametrization issue can lead to loss of predictive power since a lot of information that are not included in the estimation of the parameters become useless ((Wright, 1990).
- Non-robustness to outliers and failure of the method to include tail factors (Verdonck et al., 2009).
- Unstable prediction for recent accident years (Bornhuetter & Ferguson, 1972).

2.4 Generalized Linear Models (GLMs)

The use of Generalized linear models in loss reserving is described as parametric methods that allow assumptions to be made on the distribution of claim amounts.

Generalized linear models (Nelder & Wedderburn, 1972) were developed as a generalization of the ordinary linear regression. In general, when we have a data

of size n and p set of covariates, the ordinary linear regression model can be written mathematically as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \;\; i = 1, ..., n$$

where: Y_i and $X_{i1}, \ldots, X_{ip}$ represent the ith observation of the response and predictor variables, respectively; $\beta_0 \ldots \beta_p$ are regression parameters, and ε_i is the ith residual value which is assumed to be independent and normally distributed with mean 0 and constant variance, σ^2.

Linear regression models are easy to understand and implement. Results are also easy to interpret. However, for practical purposes, there are several drawbacks:

- The normality assumption is very restrictive. Most insurance data do not follow normal distribution.
- The constant variance assumption does not hold in real application.
- Application of a normal model is affected by issues with real data sets such as missing values, collinearity, high percentage of zeros etc.

The introduction of Generalized linear models relaxes the normality assumption imposed by the ordinary linear regression and allows the response to come from a large family of distributions called the linear exponential family which includes the normal distribution. The probability density function of random variable, Y, whose distribution is a member of the linear exponential family, is given by:

$$f_Y(y; \theta, \phi) = c(y; \phi) exp\Big(\frac{y\theta - a(\theta)}{\phi}\Big),$$

where θ is referred to as canonical parameter; ϕ, a dispersion parameter, $a(\cdot)$, a log-partition function and $c(\cdot)$, a base function.

Members of the linear exponential family such Bernoulli, Binomial, Poisson, gamma, normal, and inverse-gaussian distributions have desirable statistical properties. For instance, the mean and variance of the response variable can be specified, respectively, as;

$$\begin{aligned} E[Y] &= a'(\theta) = \mu \\ Var[Y] &= \phi a''(\theta). \end{aligned}$$

Generalized linear models allow for the choice of a link function, g, whose inverse transforms the linear predictor to get the target response variable. That is, an appropriate link function, not necessarily linear, can be chosen under GLM approach to expresses a relationship between the linear predictor and the mean of the response. Thus, we have,

$$g(E[Y_i]) = g(a'(\theta_i)) = \mathbf{X}_i'\boldsymbol{\beta}.$$

In other words:

$$E[Y_i] = g^{-1}(\mathbf{X}_i'\boldsymbol{\beta}),$$

where $\boldsymbol{\beta}$ represents regression parameters and $\mathbf{X}_i'$ represents a vector of explanatory variables for the ith observation. Therefore, the mean of the response vary with the characteristics of the individual observations. This implies that, unlike the classical linear regression, the variance of the response under GLM, is not constant but a function of the mean which may vary with regressors.

The estimation of parameters under GLM follow maximum likelihood procedures. The log-likelihood function is given by:

$$\begin{aligned} l(\mathbf{y};\boldsymbol{\beta},\phi) &= \sum_{i=1}^{n} ln\Big(f(y_i,\boldsymbol{\beta},\phi)\Big) \\ &= \sum_{i=1}^{n} ln\Big(c(y_i,\phi)\Big) + \Big(\frac{y_i\theta - a(\theta)}{\phi}\Big). \end{aligned}$$

When the first derivative of the log-likelihood, $l(\mathbf{y};\boldsymbol{\beta},\phi)$ is taken with respect to the parameters and set to zero, we arrive at the system of estimating equations given by:

$$\frac{\partial l(\mathbf{y};\boldsymbol{\beta},\phi)}{\partial \beta_t} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{g'(\mu_i)a''(\theta_i)} x_{i,t} = \mathbf{0}, \quad t = 0, 1, \ldots, p.$$

Usually, it is not convenient to solve the equations by hand. They are solved using numerical methods. The inverse of the Fisher information matrix gives the variance of the parameters as follows:

$$Var[\widehat{\boldsymbol{\beta}}] \;=\; \mathbf{I}(\widehat{\boldsymbol{\beta}})^{-1} = \sum_{i=1}^{n} \frac{1}{g(\mu_i)^2 Var[Y_i]^2} x_{i,t} x_{i,u},$$

where $t = 0, 1, \ldots, p$ and $u = 0, 1, \ldots, p$. It has been shown that, by the law of large numbers, the estimator for $\boldsymbol{\beta}$ asymptotically follows a normal distribution with mean zero and variance equal to the inverse of the fisher information matrix. That is:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \longrightarrow N(0, \mathbf{I}^{-1}(\boldsymbol{\beta})), \quad \text{as} \quad n \longrightarrow \infty.$$

GLM is used as a collective approach for loss reserving by making a distribution assumption on the incremental loss payments, $Y_{i,j}$. For instance, it is popular to assume that $Y_{i,j}$ follows a Poisson distribution which is a member of the linear exponential family. As seen in the previous classical models, the accident years, i, and development year, j, are used as predictor variables to explain the response variable, $Y_{i,j}$ under the hypothesis below:

Definition 2.4.0.1 (Generalized Linear Model for increments)**.**

(GLM1): *The random variable $Y_{i,j}$ for different accident years and/or development years are independent.*

(GLM2): *The probability density function of random variable* $Y_{i,j}$ *,is*

$$f(y;\theta_{i,j},\phi_{i,j}) = c(y;\phi_{i,j})exp\Big(\frac{y\theta_{i,j} - a(\theta_{i,j})}{\phi_{i,j}}\Big),$$

where θ *is referred to as canonical parameter;* ϕ*, a dispersion parameter,* $a(\cdot)$*, a log-partition function and* $c(\cdot)$*, a base function.*

We summarize the structure of the stochastic models for claim reserving in the framework of generalized linear models as follows:

- The incremental claim amounts $Y_{i,j}$ belongs to the exponential family.
- $E[Y_{i,j}] = \mu_{i,j}$ is the mean function.
- $\eta_{i,j} = g(\mu_{i,j})$, where $g(\cdot)$ is the link function.
- The linear predictor $\eta_{i,j}$, is a linear combination of factor effects i and j.

We consider a multiplicative structure with parameter for each row, i, and each column j as given below:

$$E[Y_{ij}] = \alpha_i \times \psi_j.$$

Here, the factor effect, α_i, denote expected exposure of the accident year while ψ_j denote expected cashflow of the development year. The multiplicative structure is linearized naturally by the choice of a logarithmic link function. With the logarithm link given as $g(\cdot) = ln(\cdot)$. Therefore,

$$\eta_{i,j} = ln(E[Y_{i,j}]) = ln(\alpha_i) + ln(\psi_j).$$

The model can be made identifiable by imposing additional constraints such as $\sum_j \psi_j = 1$ and $\alpha_1 = 1$. Then we have $2J - 1$ parameters to estimate given by:

$$\boldsymbol{\beta} = [ln(\alpha_2), \ldots, ln(\alpha_I), ln(\beta_0), \ldots, ln(\psi_J)]'.$$

A regression problem comes out when we define a the design matrix for the linear predictor.

$$\begin{aligned} Z_{1,j} &= [0 \ldots 0\,0\,0\,0 \ \ldots 0\ 1 \ \ldots]' \\ Z_{i,j} &= [0 \ldots 1\,0\,0\,0 \ \ldots 0\ 1 \ \ldots]', \end{aligned}$$

where the 1's correspond to positions where i and j are observed respectively.

Thus, we can write the mean in a convenient form as follows:

$$E[Y_{ij}] = e^{Z_{i,j}\boldsymbol{\beta}}.$$

The estimates for the parameters are obtained using maximum likelihood procedures. Then, the estimates of future increment payments is given by

$$\widehat{Y}_{i,j} = g^{-1}(\widehat{\beta}_0 + \widehat{\alpha}_i + \widehat{\psi}_j).$$

Example 2.4.1. *With reference to the increment table in Example 2.2.1, the GLM layout will be of the form in Table 2.7.*

Table 2.6 Layout of Incremental Loss triangle under GLM

	Development year			
Acc. year	1	2	3	4
1995	4,050	700	650	420
1996	3,900	950	600	$g^{-1}(\widehat{\beta}_0 + \widehat{\alpha}_2 + \widehat{\psi}_4)$
1997	2,800	1,500	$g^{-1}(\widehat{\beta}_0 + \widehat{\alpha}_3 + \widehat{\psi}_3)$	$g^{-1}(\widehat{\beta}_0 + \widehat{\alpha}_3 + \widehat{\psi}_4)$
1998	3,200	$g^{-1}(\widehat{\beta}_0 + \widehat{\alpha}_4 + \widehat{\psi}_2)$	$g^{-1}(\widehat{\beta}_0 + \widehat{\alpha}_4 + \widehat{\psi}_3)$	$g^{-1}(\widehat{\beta}_0 + \widehat{\alpha}_4 + \widehat{\psi}_4)$

We estimate the unknown parameters α_i and ψ_j from the triangle of known data with the maximum likelihood method. The estimate of reserve is the sum of

predicted claims for the unknown part of the triangle. In the following, we use the notation $\triangledown$ for the lower triangle of data, i.e., the set of all (i, j), where $Y_{i,j}$ is unknown but predicted using the GLM parameter estimates. Then,

$$\hat{R} = \sum_{i,j \in \triangledown} \widehat{Y}_{i,j}.$$

Example 2.4.2. *In this example we illustrate how to obtain the parameters of a GLM for the data given in Example 2.2.1. First, we assume that the increments follow a Poisson distribution which is a member of the linear exponential family. With a multiplicative structure for the row effects and column effects, a natural choice of link function the logarithmic function. Using R package, we implement the GLM estimation by specifying a data frame which consist of the increment payments and their corresponding accident years, i, and development years, j. The results of estimated parameters from the R package are given in Table 2.8.*

Table 2.7 Estimate of parameters of GLM Poisson using R package

Parameter	Estimate
Intercept: $\widehat{\beta}_0$	8.214769
$\widehat{\alpha}_2$	0.009217
$\widehat{\alpha}_3$	-0.105382
$\widehat{\alpha}_4$	-0.143863
$\widehat{\psi}_2$	-1.227503
$\widehat{\psi}_3$	-1.781636
$\widehat{\psi}_4$	-2.174514

The estimated parameter are used to predict the unknown incremental values following the structure in Table 2.7. The sum of the predicted values gives the esti-

mate of reserves.

$$
\begin{aligned}
\widehat{Y}_{2,4} &= exp(\widehat{\beta}_0 + \widehat{\alpha}_2 + \widehat{\psi}_4) = exp(8.21477 + 0.00922 + -2.17451) = 423.8889 \\
\widehat{Y}_{3,3} &= exp(\widehat{\beta}_0 + \widehat{\alpha}_3 + \widehat{\psi}_3) = exp(8.21477 + -0.10538 + -1.7816) = 559.8958 \\
\widehat{Y}_{3,4} &= exp(\widehat{\beta}_0 + \widehat{\alpha}_4 + \widehat{\psi}_4) = 377.9919 \\
\widehat{Y}_{4,2} &= exp(\widehat{\beta}_0 + \widehat{\alpha}_4 + \widehat{\psi}_2) = 937.6744 \\
\widehat{Y}_{4,3} &= exp(\widehat{\beta}_0 + \widehat{\alpha}_4 + \widehat{\psi}_3) = 538.7597 \\
\widehat{Y}_{4,4} &= exp(\widehat{\beta}_0 + \widehat{\alpha}_4 + \widehat{\psi}_4) = 363.7227 \\
\widehat{R}^{GLM} &= \widehat{Y}_{2,4} + \widehat{Y}_{3,3} + \widehat{Y}_{3,4} + \widehat{Y}_{4,2} + \widehat{Y}_{4,3} + \widehat{Y}_{4,4} = \mathbf{3{,}201.933}.
\end{aligned}
$$

It is interesting to find that the estimate of reserves for the GLM-Poisson model coincides with that of the chain-ladder and Mack's model. The structure of the chain-ladder estimator and that of the Poisson model give an intuitive explanation to why they lead to the same amount of reserve. For the chain-ladder method, the development factors which are also known as link ratios, λ_j, are unknown and estimated by the expression below:

$$\widehat{\lambda}_j = \frac{\sum_{i=1}^{J-j} C_{i,(j+1)}}{\sum_{i=1}^{J-j} C_{i,j}} \quad \text{for} \quad j = 1, \ldots, J-1.$$

Thus, the development factors are outcome of certain ratios. Similarly, when a multiplicative structure is adopted for the GLM-Poisson model, the incremental values are product of certain ratios. That is, considering a multiplicative structure with parameter for each row, i, and each column j as given below:

$$E[Y_{ij}] = \alpha_i \times \psi_j.$$

Here, the factor effect, α_i, denote expected exposure of the accident year. In other words, the expected ultimate claims amount (up to the latest development year so far observed). ψ_j denote expected cashflow of the development year (the

proportion of ultimate claims to emerge in each development year). According to (Wüthrich & Merz, 2008), the maximum likelihood estimators for α_i and ψ_j are given by:

$$\begin{aligned} \widehat{\alpha}_i &= \frac{\sum_{j=1}^{I-i} X_{i,j}}{\sum_{j=1}^{I-i} \psi_j} \\ \widehat{\psi}_j &= \frac{\sum_{i=1}^{I-j} X_{i,j}}{\sum_{i=1}^{I-j} \alpha_i} \end{aligned}$$

for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

Thus, the proportion factors ψ_j express the ratio of the sum of observed incremental values for certain development year j with respect to certain ultimate claims, i.e., ψ_j denotes the proportion of claims reported in development year j. The parameters α_i refer to the ratio of the sum of observed incremental values for a certain origin year i to corresponding proportion factors. In other words, if the incremental claim amounts and respective proportions factors are known, it is simple to derive the corresponding ultimate claim α_i for origin year i. One can note the principal similarities with the chain-ladder technique, where development factors are also the outcomes of certain ratios. Readers who are interested in the formal proof that the chain-ladder estimator and the estimator in the Poisson model lead to the same reserve should see (Wüthrich & Merz, 2008).

The use of GLM is important and more practical in general insurance because insurance data is rarely normal. Availability of literature and applicable software make the implementation of GLM very feasible and convenient. In addition to the the estimate of reserves, one can obtain measures of variability and confidence intervals.

2.4.1 Model selection and adequacy checks

In order to fit a GLM, one has to identifying a subset of predictors that are associated with response variable. Effective variable selection can also lead to parsimonious models with better prediction accuracy and easier interpretation. Some of the methods for selecting best subset of predictors are best subset and stepwise model selection procedures. To perform best subset selection, we fit all possible combinations of the predictors and look at the resulting models with the goal of identifying the one that is best. The problem with best subset selection is that it cannot be applied with large number of predictors. Therefore, stepwise methods which explore a far more restricted set of models are attractive alternatives to best subset selection. The three main types of stepwise methods are forward selection, backward selection and hybrid version. Forward selection begins with a model containing no predictors and then adds predictors to the model one-at-a time until all of the predictors are in the model. At each step the variable that gives the greatest additional improvement to the fit is added to the model. Backward stepwise selection begins with the full least squares model containing all predictors and then iteratively removes the least useful predictor. In the hybrid version, variables are added to the model sequentially, in analogy to forward selection. After adding each new variable, the method may remove any variable that no longer provide an improvement in the model fit.

One way of choosing an optimal model is to use a criterion to estimate out-of-sample prediction error and thereby relative quality of statistical models for a given set of data. The Akaike Information Criterion (AIC) and the Schwarz or Bayesian Information Criterion (BIC) are examples of such criterion based procedures.

$$\begin{aligned} AIC &= -2\mathcal{L} + 2p \\ BIC &= -2\mathcal{L} + p(log(n)) \end{aligned}$$

where $\mathcal{L}$ is the log-likelihood, p is the number of predictors in the model and n is the number of observation. With AIC and BIC criteria, the model with the smallest value of computed AIC or BIC is selected as the optimal model.

Deviance is another important idea associated with a fitted GLMs. It can be used to test the fit of the link function and linear predictor to the data, or to test the significance of a particular predictor variable (or variables) in the model. We define the deviance or likelihood ratio statistic, D, as

$$D = 2[\mathcal{L}_s(\hat{\theta}) - \mathcal{L}_m(\hat{\beta})]$$

where $\hat{\theta}$ and $\hat{\beta}$ are the MLEs of the saturated and proposed model, respectively. Also, $\mathcal{L}_s(\hat{\theta})$ and $\mathcal{L}_m(\hat{\beta})$ are the log-likelihoods corresponding to the saturated model and the proposed model respectively. The saturated model is an unconstrained model which has number of parameters equal to the number of observations. The proposed model is the model of interest with p number of parameters. Under some conditions,

$$D \sim \chi^2_{n-p}.$$

If the proposed model is poor, D, will be larger than value predicted by the χ^2_{n-p} distribution.

As an alternative to the approaches just discussed, we can directly estimate the test error using the validation set and cross-validation methods. We can compute the validation set error or the cross-validation error for each model under consideration, and then select the model for which the resulting estimated test error is smallest. Under the validation approach, the whole data set is divided into training set and test set. The model is trained using the training set and predictions are made on the test set. This procedure can be repeated several times by using cross-validation. The validation approach has an advantage relative to AIC and BIC in that, it provides a direct estimate of the test error, and makes

fewer assumptions about the true underlying model. It can also be used in a wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance. For the reasons mentioned above, we shall use more of validation method for model selection in this work. For an advanced discussion on the topic of model selection and adequacy tests, we refer readers to (James et al., 2013).

Some concerns on the use of GLM are as follows:

- The assumption concerning the distribution of the claims may not hold.
- The link or variance function may not be appropriate. This may affect the predictiveness of the model.
- GLM can suffer substantial precision losses if the right predictor variables are not used in the model.

CHAPTER III

INTRODUCTION OF STATISTICAL/MACHINE LEARNING TOOLS

In this chapter, we shall review some fundamental statistical/machine learning tools that have become very useful for loss reserving in general insurance. The review will be followed by a detailed discussion on neural networks.

3.1 Literature review

There has been a growing interest in the use of statistical/machine learning tools such as the decision trees and neural networks in actuarial literature due to increased amount of data and computing powers.

According to (Quan & Valdez, 2018), decision trees are useful for modelling insurance data because they do not require any probability distribution about the response and can handle missing data problems usually associated with real datasets. They are also useful for detecting non-linear effects and possible interactions among explanatory variables.

The very first regression tree algorithm known as the Automatic Interaction Detection (AID) was developed by (Morgan & Sonquist, 1963). More details about the historical development of decision trees can be found in (Loh, 2014). A popular algorithm known as Classification and Regression tress (CART) was introduced

by (Breiman et al., 1984). CART algorithm included additional features into the existing tree models such as pruning trees instead of using stopping rules, selecting trees by cross-validation, handling missing values by surrogate splits and obtaining linear splits by random search. (Breiman, 2001) and (Friedman, 2002) have shown that the use of random forests and gradient boosting further improves the predictive performance decision trees by minimizing over-fitting and bias. In addition to that, random forests have embedded algorithms that compute variable importance measures which are used for variable selection in machine learning.

There have been some interesting applications of decision trees in actuarial science. For instance, (Olbricht, 2012) and (Lopez et al., 2016) used tree based models for mortality studies. In the context of loss reserving, (Wüthrich, 2018a) modeled the number of payments using regression trees in a discrete time framework based on individual claims data. Furthermore, (Guelman, 2012), (Lee & Lin, 2018) and (Duval & Pigeon, 2019) implemented gradient boosting algorithms to estimate loss cost and insurance reserves.

Apart from decision trees, neural networks have also garnered increasing interest in recent years. The first step to neural networks was made by (McCulloch & Pitts, 1943). They modeled a simple neural network called the Threshold Logic Unit(TLU) using an electrical circuit. Over the years, the basic network has developed into complex multi-layer networks with applications of such networks cutting across many research areas. Readers who are interested in the history of neural networks can refer to (LeCun et al., 2015), (Goodfellow et al., 2016), and (Shmueli et al., 2017).

The concept and application of neural network is relatively new in insurance. However, there have been many successful applications in other fields such as finance. (Trippi & Turban, 1992) presented some interesting financial applications

of neural networks such as bankruptcy predictions, currency market trading, picking stocks, detecting fraud in credit card and monetary transactions and customer relationship management.

An early application of neural networks in loss reserving was given by (Mulquiney, 2006), who used artificial neural networks in insurance loss reserving. Also, (Wüthrich, 2018b) improved the traditional chain-ladder method to incorporate claims specific features using neural networks. (Gabrielli & Wüthrich, 2018) developed a stochastic simulation machine that generates individual claims histories of non-life insurance claims based on neural networks. This simulation machine provides a fully calibrated synthetic insurance portfolio of individual claims histories for back-testing preferred claims reserving methods. Furthermore, (Gabrielli et al., 2019) and (Kuo, 2019) used a deep learning approach to embed classical parametric models such as the over-dispersed Poisson model into neural networks.

3.2 Neural networks

3.2.1 Introduction

Neural networks (NN) are flexible data-driven machine learning tools that assume underlying functions that are much more complex than those used in traditional regressions. They are able to solve classification and regression problems by combining non-linear transformations of predictors.

The architecture and operations of neural networks draw inspiration from the biological activity of the human brain. The human brain learns from experience by processing huge amounts of information sent by human senses through interconnected neurons. In a same manner, neural networks connect nodes and layers into structures that make predictions from data. When large number of layers are

used train a neural network, the model is called Deep learning (DL).

The interest in neural network models for loss reserving in general insurance is currently growing. However, there have been successful application of neural networks in other field like finance and engineering. We give an overview of neural networks in this sections with references made to (Bishop, 2006) and (Shmueli et al., 2017).

3.2.2 Structure/architecture of a neural network

Conceptually, neural networks are structured to combine inputs in a flexible manner that captures complicated relationship among these variables as well as between them and the outcome variable.

There are various neural network structures that have been studied by researchers, but the most successfully applied neural network structure in data mining is called the **feed-forward neural network**. This network is a fully connected network with a one-way flow and no cycles (Shmueli et al., 2017). To make a prediction, one starts by entering information through an input layer and "go forward" through the network by sequentially applying transformations in one or more hidden layers, and finally getting a prediction through and output layer. The basic structure a feed-forward neural network is shown in Figure 3.1.

The key components of the feed-forward structure are summarised as follows:

- The **input layer**: this layer is made up of neurons or nodes that accept the predictor values. Usually, there is one node for each predictor variable.
- One or more **hidden layers**: output from the input layers are received as inputs by the neurons in the hidden layers. The number of hidden layers

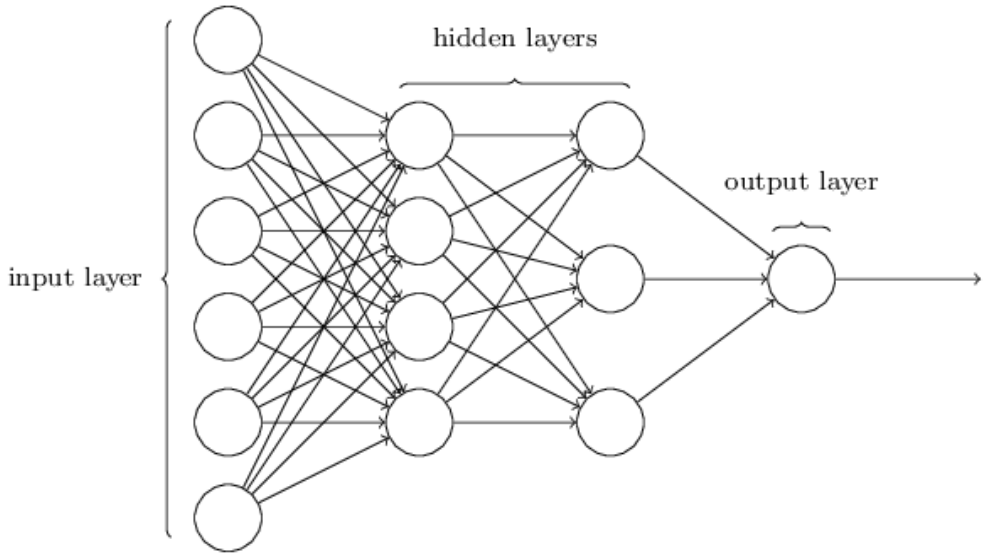

Figure 3.1 A four-layer feed-forward neural network with two hidden layers adapted from (Nielsen, 2015).

are usually depends on the structure selected by the user.

- The **output layer**: this layer returns the predictions after receiving input of data transformed by the input and hidden layers. The number of nodes depends on whether the problem is classification or regression.

From one layer to another, the data go through some transformation until the final prediction is out. The choice for the number of hidden layers is not subject to any rule. However, the most popular choice in literature and applications is one.

Fitting a neural network to data involves transformations of data in the neurons found at each layer of the network. These transformations are summarized as follows:

- Input unit transformation: a linear transformation is applied on predictors

by attaching weight and bias parameters computed by the neurons connecting from the input layer to the hidden layer of the network. A non-linear transformation is applied to generate output of the input layers.

- Hidden unit transformation: the transformed inputs pass through sequence of linear and non-linear transformations using weights and bias parameters computed by the neurons in the hidden units of the network. Hidden neurons generate output to serve the output neurons.

- Output prediction: the output neurons receive the input from the hidden unit and apply final transformation to generate predictions.

3.2.3 Formal structure of the feed-forward neural networks

The feed-forward neural network can be described more formally as follows:
First, we denote the $(p+1)$-dimensional predictors and n-dimensional responses as follows:

$$\mathbf{X} = \begin{pmatrix} \mathrm{X}^{(0)} \\ \mathrm{X}^{(1)} \\ \vdots \\ \mathrm{X}^{(p)} \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} \mathrm{Y}_1 \\ \mathrm{Y}_2 \\ \vdots \\ \mathrm{Y}_n \end{pmatrix}$$

where $\mathrm{X}^{(0)} = 1$ for all observations $i = 1, 2, , \ldots, n$. We shall omit the subscript, i, in the notations of predictors for simplicity. Considering a two-layer feed-forward neural network, we define two parameter matrices. One being an $M \times (p+1)$ matrix of first layer parameters, and the other being a $K \times (M+1)$ matrix of second layer parameters. Here, M and K represent the number of nodes in the hidden and output layers respectively.

$$\mathbf{w^{(1)}} = \begin{pmatrix} w_{1,0}^{(1)} & \cdots & w_{1,p}^{(1)} \\ w_{2,0}^{(1)} & \cdots & w_{2,p}^{(1)} \\ \vdots & \vdots & \vdots \\ w_{M,0}^{(1)} & \cdots & w_{M,p}^{(2)} \end{pmatrix}, \mathbf{w^{(2)}} = \begin{pmatrix} w_{1,0}^{(2)} & \cdots & w_{1,M}^{(2)} \\ w_{2,0}^{(2)} & \cdots & w_{2,M}^{(2)} \\ \vdots & \vdots & \vdots \\ w_{K,0}^{(2)} & \cdots & w_{K,M}^{(2)} \end{pmatrix}.$$

The first layer parameters are used to perform a first linear transformation on the predictor variables. This gives an M-dimensional vector given by:

$$\mathbf{a^{(1)}} = \mathbf{w^{(1)}}\mathbf{X}.$$

More precisely,

$$\mathbf{a^{(1)}} = \begin{pmatrix} a_1^{(1)} \\ a_2^{(1)} \\ \vdots \\ a_M^{(1)} \end{pmatrix} = \begin{pmatrix} \sum_{j=0}^{p} w_{1,j}^{(1)} X^{(j)} \\ \sum_{j=0}^{p} w_{2,j}^{(1)} X^{(j)} \\ \vdots \\ \sum_{j=0}^{p} w_{M,j}^{(1)} X^{(j)} \end{pmatrix}.$$

The vector, $\mathbf{a^{(1)}}$, is known as input unit activation. Then a suitable non-linear function, $h : \mathbb{R} \to (0, 1)$, known as an activation function or a transfer function, is applied to all transformed predictors:

$$Z_m^* = h\left(a_m^{(1)}\right), \quad m = 1, \ldots, M.$$

Thus, we have a vector given below:

$$\mathbf{Z}^* = \begin{pmatrix} Z_1^* \\ Z_2^* \\ \vdots \\ Z_M^* \end{pmatrix}.$$

The vector, $\mathbf{Z}^*$, is augmented by including a constant predictor $\mathbf{Z}_0 = 1$ to obtain another vector in the hidden units of the neural network. The new vector, $\mathbf{Z}$, is defined below:

$$\mathbf{Z} = \begin{pmatrix} 1 \\ \mathbf{Z}^* \end{pmatrix} = \begin{pmatrix} \mathrm{Z}_0 \\ \mathrm{Z}_1^* \\ \vdots \\ \mathrm{Z}_M^* \end{pmatrix}.$$

Then a second linear transform $\mathbf{a^{(2)}}$ is applied to the vector in the hidden units of the network as follows:

$$\mathbf{a^{(2)}} = \mathbf{w^{(2)}}\mathbf{Z}.$$

More precisely,

$$\mathbf{a^{(2)}} = \begin{pmatrix} \mathrm{a}_1^{(2)} \\ \mathrm{a}_2^{(2)} \\ \vdots \\ \mathrm{a}_K^{(2)} \end{pmatrix} = \begin{pmatrix} \sum_{m=0}^{M} \mathrm{w}_{1,m}^{(2)} Z_m \\ \sum_{m=0}^{M} \mathrm{w}_{2,m}^{(2)} Z_m \\ \vdots \\ \sum_{m=0}^{M} \mathrm{w}_{K,m}^{(2)} Z_m \end{pmatrix}.$$

This K-dimensional vector, $\mathbf{a^{(2)}}$, is known as an output unit activations. The prediction for the responses is obtained as expressed below after a final activation function is applied to the output unit activations, $\mathbf{a^{(2)}}$. That is,

$$\widehat{f}(\mathbf{X}) = g\left(\mathbf{a^{(2)}}\right),$$

where $g(\cdot)$ is the final activation function. The neural network prediction can be written more compactly as:

$$\widehat{f}(\mathbf{X}) = g\left(\mathrm{w}_{k,0}^{(2)} + \sum_{m=1}^{M} \mathrm{w}_{k,m}^{(2)} h\left(\mathrm{w}_{m,0}^{(1)} + \sum_{j=1}^{p} \mathrm{w}_{m,j}^{(1)} X^{(j)}\right)\right).$$

The parameters, $\mathrm{w}_{m,0}^{(1)}$ and $\mathrm{w}_{k,0}^{(1)}$, that are associated with the constant predictor are referred to as **biases**. While the other parameters, $\mathrm{w}_{m,j}^{(1)}$ for $j \neq 0$ and $\mathrm{w}_{k,m}^{(2)}$ for $m \neq 0$, are referred to as **weights**.

The two-layer structure could be generalized to incorporate additional layers. For instance, a third matrix of parameters, $\mathbf{w}^{(3)}$, can be defined to generate another

hidden layer activation. Thus, a hidden layer activation function , $h(\cdot)$, is applied on $\mathbf{a}^{(2)}$ instead of the output layer activation function, $g(\cdot)$. Then, by following procedures described previously we have,

$$\begin{aligned} \mathbf{Z}^{*(2)} &= h\left(\mathbf{a}^{(\mathbf{2})}\right) \\ \mathbf{Z}^{(2)} &= \begin{pmatrix} 1 \\ \mathbf{Z}^{*(2)} \end{pmatrix} \\ \mathbf{a}^{(3)} &= \mathbf{w}^{(3)}\mathbf{Z}^{(2)}. \end{aligned}$$

The prediction for the three layer neural network could then be obtained by applying the final activation function on $\mathbf{a}^{(3)}$ as follows:

$$\widehat{f}(\mathbf{X}) = g\left(\mathbf{a}^{(\mathbf{3})}\right).$$

An arbitrary number of additional layers could be included. For example, a feed-forward neural network with L layers would involve the following sequence of transformations:

$$\mathbf{X} \xrightarrow{\mathbf{w}^{(1)}} \mathbf{a}^{(1)} \xrightarrow{h} \mathbf{Z}^{(1)} \xrightarrow{\mathbf{w}^{(2)}} \mathbf{a}^{(2)} \xrightarrow{h} \mathbf{Z}^{(2)} \ldots \xrightarrow{\mathbf{w}^{(L)}} \mathbf{a}^{(L)} \xrightarrow{g} \widehat{f}(\mathbf{X}).$$

3.2.4 Activation functions

The three most popular activation functions used in neural networks are the sigmoid/logistic , hyperbolic tangent and identity functions.

- The sigmoid or logistic function has an S-shape. The input to the function is transformed into a value between 0 and 1. It is generally suitable for models where probabilities are predicted as output. The function is given by:

$$h(x) = \frac{1}{1 + e^{-x}}, \quad \text{where} \quad x \in \mathbb{R}.$$

- The hyperbolic tangent function also has an S-shape. However, the input to the function is transformed into a value between -1 and 1. The function ensures that negative inputs are mapped strongly negative and zero inputs are be mapped near zero. It generally suitable for classification between two classes. The function is given by:

$$h(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \quad \text{where} \quad x \in \mathbb{R}.$$

- The identity function has a linear shape. It has range from $-\infty$ to ∞. The function is suitable for regression problems since it does not confine output to any range. The function is given by:

$$h(x) = x, \quad \text{where} \quad x \in \mathbb{R}.$$

The graphical display of these three activation functions are shown in Figure 3.2. In each plot, the domain of the activation function, $h(x)$, is the set of real numbers from -4 to 4. The plot in the right panel indicates that the logistic activation function transforms the domain into a range of real numbers from 0 to 1 with an S-shape. The plot in the middle panel also indicates that the hyperbolic tangent activation function transforms the domain into a range of real numbers from -1 to 1 with an S-shape. Finally, the plot in the left panel indicates that the identity activation function is linear in shape and transforms the domain into a range of real numbers from $-\infty$ to ∞.

3.2.5 Estimation of parameters

Neural network parameters are made up of weights and biases (i.e $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(L)}$). These parameters are usually estimated by gradient descent and back propagation methods. We denote the set of all the neural network parameters by

$\boldsymbol{\beta} = (\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(L)})^T$. The parameters are the solution to an optimization problem given by:

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, \Omega\Big(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(L)}\Big)$$

where

$$\Omega\Big(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(L)}\Big) = \sum_{i=1}^{n} ||Y_i - \widehat{f}(\mathbf{X}_i, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(L)})||^2.$$

For the ith observation, Y_i and $\mathbf{X}_i$, are the response variable and vector of explanatory variables, respectively.

The gradient descent method produces a sequence of parameters for which the objective function, Ω, is decreasing. The sequence of parameter vectors is determined by a gradient descent function giving by:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta \bigtriangledown \Omega\Big(\boldsymbol{\beta}^{(t)}\Big)$$

where

- η is the constant learning rate which characterizes the speed at which parameters move to reach the optimum. The value is decided before running the algorithm.
- $\bigtriangledown\Omega\Big(\beta^{(t)}\Big)$ is the gradient of Ω with respect to $\boldsymbol{\beta}$ evaluated at $\boldsymbol{\beta}^{(t)}$. That is,

$$\begin{aligned}
\bigtriangledown\Omega\Big(\boldsymbol{\beta}^{(t)}\Big) &= \frac{\partial\Omega}{\partial\boldsymbol{\beta}}\Big(\boldsymbol{\beta}^{(t)}\Big) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial\boldsymbol{\beta}} ||Y_i - \widehat{f}(\mathbf{X}_i, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(L)})||^2 |_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}} \\
&= \sum_{i=1}^{n} \bigtriangledown\Omega^i\Big(\boldsymbol{\beta}^{(t)}\Big)
\end{aligned}$$

where

$$\bigtriangledown\Omega^i\Big(\boldsymbol{\beta}^{(t)}\Big) = \frac{\partial}{\partial\boldsymbol{\beta}} ||Y_i - \widehat{f}(\mathbf{X}_i, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(L)})||^2 |_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}.$$

In the gradient descent function, updating the current value of parameters by $-\eta \bigtriangledown \Omega\left(\boldsymbol{\beta}^{(t)}\right)$ entails moving the parameter values in the direction where the objective function decreases the most (steepest slope). Eventually, as the sequence $\{\boldsymbol{\beta}^{(t)}\}$ approaches a critical point, the gradient will converge to zero and the sequence will converge.

Gradient descent method is the most common and established way of optimizing neural network loss functions. However, it suffers these two problems:

- The sequence might not necessarily converge to a global minimum as a local minimum or a saddle point can be reached. This is very typical of neural network objective functions which are notoriously non-convex due to the non-linear transforms embedded through the activation functions.
- The method leads to slower convergence since it updates parameters by going through all observations.

The inability of the gradient descent to escape local minima can be problematic for neural networks. An approach called stochastic gradient descent was developed to allow the optimization to escape local minima. This approach involves using a single observation to update parameters on each iteration:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta \bigtriangledown \Omega^{k}\left(\boldsymbol{\beta}^{(t)}\right)$$

where $k = t \mod n$. An extension of the stochastic gradient approach called "batch" gradient involves using a subset of observations instead of a single one to update parameter estimates:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta \bigtriangledown \sum_{k \in K^{(t)}} \Omega^{k}\left(\boldsymbol{\beta}^{(t)}\right)$$

where $K^{(t)} \subseteq 1, \ldots, n$.

The backpropagation method offers a convenient recursive formula to compute the gradient in neural network. We note that the predictions from neural network leads to the following equations:

$$\begin{aligned}
\Omega^{(i)} &= \left(Y_i - \widehat{f}(\mathbf{X}_i, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(L)})\right)^2 \\
\widehat{f}(\mathbf{X}_i, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(L)}) &= g\left(\mathbf{a}^{(L)}\right) \\
\mathbf{a}^{(k)} &= \mathbf{w}^{(k)}\left[1, h\left(\mathbf{a}^{(k-1)}\right)\right]^T \\
\mathbf{a}^{(1)} &= \mathbf{w}^{(1)}\mathbf{X}.
\end{aligned}$$

$\Omega^{(i)}$ denotes the objective function for the ith observation which is obtained from the quadratic loss function.

Let $J \in 1, \ldots, L$ represent the layers in the neural network. Then, in a layer J, the derivative of $\Omega^{(i)}$ with respect to parameters in the network is given by:

$$\frac{\partial \Omega^{(i)}}{\partial \mathbf{w}^J_{j,k}} = -2\left(Y_i - \widehat{f}(\mathbf{X}_i, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(L)})\right) g'\left(\mathbf{a}^{(L)}\right) \frac{\partial \mathbf{a}^{(L)}}{\partial \mathbf{w}^J_{j,k}}.$$

The outline of how the backpropagation computes the partial derivative, $\frac{\partial \mathbf{a}^{(L)}}{\partial \mathbf{w}^J_{j,k}}$, recursively is given in the Appendix C1.

It must be noted that in this research, the methodology to predict future cashflows for loss reserving is a regression problem so the gradient descent method is used to minimize a quadratic loss function as follows:

$$L(\mathbf{Y}, \widehat{f}(\mathbf{X})) = (Y - \widehat{f}(\mathbf{X})^2.$$

For classification problems, the quadratic loss function is not used. A common loss function used with classification problems is the cross-enthropy loss function and is used with logistic activation functions. The cross-entropy loss function for a classification problem with m classes is defined as follows:

$$L(p, \widehat{p}) = -\sum_{i=1}^{m} p_i(x) log \widehat{p_i}(x)$$

where $p_i(x)$ is the ground truth label of the observation and $\widehat{p}_i(x)$ is the classifiers prediction score. Cross-entropy loss, or log loss, measures the performance of a classification model whose output is a probability value between 0 and 1. Cross-entropy loss increases as the predicted probability diverges from the actual label showing higher level of uncertainty inherent in the variables possible outcome. For further details on the gradient descent, backpropagation methods and how they work with classification problems, readers can refer to (Bishop, 2006) and (Nielsen, 2015).

3.2.6 R packages for neural networks

The most common R packages for neural networks are **nnet** and **neuralnet**. The **nnet** implements feed-forward neural networks with a single hidden layer and has no plotting option. In this study, we used the **neuralnet** package which enables us to implement multi-layer networks and obtain neural network plots. It trains the neural networks using backpropagation method and allows flexible settings through custom choice of error function, stopping criteria, activation function and learning rate.

3.2.7 An illustrative example

Example 3.2.1. *We shall use the fictitious data given in example 2.2.1 to give an illustration of a neural network. Table 3.1 is a recall of the incremental loss triangle obtained from from example 2.2.1.*

We shall consider predicting the payment amount in the lower part of the incremental triangle as a regression problem where the accident and development years are the covariates, denoted by i and j, respectively, while the response variable is the amount paid, also denoted by $Y_{i,j}$.

Table 3.1 Incremental Loss triangle

	Development year			
Acc. year	1	2	3	4
1995	4,050	700	650	420
1996	3,900	950	600	
1997	2,800	1,500		
1998	3,200			

For this collective loss reserving problem, we shall use a simple two-layer neural network with three nodes in the hidden layer to predict the unknown amount of payments. The training set obtained from the loss triangle is shown in Table 3.2.

Table 3.2 Training set of the fictitious data

i	1	1	1	1	2	2	2	3	3	4
j	1	2	3	4	1	2	3	1	2	1
$Y_{i,j}$	4050	700	650	420	3900	950	600	2800	1500	3200

Using the R package ***neuralnet****, the estimated parameters for the first layer is a* 3×3 *matrix,* $\widehat{w}^{(1)}$ *while the estimated parameters in the second layer is a* 4*-dimensional vector,* $\widehat{w}^{(2)}$*. The first row of* $\widehat{w}^{(1)}$ *are the estimate of bias parameter for the three nodes in the hidden layer. The second row of* $\widehat{w}^{(1)}$ *are the estimated weights assigned to the inputs of the variable,* i*, while the third row of* $\widehat{w}^{(1)}$ *are the estimated weights assigned to the inputs of variable* j*. The first element of the vector,* $\widehat{w}^{(2)}$*, is the estimated bias for the output layer while the other elements are the estimated weights assigned to the results of the three nodes hidden layer. Figure 3.2 is a plot that shows the estimated weight and bias parameters of the neural network. It is worth mentioning that depending on the network architecture*

and dimension of data, a neural network plot can be complex. However, the plot is meant to give a visual impression of the nature of the network and summary of activities that occurred in its layers.

$$\widehat{w}^{(1)} = \begin{pmatrix} 0.3147 & -0.8356 & 0.1078 \\ 0.4157 & -0.3874 & -0.6106 \\ 5.1015 & -2.4955 & -6.6790 \end{pmatrix}$$

$$\widehat{w}^{(2)} = \begin{pmatrix} 0.9004 \\ -0.8604 \\ -0.5418 \\ 1.3884 \end{pmatrix}.$$

After training the neural network, predictions are made on the test set which represent the values of covariates in the lower part of the loss triangle. The test set and the predicted amount of payments are given in Table 3.3.

Table 3.3 Estimated payments of the test set

i	2	3	3	4	4	4
j	4	3	4	2	3	4
$\widehat{Y}_{i,j}$	459.4258	517.4275	462.3031	1,467.5526	504.6061	464.7916

For collective loss reserving methods, estimated reserve is obtained as follows:

$$\begin{aligned} \widehat{R}^{NN} &= \widehat{Y}_{2,4} + \widehat{Y}_{3,3} + \widehat{Y}_{3,4} + \widehat{Y}_{4,2} + \widehat{Y}_{4,3} + \widehat{Y}_{4,4} \\ &= \mathbf{3{,}876.107.} \end{aligned}$$

Compared to the amount of 3,201.993 *predicted by the Mack and GLM-Poisson model, we observe that the neural networks predicted a higher loss reserve. However, we have demonstrated in this example that neural networks as a machine learning tool can be useful for loss reserving.*

3.2.8 Strengths and weaknesses

The most prominent advantage of neural networks is their good predictive performance. They are known to have high tolerance to noisy data and the ability to capture highly complicated relationships between the predictors and an outcome variable. Neural networks relax many distributional assumptions made on data. Some drawbacks in using neural networks are as follows:

- Neural networks do not have a built-in variable selection mechanism. This means that there is a need to identify key predictors using classification and regression trees or other dimension reduction techniques such as principal components analysis.
- The computational time for neural networks grows greatly when the number of predictors is increased (as there will be many more weights to compute). Therefore, the run-time should always be measured to make sure that it does not cause unacceptable delay in the decision-making.
- Due to several parameters involved in the model, neural networks may suffer from over-fitting if not regulated properly. This requires the use of regularisation methods such as controlling the learning rate and momentum to ensure effective training of the model.
- A technical problem is the risk of obtaining weights that lead to a local optimum rather than the global optimum. In that case, the weights may converge to values that do not provide the best fit to the data.

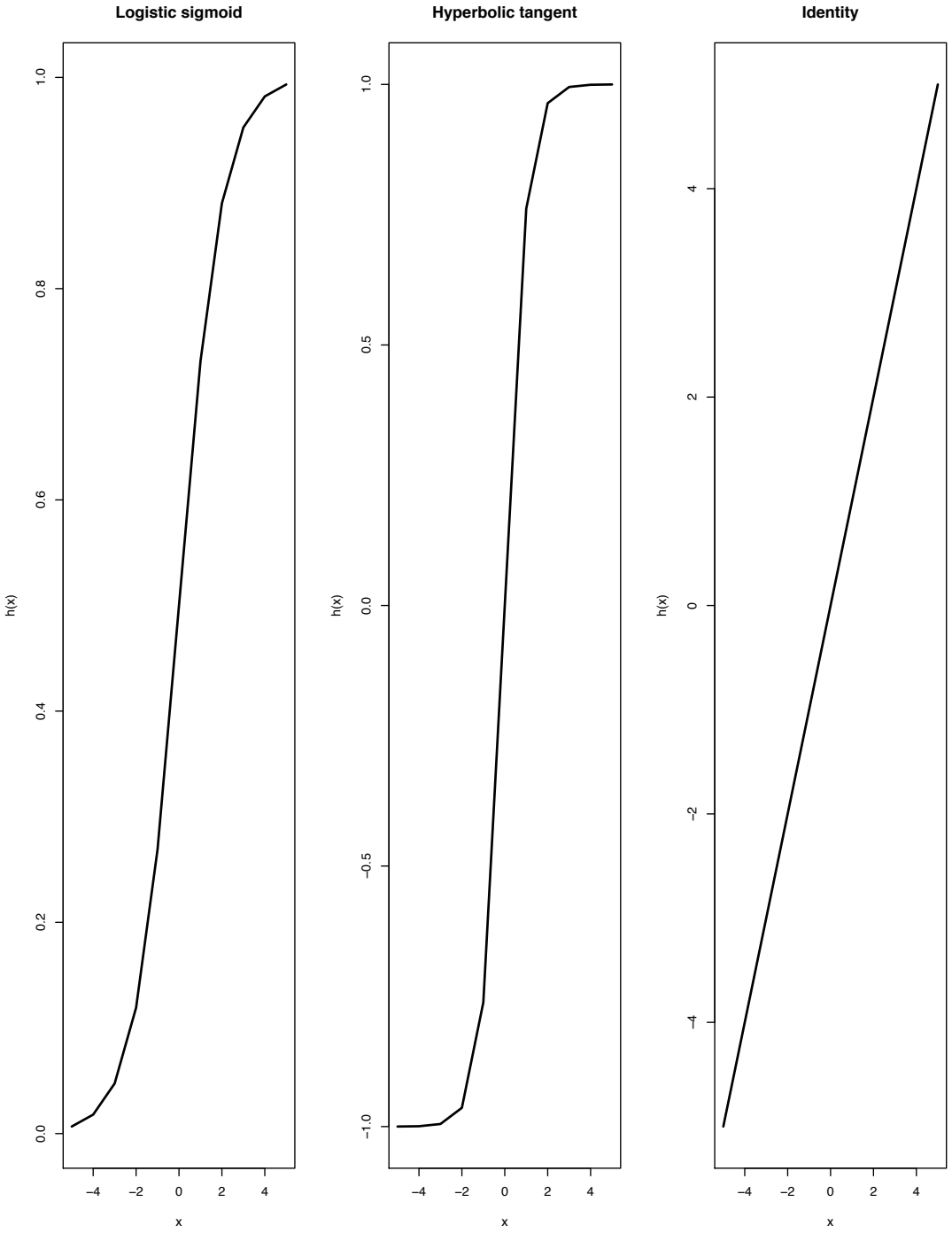

Figure 3.2 Plot of three most common activation functions used in neural networks.

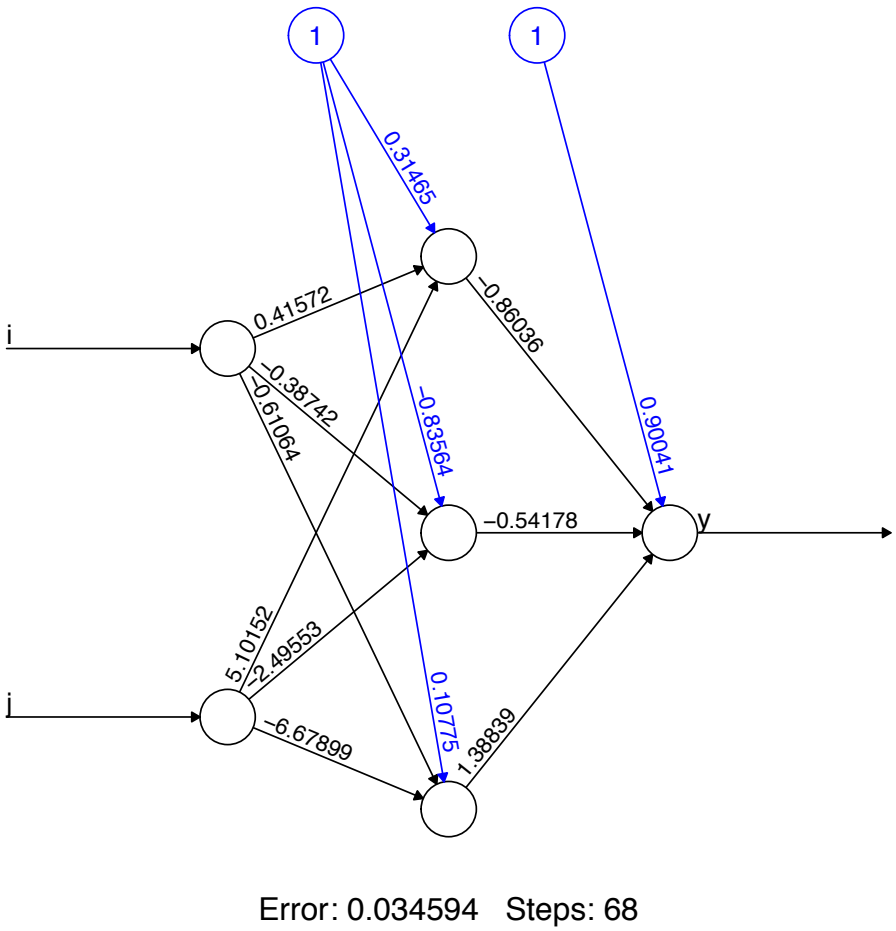

Figure 3.3 Estimated parameters of a two-layer feed-forward neural network.

CHAPTER IV

DESCRIPTION OF THE NEURAL NETWORK MODEL FOR LOSS RESERVING

In this study, we have proposed neural networks model for individual loss reserving in general insurance. The purpose of this chapter is to give readers the background and description of our model.

4.1 Background

In recent times, the development in loss reserving which has drawn the attention of many researchers is individual or micro-level loss reserving. The development hinges on some limitations of macro-level models. The observed data in a run-off triangle is typically small leading to large prediction error (England & Verrall, 2002). Also, macro-level models undermines heterogeneity in claims by their inability to incorporate any information about the individual claims. The pattern of claims do not remain the same for the future as it is assumed in some macro-level models such as the chain-ladder (Friedland, 2010). Associated with this issue, (England & Verrall, 2002) questioned the continue use of aggregate data when the underlying extensive micro-level information is available and the computation is feasible. Recently, (Charpentier & Pigeon, 2016) pointed out that the results of aggregate models will not be enough when modern solvency laws call for full

conditional distribution of future cash flows.

The new stream of literature on micro-level model for loss reserving started coming out when (Arjas, 1989), (Norberg, 1993) and (Norberg, 1999) developed a mathematical framework in continuous time for individual loss reserving. Based on this theoretical framework, several researchers contributed to the implementation of individual level loss reserving models and used case studies for illustrations. (Taylor et al., 2008) modeled individual claims with GLMs by incorporating various types of covariates. (Zhao & Zhou, 2010) applied semi-parametric and copula models to individual loss reserving methods. (Pigeon et al., 2013) proposed an individual loss model in the framework of multivariate skew normal distribution. (Antonio & Plat, 2014) laid a foundation for micro-level stochastic loss reserving which has gained considerable attention. (Charpentier & Pigeon, 2016) discussed micro and macro methods in property and casualty claims reserving using linear model, Poisson and quasi-Poisson models. Literature on micro-level loss reserving models is still growing and it has become the direction of many current research works in loss reserving for general insurance.

4.2 The individual framework for loss reserving

Contrary to the collective approach of loss reserving, the individual approach of loss reserving make use of the series of payments contained in each cell together with available information on claims and policyholders to model payments and predict loss reserves. This approach helps to estimate both RBNS and IBNR reserves. The total reserves is the sum of the two estimates.

Let $Y_{i,j}^{(k)}$ be the amount of payment for claim k related to the accident and development years i and j respectively. For claims that are still opened at the valuation date, the first part of total reserves known as RBNS reserves is estimated as

follows:

$$\widehat{R}^{RBNS} = \sum_{i=2}^{I} \sum_{j=I+2-i}^{I} \sum_{k \in \kappa_i} \widehat{Y}_{i,j}^{(k)}, \quad i, j = 1 \dots I, \tag{4.1}$$

where κ_i is the set of all observed that are still not settled at the valuation date. It is important to note that we assume $I = J$ but it is not essential for the model. The second part of total reserve is made up of prediction for IBNR claims. If we denote κ_i^{unobs} as the set of all incurred but not reported claims in accident year i, the estimate of IBNR reserve in the individual framework is given as follows:

$$\widehat{R}^{IBNR} = \sum_{i=2}^{I} \sum_{j=I+2-i}^{I} \sum_{k \in \kappa_i^{unobs}} \widehat{Y}_{i,j}^{(k)}. \tag{4.2}$$

In equation 4.1 and 4.2, we assume that all claims are paid for the earliest occurrence period ($i = 1$). In this work, we adopt this individual framework for loss reserving and mainly focus on the estimation of RBNS reserve. Obtaining the IBNR reserve will require a model for the occurrence of a claim and the delay of its declaration to the insurer in addition to more assumptions about the composition of the portfolio. We shall reserve that for future studies. It will be seen in the next chapter that, unlike the collective models, all relevant predictors will be incorporated in our model to estimate future claim payments, $\widehat{Y}_{i,j}^{(k)}$, without restricting the predictors to only accident and development years.

4.3 Frequency-severity model

In actuarial literature, a popular approach to analyse joint distribution of frequency and severity models is to decompose the joint distribution into a product of marginal and conditional distributions. That is

$$f_Y(y) = f_{N,X}(n, x) = f_N(n) \times f_{X|N}(x|n),$$

where $f_N(n)$ is the frequency model specifying the probability distribution of having $N = n$ claims, and $f_{X|N}(x|n)$ is the conditional severity distribution of having

payment amount X given n claims. Details of the frequency-severity model can be found in (Panjer & Willmot, 2008). The decomposition allows each components to be studied separately so that one can investigate factors that affect each component.

In this work, we follow the frequency-severity approach in order to model payment of claims. First, we shall introduce the general structure of the frequency-severity model. Let Y be a random variable representing the total amount of payment, N be a random variable representing the frequency of payment and X_i, for $i = 1, \ldots, N$, be a random variable representing the number of individual payments. Here, we assume N and Y are independent. Also the payments, X_i's, are independent. The general structure of the frequency-severity model is written as follows:

$$Y = \begin{cases} 0, & N = 0 \\ \sum_{i=1}^{N} X_i, & N > 0. \end{cases}$$

.

Now, in the individual framework, the structure of the frequency-severity model is given by:

$$Y_{i,j}^{(k)} = \begin{cases} 0, & N_{i,j}^{(k)} = 0 \\ \sum_{l=1}^{N_{i,j}^{(k)}} X_{i,j,l}^{(k)}, & N_{i,j}^{(k)} > 0, \end{cases}$$

where $X_{i,j,l}^{(k)}$ represent payment l for claim k and $N_{i,j}^{(k)}$ represent the number of payment of the claim k in cell (i, j) of the development triangle. Thus, for each claim, the frequency-severity model calls for a frequency model to predict the number of future payments and a severity model to predict the amount of payments. This structure of the frequency-severity model has been presented and developed in several ways by many researchers including (Norberg, 1993) and (Norberg, 1999) which are very notable for loss reserving. It is worth mentioning that the choice of distributions for the frequency and severity models differ from one researcher

to another.

4.3.1 Frequency model

The frequency model for claims is a very important factor to consider as far as the total cost of insurance is concerned. Various models have been used by researchers to analyse the frequency of claims. For instance, (Denuit & Trufin, 2017) used Poisson distribution to model claim frequency while (Wüthrich, 2003) and (Boucher & Davidov, 2011) used Tweedie distribution. The negative binomial distribution is another popular distribution used by (Schiegl, 2002). Mixed frequency distributions and random effect models have also been used in some actuarial literature. However, in this work, we shall perform statistical analyses on Poisson, negative binomial, geometric and zero-inflated regression models to select the best frequency model. We shall give a brief theoretical background of these frequency models below. We will note here that the domain of all the frequency models discussed in this section is non-negative integers. That is, $N_{i,j}^{(k)} \geq 0$ for all the models.

The Poisson regression model is a popular count model due to the following reasons:

- The Poisson model is a member of linear exponential family. Therefore, it has desirable properties to support generalized linear regression.
- Poisson regression can be implemented to incorporate micro-level covariates.
- The model is suitable for actuarial data and can be compared easily to other models.

When the frequency of payments of claims, $N_{i,j}^{(k)}$, follows a Poisson regression

model, we have the following:

$$\begin{aligned} N_{i,j}^{(k)} &\sim P(\lambda_{i,j}^{(k)}) \\ E[N_{i,j}^{(k)}] &= \lambda_{i,j}^{(k)} = \exp(\mathbf{x}_{i,j}^{\prime(k)}\boldsymbol{\beta}) \\ Var[N_{i,j}^{(k)}] &= \lambda_{i,j}^{(k)}, \end{aligned}$$

where $\mathbf{x}_{i,j}$ is a p-dimensional vector of explanatory variables and $\boldsymbol{\beta}$ is a p-dimensional vector of regression parameter. The maximum likelihood estimator of $\boldsymbol{\beta}$ is the solution of the equation:

$$\sum_{i,j} \left(n_{i,j}^{(k)} - \exp(\mathbf{x}_{i,j}^{\prime(k)}\boldsymbol{\beta})\right)\mathbf{x}_{i,j}^{(k)} = \mathbf{0}.$$

The assumptions of the Poisson model imply that variance of the frequency variable is the same as the mean. This is sometimes seen as too restrictive. The use of quasi-likelihood approach relax the variance assumption to cater for real situations where data show variations greater than expected values. Quasi models are characterized by their first two moments (see (McCullagh & Nelder, 1989) for further details).

Negative binomial regression is another popular model used to model over-dispersed count data. When the conditional variance of the count data exceeds the conditional mean, the count data is said to be over-dispersed. The Negative binomial regression is considered as a generalization of Poisson regression since it has the a mean structure which is similar to the Poisson regression. However, the negative binomial model has an extra parameter to model over-dispersion of the data. When the frequency of payments of claims, $N_{i,j}^{(k)}$, follows a Negative binomial regression, we define a linear predictor as follows:

$$\mu_{i,j}^{(k)} = ln(\lambda_{i,j}^{(k)}) = \mathbf{x}_{i,j}^{\prime(k)}\boldsymbol{\beta}$$

where $\mathbf{x}_{i,j}$ is a p-dimensional vector of explanatory variables and $\boldsymbol{\beta}$ is a p-dimensional vector of regression parameter. Using α as a dispersion parameter, the negative

binomial regression model can be written as follows:

$$Pr(N_{i,j}^{(k)} = n_{i,j}^{(k)} | \mu_{i,j}^{(k)}, \alpha) = \frac{\Gamma\left(n_{i,j}^{(k)} + \alpha^{-1}\right)}{\Gamma\left(n_{i,j}^{(k)} + 1\right)\Gamma\left(\alpha^{-1}\right)} \left(\frac{1}{1 + \alpha\mu_{i,j}^{(k)}}\right)^{\alpha^{-1}} \left(\frac{\alpha\mu_{i,j}^{(k)}}{1 + \alpha\mu_{i,j}^{(k)}}\right)^{\mu_{i,j}^{(k)}}. \quad (4.3)$$

We would like to point it out to readers that the negative binomial model used in this research refers directly to negative binomial 2 (NB2). It should not be confused with negative binomial 1 (NB1) which is a different parameterization of negation binomial.The NB1 is usually less flexible in capturing the variance since it has a non-zero covariance between the estimators of the parameters of the model. Interested readers are referred to (Cameron & Trivedi, 2013) for additional information about this parameterization. In order to avoid the complexity associated with NB1 in modelling the variance, we chose NB2 but kept the name negative binomial (NB) for simplicity.

The regression coefficients are estimated using the method of maximum likelihood. The logarithm of the likelihood function is given by:

$$\begin{aligned} \mathcal{L} &= \sum_{i,j} \left\{ \left(\sum_{w=0}^{n_{i,j}^{(k)}-1} ln(w + \alpha^{-1}) \right) - ln\left(\Gamma(n_{i,j}^{(k)} + 1\right) - (n_{i,j}^{(k)} + \alpha^{-1}) ln(1 + \alpha\mu_{i,j}^{(k)}) \right. \\ &+ \left. n_{i,j}^{(k)} ln(\mu_{i,j}^{(k)}) + n_{i,j}^{(k)} ln(\alpha) \right\}. \end{aligned}$$

To estimate the parameters of the model, first partial derivatives of the likelihood with respect to $\boldsymbol{\beta}$ and α are found. Equating the first partial derivatives to zero gives the following set of likelihood equations for estimating the $\boldsymbol{\beta}$ and α, respectively:

$$\sum_{i,j} \frac{x_{ij}^{(k)}(n_{i,j}^{(k)} - \mu_{i,j}^{(k)})}{1 + \alpha\mu_{i,j}^{(k)}} = 0,$$

$$\sum_{i,j} \left\{ \alpha^{-2}\left(ln(1 + \alpha\mu_{i,j}^{(k)}) - \sum_{w=0}^{n_{i,j}^{(k)}-1} \frac{1}{w + \alpha^{-1}} \right) + \frac{n_{i,j}^{(k)} - \mu_{i,j}^{(k)}}{\alpha(1 + \alpha\mu_{i,j}^{(k)})} \right\} = 0.$$

According to (Cameron & Trivedi, 2013), the asymptotic distribution of the maximum likelihood estimates as multivariate normal as follows:

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\alpha} \end{bmatrix} \sim N \left[\begin{array}{c|cc} \boldsymbol{\beta} & \text{Var}(\widehat{\boldsymbol{\beta}}) & \text{Cov}(\widehat{\boldsymbol{\beta}}, \widehat{\alpha}) \\ \alpha & \text{Cov}(\widehat{\boldsymbol{\beta}}, \widehat{\alpha}) & \text{Var}(\widehat{\alpha}) \end{array} \right]$$

where

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = \left[\sum_{i,j} \frac{\mu_{i,j}^{(k)}}{1 + \alpha \mu_{i,j}^{(k)}} \mathbf{x}_{i,j}^{(k)} \mathbf{x}_{i,j}^{\prime(k)} \right]^{-1},$$

$$\text{Var}(\widehat{\boldsymbol{\alpha}}) = \sum_{i,j} \left\{ \alpha^{-4} \Big(ln(1 + \alpha \mu_{i,j}^{(k)}) - \sum_{w=0}^{n_{i,j}^{(k)}-1} \frac{1}{w + \alpha^{-1}} \Big)^2 + \frac{\mu_{i,j}^{(k)}}{\alpha^2 (1 + \alpha \mu_{i,j}^{(k)})} \right\}^{-1},$$

$$Cov(\widehat{\boldsymbol{\beta}}, \widehat{\alpha}) = \mathbf{0}.$$

Readers interested in details of the partial derivatives and the asymptotic distribution of the negative binomial distribution can refer to (Cameron & Trivedi, 2013) and (Lawless, 1987).

The **MASS** package of the R software has an in-built function **glm.nb** which can be used to estimate the parameters of negative binomial regression.

It is worth noting that geometric regression is a special case of negative binomial regression in which the dispersion parameter is set to one. Our interest in negative binomial and geometric regression hinges on the fact that they relax the restrictive assumption that the variance is equal to the mean made in the Poisson regression.

In addition to the regression models discussed above, we also examine zero-inflated models of the regular models for count data. Zero-inflated models attempt to account for excess zeros in a data. They estimate two equations simultaneously, one for the count model and one for the excess zeros. References can be made to (Cameron & Trivedi, 2013) and (Hilbe, 2014) for full details of these models.

The zero-inflated Poisson (ZIP) regression is used for count data that exhibit over-dispersion and excess zeros. The data distribution combines the Poisson distribution and the logit distribution. Suppose $N_{i,j}^{(k)}$ is a count variable which follows ZIP model, then its probability distribution is given by

$$Pr(N_{i,j}^{(k)} = n_{i,j}^{(k)}) = \begin{cases} \pi_{i,j}^{(k)} + (1 - \pi_{i,j}^{(k)})\exp(-\mu_{i,j}^{(k)}), & \text{if } n_{i,j}^{(k)} = 0 \\ (1 - \pi_{i,j}^{(k)})\frac{(\mu_{i,j}^{(k)})^{n_{i,j}^{(k)}}\exp(-\mu_{i,j}^{(k)})}{n_{i,j}^{(k)}!}, & \text{if } n_{i,j}^{(k)} > 0. \end{cases}$$

Here, $\mu_{i,j}^{(k)}$ follow the same definition as before. Also, $\pi_{i,j}^{(k)}$ is a logistic function given by

$$\pi_{i,j}^{(k)} = \frac{\psi_{i,j}^{(k)}}{1 + \psi_{i,j}^{(k)}},$$

where

$$\psi_{i,j}^{(k)} = \mathbf{z}_{i,j}^{\prime(k)}\boldsymbol{\gamma}.$$

Here, $\mathbf{z}_{i,j}^{(k)}$ is an m-dimensional vector of predictors which may not be the same as $\mathbf{x}_{i,j}^{(k)}$ and $\boldsymbol{\gamma}$ is an m-dimensional vector of regression coefficients.

The regression coefficients are estimated using the method of maximum likelihood. The logarithm of the likelihood function is given by:

$$\mathcal{L} = L1 + L2 - L3$$

where

$$\begin{aligned} L1 &= \sum_{\{(i,j):n_{i,j}^{(k)}=0\}} ln\left[\psi_{i,j}^{(k)} + \exp(-\mu_{i,j}^{(k)})\right] \\ L2 &= \sum_{\{(i,j):n_{i,j}^{(k)}>0\}} \left\{n_{i,j}^{(k)} ln(\mu_{i,j}^{(k)}) - \mu_{i,j}^{(k)} - ln(n_{i,j}^{(k)}!)\right\} \\ L3 &= \sum_{i,j} ln\left(1 + \psi_{i,j}^{(k)}\right). \end{aligned}$$

The first partial derivatives of $\mathcal{L}$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are given by:

$$\frac{\partial \mathcal{L}}{\partial \beta_r} = \sum_{\{(i,j):n_{i,j}^{(k)}=0\}} \left[\frac{-x_{ir}\mu_{i,j}^{(k)}}{\psi_{i,j}^{(k)}\exp(-\mu_{i,j}^{(k)})+1}\right] + \sum_{\{(i,j):n_{i,j}^{(k)}>0\}} (n_{i,j}^{(k)} - \mu_{i,j}^{(k)})x_{ir}, \quad r = 1,2,\ldots,p$$

$$\frac{\partial \mathcal{L}}{\partial \gamma_r} = \sum_{\{(i,j):n_{i,j}^{(k)}=0\}} \left[\frac{z_{ir}\psi_{i,j}^{(k)}\exp(-\mu_{i,j}^{(k)})}{\psi_{i,j}^{(k)}\exp(-\mu_{i,j}^{(k)})+1}\right] - \sum_{\{(i,j):n_{i,j}^{(k)}>0\}} \left[\frac{\psi_{i,j}^{(k)}}{1+\psi_{i,j}^{(k)}}\right] z_{ir}, \quad r = 1,2,\ldots,m.$$

The gradients given above are set to zero to obtain estimating equations which are solved numerically for the regression coefficients. It has been shown in the literature that the asymptotic distribution of the maximum likelihood estimates the coefficients follows multivariate normal distribution.

The zero-inflated negative binomial (ZINB) regression is also used for count data that exhibit over-dispersion and excess zeros. The distribution combines the negative binomial distribution and the logit distribution. Suppose $N_{i,j}$ is a count variable which follows ZINB model, then its probability distribution is given by

$$Pr(N_{i,j}^{(k)} = n_{i,j}^{(k)}) = \begin{cases} \pi_{i,j}^{(k)} + (1-\pi_{i,j}^{(k)})g(n_{i,j}^{(k)}=0), & \text{if } n_{i,j}^{(k)} = 0 \\ (1-\pi_{i,j}^{(k)})g(n_{i,j}^{(k)}), & \text{if } n_{i,j}^{(k)} > 0, \end{cases}$$

where $g(n_{i,j}^{(k)})$ is the negative binomial distribution given by equation 4.3.

Also, the details of $\mu_{i,j}^{(k)}$ and $\pi_{i,j}^{(k)}$ will be kept the same as before.

Readers must note that, similar to what we did for the standard negative binomial, we chose the parametrization of zero-inflated negative binomial 2 (ZI-NB2) but kept the name zero-inflated negative binomial (ZI-NB) for simplicity. This was done to avoid the complexity associated with zero-inflated negative binomial 1 (ZI-NB1) in modelling the variance (Cameron & Trivedi, 2013).

The regression coefficients are estimated using the method of maximum likelihood.

The logarithm of the likelihood function is given by:

$$\mathcal{L} = L1 + L2 + L3 - L4$$

where

$$\begin{aligned}
L1 &= \sum_{\{(i,j):n_{i,j}^{(k)}=0\}} ln\left[\psi_{i,j}^{(k)} + (1+\mu_{i,j}^{(k)})^{-\alpha^{-1}}\right] \\
L2 &= \sum_{\{(i,j):n_{i,j}^{(k)}>0\}} \sum_{w=0}^{n_{i,j}^{(k)}-1} ln(w+\alpha^{-1}) \\
L3 &= \sum_{\{(i,j):n_{i,j}^{(k)}>0\}} \left\{ -ln(n_{i,j}^{(k)}!) - (n_{i,j}^{(k)}+\alpha^{-1})ln(1+\alpha\mu_{i,j}^{(k)}) + n_{i,j}^{(k)} ln\alpha + n_{i,j}^{(k)} ln(\mu_{i,j}^{(k)}) \right\} \\
L4 &= \sum_{i,j} ln(1+\psi_{i,j}^{(k)}).
\end{aligned}$$

The first partial derivatives of $\mathcal{L}$ with respect to β's and γ's and α are given by:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \beta_r} &= \sum_{\{(i,j):n_{i,j}^{(k)}=0\}} \left[\frac{-\mu_{i,j}^{(k)}(1+\alpha\mu_{i,j}^{(k)})^{-1-\alpha^{-1}}}{\psi_{i,j}^{(k)} + (1+\alpha\mu_{i,j}^{(k)})^{-\alpha^{-1}}}\right] x_{ir} + \sum_{\{(i,j):n_{i,j}^{(k)}>0\}} \left[\frac{n_{i,j}^{(k)} - \mu_{i,j}^{(k)}}{1+\alpha\mu_{i,j}^{(k)}}\right] x_{ir}, \\
\frac{\partial \mathcal{L}}{\partial \gamma_r} &= \sum_{\{(i,j):n_{i,j}^{(k)}=0\}} \left[\frac{\psi_{i,j}^{(k)}}{\psi_{i,j}^{(k)} + (1+\alpha\mu_{i,j}^{(k)})^{-\alpha^{-1}}}\right] z_{ir} - \sum_{i,j} \left[\frac{\psi_{i,j}^{(k)}}{1+\psi_{i,j}^{(k)}}\right] z_{ir}, \\
\frac{\partial \mathcal{L}}{\partial \alpha} &= \sum_{\{(i,j):n_{i,j}^{(k)}=0\}} \left\{\frac{(1+\alpha\mu_{i,j}^{(k)})ln(1+\alpha\mu_{i,j}^{(k)}) - \alpha\mu_{i,j}^{(k)}}{\alpha^2[\psi_{i,j}^{(k)}(1+\alpha\mu_{i,j}^{(k)})\alpha^{-1}+1]}\right\} \\
&\quad + \sum_{\{(i,j):n_{i,j}^{(k)}>0\}} \left\{ \sum_{w=0}^{n_{i,j}^{(k)}-1} \frac{-1}{\alpha^2 w + \alpha} + \frac{ln(1+\alpha\mu_{i,j}^{(k)})}{\alpha^2} + \frac{n_{i,j}^{(k)} - \mu_{i,j}^{(k)}}{\alpha(1+\alpha\mu_{i,j}^{(k)})} \right\}
\end{aligned}$$

where $r = 1, 2, \ldots, p$ and $r = 1, 2, \ldots, m$ in the first two derivatives respectively.

The gradients given above are set to zero to obtain estimating equations which are solved numerically for the regression coefficients. It has been shown in literature that the asymptotic distribution of the maximum likelihood estimates the coefficients follows multivariate normal distribution.

In the special case where the parameter, α, is set to 1 we have the zero-inflated geometric (ZIG) model which is also another model we considered as a possible frequency model in this study.

The zero-inflated regression models can be implemented using the the function **zeroinfl()** from the package **pscl** available in the R software.

4.3.2 Severity model

Severity model estimates the amount of claim payment once the claim frequency has been predicted. Just like the frequency model, the choice of distributions for severity models vary widely among researchers. For instance, (Schiegl, 2002) used Pareto distribution, (Denuit & Trufin, 2017) used mixture of gamma and pareto distributions while (Antonio & Plat, 2014) used lognormal distribution.

Instead of restricting ourselves to one particular distribution, we shall use the machine learning tool to predict loss amounts.

The description of our severity model follows from the neural network model discussed in Chapter 3. We saw that for an n-dimensional vector of response, $\mathbf{Y}$, and $(p+1)$-dimensional vector of covariates, $\mathbf{X}$, the feed-forward neural network model predicts responses by the function given by:

$$\widehat{f}(\mathbf{X}) = g\left(\mathrm{w}_{k,0}^{(2)} + \sum_{m=1}^{M} \mathrm{w}_{k,m}^{(2)} h\left(\mathrm{w}_{m,0}^{(1)} + \sum_{j=1}^{p} \mathrm{w}_{m,j}^{(1)} X^{(j)}\right)\right).$$

The parameters $\mathrm{w}_{m,0}^{(1)}$ and $\mathrm{w}_{k,0}^{(2)}$ are referred to as **biases**. While the other parameters $\mathrm{w}_{m,j}^{(1)}$ for $j \neq 0$ and $\mathrm{w}_{k,m}^{(2)}$ for $m \neq 0$ are referred to as **weights**. Also, the function, h, is referred to as the input activation function while, g, is an output activation function.

The set of all the network parameter are found so as to minimize the loss function

given by:

$$L(\mathbf{Y}, \widehat{f}(\mathbf{X})) = (Y - \widehat{f}(\mathbf{X})^2.$$

We propose the feed-forward neural network model for our severity model. For each claim, define the statistical model below for amount of payment:

$$Y_{i,j}^{(k)} = f(\mathbf{x}_{i,j}^{(k)}) + \epsilon_{i,j}^{(k)}$$

where $Y_{i,j}^{(k)}$ is the amount of payment for the claim k in cell (i, j) of the development triangle, $\mathbf{x}_{i,j}^{(k)}$ are the individual level covariates associated with the claim payment, and $\epsilon_{i,j}^{(k)}$ is a residual term with zero mean.

The neural network model gives estimate of claim payments, $\widehat{Y}_{i,j}^{(k)} = \widehat{f}(\mathbf{x}_{i,j}^{(k)})$, that minimize the squared loss function.

Implementation of neural networks will be done using **neuralnet** package in the R software which uses backpropagation method to estimate the parameter of the neural network.

By using a feed-forward neural network structure already discussed in the previous chapter, we have adopted a non-linear method of predicting loss amounts. Some key features of modeling severity of loss using neural networks can be summarized as follows.

- Neural network is flexible and more computational than theoritical. It learns from data through the internal structures of the network.
- Micro level covariates can be incorporated in the predictions.
- It involves non-linear transformation of predictors. The data go through series of transformations before predictions comes out.

4.4 Predicting from the frequency-severity model

In the context of individual actuarial reserves, the concept of frequency-severity model was implemented for the first time both in (Norberg, 1993) and further developed in (Norberg, 1999). In this article, the two variables frequency (N) and severity (X) are modeled using a Poisson distribution. Subsequently, several researchers took up the concept to propose different ways of modeling N and X including, in particular, (Schiegl, 2002). This article proposes a model where the frequency (N) follows a negative binomial distribution and the severity (X) follows a Pareto sistribution. One year later (Wüthrich, 2003) proposed in turn a Tweedie model where the frequency follows a Poisson distribution and the severity follows a Gamma law using the occurrence years (i) and the years of development (j) as explanatory variables. In 2011, (Boucher & Davidov, 2011) use a Tweedie model, this time using a double GLM (DGLM), that is to say, they allowed the expectation and the variance to depend of cell (i, j). In 2017, (Denuit & Trufin, 2017) proposed a model where the frequency is modeled using the Poisson distribution and severity is modeled by a mixture of Gamma and Pareto distributions. Thus, the Gamma models models severity of payments made in the early years of development and the Pareto distribution models the payments made during the last years of development. Those who wish to complete this brief overview of the use of the frequency-severity approach in modeling reserves can be consulted (Wüthrich & Merz, 2008) and (England & Verrall, 2002).

The model we propose in this work brings together a best frequency model from generalized linear models to predict the number of payments and a feed-forward neural network model to predict the severity of payments. Using the frequency-severity model, predictions are made as follows:

- Predict $N_{i,j}^{(k)}$, the number of future payments of claim k using the selected

frequency model.

- Predict $X_{i,j,1}^{(k)}, \ldots, X_{i,j,N_{i,j}^{(k)}}^{(k)}$, the amount of each future payment using the neural network severity model.
- Compute $\widehat{Y}_{i,j}^{(k)} = \sum_{l=1}^{N_{i,j}^{(k)}}$ which is the total payment for claim k.
- Use $\widehat{Y}_{i,j}^{(k)}$ to estimate the RBNS reserve.

For simulations to obtain the empirical distribution, we implement the above steps for large number of times (say 10000). The empirical mean, standard error and quantiles are found to complete the process. In this work, the empirical studies on our model will be based on case resampling whereby bootstrap samples will be created by randomly selecting observations (with replacement) from the original data.

The choice of structure of our model is motivated by our intention to investigate the components of total amount of claims separately. By considering a separate model for each component we have the advantage of using appropriate statistical tools to determine the factors that influence each component and distinguish them from each other. One successful implementation of this strategy was done by (Frees & Valdez, 2008) in the context of rate making. Furthermore, separating the analysis for frequency and severity enable us to harness the strength of two different powerful statistical tools namely generalized linear models (GLM) and neural networks. Each tool has its strengths and weaknesses. One major strength of GLMs is the ability to capture mass probability of zero in frequency model whereas neural network is very prominent for its predictive power. We therefore exploit the unique strength of GLMs for our frequency model and harness the predictive power of neural networks for our severity model. Using just one neural network architecture for the whole frequency-severity model will deny us the

opportunity to obtain useful information about the each components of the total amount of claims.

CHAPTER V

NUMERICAL ANALYSES

In this chapter, we shall discuss the application of classical models and our proposed model to a real dataset and present summary of the results we obtained from the data analyses.

5.1 Description and cleaning of data

The dataset used in this research is obtained from a Canadian automobile insurer. The insurer offers three major general insurance products: personal auto products, corporate commercial auto products and SR4 commercial auto products. (SR4 means Smart Reefer 4; it is a technological unit used to control temperature in cargo vehicles such as vehicles with refrigerated trailers). Personal auto products are designed for insuring private-passenger-type autos owned by individuals. The policy may be structured to provide a combination of liability, personal injury protection (PIP), medical payments, uninsured and underinsured motorists (UM/UIM), and physical damage coverages. Corporate commercial auto products cover vehicles such as cars, trucks, vans and commercial use trailers. They offer business protection against the financial impact resulting from damage or loss of your business vehicles as well as damage caused to vehicles or property owned by others through the use of your commercial vehicle. The SR4 commercial auto

products is a special commercial insurance product designed for temperature controlled cargo vehicles such as vehicles with refrigerated trailers. They protect business owners against loss of the vehicle or cargo as well as damage to vehicles and properties of other people that result from the operation of the cargo vehicle. Thirty different loss kinds are covered by the insurer including collision, fire, hail, explosion, damages by water etc. The insurer has six valuations regions across the provinces and territories in Canada and few states in USA.

The full dataset contains information on accidents occurring from the year 1978 through to the year 2017. However, we restricted our analyses only to the data for the most recent years from 2005 through to 2017 in order to ensure that the dataset has a manageable size. In cleaning the data for analyses, we removed $1,255$ out of the total observations of 9,337,420 because they contained missing information. In addition to that, we got rid of observations with payments less than zero for convenience. Furthermore, we removed certain information peculiar to the insurer and variables in the dataset that were found not relevant for loss reserving purposes such as underwriter, source system, policy key etc.

For the purpose of analyses, we used claims occurring between the years 2005 to 2012 inclusive as training set and those occurring between the years 2013 to 2017 inclusive as validation set. The training set has 522,544 insurance claims, 2,351,237 observations for the interval considered and 17 variables. The description of the variables is given in Appendix A. We did not use discounted values for payments amounts and neither did we apply any form of transformations to variables in the data.

5.2 Descriptive statistics

In the preliminary analyses of data, we were interested in observing patterns in the data with respect to payments that were made by the insurer from 2005 to 2012 inclusive as well as how the payments are affected by other variables in the data.

Over the period from year 2005 through to 2012, it was observed that payments made by the insurer range from 0 to \$7,207,200 with mean payment of $\$1,049$. In addition to that, Figure 5.1 shows that the proportion of zero payments for claims is about 77.23%. This signifies a large mass probability for the frequency of payment of claims. It was also found that the number of claim payments ranges between 0 and 298. Apart from the high occurrence of zero payments, the most observed number of claim payments were between one and four since the proportion of one to four number of claim payments was 21.51%.

Table 5.1 shows the distribution of the payment of claims according to the insurance products. It is observed that 98.68% of the payments made by the insurer are associated with claims from auto insurance policies. According to Table A.6 shown in Appendix A, the loss kind names: not available, multi-vehicle and glass windshield caused majority of the payments for Auto products while the loss kind names: not available, multi-vehicle and single-vehicle caused majority of payments for Corporate Commercial Auto and SR4 Commercial Auto products. Earthquake and collision were the loss kinds that caused the least payments in all the three products.

It is interesting to note that the insurer's payments of claims vary widely among the provinces. As shown in Figure C.4, very high payment of claims were observed for the provinces of Ontario and Alberta while Manitoba, Illinois, British

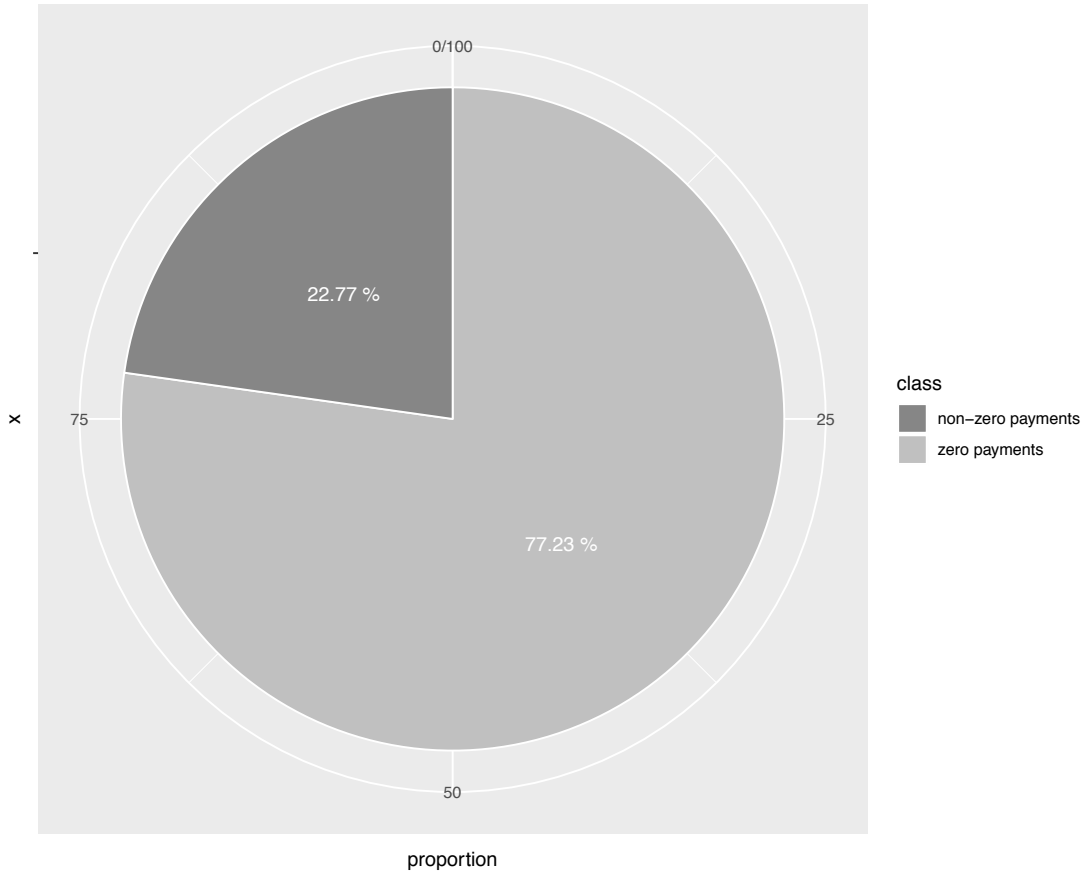

Figure 5.1 Proportion of zero and non-zero payments made by the insurer

Table 5.1 Distribution of payments by insurance product

Product type	Number of Payment	Percentage
Auto	2,320,163	98.68
Corporate Commercial Auto	10,135	0.43
SR4 Commercial Auto	20,939	0.89
Total	2,351,237	100.00

Columbia and Not Available provinces experienced very low payment of claims.

A cross-tabulation of the variables Product type and Province shown in Table A.5 of Appendix A also revealed that payments for personal auto claims are highly predominant in the Province of Ontario while the province of Alberta is also leading in the payment of Corporate Commercial Auto and SR4 Commercial Auto insurance payments. This supports a further analyses that, by valuation regions, the payments for auto products are high in the Ontario region while the payments for Commercial Auto and SR4 Commercial Auto products are high in the west region.

Furthermore, the preliminary analysis revealed that from 2005 to 2012, about 75% of total claim payments made by the insurer occurred in development years 1 to 2 alone. The plot in Figure 5.3 indicated that the proportion of payments in latter development years decreases. In addition to that, we observed from the plot in Figure 5.4 that there were a lot of closed claims from year 2005 to 2009. This suggest that claims are settled more quickly by the insurer. However, there is not much difference in the number of open claim in all the accident years.

We were also interested in investigating how quickly claims are reported to the insurer. Figure 5.5 shows that claims are declared mostly within one year.

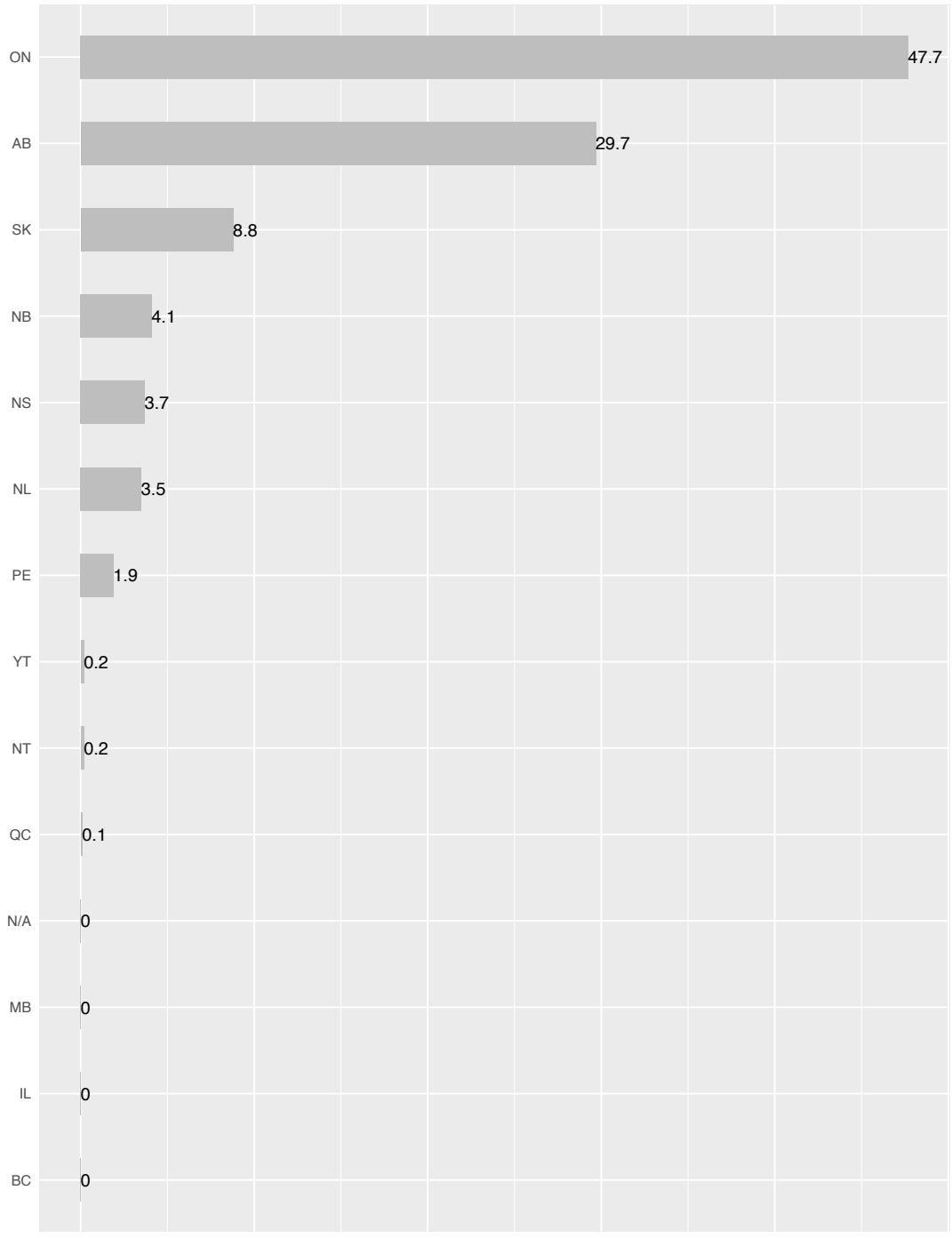

Figure 5.2 Percentages of amount paid by provinces

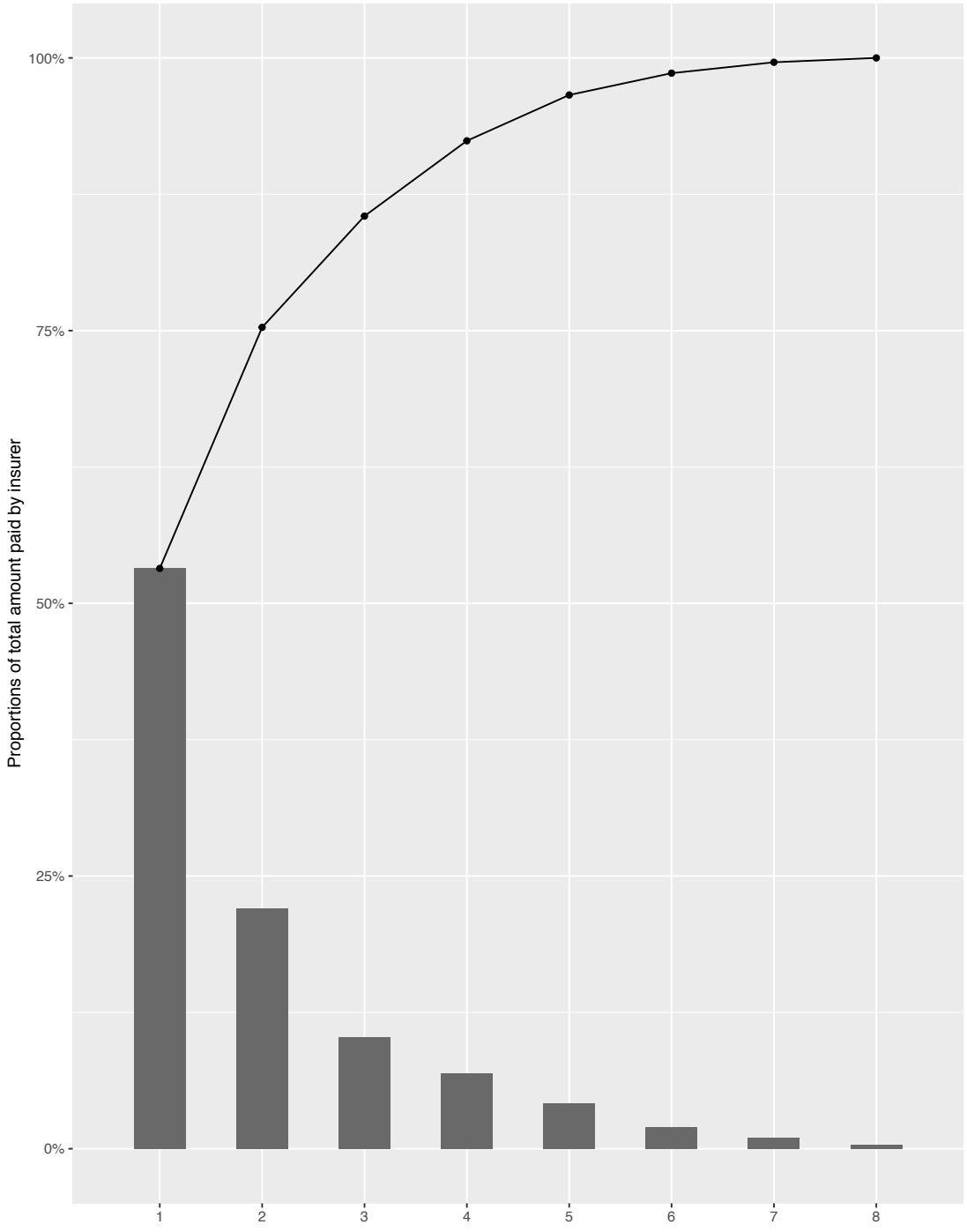

Figure 5.3 Proportions of total amount paid by development years

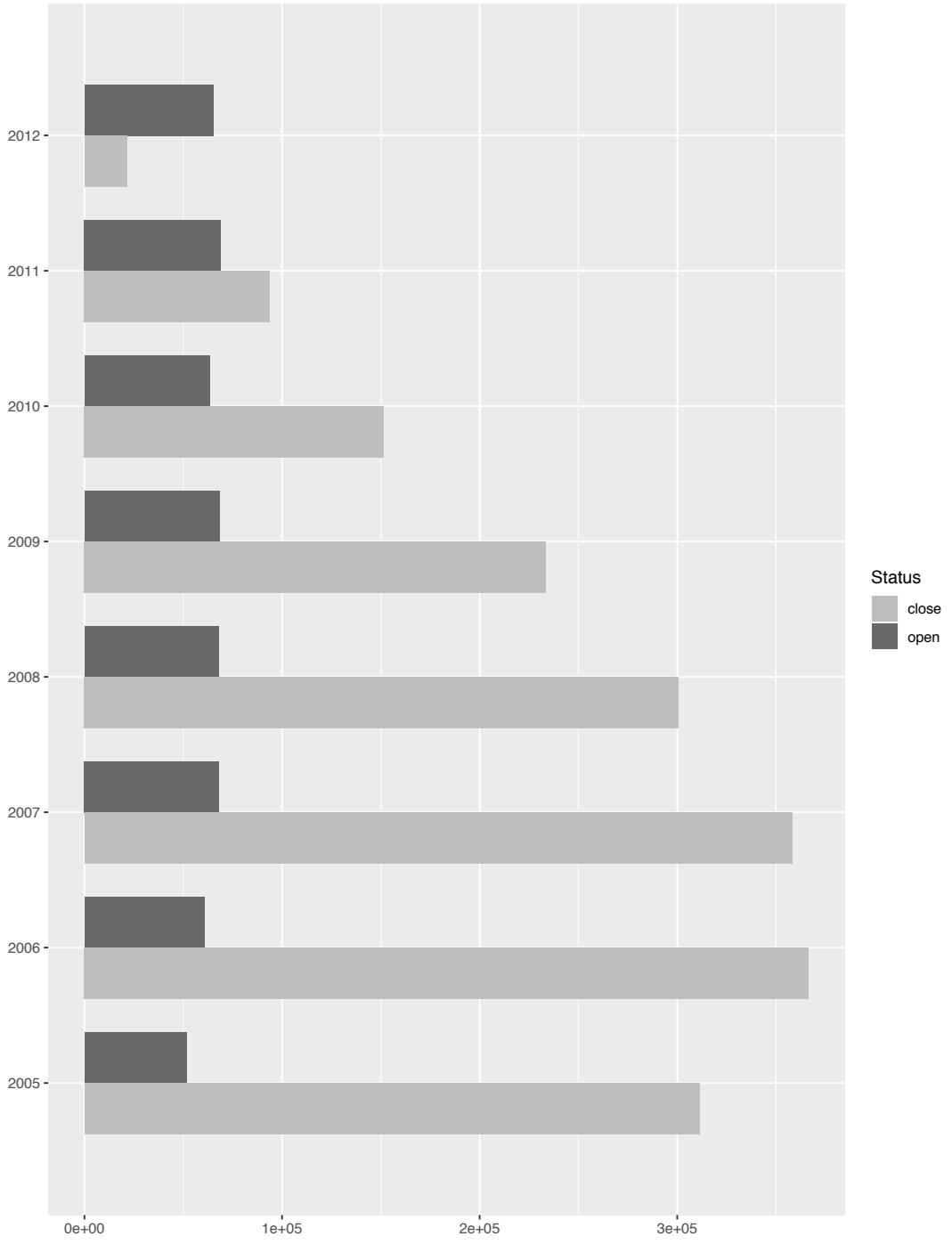

Figure 5.4 Open and closed claims by accident years

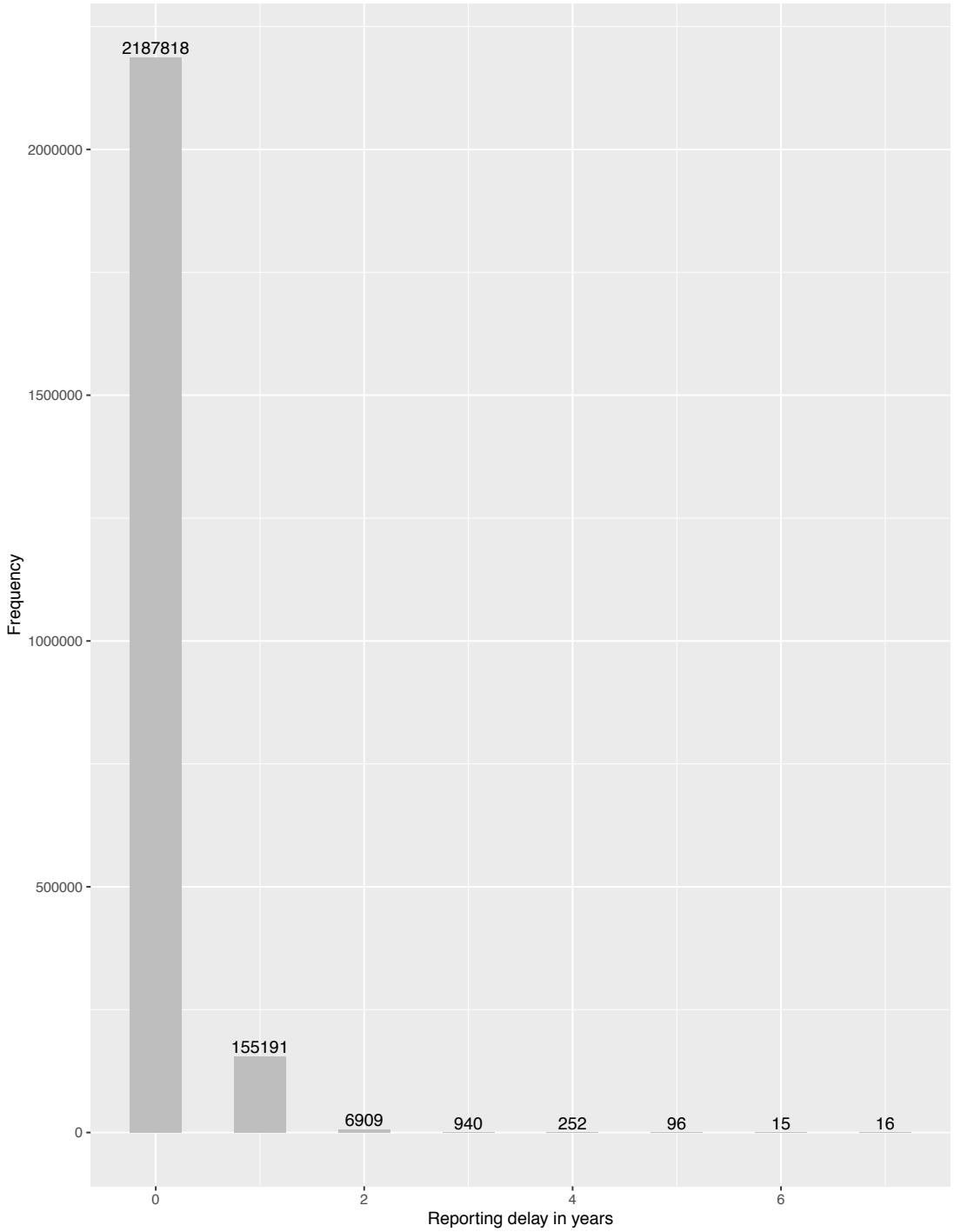

Figure 5.5 Chart showing frequency of reporting delay in years

Some descriptive statistics of payments by accident years and development years are given by Table A.7 and Table A.8 in Appendix A. An interesting trend from the summary of total payments by accident and development years is shown in Figure 5.6. From the graph on the left panel, it is observed that total payments increased rapidly for the first four accident years and begins to drop mildly afterwards. Then from the graph on the right panel, the total payment for development year one was the highest and it decreased at an increasing rate for the first four development years and decreased at a slower rate to development year eight. Therefore, one can deduce that as claims are reported early by the insureds, consequently, the insurer makes payments to settle the claims as early as possible.

5.3 Application of classical models

In order to apply classical models to the dataset, it was necessary to extract the run-off triangle from the training set. This was done by using a function in R called **aggregate** to group the data by accident and developing years. Using the function in R, we obtained the incremental run-off triangle as shown by the upper triangle in Table 5.2. In the same table, the figures in bold character are figures from the validation set captured in the 8-year development table.

Table 5.2 Incremental Loss triangle from the training set

	Development year							
Accident year	1	2	3	4	5	6	7	8
2005	143,674,854	73,404,369	45,387,867	28,797,839	26,875,423	14,604,037	13,770,109	10,084,663
2006	150,629,123	76,897,679	45,450,272	33,767,706	27,009,255	18,764,257	11,829,158	**6,522,013.9**
2007	180,275,512	89,379,356	40,223,063	34,151,530	19,876,707	15,223,309	**15,436,512.6**	**6,670,408.4**
2008	179,645,909	95,758,172	51,223,617	37,465,949	29,039,994	**17,177,468.9**	**9,953,976.0**	**7,150,389.9**
2009	180,714,549	79,066,092	37,297,565	36,226,347	**33,483,307.2**	**20,343,396.6**	**14,923,983.2**	**3,436,144.3**
2010	156,574,090	75,089,362	33,220,399	**31,889,199.1**	**22,525,294.9**	**19,394,771.0**	**8,169,391.0**	**3,378,133.1**
2011	163,510,814	53,847,566	**31,419,751.2**	**26,113,614.1**	**25,215,939.4**	**15,054,852.4**	**14,715,671.1**	
2012	157,885,214	**53,610,637.7**	**25,983,536.2**	**24,590,977.5**	**18,886,468.4**	**12,330,807.8**		

The classical models we applied to the run-off triangle are the following;

- Mack model
- GLM-Poisson model.

We would like to emphasize that even though collective models are not our main focus in this thesis, we have chosen the Mack model and the GLM-Poisson model as collective models to strengthen the background we built for this thesis and to demonstrate to our readers, with practical example, the connection between these two models which was discussed in the latter part of Chapter 2. Applications of other GLMs will be discussed under the individual loss reserving approach.

As already discussed in Chapter 2, the main covariates for the collective models are the accident years and development years. In this data, the accident years are the years 2005 through to 2012; and we have 8 development years.

The classical models were applied to estimate total reserves. In addition to that, 10,000 simulations, were used to obtained the predictive distributions for the models. The simulations were based on bootstrap samples drawn from the data. Usually, insurance companies use high quantiles of the predictive distribution as reserve amount. Therefore, we were also interested to find estimates for the 90^{th}, 95^{th} and 99^{th} quantiles of the predictive distributions. In Table 5.3, we present the summary of results for the classical models. The table shows the estimate of total reserves and selected quantiles of the predictive distribution. Figure 5.7 shows the graph of the predictive distributions. Other details of the results from the classical model are presented in Appendix B.

The observed amount of RBNS reserve obtained from the validation set which contained data for years 2013 through to 2017 was \$505,661,693. From available data we also observed IBNR reserve of 56,703,282, making an observed total reserve of \$562,364,975. Since the collective models considered here do not separate RBNS

Table 5.3 The summary of results for classical model

	Estimates				
Model	Mean	SE	90^{th} quantile	95^{th} quantile	99^{th} quantile
Mack	578,064,728	42,788,129	632,426,617	648,841,907	678,979,299
GLM-Poisson	578,131,075	42,714,561	632,601,809	648,690,549	676,779,315
Observed reserve	562,364,975				

and IBNR, it will make sense to compare their performance using the observed total reserve.

The results clearly show that the estimates of total reserve using the Mack's model and GLM-Poisson are very similar. The standard error was slightly lower for GLM-Poisson model than the Mack's model. However, the estimated high quantiles, which the insurance companies usually use as reserve are above $600,000,000 for both models. Such high estimates of reserves will not be favourable to the insurance company as they will have adverse impact on other operations of the company.

5.4 Application of new model

5.4.1 Analyses of Frequency models

In this section, we shall describe the analyses of the frequency models described in Chapter 4 with the aim of selecting the best ones that fit our data. The models we investigated are **standard Poisson, negative binomial and geometric distributions** as well as **zero-inflated Poisson(ZIP), zero-inflated negative binomial (ZINB) and zero-inflated geometic (ZIG)** distributions.

The response variable, number of payments, N, was studied in two main scenarios.

In the first scenario, we put N into six categories as $0, 1, 2, 3, 4$ and 5^{+}. This was meant to help us get rid of some extreme values of the variable N. In the second scenario, we used the original values of N with no categories.

In our data, seven variables were available to be used as predictors. However, in our analyses we tested the selection of three different subsets of predictor for all the 6 frequency models. The first and second subsets of predictors used in our models consist of the best three and four variables, respectively, that were identified using the random forest variable importance plot as shown in Figure 5.8 and Table 5.4.1. "IncNodePurity" measures the degree of impurities of a split at each node of a tree based on the loss function."%IncMSE" calculates the deterioration of the prediction accuracy of the random forest when permuting the values of each explanatory variable of the test set in order to break the association with the response variable. Higher values of "IncNodePurity" and "%IncMSE" represent higher variable importance. The third subset consists of six variables identified using **stepAIC** function in R that implements a stepwise regression method to select important predictors for a model. The summary of the three subset of predictors were as follows:

1. Three predictors: *Development year, Loss kind and accident year.*
2. Four predictors: *Development year, Loss kind, accident year and province.*
3. Six predictors: *Development year, Loss kind, accident year, province, product and major.*

Table 5.5 and Table 5.6 summarize the performance of the best six models based on the lowest out-of-sample Mean Square error (MSE) for the scenario when N is not categorized and the other scenario when N is categorized, respectively. The number of predictors used in the model as well as their predictions of total number

of payment are shown in the tables. Other details of the analyses of the frequency models can be found tables presented in Appendix C.

The results indicate that the **zero-inflated Poisson model** performed better than the other models in terms of predicting the the number of payments. In both scenarios, the zero-inflated Poisson had the lowest MSE and estimate of total N that was the closest to the observed values in the data. Zero-inflated geometric model was the second best frequency model. It was observed that the negative binomial models did not do well in all scenarios. For the scenario whereby N was not categorized, the total number of payment, N, that was found in the validation set was 566,269. The zero-inflated poisson model with three predictors estimated a total of 563,643 with the lowest MSE of 4.5511. The zero-inflated geometric with three predictors had the second lowest MSE of 5.0131 but estimated a slightly higher total number of payment, N, of 570,731. For the scenario whereby N was categorized, the total number of payment, N, that was found in the validation set was 534,447. The zero-inflated Poisson model with three predictors emerged to be the best model. It estimated a total of 543,163 with the lowest MSE of 1.4270. The zero-inflated geometric with three predictors did quite well by estimating a total of 559,559 with MSE of 1.5163.

It is worth mentioning that, according to the AIC criterion, the model which performed best in predicting the frequency of claims was the zero–inflated geometric model with six parameters. This model had the smallest AIC of 988,051.63 and out-of-sample MSE of 4.6160 . It is obvious that this model could not out-perform the models selected above using the validation method. The AIC model produced an out-of-sample MSE greater than the models selected by the validation method. It gave a much higher prediction of the total N in the data; a value of 590,175 instead of 566,269 observed in the data.

Based on results shown above, we shall give attention to zero-inflated Poisson and zero-inflated geometric as frequency our models in further analyses. That is, the estimation of RBNS reserves will take into consideration estimated frequencies from zero-inflated Poisson and zero-inflated geometric models.

Table 5.4 Results of random forest variable importance for number of payments

	Importance	
Variable	%IncMSE	IncNodePurity
Major	3.3815	1.3812
Product	0.7202	14.6718
Region	-0.0596	110.9096
Province	3.9630	355.3943
Year	5.7952	457.6994
Losskind	10.3422	617.9069
Dev	30.9875	780.3783

Table 5.5 Results of the 1st six frequency models for the case when N is not categorized

Rank	Model	No of predictors	Estimate of total N	MSE
1	ZI_Poisson	3	563,643	4.5511
2	ZI_Geometric	3	570,731	5.0131
3	ZI_Poisson	4	572,827	5.1257
4	ZI_Poisson	6	575,920	5.2129
5	ZI_Geometric	4	668,326	5.6890
6	ZI_Geometric	6	680,665	5.7325

Table 5.6 Results of the 1st six frequency models for the case when N is categorized

Rank	Model	No of predictors	Estimate of total N	MSE
1	ZI_Poisson	3	543,163	1.4270
2	ZI_Poisson	4	554,999	1.4655
3	ZI_Poisson	6	566,440	1.4694
4	ZI_Geometric	3	559,559	1.5163
5	ZI_Geometric	4	560,393	2.5596
6	ZI_Geometric	6	590,175	4.6160

5.4.2 Analyses of severity models and computation of RBNS reserves

The analyses began by selecting four best frequency models from the previous section. They are described as follows:

M1: zero-inflated Poisson with 3 predictors and no categories in the response variable, N.

M2: zero-inflated geometric with 3 predictors and no categories in the response variable, N.

M3: zero-inflated geometric with 3 predictors and six categories of the response variable, N.

M4: zero-inflated Poisson with 3 predictors and six categories of the response variable, N.

Selection of predictors for analyses of severity of payments was based on random forest variable importance measures shown in Table 5.7 and Figure 5.9. Five different subsets of predictors were considered in the analysed based on the ranking of the variable importance measures "%IncMSE" and "IncNodePurity". Since higher

values of these measures indicate higher importance of variables, we selected the subset as follows: **Subset1** is made up of the top 3 variable , *Development year, Loss kind and loss year*. **Subset2** is made up of the top 4 variable, *Development year, Loss kind , loss year and province*. **Subset3** is made up of the top 5 variable, *Development year, Loss kind, loss year, province and region*. **Subset4** is made up of the top 6 variable, *Development year, Loss kind, loss year, province, region and product* and **Subset5** is made up of all the seven variable.

In order to predict the amount of payment using neural networks, we used the **neuralnet** package in the R. We used one hidden layer, a threshold of 0.01, learning rate of one, and maximum number of iteration 10,000,000. Implementing this set-up helped us obtain convergence of neural network without taking too much time. In Appendix C, Figure C.3 and Figure C.4 are neural network plots when three and five predictors were used in the model. Numerical design matrices were first constructed for the predictors before the network parameters were computed. With three predictors, the numerical design matrix had 31 variables, and 45 variables when five predictors were used. Bias parameter are printed in blue colour while weight parameters are in black.

For each of subsets of predictors, RBNS reserves were computed following the description already given in Chapter 4. Furthermore, empirical analyses were made to determine the predictive distribution of the RBNS reserves for each of the selected frequency models. The predictive distributions are more useful for comparison of results and beneficial to insurance company since they set their reserve based on high quantiles of these distribution.

The performance of all the neural network models were compared to the observed total reserve from the validation set which was $\$505,661,693$. The best model was expected to give estimates close to this value with minimum variability as

possible. This is because very low estimates will cause solvency problems for the insurance company and very high estimates will hinder the company from saving money to cater for other operations and expenses.

For each subset of predictors, we present a summary table that contains the empirical mean, standard error, 90^{th} quantile, 95^{th} quantile and 99^{th} quantile for each of the four selected frequency model. We also present an RBNS predictive distribution with the following description: the vertical line is the observed RBNS found from the validation set, the black curve is the predictive distribution for **M1**, the green curve is the predictive distribution for **M2**, the red curve is the predictive distribution for **M3** and the blue curve is the predictive distribution for **M4**.

After predicting severity of payments and estimating RBNS with the models described above, we found that the neural network model with the three most importance variables indicated by the random forest variable importance measures gave the best results as shown in Table 5.8 and Figure 5.10. For all the four frequency models considered with this neural network severity model, average estimates of RBNS reserve captured the observed reserve adequately with minimum variability and the quantiles were not so extravagant to hinder the saving potentials of the company. The best record was obtained when zero-inflated poisson was used as a frequency model and the severity of payment was predicted with neural network using the best three predictors. The average RBNS of that model was \$530,319,309 which is the minimun above the observed RBNS reserve of \$505,661,693. With the smallest 99^{th} quantile of \$562,728,966, the company can set an adequate reserve that will not hinder its saving plans. The results of the other three frequency models with neural network severity model with best three predictors were very encouraging as the worst 99^{th} quantile was \$592,836,894 which does not exceed \$600,000,000.

The second best model for RBNS reserve was the neural network model with the five most important variables. The results in Table 5.9 and Figure 5.11 indicate that the estimates are quite close to the results from neural network model with three predictors but the company will save a little more with the former model than the latter one. The model with zero-inflated poisson had the mean RBNS of \$531,807,224 which was is closest to the observed RBNS and 99^{th} quantile of \$593,410,769 which does not exceed \$600,000,000. The other frequency models had mean RBNS reserve that adequately covers the observed reserves but their quantiles were slightly higher with five of them exceeding \$600,000,000.

The worst RBNS model was the one in which all the seven predictors were included in the neural network severity model. The results of this model are shown in Table 5.10 and Figure 5.12. The mean RBNS estimates were the highest with the highest variability as compared to all other models. The estimated quantiles exceeded \$600,000,000 for all the four frequency models. Even though these estimates will help the company solve solvency problem, the company will have problems saving enough fund for other managerial expenses. The results of the models with four and six predictors were found in between the best two and the worst performing model. The summary of empirical results and RBNS predictive distributions are found in Appendix C.

We observed that the neural network models with zero-inflated Poisson as the frequency distribution did better than those with zero-inflated Geometric distribution. Also, in the RBNS computation, we found that there was no much difference between the frequency models with N categorized and those with no categories in N.

Table 5.7 Results of random forest variable importance for amount of payment

	Importance	
Variable	%IncMSE	IncNodePurity
Major	-2.4942	2,379,551,390
Product	5.3325	3,458,362,812
Region	5.8513	5,314,328,084
Province	7.0042	10,889,545,247
Year	7.0413	16,101,558,664
Dev	8.0321	23,284,255,702
Losskind	23.0670	28,885,992,561

Table 5.8 The summary of empirical results from neural network models with three predictors

	Estimates				
Model	Mean	SE	90^{th} quantile	95^{th} quantile	99^{th} quantile
M1	530,319,309	11,565,460	545,495,519	548,749,735	562,728,966
M2	539,513,053	11,759,379	555,049,431	577,769,425	572,607,465
M3	558,843,634	12,159,232	574,794,804	578,275,106	592,836,894
M4	552,445,117	12,046,505	568,370,497	571,149,588	586,362,316
Observed RBNS	505,661,693				

5.4.3 Computation of RBNS reserves using classical generalized linear models

In this section, we shall present the results of using Poisson and gamma distributions to model the severity of payments instead of neural networks.

The frequency models were the same models used in the analyses with neural

Table 5.9 The summary of empirical results from neural network models with five predictors

	Estimates				
Model	Mean	SE	90^{th} quantile	95^{th} quantile	99^{th} quantile
M1	531,807,224	28,435,557	569,509,978	576,438,240	593,410,769
M2	540,993,806	29,495,414	579,546,498	586,687,560	606,694,182
M3	559,975,715	29,929,165	598,861,955	608,333,631	624,585,820
M4	554,084,138	30,527,374	593,685,766	601,816,313	621,426,401
Observed RBNS	505,661,693				

Table 5.10 The summary of empirical results from neural network models all seven predictors

	Estimates				
Model	Mean	SE	90^{th} quantile	95^{th} quantile	99^{th} quantile
M1	544,696,791	97,318,043	616,266,987	704,147,352	900,034,815
M2	547,539,995	93,383,043	616,007,857	701,382,754	886,757,581
M3	568,622,284	98,933,766	638,948,968	735,199,566	926,523,991
M4	563,906,332	98,428,369	635,514,059	725,393,100	922,394,321
Observed RBNS	505,661,693				

networks. Poisson and gamma generalized linear models with logarithm link functions were used as severity models. In order to compare to the best neural network model, the predictors used in the model were the best three variables indicated by the random forest variable important measures.

The results of RBNS estimation for a GLM-Poisson severity model are given in Table 5.11 and Figure 5.13 while that of GLM-gamma severity model are shown in Table 5.12 and Figure 5.14. Comparing the results to Table 5.8 and Figure

5.10, it is very clear that using neural network with the three predictors gave better RBNS estimates than using the two classical generalized linear models to predict the severity of payments. That is, in the case of the classical GLM severity models, the quantiles are very high with most of them above $600,000,000.

Table 5.11 The summary of empirical results from GLM-Poisson models with three predictors

	Estimates				
Model	Mean	SE	90^{th} quantile	95^{th} quantile	99^{th} quantile
M1	549,779,584	101,214,420	649,109,226	717,422,497	817,422,720
M2	558,977,871	102,511,508	662,563,340	726,408,626	830,984,191
M3	575,989,376	107,193,473	680,812,222	763,739,059	858,785,326
M4	564,338,832	105,277,059	677,560,686	746,119,262	850,368,847
Observed RBNS	505,661,693				

Table 5.12 The summary of empirical results from GLM-gamma models with three predictors

	Estimates				
Model	Mean	SE	90^{th} quantile	95^{th} quantile	99^{th} quantile
M1	573,801,311	118,784,112	729,567,388	783,409,682	991,367,345
M2	584,007,156	120,358,610	745,374,379	793,632,156	997,914,405
M3	604,757,060	125,895,804	767,593,784	829,266,394	1,055,526,917
M4	590,170,606	123,523,685	761,423,935	815,501,457	1,026,604,139
Observed RBNS	505,661,693				

5.5 Comparison of results

In this section we summarise the results of the numerical analyses by comparing the performance of the classical and the neural network loss reserving models. Based on available data, the observed RBNS amount was $505,661,693 (obtained from the validation data), but we can reasonably think that this amount would be closer $560,000,000 if we could observe more years. The observed IBNR amount was 56,703,282 for a total reserve amount of $562,364,975.

In contrast to collective approaches, individual methods used micro-covariates. This allows us to distinguish between IBNR claims and RBNS claims. As previously mentioned, in this thesis we mainly focus on the modeling of the RBNS reserve. Nevertheless, since we have both the observed IBNR and RBNS reserves in our dataset, we can reasonably evaluate the performance of both collective and individual models. For the collective models, their estimate of reserve will be compared to the observed total reserve of $562,364,975 while for the individual model, we will compare their estimates to observed RBNS. Reasonable practices, at least in North America, generally require a reserve amount given by a high quantile of the reserve's predictive distribution.

The summary of performance of models under the collective and individual approaches is given in Table 5.13. The **Mack** and **GLM-Poisson** models are both collective models. The reserve amount obtained from these are too high (between $630,000,000 and $680,000,000) compared to the observed total value of $562,364,975 which may get closer to $600,000,000 in the worst scenario. Furthermore, these models not take into consideration claim specific features.

The individual models, **ZIP-Poisson** and **ZIP-Gamma**, are the two best models obtained from the analyses when we used classical GLMs as severity models

instead of neural network. Their frequency model is zero-inflated Poisson while their severity models are standard Poisson and gamma models, respectively. The results indicate that these models are very expensive since their high quantiles are above $650,000,000 when we expect between $520,000,000 and $560,000,000 for real RBNS reserves.

For our proposed model, **ZIP-NN3** and **ZIP-NN5** performed better than the rest. Both models are individual models with zero-inflated Poisson as the frequency model. While **ZIP-NN3** uses a neural network severity model with three predictors, **ZIP-NN5** uses a neural network severity model with five predictors. The high quantile of **ZIP-NN3** are between $540,000,000 and $560,000,000 for RBNS reserves while that of **ZIP-NN5** are quite higher but below $600,000,000. Since the real RBNS may get closer to $560,000,000, we find the estimates from our model more applicable to the insurance company than those from the classical models.

Table 5.13 Comparing empirical results of classical and neural network models

	Estimates				
Model	Mean	SE	90^{th} quantile	95^{th} quantile	99^{th} quantile
Mack	578,064,728	42,788,129	632,426,617	648,841,907	678,979,299
GLM-Poisson	578,131,075	42,714,561	632,601,809	648,690,549	676,779,315
ZIP-Poisson	549,779,584	101,214,420	649,109,226	717,422,497	817,422,720
ZIP-Gamma	573,801,311	118,784,112	729,567,388	783,409,682	991,367,345
ZIP-NN3	530,319,309	11,565,460	545,495,519	548,749,735	562,728,966
ZIP-NN5	531,807,224	28,435,557	569,509,978	576,438,240	593,410,769
Observed RBNS	505,661,693				
Observed IBNR	56,703,282				
Observed Total	562,364,975				

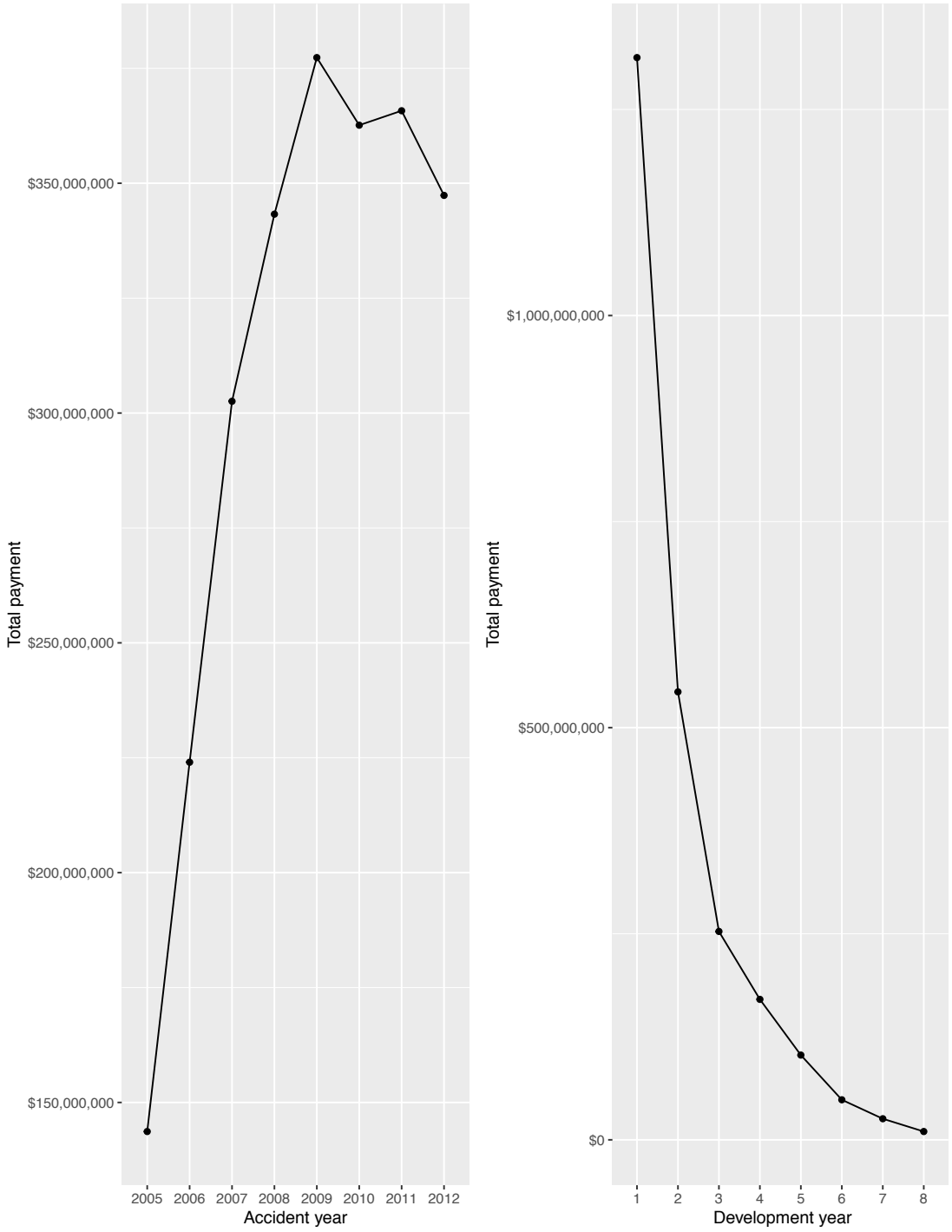

Figure 5.6 Trends of total payments in millions of dollars by accident and development years

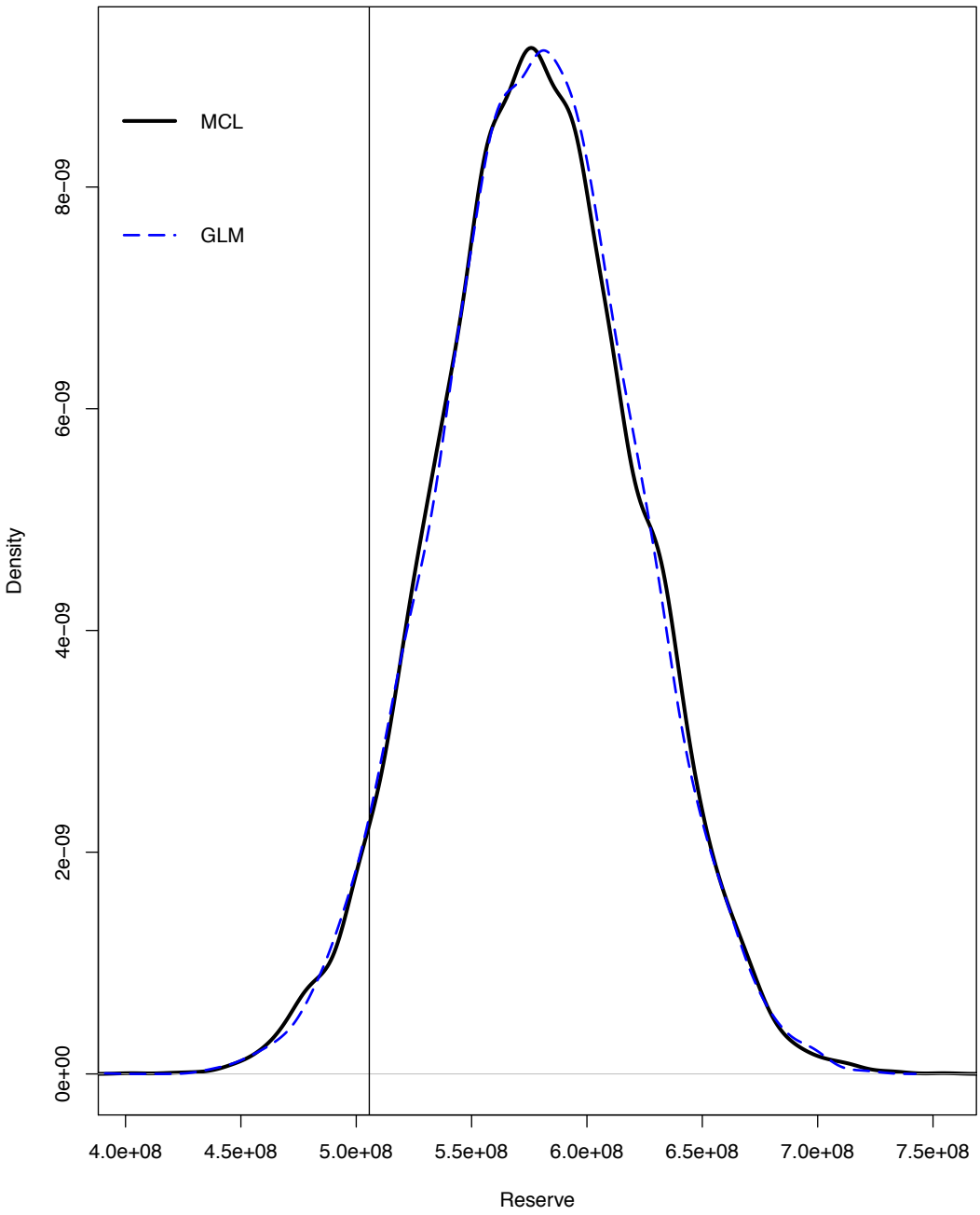

Figure 5.7 Predictive distribution of classical models

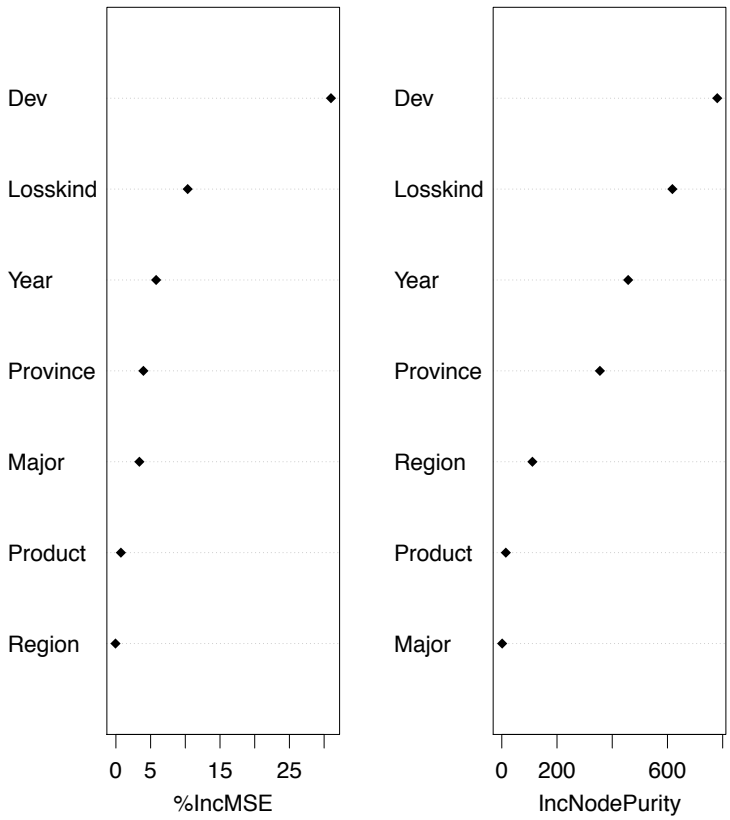

Figure 5.8 Random forest variable importance plot for number of payment

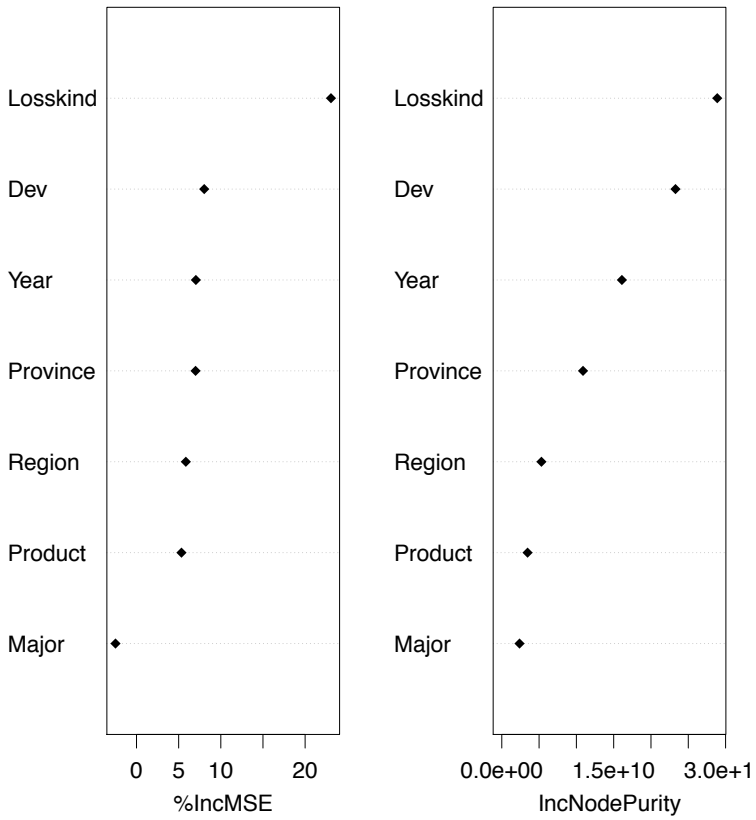

Figure 5.9 Random forest variable importance plot for amount of payment

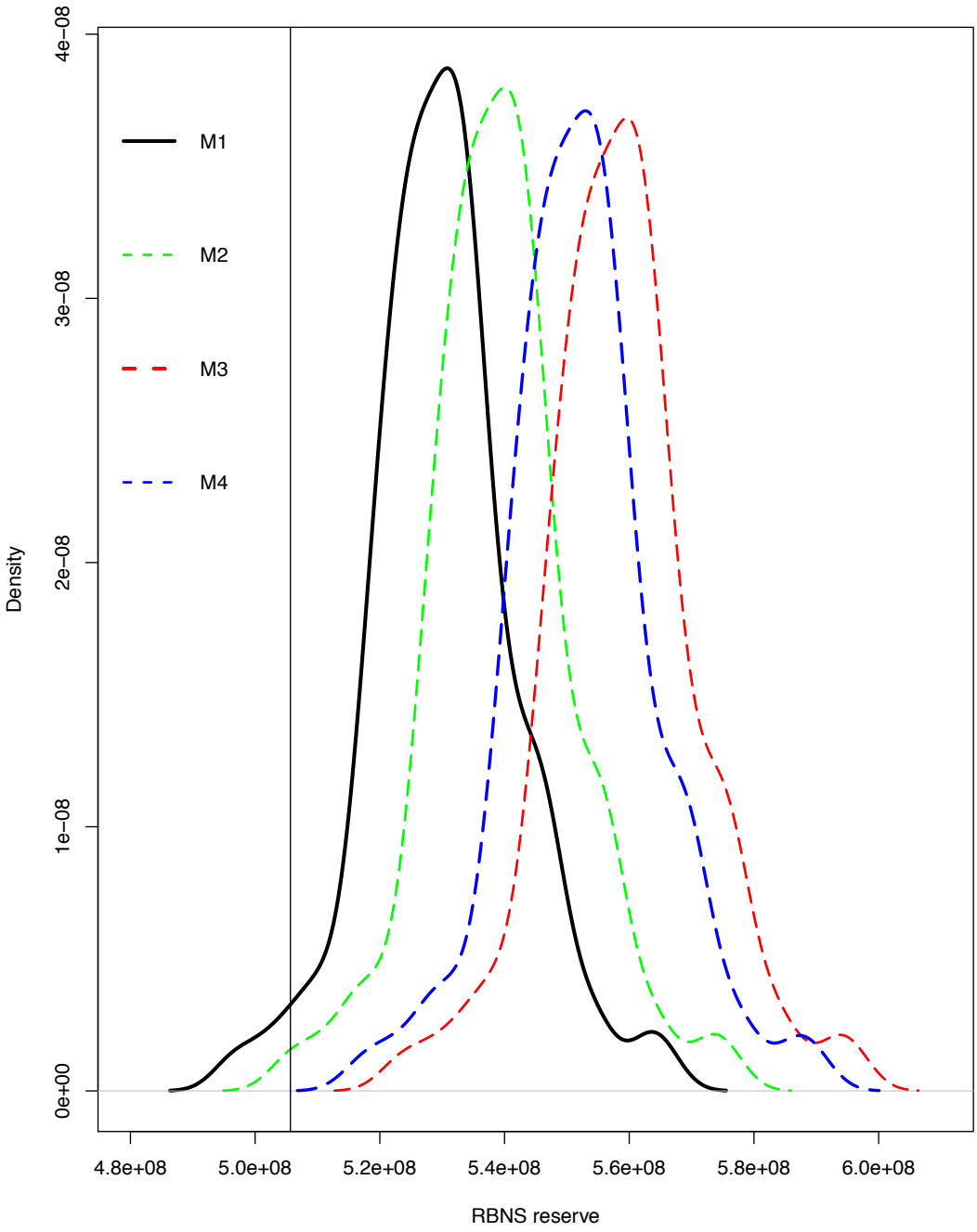

Figure 5.10 RBNS predictive distributions for neural network models with three predictors.

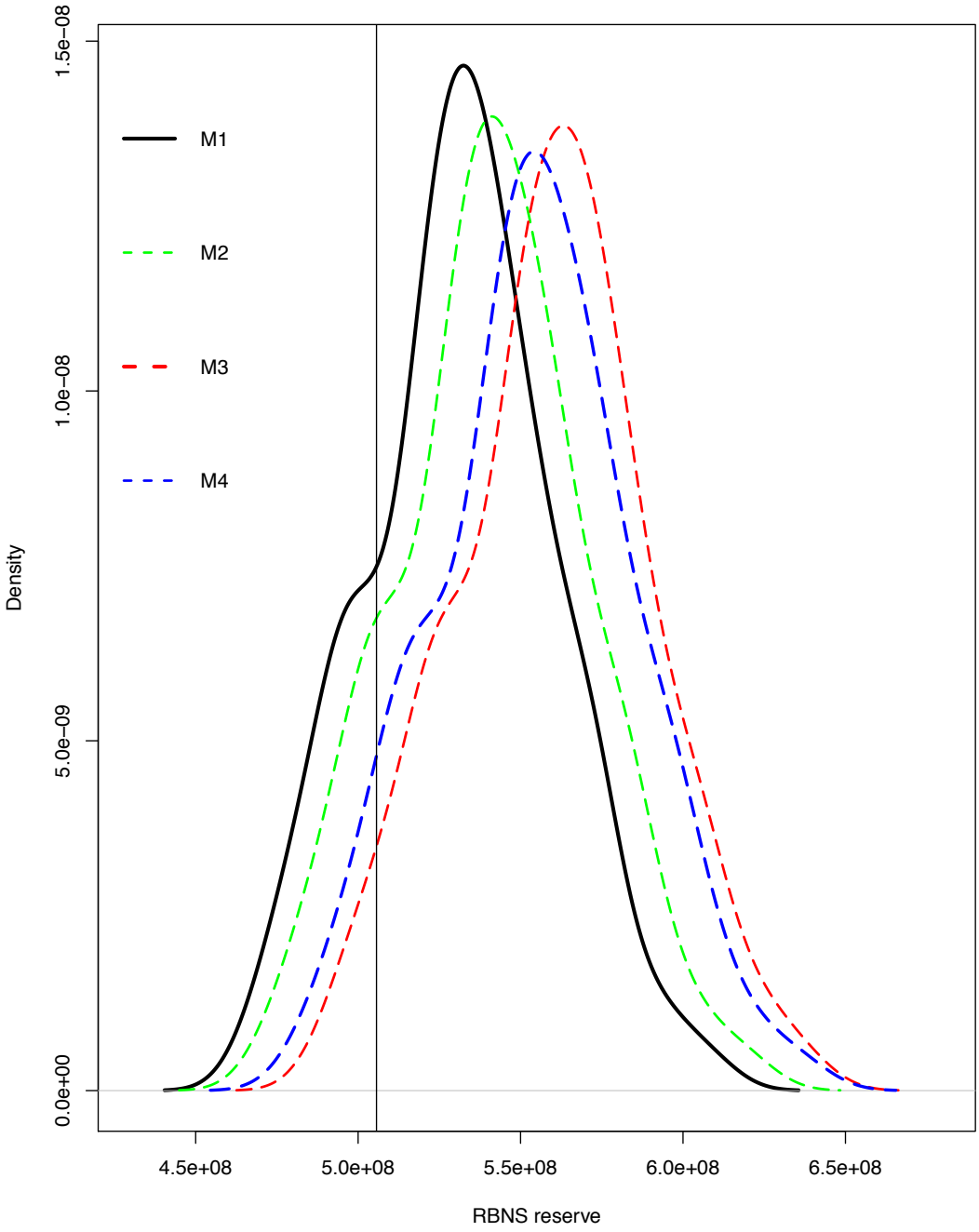

Figure 5.11 RBNS predictive distributions for neural network models with five predictors.

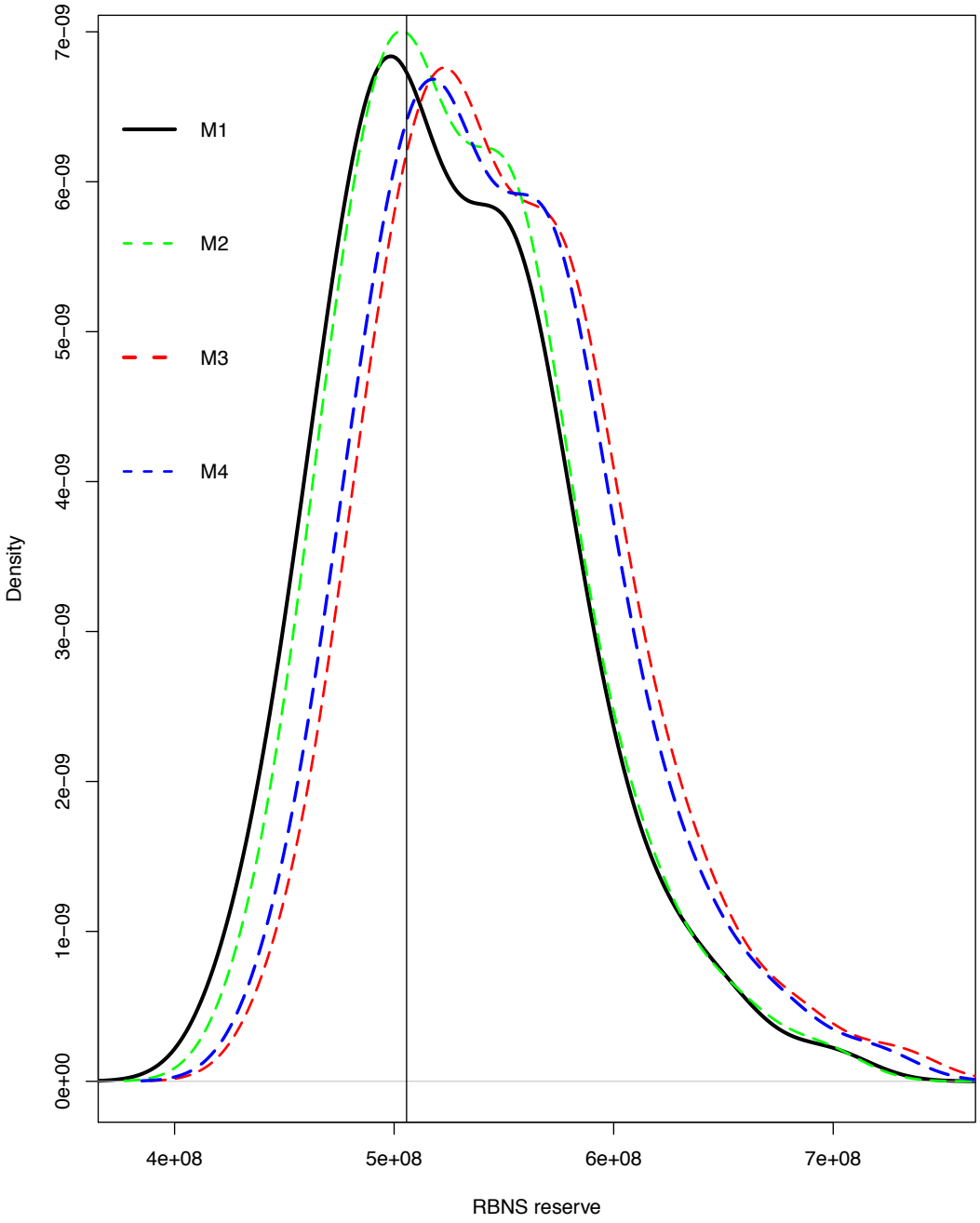

Figure 5.12 RBNS predictive distribution for neural network models with all seven predictors.

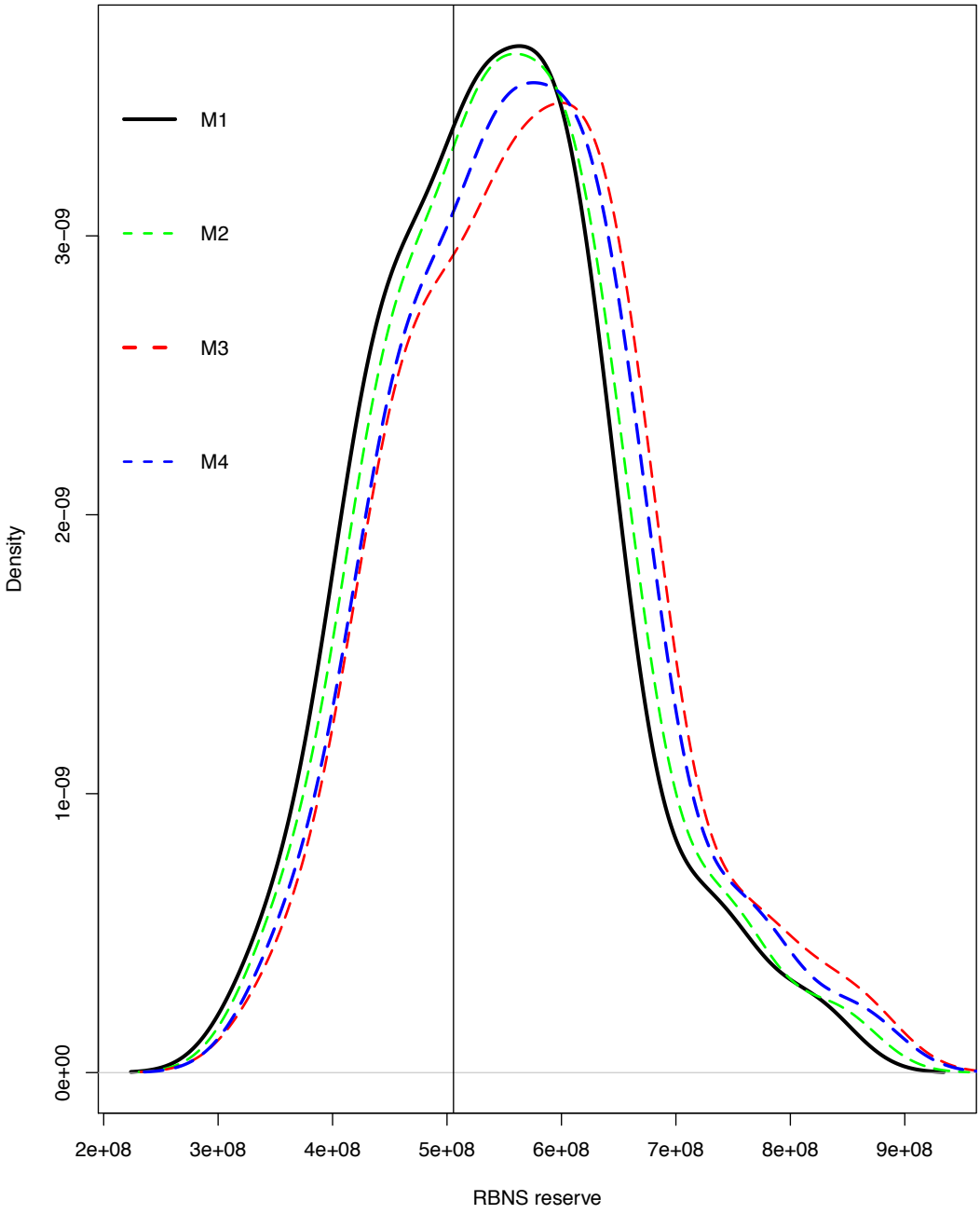

Figure 5.13 RBNS predictive distributions for GLM-Poisson severity models with three predictors.

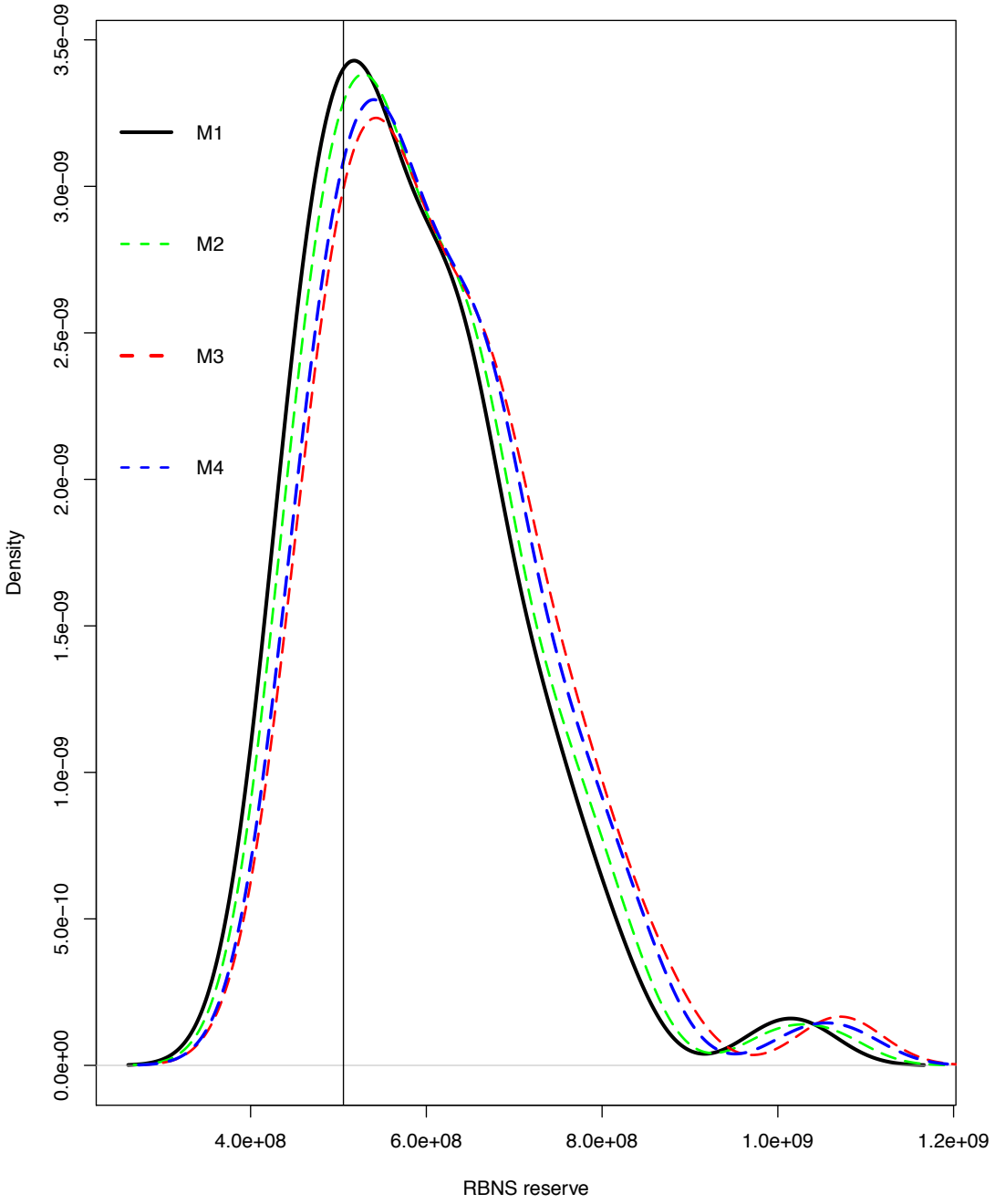

Figure 5.14 RBNS predictive distributions for GLM-gamma severity models with three predictors.

CONCLUSION

In this study, we proposed a new frequency-severity model in an individual framework for loss reserving in general insurance. Our model uses generalised linear model (GLM) to model the frequency of claims and neural networks to predict the severity of claims.

It has been shown in this study that the individual framework we adopted gives us the advantage of incorporating more information on claims in the loss reserving model as compared to the collective loss reserving models. In addition to that, the frequency-severity model allowed us to investigate the factors that influence frequency and severity of claims separately. It is interesting to note that the machine learning approach we adopted enabled us to relax the distributional assumptions that many classical models make on severity of claims.

Our analyses indicated that a zero-inflated Poisson model is the best model that fits the frequency of claims for our data. The most important factors that influence the frequency of claims are the development year, loss kind, year of the loss and the province. The zero-inflated Poisson model had the greatest ability to capture the large mass probability of zero in the data. Also, the neural network model performed better than the classical models by predicting RBNS reserves that are more realistic to the operations of the insurance company. The most important factors that influence the severity of claims are the development year, loss kind, year of the loss, the province and region. It is impressive to note that unlike some classical methods of loss reserving such as the chain-ladder, we have established reserve not only by dwelling on best estimates but also setting them

at very high quantiles of our predictive distribution in order to incorporate risk of faced by the insurer. Our approach complies with modern solvency requirements used in practise including guidelines stipulated by Office of the Superintendent of Financial Institutions(OSFI).

In future studies, we would like to consider alternative models for the frequency of claims. It will be interesting to study how decision tree algorithms such as random forest and gradient boosting can improve the prediction of frequency of claims. Another area of future work will be to study how different architecture of neural network will affect the results of our model. Therefore, we will adopt a more complex structure of neural network by increasing the number of hidden layers and interior nodes. It will be interesting to investigate more into model performance in terms of training and validation loss. We shall also consider the possibility of incorporating the dependence between frequency and severity of claims and possibly developing a neural network architecture for both of them. Furthermore, we shall pursue the second aspect of the total reserve which is obtaining the IBNR reserve and research into alternative ways of reflecting risk of the insurer into the model. This will involve finding a suitable model for the occurrence of a claim and the delay of its declaration to the insurer.

Finally, we will point out that neural networks has proven to be a great tool for prediction. However, they takes more computational time and require maximum attention to get the best out of them.

APPENDIX A

DATA DESCRIPTION AND SUMMARY STATISTICS

Table A.1 Description of variables in the training set

Variable code	Description
CLAIMNUMBER	Number assigned to claim
LOSSYEAR	Year an accident occurred
CALYEAR	Calender year an accident occurred
REPORTYEAR	Year an accident was declared
EXPOSURENUMBER	Exposure number
INTC_RISK_PROVINCE_CODE	Province code
LOSSKIND_NAME	Kind of loss
VALUATION_REGION	Valuation region
CGIC_PRODUCT	Type of general insurance product
MAJOR_EVENT	Major event
FERMEan	Year claim file was closed
ORI	Accident or occurence year
DEV	Development year
N	Number of payments
PAID	Amount of payment
claim2	Code for the payment
STATUT	Status of claim

Table A.2 Description of levels for the variable representing provinces

Code	Description
AB	Alberta
BC	British Columbia
IL	Illinois
MB	Manitoba
N/A	Not Available
NB	New Brunswick
NL	Newfoundland
NS	Nova Scotia
NT	Nunavut
ON	Ontario
PE	Prince Edward Island
QC	Quebec
SK	Saskatchewan
YT	Yukon

Table A.3 List of levels for the variable representing loss kinds

Loss kinds
Animal collision
Collision
Collision with parked vehicle
Damage by water
Earthquake
Emergency road service
Explosion
Fire
Flying
Glass other
Glass windshield
Hail
Hit and run
Hit pedestrian
In transit Transport
Lightning
Multi-vehicle

Table A.4 List of levels for the variable representing valuation regions

Valuation regions
Atlantic
Not Available
Ontario
Other
Quebec
West

Table A.5 Table showing payments made for each product type across all provinces

	Product type		
Province	Auto	Corporate commercial auto	SR4 commercial auto
AB	682,057	3,389	13,611
BC	0	77	18
IL	0	4	0
MB	0	241	6
N/A	6	0	0
NB	94,627	471	1,169
NL	80,952	206	1,240
NS	85,268	833	825
NT	5,558	87	0
ON	1,115,552	2,766	2,111
PE	44,151	326	812
QC	2,148	7	46
SK	204,632	1,723	1,101
YT	5,212	5	0
Total	2,320,163	10,135	20,939

Table A.6 Table showing payments made for each product type for all loss kinds

	Product type		
Loss kind name	Auto	Corporate Commercial Auto	SR4 Commercial Auto
Animal Collision	134,343	571	1,675
Collision	10	0	0
Collision With Anything	355	0	0
Collision With Parked Vehicle	15,477	48	187
Damage By Water	390	0	1
Earthquake	2	0	0
Emergency Road Service	453	0	0
Explosion	286	0	6
Fire	13,147	95	316
Flying/Falling Object	6,640	13	35
Glass Other	26,203	49	106
Glass Windshield	297,237	311	1,372
Hail	50,484	171	509
Hit and Run	70,470	318	287
Hit Pedestrian	2,920	12	16
In Transit	314	0	3
Lightning	513	4	1
Multi-vehicle	427,850	1,808	3,956
Not Available	773,579	4,515	8,079
Other	30,484	540	451
Riot	65	0	5
Rising Water	1,758	0	15
Single-vehicle	157,753	1,294	1,831
Smoke	63	0	0
Total Theft	65,059	133	1,208
Vandalism	73,572	172	344
Vehicle Break-In	14,323	23	131
Wind	15,729	34	123
Windshield Repair	61,951	14	100
Windshield Replace	78,733	10	182
Total	2,320,163	10,135	20,939

Table A.7 Summary of payments by accident years

	Summary			
Accident year	Freq	Total	Mean	Max
2005	362,985	143,674,854	395.8149	548,277.7
2006	427,057	224,033,492	524.5986	1,587,917.0
2007	426,368	302,561,058	709.6242	2,131,800.0
2008	368,637	343273375	931.1962	2,024,986.0
2009	301,377	338,912	1,250.2780	2,632,005.0
2010	214,989	362,628,620	1,686.7310	2,905,257.0
2011	162,851	365,774,763	2,246.0701	7,207,200.0
2012	86,546	347,356,650	4,013.5494	2,264,906.0

Table A.8 Summary of payments by development years

	Summary			
Development year	Freq	Total	Mean	Max
1	521,726	1,312,910,064	2,,516.4743	7,207,200
2	463,492	543,442,596	1,,172.4962	2,632,005
3	394,358	252,802,783	641.0490	2,905,257
4	329,729	170,409,370	516.8164	2,264,906
5	262,376	102,801,379	391.8094	2,445,469
6	194,093	48,591,603	250.3522	1,873,053
7	124,643	25,599,267	205.3807	3,150,850
8	60,820	10,084,663	165.8116	1,654,570

APPENDIX B

SUPPLEMENTARY RESULTS FOR CLASSICAL MODELS

Table B.1 Further details of results of Mack model

	Quantity					
Accident year	Latest	Dev.To.Date	Ultimate	IBNR	Mack.S.E	CV(IBNR)
2005	356,599,161	1.0000000	356,599,161	0	0	NaN
2006	364,347,450	0.9717199	374,951,108	10,603,658	2,086,343	0.19675693
2007	379,129,475	0.9367267	404,738,617	25,609,142	3,380,889	0.13201885
2008	393,133,642	0.8939633	439,764,860	46,631,218	4,961,768	0.10640442
2009	333,304,552	0.8287362	402,184,129	68,879,576	7,886,699	0.11449981
2010	264,883,851	0.7425942	356,700,666	91,816,816	7,683,640	0.08368446
2011	217,358,380	0.6343246	342,661,098	125,302,718	12,349,594	0.09855807
2012	157,885,214	0.4313658	366,012,378	208,127,164	22,775,000	0.10942829

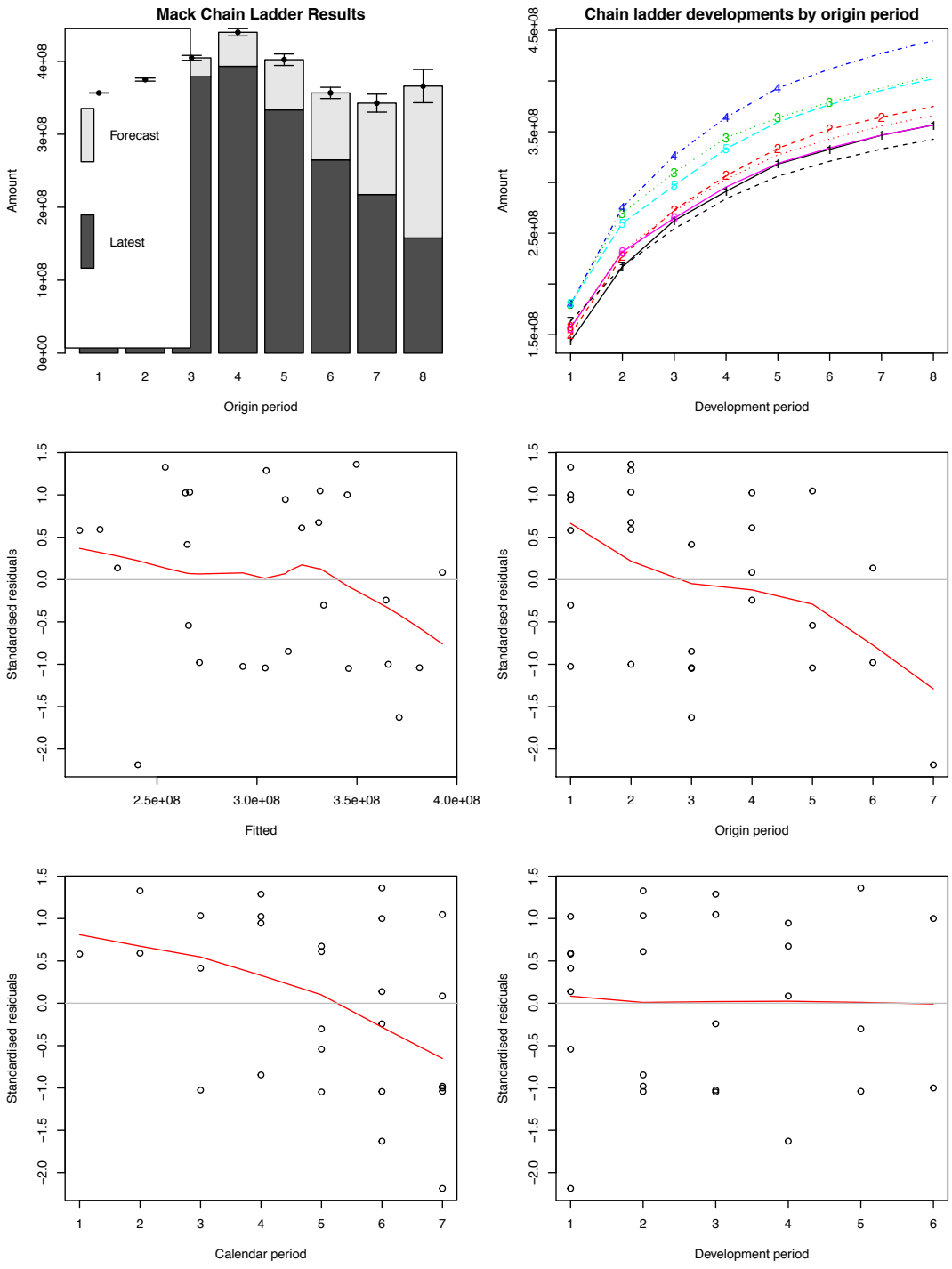

Figure B.1 Supplementary plots from the Mack model

Table B.2 Further details of results of GLM-Poisson

	Quantity					
Accident year	Latest	Dev.To.Date	Ultimate	IBNR	Mack.S.E	CV(IBNR)
2006	364,347,450	0.9717199	374,951,108	10,603,658	4,271,308	0.40281454
2007	379,129,475	0.9367267	404,738,617	25,609,142	6,461,415	0.25230890
2008	393,133,642	0.8939633	439,764,860	46,631,218	8,119,139	0.17411380
2009	333,304,552	0.8287362	402,184,128	68,879,576	10,101,302	0.14665162
2010	264,883,851	0.7425942	356,700,667	91,816,816	11,285,245	0.12291044
2011	217,358,380	0.6343246	342,661,098	125,302,718	14,389,818	0.11484043
2012	157,885,214	0.4313658	366,012,378	208,127,164	21,373,674	0.10269526

APPENDIX C

SUPPLEMENTARY RESULTS FOR NEURAL NETWORK MODEL

C.1 Backpropagation method

As mentioned in section 3.3 of Chapter 3, The backpropagation method computes the partial derivative, $\frac{\partial a^{(L)}}{\partial w^J_{j,k}}$, recursively. Let $\#J$ be the number of activations in the level-J layer.

$$\begin{aligned}
\frac{\partial a^{(L)}}{\partial w^J_{j,k}} &= \sum_{m=1}^{\#J} \frac{\partial a^{(L)}}{\partial a^{(J)}_m} \frac{\partial a^{(J)}_m}{\partial w^J_{j,k}} \\
&= \sum_{m=1}^{\#J} \frac{\partial a^{(L)}}{\partial a^{(J)}_m} \mathbb{I}_{\{j=m\}} \frac{\partial a^{(J)}_m}{\partial w^J_{j,k}} \\
&= \frac{\partial a^{(L)}}{\partial a^{(J)}_j} \frac{\partial a^{(J)}_j}{\partial w^J_{j,k}}.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\frac{\partial a^{(J)}_j}{\partial w^J_{j,k}} &= \begin{cases} X^{(k)}, & \text{if } J = 1 \\ Z^{(J-1)}_k, & \text{if } J > 1 \end{cases} \\
&= \begin{cases} X^{(k)}, & \text{if } J = 1 \\ 1 & \text{if } J > 1, k = 0 \\ h\left(a^{(J-1)}_k\right), & \text{if } J > 1, k > 0. \end{cases}
\end{aligned}$$

where $Z_k^{(J-1)}$ is the hidden unit k from the $(J-1)^{th}$ hidden layer.

For $t \geq 1$ and any m, k

$$a_m^{(t+1)} = w_{m,0}^{t+1} + \sum_{i=1}^{\#t} w_{m,i}^{t+1} h\left(a_i^{(t)}\right).$$

This implies that

$$\frac{\partial a_m^{(t+1)}}{\partial a_k^{(t)}} = w_{m,k}^{t+1} h'\left(a_k^{(t)}\right).$$

For the computation of $\frac{\partial a^{(L)}}{\partial a_j^{(J)}}$, one proceeds recursively.
For $s = 1, \ldots, L-1$ define

$$\mathcal{U}_k^{(s)} = \frac{\partial a^{(L)}}{\partial a_k^{(s)}} = \sum_\ell \frac{\partial a^{(L)}}{\partial a_\ell^{(s+1)}} \frac{\partial a_\ell^{(s+1)}}{\partial a_k^{(s)}}.$$

Thus, $\frac{\partial a^{(L)}}{\partial a_j^{(J)}} = \mathcal{U}_j^{(J)}$ which are computed recursively as follows:

$$\begin{aligned} \mathcal{U}_k^{(s)} &= \sum_\ell \frac{\partial a^{(L)}}{\partial a_\ell^{(s+1)}} \frac{\partial a_\ell^{(s+1)}}{\partial a_k^{(s)}} \\ &= \sum_\ell \mathcal{U}_\ell^{(s+1)} w_{\ell,k}^{s+1} h'\left(a_k^{(s)}\right) \\ &= h'\left(a_k^{(s)}\right) \sum_\ell \mathcal{U}_\ell^{(s+1)} w_{\ell,k}^{s+1} \end{aligned}$$

where the recursion is initialized with

$$\mathcal{U}_\ell^{(L)} = \mathbb{I}_{\{\ell=1\}}.$$

Table C.1 Results of the frequency models for the case when three predictors were used in the models

	No categories			Categories		
Model	Total_N	MSE	AIC	Total_N	MSE	AIC
Poisson	694,238	22.6713	2,660,486.75	755,478	17.018685	1,274,776.6
Neg_ Bin	1,271,341	96.8879	1,199,124.24	944,728	33.797776	1,064,957.8
Geometric	1,073,146	52.7373	1,577,537.10	885,771	25.752784	1,131,450.8
ZI_Poisson	543,643	4.5511	1,679,024.30	554,999	1.465527	999,719.6
ZI_Neg_bin	806,616	7.4278	1,096,272.41	560,869	5.521462	989,298.2
ZI_Geometric	626,731	5.0131	1,135,605.59	560,393	1.5163	989,298.4

Table C.2 Results of the frequency models for the case when four predictors were used in the models

	No categories			Categories		
Model	Total_N	MSE	AIC	Total_N	MSE	AIC
Poisson	716,136	23.6383	2,631,833	799,889	19.3509	1,271,079.08
Neg_bin	1,301,482	112.3888	1,194,393	1,077,011	39.4691	1,063,913.05
Geometric	1,043,983	54.6192	1,560,092	1,019,444	31.3420	1,129,050.93
ZI_Poisson	572,827	5.1257	1,610,757	553,163	1.4270	997,949.19
ZI_Neg_bin	1,458,030	19.5512	1,094,283	-	-	-
ZI_Geometric	668,326	5.6890	1,126,204	559,559	2.5596	988,228.51

Table C.3 Results of the frequency models for the case when six predictors were used in the models

	No categories			Categories		
Model	Total_N	MSE	AIC	Total_N	MSE	AIC
Poisson	708,692	23.2391	2,631,612.02	759,474	17.2922	1,270,642.79
Neg_bin	1,268,422	107.0670	1,194,282.30	940,108	32.86378	1,063,674.22
Geometric	1,009,461	50.3905	1,559,877.73	906,286	26.1466	1,128,691.53
ZI_Poisson	575,920	5.2129	1,610,250.06	566,440	1.4694	997,774.43
ZI_Neg_bin	-	-	-	-	-	
ZI_Geometric	680,665	5.7325	1,126,088.66	590,175	4.6160	988,051.68

Table C.4 The summary of empirical results from neural network models with four predictors

	Estimates				
Model	Mean	SE	90^{th} quantile	95^{th} quantile	99^{th} quantile
M1	534,758,333	33,697,982	579,119,568	591,906,417	613,086,257
M2	543,889,974	34,709,137	592,041,701	601,467,273	625,566,004
M3	562,610,589	34,973,903	608,961,372	624,359,476	642,542,826
M4	557,226,667	35,955,026	607,178,573	616,946,128	641,782,447
Observed RBNS	505,661,693				

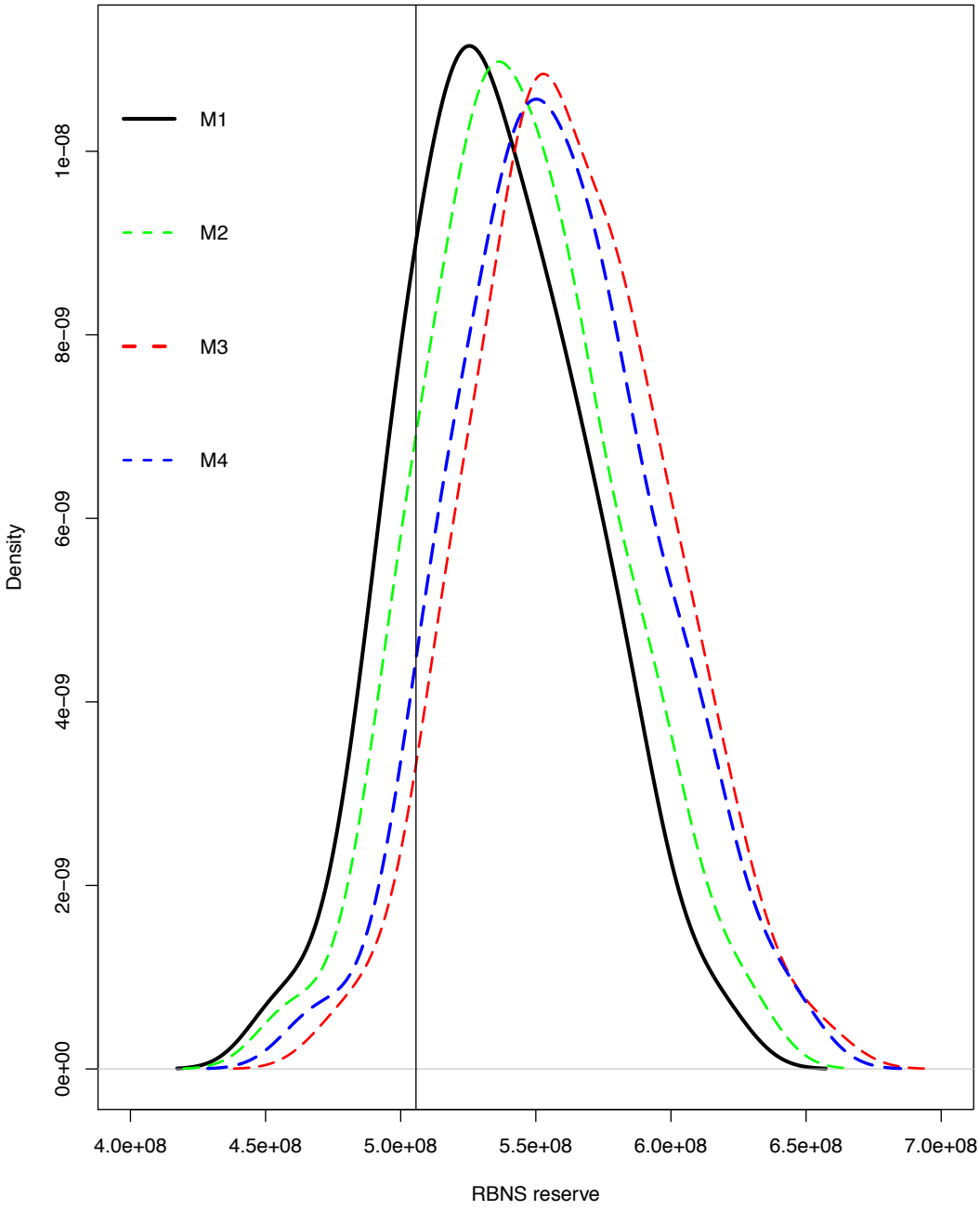

Figure C.1 RBNS predictive distributions for neural network models with four predictors.

Table C.5 The summary of empirical results from neural network models with six predictors

	Estimates				
Model	Mean	SE	90^{th} quantile	95^{th} quantile	99^{th} quantile
M1	567,698,020	33,571,852	607,271,280	624,463,133	643,142,602
M2	577,804,778	34,240,098	617,047,274	634,220,837	654,917,686
M3	597,633,295	35,076,877	638,264,291	655,426,601	677,712,163
M4	592,049,864	35,297,411	632,358,909	677,712,163	671,382,091
Observed RBNS	505,661,693				

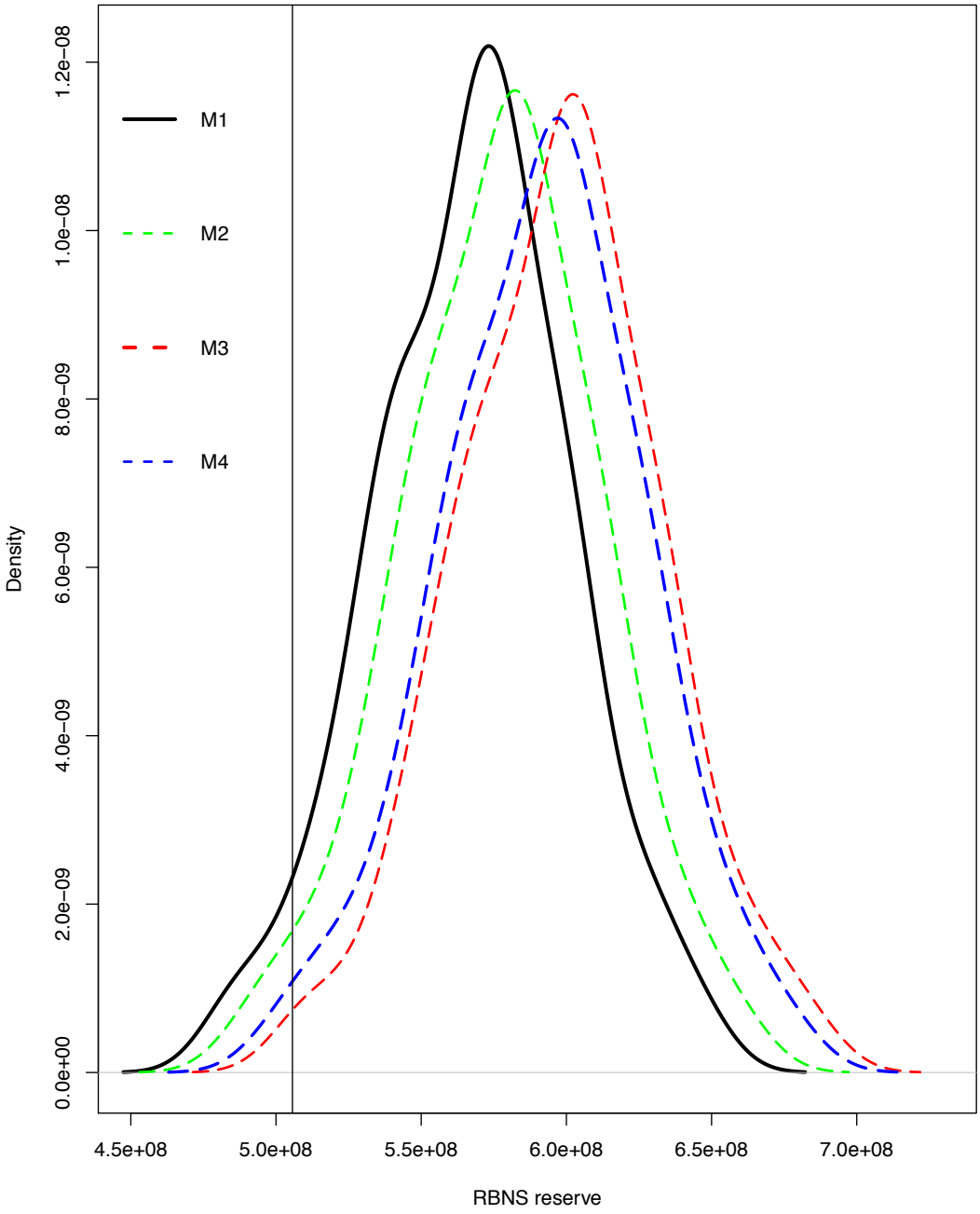

Figure C.2 RBNS predictive distribution for neural network models with six predictors.

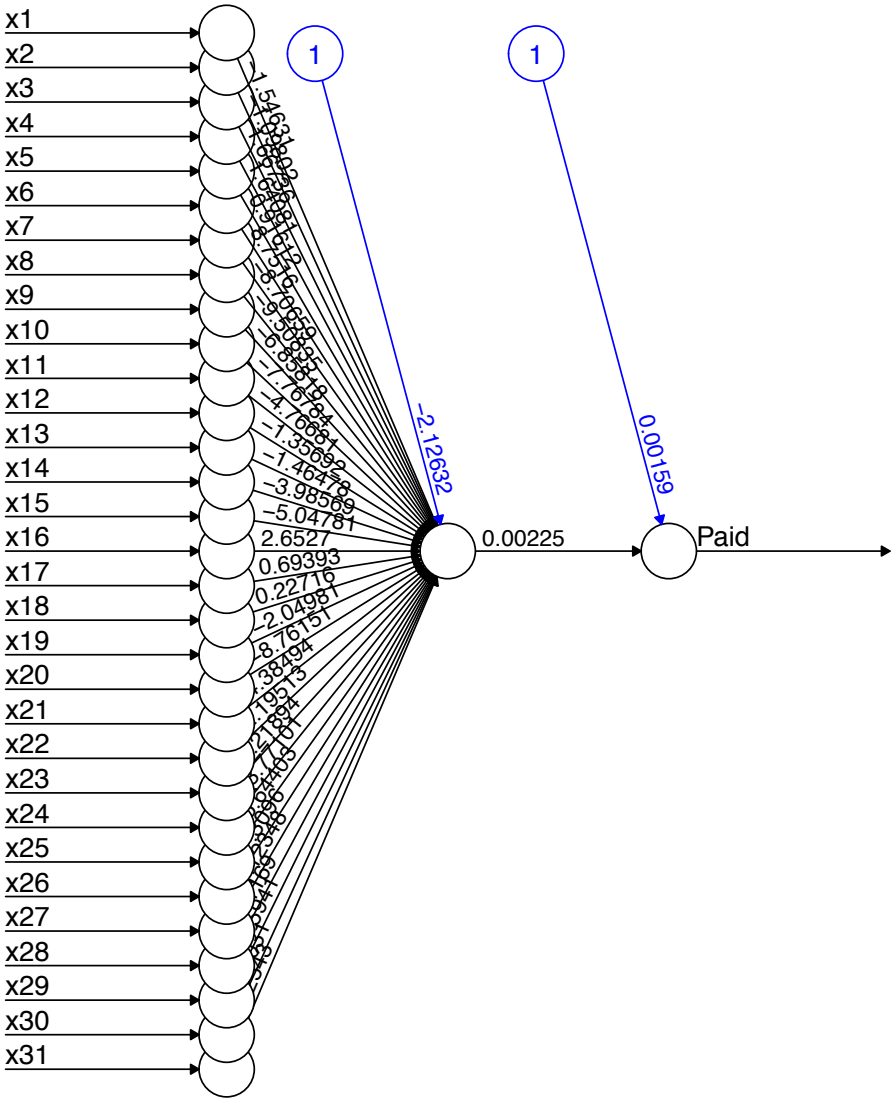

Figure C.3 Neural network plot of estimated parameters for three predictors.

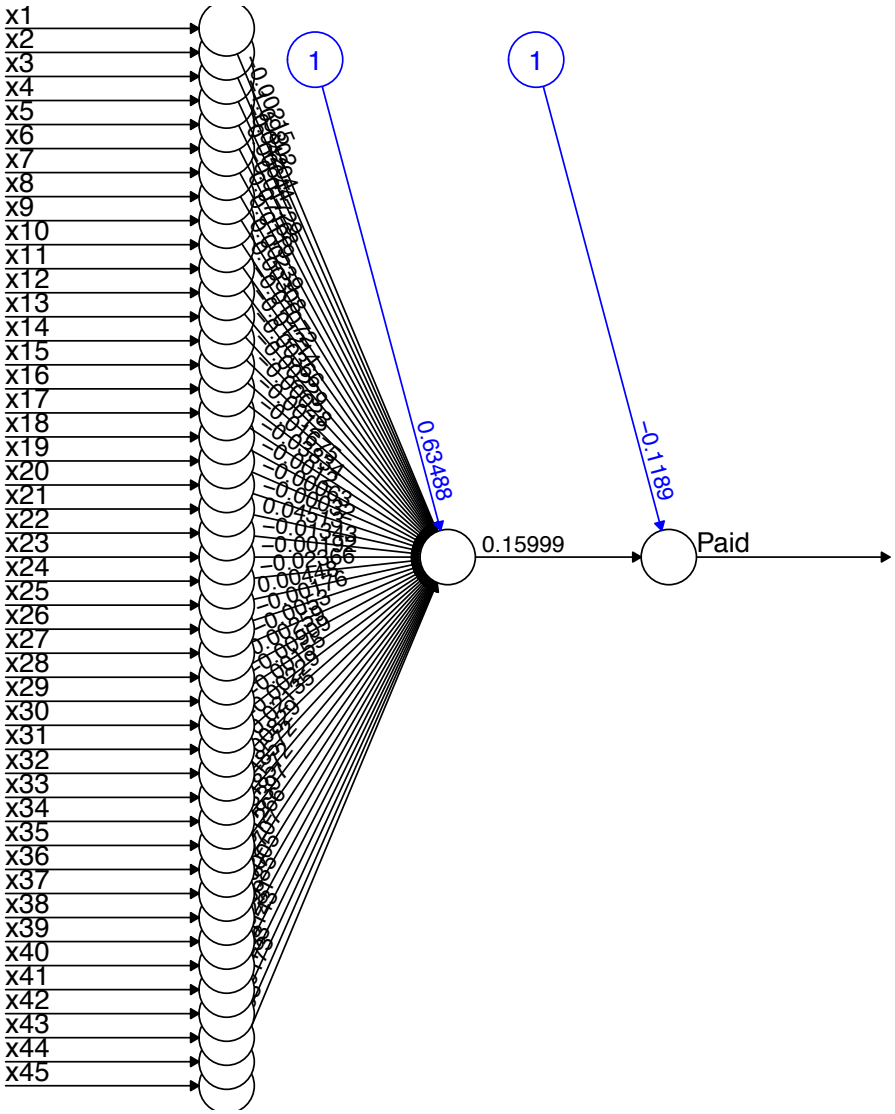

Figure C.4 Neural network plot of estimated parameters for five predictors.

BIBLIOGRAPHY

Antonio, K. & Plat, R. (2014). Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal*, *2014*(7), 649–669.

Arjas, E. (1989). The claims reserving problem in non-life insurance: Some structural ideas. *ASTIN Bulletin*, *19*(2), 139–152.

Barnett, G. & Zehnwirth, B. (2000). Best estimates for reserves. In *Proceedings of the Casualty Actuarial Society*, volume 87, pp. 245–321.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bornhuetter, R. L. & Ferguson, R. E. (1972). The actuary and ibnr. In *Proceedings of the casualty actuarial society*, volume 59, pp. 181–195.

Boucher, J.-P. & Davidov, D. (2011). On the importance of dispersion modeling for claims reserving: An application with the tweedie distribution. *Variance*, *5*(2), 158–172.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Cameron, A. C. & Trivedi, P. K. (2013). *Regression analysis of count data*, volume 53. Cambridge university press.

Charpentier, A. & Pigeon, M. (2016). Macro vs. micro methods in non-life claims reserving (an econometric perspective). *Risks*, *4*(2), 12.

Denuit, M. & Trufin, J. (2017). Beyond the tweedie reserving model: the collective approach to loss development. *North American actuarial journal*, *21*(4), 611–619.

Denuit, M. & Trufin, J. (2018). Collective loss reserving with two types of claims in motor third party liability insurance. *Journal of computational and applied mathematics*, *335*, 168–184.

Duval, F. & Pigeon, M. (2019). Individual loss reserving using a gradient boosting-based approach. *Risks*, *7*(3), 79.

England, P. D. & Verrall, R. J. (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal*, *8*(3), 443–518.

Frees, E. W. & Valdez, E. A. (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association*, *103*(484), 1457–1469.

Friedland, J. (2010). Estimating unpaid claims using basic techniques. In *Casualty Actuarial Society*, volume 201.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367–378.

Gabrielli, A., Richman, R. & Wüthrich, M. V. (2019). Neural network embedding of the over-dispersed poisson reserving model. *Scandinavian Actuarial Journal*, pp. 1–29.

Gabrielli, A. & Wüthrich, M. V. (2018). An individual claims history simulation machine. *Risks*, *6*(2), 29.

Gogol, D. (1993). Using expected loss ratios in reserving. *Insurance: Mathematics and Economics*, *12*(3), 297–299.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep learning*. MIT press.

Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, *39*(3), 3659–3667.

Halliwell, L. J. (1996). Loss prediction by generalized least squares. In *Proceedings of the Casualty Actuarial Society*, volume 83, pp. 436–489.

Hamer, M. D. (1999). Loss prediction by generalized least-squares-discussion. In *Proc. CAS*, volume 86, pp. 748–763.

Hilbe, J. M. (2014). *Modeling count data*. Cambridge University Press.

Huang, J., Qiu, C., Wu, X. & Zhou, X. (2015). An individual loss reserving model with independent reporting and settlement. *Insurance: Mathematics and Economics*, *64*, 232–245.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Kunkler, M. (2004). Modelling zeros in stochastic reserving models. *Insurance: Mathematics and Economics*, *34*(1), 23–35.

Kuo, K. (2019). Deeptriangle: A deep learning approach to loss reserving.

Risks, *7*(3), 97.

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *Canadian Journal of Statistics*, *15*(3), 209–225.

Leadbetter, D. & Dibra, S. (2008). Why insurers fail: The dynamics of property and casualty insurance insolvency in canada. *The Geneva Papers on Risk and Insurance-Issues and Practice*, *33*(3), 464–488.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.

Lee, S. C. & Lin, S. (2018). Delta boosting machine with application to general insurance. *North American Actuarial Journal*, *22*(3), 405–425.

Liu, H. & Verrall, R. (2009). Predictive distributions for reserves which separate true ibnr and ibner claims. *ASTIN Bulletin*, *39*(1), 35–60.

Loh, W. Y. (2014). Fifty years of classification and regression trees. *International Statistical Review*, *82*(3), 329–348.

Lopez, O., Milhaud, X., Thérond, P. et al. (2016). Tree-based censored regression with applications in insurance. *Electronic journal of statistics*, *10*(2), 2685–2716.

Ludwig, A., Schmeißer, C. & Thänert, K. (2011). Reserve risk estimation in a linear model. *Zeitschrift für die gesamte Versicherungswissenschaft*, *100*(4), 493–516.

Mack, T. (1991). A simple parametric model for rating automobile insurance or estimating ibnr claims reserves. *ASTIN Bulletin*, *21*(1), 93–109.

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

Mack, T. (1999). The standard error of chain-ladder reserve estimates: Recursive calculation and inclusion of a tail factor. *ASTIN Bulletin*, *29*(2), 361–366.

McCullagh, P. & Nelder, J. A. (1989). Monographs on statistics and applied probability. *Generalized linear models*, *37*.

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115–133.

Meyers, G. G. (2007). Estimating predictive distributions for loss reserve

models. *Variance*, *1*(2), 248–272.

Miranda, M. D. M., Nielsen, J. P. & Verrall, R. (2012). Double chain-ladder. *ASTIN Bulletin*, *42*(1), 59–76.

Morgan, J. N. & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, *58*(302), 415–434.

Mulquiney, P. (2006). Artificial neural networks in insurance loss reserving. In *9th Joint International Conference on Information Sciences (JCIS-06)*. Atlantis Press.

Murphy, D. M. (1994). Unbiased loss development factors. In *CAS Forum*, volume 1, pp. 183.

Nelder, J. A. & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384.

Nielsen, M. A. (2015). *Neural networks and deep learning*, volume 2018. Determination press San Francisco, CA, USA.

Norberg, R. (1993). Prediction of outstanding liabilities in non-life insurance 1. *ASTIN Bulletin*, *23*(1), 95–115.

Norberg, R. (1999). Prediction of outstanding liabilities. *Astin Bulletin*, *29*(1), 5–27.

Olbricht, W. (2012). Tree-based methods: a useful tool for life insurance. *European Actuarial Journal*, *2*(1), 129–147.

Panjer, H. H. & Willmot, G. E. (2008). *Loss models: From data to decisions*. John Wiley & Sons.

Pigeon, M., Antonio, K. & Denuit, M. (2013). Individual loss reserving with the multivariate skew normal framework. *ASTIN Bulletin*, *43*(3), 399–428.

Quan, Z. & Valdez, E. A. (2018). Predictive analytics of insurance claims using multivariate decision trees. *Dependence Modeling*, *6*(1), 377–407.

Quarg, G. & Mack, T. (2004). Munich chain-ladder. *Blätter der DGVFM*, *26*(4), 597–630.

Renshaw, A. E. (1994). Modelling the claims process in the presence of covariates. *ASTIN Bulletin*, *24*(2), 265–285.

Renshaw, A. E. & Verrall, R. J. (1998). A stochastic model underlying the

chain-ladder technique. *British Actuarial Journal*, *4*(4), 903–923.

Schiegl, M. (2002). On the safety loading for chain-ladder estimates: A monte carlo simulation study. *ASTIN Bulletin*, *32*(1), 107–128.

Schnieper, R. (1991). Separating true ibnr and ibner claims. *ASTIN Bulletin*, *21*(1), 111–127.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R. & Lichtendahl, K. C. (2017). *Data mining for business analytics: concepts, techniques, and applications in R*. John Wiley & Sons.

Tarbell, T. F. (1934). Incurred but not reported claim reserves. In *Proceedings of the Casualty Actuarial Society*, volume 20, pp. 275–280.

Taylor, G. (2000). *Loss reserving: an actuarial perspective*. Kluwer Academic Publishers.

Taylor, G., McGuire, G. & Sullivan, J. (2008). Individual claim loss reserving conditioned by case estimates. *Annals of Actuarial Science*, *3*(1-2), 215–256.

Trippi, R. R. & Turban, E. (1992). *Neural networks in finance and investing: Using artificial intelligence to improve real world performance*. McGraw-Hill, Inc.

Verdonck, T., Van Wouwe, M. & Dhaene, J. (2009). A robustification of the chain-ladder method. *North American Actuarial Journal*, *13*(2), 280–298.

Verrall, R. (2004a). Obtaining predictive distributions for reserves which incorporate expert opinion. In *CAS Forum*, pp. 283–316.

Verrall, R. J. (2000). An investigation into stochastic claims reserving models and the chain-ladder technique. *Insurance: mathematics and economics*, *26*(1), 91–99.

Verrall, R. J. (2004b). A bayesian generalized linear model for the bornhuetter-ferguson method of claims reserving. *North American Actuarial Journal*, *8*(3), 67–89.

Wenck, T. L. (1987). Insurer insolvency: Causes, effects and solutions. *The Journal of Insurance Issues and Practices*, pp. 35–48.

Wright, T. S. (1990). A stochastic method for claims reserving in general insurance. *Journal of the Institute of Actuaries*, *117*(3), 677–731.

Wüthrich, M. & Merz, M. (2008). *Stochastic claims reserving methods in insurance*, volume 435. John Wiley & Sons.

Wüthrich, M. V. (2003). Claims reserving using tweedie's compound poisson model. *ASTIN Bulletin*, *33*(2), 331–346.

Wüthrich, M. V. (2018a). Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, *2018*(6), 465–480.

Wüthrich, M. V. (2018b). Neural networks applied to chain-ladder reserving. *European Actuarial Journal*, *8*(2), 407–436.

Zhao, X. & Zhou, X. (2010). Applying copula models to individual claim loss reserving methods. *Insurance: Mathematics and Economics*, *46*(2), 290–299.